Cognitive Technologies

Georg Rehm
Andy Way *Editors*

European Language Equality

A Strategic Agenda for Digital Language Equality

OPEN ACCESS

Cognitive Technologies

Titles in this series now included in the Thomson Reuters Book Citation Index and Scopus!

The Cognitive Technologies (CT) series is committed to the timely publishing of high-quality manuscripts that promote the development of cognitive technologies and systems on the basis of artificial intelligence, image processing and understanding, natural language processing, machine learning and human-computer interaction.

It brings together the latest developments in all areas of this multidisciplinary topic, ranging from theories and algorithms to various important applications. The intended readership includes research students and researchers in computer science, computer engineering, cognitive science, electrical engineering, data science and related fields seeking a convenient way to track the latest findings on the foundations, methodologies and key applications of cognitive technologies.

The series provides a publishing and communication platform for all cognitive technologies topics, including but not limited to these most recent examples:

- Interactive machine learning, interactive deep learning, machine teaching
- Explainability (XAI), transparency, robustness of AI and trustworthy AI
- Knowledge representation, automated reasoning, multiagent systems
- Common sense modelling, context-based interpretation, hybrid cognitive technologies
- Human-centered design, socio-technical systems, human-robot interaction, cognitive robotics
- Learning with small datasets, never-ending learning, metacognition and introspection
- Intelligent decision support systems, prediction systems and warning systems
- Special transfer topics such as CT for computational sustainability, CT in business applications and CT in mobile robotic systems

The series includes monographs, introductory and advanced textbooks, state-of-the-art collections, and handbooks. In addition, it supports publishing in Open Access mode.

Georg Rehm • Andy Way
Editors

European Language Equality

A Strategic Agenda for Digital Language Equality

Springer

EUROPEAN LANGUAGE EQUALITY

Editors
Georg Rehm
German Research Centre for Artificial In
Berlin, Germany

Andy Way
ADAPT Centre
Dublin City University
Dublin, Ireland

The European Language Equality project has received funding from the European Union under grant agreement no. LC-01641480 – 101018166.

ISSN 1611-2482 ISSN 2197-6635 (electronic)
Cognitive Technologies
ISBN 978-3-031-28818-0 ISBN 978-3-031-28819-7 (eBook)
https://doi.org/10.1007/978-3-031-28819-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Europe is a mosaic of languages, cultures and peoples. The promotion and encouragement of this diversity is the reflection of our will to keep having diverse societies that live together in the use of different languages.

These languages are more than a communication tool. They are factors of identity, vectors of culture, and ways of understanding and explaining the world. Each language, regardless of its status and number of speakers, is a treasure that has been created and polished over generations. And while the means and ambition must be put in place to promote and preserve all languages, those that are in a situation of greater weakness must be the object of special attention.

The preservation of multilingualism as an expression of Europe's intrinsic diversity is therefore a political commitment that today faces significant challenges. We have built a world with very powerful uniformizing tendencies, inertias that make it increasingly difficult to protect the treasure of cultural and linguistic diversity. So much so that one language disappears every two weeks, and up to 90% of existing languages could be gone by the turn of the century.

In this sense, digital tools, although possessing many virtues, can also generate a clear concern when it comes to their impacts on linguistic diversity and equality. It has never been so fast and so easy to communicate and inform, and never have the temptation and incentives to end up doing it in just a handful of languages – the most powerful and influential – been so great. If the audience is the world, and if what counts is getting more followers, then the temptation to stop using our own languages is enormous. In this sense, preserving linguistic equality in the digital age must be an objective assumed by all EU institutions. And part of the solution can come precisely from the tools that the digital world can offer to us.

The European Parliament has long expressed its concern about the future of multilingualism in the digital age. In a landmark document, our Parliament adopted a 2018 resolution on achieving language equality in the digital age, whose rapporteur was my Welsh colleague and former MEP Jill Evans.

Building up on that report, the Panel for the Future of Science and Technology (STOA), of which I am a proud member, held a seminar in late 2022 titled "Towards full digital language equality in a multilingual European Union". This event

presented the conclusions of the European Language Equality (ELE) project, which analysed over 80 languages to develop a roadmap towards achieving full digital language equality in Europe by 2030.

It is tempting to think that multilingualism begins and ends with the languages that have a guaranteed official status; in the case of the EU, the 24 languages that appear in the treaties as the official languages of the Union. But in the EU alone there are at least 60 other languages that also deserve to be preserved and encouraged, despite the fact that they do not have official status. That is why we must welcome initiatives like the ELE project, and work together towards a Union in which all languages, especially minority ones, enjoy the same rights.

As a native Catalan speaker, who is very much aware of the pressure that technological and digital trends exert especially on lesser-used languages, and committed to the protection and promotion of these languages, I am very honoured to introduce this book, *European Language Equality: A Strategic Agenda for Digital Language Equality*, and I would like to thank and congratulate all who contributed in the ELE project and in the writing of these pages. Projects and publications like these draw the right path towards a more inclusive and diverse Union.

Brussels, January 2023 *Jordi Solé*

Preface

The origins of this book date back to 2010. Back then, under the umbrella of the EU Network of Excellence META-NET, we started preparing the White Paper Series *Europe's Languages in the Digital Age* (published in 2012)[1] and the *Strategic Research Agenda for Multilingual Europe 2020* (SRA, published in 2013), the first document of its kind for the European Language Technology (LT) field in a community-driven process.[2] The META-NET White Paper Series revealed, among others, that, back then, 21 European languages were threatened by what we called *digital language extinction*. As a direct response to this danger, the META-NET SRA provided suggestions as to how to bring about a change and how to increase the collaboration with the entire European LT community on a number of priority research themes towards the common goal of what is now known as *digital language equality in Europe*.

Especially the notion of digital language extinction but also our strategic recommendations generated a certain amount of attention. Back in 2013 and 2014, colleagues from META-NET were involved in dozens of television, radio and print interviews and there have also been several follow-up publications and EU projects as well as official questions raised in the European Parliament (EP). These eventually led to a number of workshops held in the EP and to a study commissioned by the EP's Science and Technology Options Assessment (STOA) unit. The STOA study[3] (2018) eventually paved the way for the report *Language Equality in the Digital Age*[4] jointly prepared by the EP's Committees on Culture and Education (CULT) and on Industry, Research and Energy (ITRE). These recommendations, informally known as the *Jill Evans report*, were adopted by the EP in a landslide vote in September 2018. Among other recommendations, this report suggested to the European Commission to "establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands". The

[1] http://www.meta-net.eu/whitepapers

[2] http://www.meta-net.eu/sra

[3] https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2017)598621

[4] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html

European Language Equality (ELE) proposal[5] and eventual project, described in the present volume, represented our direct response to this recommendation.

It was a pleasure to lead the ELE project and to collaborate with such a strong and dedicated team consisting of 52 partner organisations covering all European countries, academia and industry as well as all major pan-European initiatives. Like many other projects around the globe, ELE was also affected by the SARS-CoV-2 pandemic but, fortunately, our initial plans and project proposal had already been prepared during the pandemic, which meant that we were able to tailor the project to the new normal. Nevertheless, everybody involved was happy to eventually be able to attend our joint META-FORUM event, which took place in June 2022 with about 100 participants in the conference centre in Brussels and hundreds more participating remotely. After what felt like an endless succession of virtual meetings, for many of us, this was the first opportunity to meet face-to-face.

This book describes the results produced during the project's runtime; additional details are available in more than 60 project reports.[6] We would like to express our gratitude towards the consortium for its hard and dedicated work towards our goal of developing the *Strategic Research, Innovation and Implementation Agenda and Roadmap for Achieving Full Digital Language Equality in Europe by 2030*.[7] We would also like to thank all ELE colleagues wholeheartedly for the chapters they contributed, without which this book would not have been possible. Additionally, we would like to thank all initiatives ELE collaborated with, especially the European Language Grid[8] project, the results of which have also been documented in a book in the same series. Finally, we would like to thank Jordi Solé for supporting and chairing the workshop *Towards full digital language equality in a multilingual European Union*, held on 8 Nov. 2022 in the EP, and for contributing the foreword.

This volume covers the results achieved during the project's first iteration (January 2021 until June 2022). Immediately after the end of the first project, the initiative continued with the project ELE 2, which will end in June 2023. We sincerely hope that the whole ELE initiative will serve its purpose, which is to help bring about digital language equality in Europe by 2030. This book provides an analysis of the current state of play (Part I) and our recommendations for the future situation in 2030 (Part II). Proper support for the implementation of these plans would mean a quantum leap for Europe's multilingual landscape with concomitant benefits for all its citizens, regardless of the language they prefer to communicate in.

Berlin and Dublin, April 2023 *Georg Rehm*
Andy Way

Acknowledgements The European Language Equality project has received funding from the European Union under the grant agreement no. LC-01641480 – 101018166 (ELE).

[5] https://www.european-language-equality.eu

[6] https://www.european-language-equality.eu/deliverables/

[7] https://www.european-language-equality.eu/agenda/

[8] https://www.european-language-grid.eu

Contents

List of Contributors

Gilles Adda
Université Paris-Saclay, CNRS, LISN, France, gilles.adda@limsi.fr

Rodrigo Agerri
University of the Basque Country, Spain, rodrigo.agerri@ehu.eus

Eneko Agirre
University of the Basque Country, Spain, e.agirre@ehu.eus

Itziar Aldabe
University of the Basque Country, Spain, itziar.aldabe@ehu.eus

Dimitra Anastasiou
Luxembourg Institute of Science and Technology, Luxembourg,
dimitra.anastasiou@list.lu

Nora Aranberri
University of the Basque Country, Spain, nora.aranberri@ehu.eus

Victoria Arranz
Evaluations and Language Resources Distribution Agency, France, arranz@elda.org

Jose Maria Arriola
University of the Basque Country, Spain, josemaria.arriola@ehu.eus

Aitziber Atutxa
University of the Basque Country, Spain, aitziber.atutxa@ehu.eus

Ilze Auziņa
Institute of Mathematics and Computer Science, University of Latvia, Latvia,
ilze.auzina@lumii.lv

Gorka Azkune
University of the Basque Country, Spain, gorka.azkune@ehu.eus

Gerhard Backfried
HENSOLDT Analytics GmbH, Austria, gerhard.backfried@hensoldt.net

Cristian Berrio
Expert.AI, Spain, cberrio@expert.ai

Aivars Bērziņš
Tilde, Latvia, aivars.berzins@tilde.com

Joachim Van den Bogaert
CrossLang, Belgium, joachim.van.den.bogaert@crosslang.com

Kalina Bontcheva
University of Sheffield, United Kingdom, k.bontcheva@sheffield.ac.uk

Claudia Borg
University of Malta, Malta, claudia.borg@um.edu.mt

Lars Borin
University of Gothenburg, Sweden, lars.borin@svenska.gu.se

Svetla Boytcheva
Ontotext, Bulgaria, svetla.boytcheva@ontotext.com

António Branco
University of Lisbon, Portugal, antonio.branco@di.fc.ul.pt

David Cabestany
Barcelona Supercomputing Center, Spain, david.cabestany@bsc.es

Blanca Calvo
Barcelona Supercomputing Center, Spain, blanca.calvo@bsc.es

Jon Ander Campos
University of the Basque Country, Spain, jonander.campos@ehu.eus

Arantza Casillas
University of the Basque Country, Spain, arantza.casillas@ehu.eus

Sheila Castilho
Dublin City University, ADAPT Centre, Ireland, sheila.castilho@adaptcentre.ie

Khalid Choukri
Evaluations and Language Resources Distribution Agency, France,
choukri@elda.org

Tarik Ćušić
University of Sarajevo, Bosnia and Herzegovina, tarik.cusic@izj.unsa.ba

Walter Daelemans
University of Antwerp, Belgium, walter.daelemans@uantwerpen.be

Daniel Dahlmeier
SAP SE, Germany, daniel.dahlmeier@sap.com

Ingerid Løyning Dale
The National Library of Norway, Norway, ingerid.dale@nb.no

Emma Daly
Dublin City University, ADAPT Centre, Ireland, emma.daly@adaptcentre.ie

Andrea DeMarco
University of Malta, Malta, andrea.demarco@um.edu.mt

Arantza Díaz de Ilarraza
University of the Basque Country, Spain, a.diazdeilarraza@ehu.eus

Laura Docío Fernández
University of Vigo, Spain, ldocio@gts.uvigo.es

Rickard Domeij
Institute of Languages and Folklore, Sweden, rickard.domeij@isof.se

Jane Dunne
Dublin City University, ADAPT Centre, Ireland, jane.dunne@adaptcentre.ie

Wilhelmina Dyster
University of Helsinki, Finland, wilhelmina.dyster@helsinki.fi

Jens Edlund
KTH Royal Institute of Technology, Sweden, edlund@speech.kth.se

Kristine Eide
The Language Council of Norway, Norway, kristine.eide@sprakradet.no

Ainara Estarrona
University of the Basque Country, Spain, ainara.estarrona@ehu.eus

Aritz Farwell
University of the Basque Country, Spain, aritz.farwell@ehu.eus

Markus Forsberg
University of Gothenburg, Sweden, markus.forsberg@gu.se

Anželika Gaidienė
Institute of the Lithuanian Language, Lithuania, anzelika.gaidiene@lki.lt

Owen Gallagher
Dublin City University, ADAPT Centre, Ireland, owen.gallagher@adaptcentre.ie

Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia,
radovan.garabik@kassiopeia.juls.savba.sk

Mercedes García-Martínez
Pangeanic, Spain, m.garcia@pangeanic.com

Carmen García-Mateo
University of Vigo, Spain, carmen.garcia@uvigo.es

Andrés García-Silva
Expert.AI, Spain, agarcia@expert.ai

Federico Gaspari
Dublin City University, ADAPT Centre, Ireland, federico.gaspari@adaptcentre.ie

Maria Gavriilidou
Institute for Language and Speech Processing, R. C. “Athena”, Greece, maria@athenarc.gr

Maria Giagkou
Institute for Language and Speech Processing, R. C. “Athena”, Greece, mgiagkou@athenarc.gr

Iakes Goenaga
University of the Basque Country, Spain, iakes.goenaga@ehu.eus

Josu Goikoetxea
University of the Basque Country, Spain, josu.goikoetxea@ehu.eus

Koldo Gojenola
University of the Basque Country, Spain, koldo.gojenola@ehu.eus

Jose Manuel Gómez-Pérez
Expert.AI, Spain, jmgomez@expert.ai

Mark A. Greenwood
University of Sheffield, United Kingdom, m.greenwood@sheffield.ac.uk

Sara Grilo
University of Lisbon, Portugal, sara.grilo@di.fc.ul.pt

Dagmar Gromann
University of Vienna, Austria, dagmar.gromann@univie.ac.at

Annika Grützner-Zahn
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany, annika.gruetzner-zahn@dfki.de

Normunds Grūzītis
Institute of Mathematics and Computer Science, University of Latvia, Latvia, normunds.gruzitis@lumii.lv

Jan Hajič
Charles University, Czech Republic, hajic@ufal.mff.cuni.cz

Stefanie Hegele
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany, stefanie.hegele@dfki.de

Barbara Heinisch
University of Vienna, Austria, barbara.heinisch@univie.ac.at

Enikő Héja
Research Centre for Linguistics, Hungary, heja.eniko@nytud.hu

Lina Henriksen
University of Copenhagen, Denmark, linah@hum.ku.dk

Inma Hernáez
University of the Basque Country, Spain, inma.hernaez@ehu.eus

Manuel Herranz
Pangeanic, Spain, m.herranz@pangeanic.com

Maria Heuschkel
Wikimedia Deutschland, Germany, maria.heuschkel@wikimedia.de

Jaroslava Hlaváčová
Charles University, Czech Republic, hlavacova@ufal.mff.cuni.cz

Johannes Hoffart
SAP SE, Germany, johannes.hoffart@sap.com

Mikel Iruskieta
University of the Basque Country, Spain, mikel.iruskieta@ehu.eus

Kinga Jelencsik-Mátyus
Research Centre for Linguistics, Hungary, jelencsik-matyus.kinga@nytud.hu

Gareth Jones
Bangor University, United Kingdom, g.jones@bangor.ac.uk

Franciska de Jong
CLARIN ERIC, The Netherlands, franciska@clarin.eu

Martin Kaltenböck
Semantic Web Company, Austria, martin.kaltenboeck@semantic-web.com

Andre Kåsen
The National Library of Norway, Norway, andre.kasen@nb.no

Svetla Koeva
Institute for Bulgarian Language Prof. Lyubomir Andreychin,
Bulgarian Academy of Sciences, Bulgaria, svetla@dcl.bas.bg

Marek Kováč
Phonexia, Czech Republic, kovac@phonexia.com

Simon Krek
Jožef Stefan Institute, Slovenia, simon.krek@ijs.si

Cvetana Krstev
University of Belgrade, Serbia, cvetana@matf.bg.ac.rs

Gorka Labaka
University of the Basque Country, Spain, gorka.labaka@ehu.eus

Alberto Lavelli
Fondazione Bruno Kessler, Italy, lavelli@fbk.eu

Marek Łaziński
University of Warsaw, Poland, m.lazinski@uw.edu.pl

Christian Lieske
SAP SE, Germany, christian.lieske@sap.com

Krister Lindén
University of Helsinki, Finland, krister.linden@helsinki.fi

Dora Loizidou
University of Cyprus, Cyprus, loizidou.dora@ucy.ac.cy

Oier Lopez de Lacalle
University of the Basque Country, Spain, oier.lopezdelacalle@ehu.eus

Teresa Lynn
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie

Bernardo Magnini
Fondazione Bruno Kessler, Italy, magnini@fbk.eu

Katrin Marheinecke
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany, katrin.marheinecke@dfki.de

Diana Maynard
University of Sheffield, United Kingdom, d.maynard@sheffield.ac.uk

Maite Melero
Barcelona Supercomputing Center, Spain, maite.melero@bsc.es

Marcin Miłkowski
Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland, mmilkows@ifispan.edu.pl

Kadri Muischnek
University of Tartu, Estonia, kadri.muischnek@ut.ee

Eva Navas
University of the Basque Country, Spain, eva.navas@ehu.eus

Sally O'Connor
KantanMT, Ireland, sallyoc@kantanai.io

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences, Poland, maciej.ogrodniczuk@ipipan.waw.pl

Sussi Olsen
University of Copenhagen, Denmark, saolsen@hum.ku.dk

Maite Oronoz
University of the Basque Country, Spain, maite.oronoz@ehu.eus

Arantxa Otegi
University of the Basque Country, Spain, arantza.otegi@ehu.eus

Vasile Păiş
Research Institute for Artificial Intelligence "Mihai Drăgănescu",
Romanian Academy, Romania, vasile@racai.ro

Pablo Peñarrubia
Barcelona Supercomputing Center, Spain, pablo.penarrubia@bsc.es

Alicia Pérez
University of the Basque Country, Spain, alicia.perez@ehu.eus

Olatz Perez de Viñaspre
University of the Basque Country, Spain, olatz.perezdevinaspre@ehu.eus

Piotr Pęzik
University of Łódź, Poland, piotr.pezik@uni.lodz.pl

Mārcis Pinnis
Tilde, Latvia, marcis.pinnis@tilde.com

Stelios Piperidis
Institute for Language and Speech Processing, R. C. "Athena", Greece,
spip@athenarc.gr

Peter Polák
Charles University, Czech Republic, polak@ufal.mff.cuni.cz

Martin Popel
Charles University, Czech Republic, popel@ufal.mff.cuni.cz

Maja Popović
Dublin City University, ADAPT Centre, Ireland, maja.popovic@adaptcentre.ie

Antonia Popp
University of Zurich, Switzerland, popp@cl.uzh.ch

Delyth Prys
Bangor University, United Kingdom, d.prys@bangor.ac.uk

José Manuel Ramírez Sánchez
University of Vigo, Spain, jmramirez@gts.uvigo.es

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany,
georg.rehm@dfki.de

Natalia Resende
Dublin City University, ADAPT Centre, Ireland, natalia.resende@adaptcentre.ie

Artem Revenko
Semantic Web Company, Austria, artem.revenko@semantic-web.com

German Rigau
University of the Basque Country, Spain, german.rigau@ehu.eus

Annette Rios
University of Zurich, Switzerland, rios@cl.uzh.ch

Mar Rodríguez
Barcelona Supercomputing Center, Spain, mar.rodriguez@bsc.es

Eiríkur Rögnvaldsson
Árni Magnússon Institute for Icelandic Studies, Iceland, eirikur@hi.is

Johan Rohdin
Phonexia, Czech Republic, rohdin@phonexia.com

Rudolf Rosa
Charles University, Czech Republic, rosa@ufal.mff.cuni.cz

Michael Rosner
University of Malta, Malta, mike.rosner@um.edu.mt

Ander Salaberria
University of the Basque Country, Spain, ander.salaberria@ehu.eus

Jon Sanchez
University of the Basque Country, Spain, jon.sanchez@ehu.eus

Bolette Sandford Pedersen
University of Copenhagen, Denmark, bspedersen@hum.ku.dk

Kepa Sarasola
University of the Basque Country, Spain, kepa.sarasola@ehu.eus

Ibon Saratxaga
University of the Basque Country, Spain, ibon.saratxaga@ehu.eus

Felix Sasaki
SAP SE, Germany, felix.sasaki@sap.com

Petr Schwarz
Phonexia, Czech Republic, schwarz@phonexia.com

João Silva
University of Lisbon, Portugal, joao.silva@di.fc.ul.pt

Inguna Skadiņa
Institute of Mathematics and Computer Science, University of Latvia and Tilde, Latvia, inguna.skadina@tilde.com

Marcin Skowron
HENSOLDT Analytics GmbH, Austria, marcin.skowron@hensoldt.net

Aitor Soroa
University of the Basque Country, Spain, a.soroa@ehu.eus

Manuela Speranza
Fondazione Bruno Kessler, Italy, manspera@fbk.eu

Ranka Stanković
University of Belgrade, Serbia, ranka@rgf.bg.ac.rs

Frieda Steurs
Dutch Language Institute, The Netherlands, frieda.steurs@ivdnt.org

Riccardo Superbo
KantanMT, Ireland, riccardos@kantanai.io

Marko Tadić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia, marko.tadic@ffzg.hr

Aurelija Tamulionienė
Institute of the Lithuanian Language, Lithuania, aurelija.tamulioniene@lki.lt

Dan Tufiș
Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Romania, tufis@racai.ro

Baiba Valkovska
Institute of Mathematics and Computer Science, University of Latvia, Latvia, baiba.valkovska@lumii.lv

Vincent Vandeghinste
Dutch Language Institute, The Netherlands, vincent.vandeghinste@ivdnt.org

Tamás Váradi
Research Centre for Linguistics, Hungary, varadi.tamas@nytud.hu

Zsófia Varga
Research Centre for Linguistics, Hungary, varga.zsofia@nytud.hu

Ioana Vasilescu
Université Paris-Saclay, CNRS, LISN, France, ioana.vasilescu@limsi.fr

Andrejs Vasiļjevs
Tilde, Latvia, andrejs.vasiljevs@tilde.com

Marta Villegas
Barcelona Supercomputing Center, Spain, marta.villegas@bsc.es

Tereza Vojtěchová
Charles University, Czech Republic, vojtechova@ufal.mff.cuni.cz

Martin Volk
University of Zurich, Switzerland, volk@cl.uzh.ch

Gareth Watkins
Bangor University, United Kingdom, g.watkins@bangor.ac.uk

Andy Way
Dublin City University, ADAPT Centre, Ireland, andy.way@adaptcentre.ie

Joanna Wright
University of Sheffield, United Kingdom, j.wright@sheffield.ac.uk

François Yvon
Université Paris-Saclay, CNRS, LISN, France, francois.yvon@limsi.fr

Acronyms

ABSA	Aspect-based Sentiment Analysis
ACL	Association for Computational Linguistics
ADRA	AI, Data and Robotics Asociation
AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
API	Application Programming Interface
ASR	Automatic Speech Recognition
BART	Bidirectional Auto-Regressive Transformers
BDVA	Big Data Value Association
BERT	Bidirectional Encoder Representations from Transformers
BLARK	Basic Language Resource Kit
BLEU	Bilingual Evaluation Understudy
CA	Conversational Agent
CEF	Connecting Europe Facility
CF	Contextual Factor
CL	Computational Linguistics
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CMS	Content Management System
CNN	Convolutional Neural Network
CPAI	Coordinated Plan on Artificial Intelligence
CRACKER	Cracking the Language Barrier
CSA	Coordination and Support Action
CULT	Committee on Culture and Education
DAIRO	Data, AI and Robotics
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DCAT	Data Catalogue Vocabulary
DG CNECT	Directorate-General for Communications Networks, Content and Technology (European Commission)
DGA	Data Governance Act
DH	Digital Humanities
DIN	Deutsches Institut für Normung (German Institute for Standardisation)

DL	Deep Learning
DLE	Digital Language Equality
DNN	Deep Neural Networks
DPP	Data Protection and Privacy
DSM	Digital Single Market
E2E	End-to-End System
EC	European Commission
ECPAI	EU Coordinated Plan on Artificial Intelligence
ECRML	European Charter for Regional or Minority Languages
ECSPM	European Civil Society Platform for Multilingualism
EEA	European Economic Area
EFNIL	European Federation of National Institutions for Language
EL	Entity Linking
ELE	European Language Equality EU Project
ELEN	European Language Equality Network
ELEXIS	European Lexicographic Infrastructure
ELG	European Language Grid
ELIS/EUATC	European Language Industry Survey
ELISE	European Learning and Intelligent Systems Excellence
ELITR	European Live Translator
ELLIS	European Laboratory for Learning and Intelligent Systems
ELM	European Language Monitor
ELRA	European Language Resources Association
ELRC	European Language Resource Coordination
ELT	European Language Technology
EMM	Enterprise Metadata Management
EOSC	European Open Science Cloud
EP	European Parliament
ESF	European Social Fund
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
EUDAT CDI	EUDAT Collaborative Data Infrastructure
EVALITA	Evaluation of NLP and Speech Tools for Italian
FAIR	Findable, Accessible, Interoperable, Reusable Principles
GAN	Generative Adversarial Networks
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Model
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
H2H	Human-to-Human Communication
H2M	Human-to-Machine Communication
HCI	Human-Computer Interaction
HLT	Human Language Technology
HMM	Hidden Markov Models

HPC	High Performance Computing
HTML	Hypertext Markup Language
IA	Innovation Action
ICT	Information and Communication Technology
IDSA	International Data Spaces Association
IE	Information Extraction
IEEE	Institute of Electrical and Electronics Engineers
IPR	Intellectual Property Rights
IR	Information Retrieval
ISCA	International Speech Communication Association
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
KG	Knowledge Graph
LDC	Linguistic Data Consortium
LDS	Language Data Space
LEAM	Large European AI Models
LIBER	Association of European Research Libraries
LID	Language Identification
LLM	Large Language Model
LM	Language Model
LOD	Linked Open Data
LR	Language Resource
LRT	Language Resources and Technologies
LSC	Catalan Sign Language
LSE	Spanish Sign Language
LT	Language Technology
LTPI	Language Technology Programme for Icelandic
MAPA	Multilingual Anonymisation for Public Administrations
MARCELL	Multilingual Resources for CEF.AT in the Legal Domain
META	Multilingual Europe Technology Alliance
META-NET	A Network of Excellence forging META
MFF	Multiannual Financial Framework
ML	Machine Learning
MLLM	Multilingual Language Model
MMT	Multimodal Machine Translation
MNMT	Multilingual Neural Machine Translation
MT	Machine Translation
MWE	Multiword Expression
NE	Named Entity
NED	Named Entity Disambiguation
NEM	New European Media Initiative
NER	Named Entity Recognition
NFDI	Nationale Forschungsdateninfrastruktur (German National Research Data Infrastructure)
NLG	Natural Language Generation

NLP	Natural Language Processing
NLTP	National Language Technology Platform
NLU	Natural Language Understanding
NMT	Neural Machine Translation
NN	Neural Network
NTEU	Neural Translation for the European Union
OCR	Optical Character Recognition
OECD	Organisation for Economic Co-operation and Development
OIE	Open Information Extraction
PII	Personal Identifiable Information
POS	Part-of-Speech
PRINCIPLE	Providing Resources in Irish, Norwegian, Croatian and Icelandic for Purposes of Language Engineering
QA	Question Answering
RDA	Research Data Alliance
RE	Relation Extraction
RI	Research Infrastructures
RIA	Research and Innovation Action
RML	Regional and Minority Languages
RNN	Recurrent Neural Network
ROI	Return On Investment
SER	Speech Emotion Recognition
SID	Speaker Identification
SME	Small and Medium-size Enterprises
SR	Speaker Recognition
SRA	Strategic Research Agenda
SRIA	Strategic Research, Innovation and Implementation Agenda
SRL	Semantic Role Labelling
SSH	Social Sciences and Humanities
SSHOC	Social Sciences and Humanities Open Cloud
SSL	Swedish Sign Language
ST	Speech Technology
STOA	Science and Technology Options Assessment
TA	Text Analysis
TF	Technological Factor
TLD	Top-Level Domain
TM	Text Mining
TTS	Text-to-Speech Synthesis
UD	Universal Dependencies
UN	United Nations
UNESCO	UN Educational, Scientific and Cultural Organization
VQA	Visual Question Answering
WER	Word Error Rate
WP	Work Package
WSD	Word Sense Disambiguation

Chapter 1
European Language Equality: Introduction

Georg Rehm and Andy Way

Abstract This chapter provides an introduction to the EU-funded project *European Language Equality* (ELE). It motivates the project by taking a general look at multilingualism, especially with regard to the political equality of all languages in Europe. Since 2010, several projects and initiatives have developed the notion of utilising sophisticated language technologies to unlock and enable multilingualism technologically. However, despite a landmark resolution that was adopted by the European Parliament in 2018, no significant progress has been made. Together with the whole European LT community, and making use of a concerted community consultation process, the ELE project produced strategic recommendations that specify how to bring about full digital language equality in Europe and reach the scientific goal of Deep Natural Language Understanding by 2030, not only addressing but eventually solving the problem of *digital inequality* of Europe's languages.

1 Overview and Context

In Europe's multilingual setup, all 24 official EU languages are granted equal status by the EU Charter and the Treaty on EU. Furthermore, the EU is home to over 60 regional and minority languages which have been protected and promoted under the European Charter for Regional or Minority Languages (ECRML) treaty since 1992, in addition to various sign languages and the languages of immigrants as well as trade partners. Additionally, the Charter of Fundamental Rights of the EU under Article 21 states that, "[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

Andy Way
Dublin City University, ADAPT Centre, Ireland, andy.way@adaptcentre.ie

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_1

Unfortunately, language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across language communities and continue to be unbreachable in many situations. While multilingualism is one of the key cultural cornerstones of Europe and signifies part of what it means to be and to feel European, no EU policy has been proposed to address the problem of language barriers.

Artificial Intelligence (AI), Natural Language Processing (NLP), Natural Language Understanding (NLU), Language Technologies (LTs), and Speech Technologies (STs) have the potential to enable multilingualism technologically but, as the META-NET White Paper Series *Europe's Languages in the Digital Age* (Rehm and Uszkoreit 2012) found in 2012, our languages suffer from an extreme imbalance in terms of technological support. English is very well supported through technologies, tools, datasets and corpora, for example, but languages such as Maltese, Estonian or Icelandic have hardly any support at all. In fact, the 2012 study assessed *at least 21 European languages to be in danger of digital extinction*. If, as mentioned above, all European languages are supposed to be on an equal footing in general, technologically, they clearly are not (Kornai 2013).

After the findings of the META-NET study and a set of follow-up projects, studies and recommendations (e. g., Rehm and Uszkoreit 2013; STOA 2018), the joint CULT/ITRE report *Language Equality in the Digital Age* (European Parliament 2018) was eventually passed with an overwhelming majority by the European Parliament on 11 September 2018. It concerns the improvement of the institutional framework for LT policies at the EU level, EU research and education policies to improve the future of LTs in Europe, and the extension of the benefits of LTs for both private companies and public bodies. The resolution also recognises that there is an imbalance in terms of technology support of Europe's languages, that there has been a substantial amount of progress in research and technology development and that a large-scale, long-term funding programme should be established to ensure full technology support for all of Europe's languages. The goal is to enable multilingualism technologically since "the EU and its institutions have a duty to enhance, promote and uphold linguistic diversity in Europe" (European Parliament 2018).

While the resolution was a important milestone for the idea of enabling Europe's multilingualism technologically and bringing every language in Europe to the same level of technology support, there has been no concrete follow-up action along the lines laid out in the resolution, i. e., to set up "a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands". In the meantime, however, many highly influential breakthroughs in the area of language-centric AI have been achieved, mostly by large enterprises in the US and Asia, especially approaches and technologies concerning large language models (LLMs such as BERT or ChatGPT).[1]

Due to a lack of action over the last five to seven years, Europe has mostly been playing "second fiddle" in the area of language-centric AI and Language Technolo-

[1] ChatGPT was released in Nov. 2022, https://chat-gpt.org. Most chapters of this book were written by mid-2022, which is why they do not reflect the widespread impact and subsequent recognition of this novel application.

gies. Driven by the "European Strategy for data", the EU is currently concentrating on setting up a number of sectorial data spaces to enable and support the data economy and to boost its digital sovereignty.[2] These, fortunately, also include a dedicated language data space with a focus on stakeholders from industry. *But, simply put, language is much more than data*. In addition to the complex and long-term activity of constructing the aforementioned data spaces, the EU also invests in AI-related actions that include language, albeit with limited budgets. However, much more needs to be done to properly address the challenge of Europe's multilingualism with meaningful and long-lasting solutions.

With a consortium of 52 partners, the EU project *European Language Equality* (ELE; Jan. 2021 – June 2022) and its follow-up project ELE 2 (July 2022 – June 2023) developed, through a large-scale, community-driven process, a *Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030* to address this major issue by means of a coordinated, pan-European research, development and innovation programme.[3] This book is the definitive documentation of the EU project ELE. It describes the current situation of technology support for Europe's languages and our overall recommendations of what more needs to be done to achieve Digital Language Equality (DLE) in Europe by 2030.

2 The European Language Equality Project

The original proposal for the EU project "European Language Equality" was prepared by a consortium of 52 partners[4] (see Figure 1) and submitted on 29 July 2020, responding to the European Commission call topic PPA-LANGEQ-2020 ("Developing a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030").[5] The ELE project started in January 2021 and finished in June 2022. Immediately after the end of the first ELE project, the one-year ELE 2 project began with a reduced consortium of seven partners, continuing some of the work strands of the first project.

Developing a strategic agenda and roadmap for achieving full DLE in Europe by 2030 involves many stakeholders, which is why the process of preparing the different parts of the strategic agenda and roadmap – the key objective and result of the project – was carried out together with all 52 partners of the consortium and the wider European LT community. We concentrated on two distinct but related aspects: 1. describing the current state of play (as of 2021/2022) of LT support for the languages under investigation; and 2. strategic and technological forecasting, i. e., estimating and envisioning the future situation ca. 2030. Furthermore, we distinguished between two main stakeholder groups: 1. *LT developers* (industry and research) and

[2] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

[3] https://european-language-equality.eu

[4] https://european-language-equality.eu/consortium/

[5] https://ec.europa.eu/research/participants/data/ref/other_eu_prog/other/pppa/wp-call/call-fiche_pppa-langeq-2020_en.pdf

Fig. 1 Members of the ELE consortium at META-FORUM 2022 in Brussels (9 June 2022)

2. *LT users and consumers*. Both groups were represented in ELE with several networks, initiatives and associations who produced one report each, highlighting their own individual needs, wishes and demands towards DLE. The project's industry partners produced four in-depth reports compiling the needs, wishes and visions of the European LT industry. We also organised a larger number of surveys (inspired by Rehm and Hegele 2018) and consultations with stakeholders not directly represented in the consortium.

With the development of the strategic agenda, the project followed two complementary goals. 1. The *socio-political goal* was the preparation of a strategic agenda explaining how Europe can bring about full digital language equality by 2030. This objective and the need for a corresponding large-scale, long-term programme have been recognised already by the EU (European Parliament 2018). 2. Additionally, the strategic agenda and the eventual large-scale, long-term funding programme are also meant to pursue a *scientific goal*, i. e., reaching *Deep Natural Language Understanding* by 2030. As briefly mentioned, Europe is currently lagging behind the breakthroughs achieved on other continents, which is why the dedicated large-scale, long-term funding programme we envision can and must achieve both objectives: develop resources and technologies to fully unlock and benefit from multilingualism technologically and also put Europe back into the pole position in the area of LT, NLP and language-centric AI research.

Operationally, the project was structured into five work packages (see Figure 2). In WP1, "European Language Equality: Status Quo in 2020/2021", a definition of the concept of DLE was prepared and the current state-of-the-art in the research area of LT and language-centric AI was documented in a report. The heart of WP1 was the preparation of more than 30 language reports, each documenting one European language and the level of technology support it had as of 2022. While WP1 examined the status quo, WP2, "European Language Equality: The Future Situation in 2030" looked into the future. Operationalised through a complex community consultation

process, we collected and analysed the demands, needs, ideas and wishes of European LT developers (industry and research), European LT users and consumers as well as European citizens. Four technical deep dives took a detailed look at the four main areas of LT (Machine Translation, Speech, Text Analytics and Data). The results of WP1 and WP2 were fed to WP3, "Development of the Strategic Agenda and Roadmap", in which the overall strategic agenda was developed based on the collected findings of WP1 and WP2, including an additional feedback loop with the wider community. WP4, "Communication – Dissemination – Exploitation – Sustainability" organised a number of events, including META-FORUM 2022[6] in Brussels (see Figure 1) and a workshop in the European Parliament.[7] WP4 also set up and managed our social media channels and a newsletter under the umbrella brand "European Language Technology".[8] WP5 took care of managing the large consortium of 52 partners. Figure 3 shows the overall timeline of the project.

Our methodology was, thus, based on a number of stakeholder-specific surveys as well as collaborative document preparation that also involved technology forecasting. Both approaches were complemented through the collection of additional input and feedback through various online channels. The two main stakeholder groups (LT developers and LT users/consumers) differ in one substantial way: while the group of commercial or academic LT developers is, in a certain way, *closed* and well represented through relevant organisations, networks and initiatives in the ELE consortium, the group of LT users is an *open* set of stakeholders that is only partially represented in our consortium. Both stakeholder groups have been addressed with targeted and stakeholder-specific surveys.

The ELE project resulted in around 70 deliverables, of which the public ones are available online.[9] In addition, a number of reports were prepared pro bono by collaborators who supported the goals of the project, including language reports on Bosnian, Serbian, West Frisian, the Nordic minority languages and Europe's sign languages. All reports are available on the ELE website.

3 Beyond the ELE Project

While forecasting the future of the field of LT and language-centric AI is surely an enormous challenge, we can confidently predict that even greater advances will be achieved in all LT research areas and domains in the near future (Rehm et al. 2022). However, despite claims of human parity in many LT tasks, *Deep Natural Language Understanding*, the main scientific goal of the ELE Programme, is still an *open research problem* far from being solved since all current approaches have

[6] https://www.european-language-grid.eu/events/meta-forum-2022

[7] https://www.europarl.europa.eu/stoa/en/events/details/towards-full-digital-language-equality-i/20220711WKS04301

[8] The social media channels and the newsletter were organised in close collaboration with ELE's sister project European Language Grid (ELG, Rehm 2023).

[9] https://www.european-language-equality.eu/deliverables

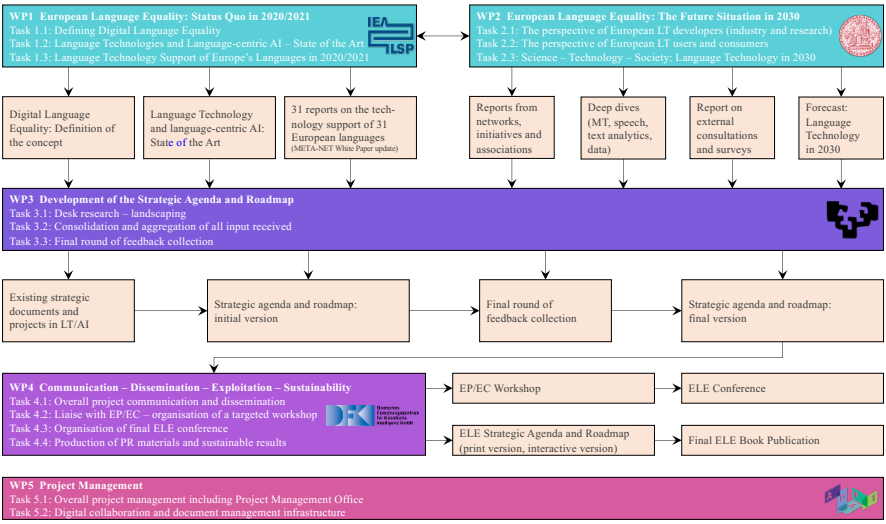

Fig. 2 Work packages and tasks of the ELE project

severe limitations (Bender et al. 2021). Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models and self-supervised systems opens up the way to leverage LT for less-developed languages. For the first time, a single multilingual model recently outperformed the best specially trained bilingual models on news translations, i. e., one multilingual model provided the best translations for both low- and high-resource languages, indicating that the multilingual approach appears to be the future of MT (Tran et al. 2021). However, the development of these new systems would not be possible without sufficient resources (experts, data, compute facilities, etc.), including the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application.

Unfortunately, as of now, there is no equality in terms of tool, resource and application availability across languages and domains. Although LT has the potential to overcome the linguistic divide in the digital sphere, most languages are neglected for various reasons, including an absence of institutional engagement from decision-makers and policy stakeholders, limited commercial interest and insufficient resources. For instance, Joshi et al. (2020) and Blasi et al. (2022) look at the relation between the types of languages, resources and their representation in NLP conferences over time. As expected, but also disappointingly, only a very small number of the over 6,000 languages of the world are represented in the rapidly evolving field of LT. A growing concern is that due to unequal access to digital resources and financial support, only a small group of large enterprises and elite universities are in a position to lead further development in this area (Ahmed and Wahed 2020).

To unleash the full potential of LT in Europe and ensure that no users of these technologies are disadvantaged in the digital sphere *simply due to the language they*

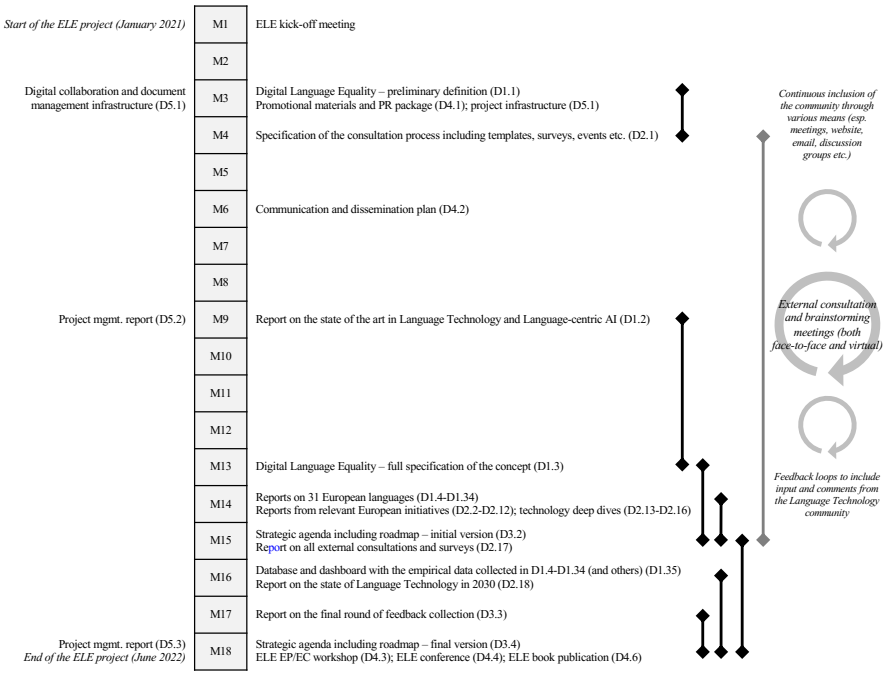

Fig. 3 Overall timeline of the ELE project

speak, we argue that there is a pressing need to facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical, fair and unbiased language understanding and communication. In short, we must ensure DLE in all areas of society, from government to business to citizens.

4 Summary of this Book

This book is structured into two main parts. Part I examines the *current state of play* of technology support for Europe's languages. Part II outlines the *future situation* in 2030 and beyond, as specified through the community consulting and forecasting process of the ELE project. Below we include short summaries of the two parts.

4.1 Part I: European Language Equality – Status Quo in 2022

Part I concentrates on the current situation as of 2022. First, Chapter 2 examines the state-of-the-art in LT, NLP and language-centric AI. It provides the technical foundation of all subsequent chapters. Chapter 3 defines the DLE metric, developed

within the project, with its technological (Gaspari et al. 2022) and contextual factors (Grützner-Zahn and Rehm 2022). This chapter also describes the interactive DLE dashboard, which was implemented as an additional component of the European Language Grid cloud platform (ELG, Rehm 2023). Assuming that the ELG catalogue of resources, tools and services contains, at any given point in time, a representative picture of the technology support of Europe's languages, the dashboard can be used to visualise the overall situation in different ways, including comparisons of multiple languages along various dimensions. Chapter 4 summarises the findings and provides an answer to the question of how Europe's languages compare technologically ca. 2022. The chapter describes the methodology of basing the computation of the DLE scores on the contents of the ELG repository, which has been substantially expanded by the ELE project with more than 6,000 additional resources, and highlights the current situation using a number of graphs. Chapters 5 to 37 contain extended high-level summaries of the 33 language reports produced by the ELE project. These reports can be conceptualised as updates, ten years on, of the META-NET White Papers (Rehm and Uszkoreit 2012), especially as many of them were written by the original authors.

4.2 Part II: European Language Equality – The Future Situation in 2030 and beyond

Part II outlines the future situation in 2030 and beyond, making use of the collected and synthesised results of the community consultation process. First, Chapter 38 describes the community consultation process on a general level, primarily with regard to the different surveys used in the project vis-à-vis European LT developers, European LT users and consumers as well as European citizens. The chapter also summarises the approach regarding the four technology deep dives as well as the dissemination and feedback collection activities in the project. Chapter 39 summarises the results of the three main surveys. The following four chapters highlight the main findings of the four technology deep dives on the four main areas of LT research and development: Machine Translation (Chapter 40), Speech Technologies (Chapter 41), Text Analytics (Chapter 42) as well as Data and Knowledge (Chapter 43). The penultimate Chapter 44 presents the strategic plans and projects in LT and AI from an international, European and national perspective. It contextualises the strategic recommendations of the project. Finally, Chapter 45, provides an extended summary of the stand-alone document of the *Strategic Research, Innovation and Implementation Agenda and Roadmap* the ELE project has developed.[10] On the whole, the present book can be conceptualised as the collective findings and recommendations of the ELE project, and as such it reflects years of work based on the distilled input and collaboration of hundreds of experts and stakeholders from across the European LT and language-centric AI community.

[10] https://european-language-equality.eu/agenda/

References

Ahmed, Nur and Muntasir Wahed (2020). "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research". In: *CoRR* abs/2010.15581. https://arxiv.org/abs/2010.15581.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada, pp. 610–623.

Blasi, Damian, Antonios Anastasopoulos, and Graham Neubig (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5486–5505. DOI: 10.18653/v1/2022.acl-long.376. https://aclanthology.org/2022.acl-long.376.

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way (2022). "Introducing the Digital Language Equality Metric: Technological Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 1–12. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf.

Grützner-Zahn, Annika and Georg Rehm (2022). "Introducing the Digital Language Equality Metric: Contextual Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 13–26. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. https://aclanthology.org/2020.acl-main.560.

Kornai, Andras (2013). "Digital Language Death". In: *PLoS ONE* 8.10. DOI: 10.1371/journal.pone.0077056. https://doi.org/10.1371/journal.pone.0077056.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Federico Gaspari, German Rigau, Maria Giagkou, Stelios Piperidis, Annika Grützner-Zahn, Natalia Resende, Jan Hajic, and Andy Way (2022). "The European Language Equality Project: Enabling digital language equality for all European languages by 2030". In: *The Role of National Language Institutions in the Digital Age – Contributions to the EFNIL Conference 2021 in Cavtat*. Ed. by Željko Jozić and Sabine Kirchmeier. Budapest, Hungary: Nyelvtudományi Kutatóközpont, Hungarian Research Centre for Linguistics, pp. 17–47.

Rehm, Georg and Stefanie Hegele (2018). "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs". In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA, pp. 3282–3289. https://aclanthology.org/L18-1519.pdf.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg etc.: Springer. http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. https://data.europa.eu/doi/10.2861/136527.

Tran, Chau, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan (2021). "Facebook AI's WMT21 News Translation Task Submission". In: *Proceedings of the Sixth Conference on Machine Translation (WMT 2021)*. Online: Association for Computational Linguistics, pp. 205–215. https://aclanthology.org/2021.wmt-1.19.

Part I
European Language Equality: Status Quo in 2022

Chapter 2
State-of-the-Art in Language Technology and Language-centric Artificial Intelligence

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Jon Ander Campos, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernáez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Ander Salaberria, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa

Abstract This chapter landscapes the field of Language Technology (LT) and language-centric AI by assembling a comprehensive state-of-the-art of basic and applied research in the area. It sketches all recent advances in AI, including the most recent deep learning neural technologies. The chapter brings to light not only where language-centric AI as a whole stands, but also where the required resources should be allocated to place European LT at the forefront of the AI revolution. We identify key research areas and gaps that need to be addressed to ensure LT can overcome the current inequalities.[1]

1 Introduction

Interest in the computational processing of human languages led to the establishment of specialised fields known as Computational Linguistics (CL), Natural Language Processing (NLP) and Language Technology (LT). CL is more informed by linguis-

Rodrigo Agerri · Eneko Agirre · Itziar Aldabe · Nora Aranberri · Jose Maria Arriola · Aitziber Atutxa · Gorka Azkune · Jon Ander Campos · Arantza Casillas · Ainara Estarrona · Aritz Farwell · Iakes Goenaga · Josu Goikoetxea · Koldo Gojenola · Inma Hernaez · Mikel Iruskieta · Gorka Labaka · Oier Lopez de Lacalle · Eva Navas · Maite Oronoz · Arantxa Otegi · Alicia Pérez · Olatz Perez de Viñaspre · German Rigau · Ander Salaberria · Jon Sanchez · Ibon Saratxaga · Aitor Soroa
University of the Basque Country, Spain,
rodrigo.agerri@ehu.eus, e.agirre@ehu.eus, itziar.aldabe@ehu.eus, nora.aranberri@ehu.eus, josemaria.arriola@ehu.eus, aitziber.atutxa@ehu.eus, gorka.azkune@ehu.eus, jonander.campos@ehu.eus, arantza.casillas@ehu.eus, ainara.estarrona@ehu.eus, aritz.farwell@ehu.eus, iakes.goenaga@ehu.eus, josu.goikoetxea@ehu.eus, koldo.gojenola@ehu.eus, inma.hernaez@ehu.eus, mikel.iruskieta@ehu.eus, gorka.labaka@ehu.eus, oier.lopezdelacalle@ehu.eus, eva.navas@ehu.eus, maite.oronoz@ehu.eus, arantza.otegi@ehu.eus, alicia.perez@ehu.eus, olatz.perezdevinaspre@ehu.eus, german.rigau@ehu.eus, ander.salaberria@ehu.eus, jon.sanchez@ehu.eus, ibon.saratxaga@ehu.eus, a.soroa@ehu.eus

[1] This chapter is an abridged version of Agerri et al. (2021).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_2

tics and NLP by computer science, LT is a more neutral term. In practice, these communities work closely together, sharing the same publishing venues and conferences, combining methods and approaches inspired by both, and together making up language-centric AI. In this chapter we treat them interchangeably.

Over the years, LT has developed different methods to make the information contained in written and spoken language explicit or to generate or synthesise written or spoken language. Despite the inherent difficulties in many of the tasks performed, current LT support allows many advanced applications which were unthinkable only a few years ago. LT is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, text editors, text predictors, automatic translation systems, automatic subtitling, automatic summarisation and inclusive technology. Its recent accelerated development promises even more encouraging and exciting results in the near future.

This state-of-the-art in LT and language-centric AI begins with a brief historical account in Section 2 on the development of the field from its inception through the current deep learning era. The following three sections are neural language models (Section 3), research areas (Section 4) and LT beyond language (Section 5). They offer a survey that maps today's LT and language-centric AI landscape. Finally, a discussion and various conclusions are outlined in Section 6.

2 Language Technology: Historical Overview

2.1 A Brief History

The 1950s mark the beginning of Language Technology as a discipline. In the middle of the 20th century, Alan Turing proposed his famous test, which defines a criterion to determine whether a machine can be considered intelligent (Turing 1950). A few years later, Noam Chomsky laid the foundations to formalise, specify and automate linguistic rules with his generative grammar (Chomsky 1957). For a long period of time, the horizon defined by Turing and the instrument provided by Chomsky influenced the majority of NLP research.

The early years of LT were closely linked to Machine Translation (MT), a well-defined task, and also relevant from a political and strategic point of view. In the 1950s it was believed that a high-quality automatic translator would be available soon. By the mid-1960s, however, the Automatic Language Processing Advisory Committee (ALPAC) report revealed the true difficulty of the task and NLP in general. The following two decades were heavily influenced by Chomsky's ideas, with increasingly complex systems of handwritten rules. At the end of the 1980s, a revolution began which irreversibly changed the field of NLP. This change was driven mainly by four factors: 1. the clear definition of individual NLP tasks and corresponding rigorous evaluation methods; 2. the availability of relatively large amounts of data; 3. machines that could process these large amounts of data; and 4. the gradual

introduction of more robust approaches based on statistical methods and machine learning (ML), that would pave the way for subsequent major developments.

Since the 1990s, NLP has moved forward with new resources, tools and applications. An effort was made to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc., from which WordNet (Miller 1992) is one of the main results. Data-driven systems displaced rule-based systems, leading to the almost ubiquitous presence of ML components in NLP systems. In the 2010s we observed a radical technological shift in NLP. Collobert et al. (2011) presented a multilayer neural network (NN) adjusted by backpropagation that solved various sequential labeling problems. Word embeddings gained particular relevance due to their role in the incorporation of pre-trained external knowledge into neural architectures (Mikolov et al. 2013). Large volumes of unannotated texts, together with progress in self-supervised ML and the rise of high-performance hardware (Graphics Processing Units, GPU), enabled highly effective deep learning systems to be developed across a range of application areas. These and other breakthroughs helped launch today's Deep Learning Era.

2.2 The Deep Learning Era

Today, LT is moving away from a methodology in which a pipeline of multiple modules is utilised to implement solutions to architectures based on complex neural networks trained on vast amounts of data. Four research trends are converging: 1. mature deep neural network technology, 2. large amounts of multilingual data, 3. increased High Performance Computing (HPC) power, and 4. the application of simple but effective self-learning approaches (Devlin et al. 2019; Yinhan Liu et al. 2020). These advancements have produced a new state-of-the-art through systems that are claimed to obtain human-level performance in laboratory benchmarks on difficult language understanding tasks. As a result, various large IT enterprises have started deploying large language models (LLMs) in production.

Despite their notable capabilities, however, LLMs have certain drawbacks that will require interdisciplinary collaboration and research to resolve. First, we have no clear understanding of how they work, when they fail, or what emergent properties they present. Indeed, some authors call these models "foundation models" to underscore their critically central yet incomplete character (Bommasani et al. 2021). Second, the systems are very sensitive to phrasing and typos, are not robust enough, and perform inconsistently (Ribeiro et al. 2019). Third, these models are expensive to train, which means that only a limited number of organisations can currently afford their development (Ahmed and Wahed 2020). Fourth, large NLP datasets used to train these models have been 'filtered' to remove targeted minorities (Dodge et al. 2021). In addition, LLMs can sometimes produce unpredictable and factually inaccurate text or even recreate private information. Finally, computing large pre-trained models comes with a substantial carbon footprint (Strubell et al. 2019).

The implications of LLMs may extend to questions of language-centred AI sovereignty. Given the impact of LT in everyone's daily lives, many LT practitioners are particularly concerned by the need for digital language equality (DLE) across all aspects of our societies. As expected, only a small number of the world's more than 6,000 languages are represented in the rapidly evolving LT field. This disproportionate representation is further exacerbated by systematic inequalities in LT across the world's languages (Joshi et al. 2020). Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models, prompt learning and self-supervised systems opens a path to leverage LT for less-developed languages. However, the development of these new LT systems will require resources along with carefully designed evaluation benchmarks and annotated datasets for every language and domain of application.

Forecasting the future of LT and language-centric AI is a challenge. It is, nevertheless, safe to assume that many more advances will be achieved utilising pre-trained language models and that they will substantially impact society. Future users are likely to discover novel applications and wield them positively or negatively. In either case, as Bender et al. (2021) argue, it is important to understand the current limitations of LLMs, which they refer to as "stochastic parrots". Focusing on state-of-the-art results exclusively with the help of leaderboards, without encouraging deeper understanding of the mechanisms by which they are attained, can give rise to misleading conclusions. These, in turn, may direct resources away from efforts that would facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication.

3 Neural Language Models

LT is undergoing a paradigm shift with the rise of neural language models that are trained on broad data at scale and are adaptable to a wide range of monolingual and multilingual downstream tasks (Devlin et al. 2019; Yinhan Liu et al. 2020). These models are based on standard self-supervised deep learning and transfer learning, but their scale results in emergent and surprising capabilities. One of the advantages is their ability to alleviate the feature engineering problem by using low-dimensional and dense vectors (distributed representation) to implicitly represent the language examples (Collobert et al. 2011). In self-supervised learning, the language model is derived automatically from large volumes of unannotated language data (text or voice). There has been considerable progress in self-supervised learning since word embeddings associated word vectors with context-independent vectors.

With transfer learning, the learning process starts from patterns that have been learned when solving a different problem, i. e., leveraging previous learning to avoid starting from scratch. Within deep learning, pre-training is the dominant approach to transfer learning: the objective is to pre-train a deep Transformer model on large amounts of data and then reuse this pre-trained language model by fine-tuning it on small amounts of (usually annotated) task-specific data. Recent work has shown that

pre-trained language models can robustly perform tasks in a few-shot or even zero-shot fashion when given an adequate task description in its natural language prompt (Brown et al. 2020). Unlike traditional supervised learning, which trains a model to take in an input and predict an output, prompt-based learning or in-context learning is based on exploiting pre-trained language models to solve a task using text directly. This framework is very promising since some NLP tasks can be solved in a fully unsupervised fashion by providing a pre-trained language model with task descriptions in natural language (Raffel et al. 2020). Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al. 2021).

Multilingual Large Language Models (MLLMs) such as mBERT (Devlin et al. 2019), XLM-R (Conneau et al. 2020), mBART (Yinhan Liu et al. 2020), mT5 (Xue et al. 2021), etc. have emerged as viable options for bringing the power of pre-training to a large number of languages. For example, mBERT is pre-trained on Wikipedia corpora in 104 languages. mBERT can generalise cross-lingual knowledge in zero-shot scenarios. This indicates that even with the same structure of BERT, using multilingual data can enable the model to learn cross-lingual representations. The surprisingly good performance of MLLMs in cross-lingual transfer as well as bilingual tasks suggests that these language models are learning universal patterns (Doddapaneni et al. 2021). Thus, one of the main motivations of training MLLMs is to enable transfer from high-resource languages to low-resource languages.

New types of processing pipelines and toolkits have arisen in recent years due to the fast-growing collection of efficient tools. Libraries that are built with NN components are increasingly common, including pre-trained models that perform multilingual NLP tasks. Neural language models are adaptable to a wide spectrum of monolingual and multilingual tasks. These models are currently often considered black boxes, in that their inner mechanisms are not clearly understood. Nonetheless, Transformer architectures may present an opportunity to offer advances to the broader LT community if certain obstacles can be successfully overcome. One is the question of the resources needed to design the best-performing neural language models, currently done almost exclusively at large IT companies. Another is the problem of stereotypes, prejudices and personal information within the corpora used to train the models. The predominance of English as the default language in NLP can be successfully addressed if there is sufficient will and coordination. The continued consolidation of large infrastructures will help determine how this is accomplished in the near future. Their successful implementation would mark a crucial first step towards the development, proliferation and management of language resources for *all* European languages. This capability would, in turn, enable Europe's languages to enjoy full and equal access to digital language technology.

4 Research Areas

Section 4 introduces some of the more prominent research areas in the field: Language Resources (Section 4.1), Text Analysis (Section 4.2), Speech Processing (Section 4.3), Machine Translation (Speech 4.4), Information Extraction and Retrieval (Section 4.5), NLG and Summarisation (Section 4.6) as well as HCI (Section 4.7).

4.1 Language Resources

The term Language Resource (LR) refers to a set of speech or written data and descriptions in machine readable form. These are utilised for building, improving or evaluating text- and speech-based algorithms or systems. They also serve as resources for the software localisation and language services industries, language studies, digital publishing, international transactions, subject-area specialists and end users. Although no widely standardised typology of LRs exists, they are usually classified as: 1. Data (i. e., corpora and lexical/conceptual resources); 2. Tools/Services (i. e., linguistic annotations; tools for creating annotations; search and retrieval applications; applications for automatic annotation) and 3. Metadata and vocabularies (i. e., vocabularies or repositories of linguistic terminology; language metadata). In this section we will focus on the first two categories.

A main objective of the LR community is the development of infrastructures and platforms for presenting and disseminating LRs. There are numerous repositories in which resources for each language are documented. Among the major European catalogues are European Language Grid (ELG, Rehm 2023),[2] ELRC-SHARE, [3] European Language Resources Association (ELRA), [4] Common Language Resources and Technology Infrastructure (CLARIN)[5] and META-SHARE.[6] The Linguistic Data Consortium,[7] which operates outside of Europe, should also be highlighted.

In addition, there are several relevant multilingual public domain initiatives. Among these are the Common Voice Project,[8] designed to encourage the development of ASR systems; the M-AILABS Speech Dataset,[9] for text-to-speech synthesis; the Ryerson Audio-Visual Database of Emotional Speech and Song,[10] for research

[2] https://www.european-language-grid.eu
[3] http://www.elrc-share.eu
[4] http://catalogue.elra.info
[5] https://www.clarin.eu/content/language-resources
[6] http://www.meta-share.org
[7] https://catalog.ldc.upenn.edu
[8] https://commonvoice.mozilla.org
[9] https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/
[10] https://zenodo.org/record/1188976

on emotional multimedia content; and LibriVox,[11] an audiobook repository that can be used in different research fields and applications.

A cursory glance at these repositories not only gives us an idea of the amount of resources available for Europe's languages, but also reveals the clear inequality between official and minority languages. Moreover, although the four European languages with the most resources are English, French, German and Spanish, English is far ahead of the rest, with more than twice as many resources as the next language (see Figure 1, p. 50). At the same time, the languages without official status trail significantly behind in terms of LR development, demonstrating the critical impact that official status has on the extent of available resources.

4.2 Text Analysis

Text Analysis (TA) aims to extract relevant information from large amounts of unstructured text in order to enable data-driven approaches to manage textual content. In other words, its purpose is to create structured data out of unstructured text content by identifying entities, facts and relationships that are buried in the textual data. TA employs a variety of methodologies to process text. It is crucial for establishing "who did what, where and when," a technology that has proven to be key for applications such as Information Extraction, Question Answering, Summarisation and nearly every linguistic processing task involving semantic interpretation, including Opinion Mining and Aspect-based Sentiment Analysis (ABSA).

The best results for TA tasks are generally obtained by means of supervised, corpus-based approaches. In most cases, manually annotating text for every single specific need is extremely time-consuming and not affordable in terms of human resources and economic costs. To make the problem more manageable, TA is addressed in several tasks that are typically performed in order to preprocess the text to extract relevant information. The most common tasks currently available in state-of-the-art NLP tools and pipelines include Part-of-Speech (POS) tagging, Lemmatisation, Word Sense Disambiguation (WSD), Named Entity Recognition (NER), Named Entity Disambiguation (NED) or Entity Linking (EL), Parsing, Coreference Resolution, Semantic Role Labelling (SRL), Temporal Processing, ABSA and, more recently, Open Information Extraction (OIE).

Today, all these tasks are addressed in an end-to-end manner, i. e., even for a traditionally complex task such as Coreference Resolution (Pradhan et al. 2012), current state-of-the-art systems are based on an approach in which no extra linguistic annotations are required. These systems typically employ LLMs. Similarly, most state-of-the-art TA toolkits, including AllenNLP and Trankit, among others (Gardner et al. 2018; M. V. Nguyen et al. 2021), use a highly multilingual end-to-end approach. Avoiding intermediate tasks has helped to mitigate the common cascading errors problem that was pervasive in more traditional TA pipelines. As a consequence, the

[11] https://librivox.org

appearance of end-to-end systems has helped bring about a significant jump in performance across every TA task.

4.3 Speech Processing

Speech processing aims at allowing humans to communicate with digital devices through voice. This entails developing machines that understand and generate not only oral messages, but also all the additional information that we can extract from the voice, like who is speaking, their age, their personality, their mood, etc. Some of the main areas in speech technology are text-to-speech synthesis (TTS), automatic speech recognition (ASR) and speaker recognition (SR).

TTS attempts to produce the oral signal that corresponds to an input text with an intelligibility, naturalness and quality similar to a natural speech signal. Statistical parametric speech synthesis techniques generate speech by means of statistical models trained to learn the relation between linguistic labels derived from text and acoustic parameters extracted from speech by means of a vocoder. HMM (Hidden Markov Models) and more recently DNN (Deep Neural Networks) have been used as statistical frameworks. Various architectures have been tested, such as feed-forward networks (Qian et al. 2014), recurrent networks (Y. Fan et al. 2014) and WaveNet (Oord et al. 2016). Among the criteria used for training, the most common is minimum generation error (Z. Wu and King 2016), although recently new methods based on Generative Adversarial Networks (GAN, Saito et al. 2017) have been proposed with excellent results in terms of naturalness of the produced voice.

ASR, producing a transcription from a speech signal, has been long sought after. The intrinsic difficulty of the task has required a step-by-step effort, with increasingly ambitious objectives. Only in the last two decades has this technology jumped from the laboratory to production. The first commercial systems were based on statistical models, i. e., HMMs (Juang and Rabiner 2005; Gales and Young 2008). While this technology was the standard during the first decade of the century, in the 2010s, the increase in computing power and the ever-growing availability of training data allowed for the introduction of DNN techniques for ASR.

More recently, end-to-end or fully differentiable architectures have appeared that aim to simplify a training process that is capable of exploiting the available data. In these systems, a DNN maps the acoustic signal in the input directly to the textual output. Thus, the neural network models the acoustic information, the time evolution and some linguistic information, learning everything jointly. New architectures, in the form of Transformers (Gulati et al. 2020; Xie Chen et al. 2021) and teacher-student schemes (Z. Zhang et al. 2020; Jing Liu et al. 2021), have been applied to ASR with great success. Recently, Whisper, a Transformer sequence-to-sequence model trained on very large amounts of data that can perform several tasks such as multilingual ASR, translation and language identification, has been developed by OpenAI (Radford et al. 2022) showing the potential of weakly supervised systems.

A similar evolution has taken place in the area of SR. Part of the widespread emergence of biometric identification techniques, exemplified by the now commonplace ability to unlock a smartphone with a fingerprint or an iris, speaker recognition involves the automatic identification of people based on their voice. Nowadays, the classical systems have been outperformed by end-to-end neural network based systems, which are being improved using widespread databases (Nagrani et al. 2017) and enforcing research (Nagrani et al. 2020), obtaining better recognition rates by means of new network architectures and techniques (Safari et al. 2020; H. Zhang et al. 2020; R. Wang et al. 2022).

4.4 Machine Translation

Machine Translation (MT) is the automatic translation from one natural language into another. Since its first implementation (Weaver 1955) it has remained a key application in LT/NLP. While a number of approaches and architectures have been proposed and tested over the years, Neural MT (NMT) has become the most popular paradigm for MT development both within the research community (Vaswani et al. 2018; Yinhan Liu et al. 2020; Zhu et al. 2020; Sun et al. 2022) and for large-scale production systems (Y. Wu et al. 2016). This is due to the good results achieved by NMT systems, which attain state-of-the-art results for many language pairs (Akhbardeh et al. 2021; Adelani et al. 2022; Min 2023). NMT systems use distributed representations of the languages involved, which enables end-to-end training of systems. If we compare them with classical statistical MT models (Koehn et al. 2003), we see that they do not require word aligners, translation rule extractors, and other feature extractors; the *embed – encode – attend – decode* paradigm is the most common NMT approach (Vaswani et al. 2017; You et al. 2020; Dione et al. 2022).

Thanks to current advances in NMT it is common to find systems that can easily incorporate multiple languages simultaneously. We refer to these types of systems as Multilingual NMT (MNMT) systems. The principal goal of an MNMT system is to translate between as many languages as possible by optimising the linguistic resources available. MNMT models (Aharoni et al. 2019; B. Zhang et al. 2020; Emezue and Dossou 2022; Siddhant et al. 2022) are interesting for several reasons. On the one hand, they can address translations among all the languages involved within a single model, which significantly reduces training time and facilitates deployment of production systems. On the other hand, by reducing operational costs, multilingual models achieve better results than bilingual models for low- and zero-resource language pairs: training is performed jointly and this generates a positive transfer of knowledge from high(er)-resource languages (Aharoni et al. 2019; Arivazhagan et al. 2019). This phenomenon is known as translation knowledge transfer or transfer learning (Zoph et al. 2016; T. Q. Nguyen and Chiang 2017; Hujon et al. 2023).

For instance, A. Fan et al. (2021) have created several MNMT models by building a large-scale many-to-many dataset for 100 languages. They significantly reduce the complexity of this task, employing automatic building of parallel corpora (Artetxe

and Schwenk 2019; Schwenk et al. 2021) with a novel data mining strategy that exploits language similarity in order to avoid mining all directions. The method allows for direct translation between 100 languages without using English as a pivot and it performs as well as bilingual models on many competitive benchmarks. Additionally, they take advantage of backtranslation to improve the quality of their model on zero-shot and low-resource language pairs.

4.5 Information Extraction and Information Retrieval

Deep learning has had a tremendous impact on Information Retrieval (IR) and Information Extraction (IE). The goal of IR is to meet the information needs of users by providing them with documents or text snippets that contain answers to their queries. IR is a mature technology that enabled the development of search engines. The area has been dominated by classic methods based on vector space models that use manually created sparse representations such as TF-IDF or BM25 (Robertson and Zaragoza 2009), but recent approaches that depend on dense vectors and deep learning have shown promising results (Karpukhin et al. 2020; Izacard and Grave 2021). Dense representations are often combined with Question Answering (QA) to develop systems that are able to directly answer specific questions posed by users, either by pointing at text snippets that answer the questions (Karpukhin et al. 2020; Izacard and Grave 2021) or by generating the appropriate answers themselves (P. Lewis et al. 2021).

IE aims to extract structured information from text. Typically, IE systems recognise the main events described in a text, as well as the entities that participate in those events. Modern techniques mostly focus on two challenges: learning textual semantic representations for events in event extraction (both at sentence and document level) and acquiring or augmenting labeled instances for model training (K. Liu et al. 2020). Regarding the former, early approaches relied on manually coded lexical, syntactic and kernel-based features (Ahn 2006). With the development of deep learning, however, researchers have employed neural networks, including CNNs (Y. Chen et al. 2015), RNNs (T. H. Nguyen and Grishman 2016) and Transformers (Yang et al. 2019). Data augmentation has been typically performed by using methods such as distant supervision or employing data from other languages to improve IE on the target language, which is especially useful when the target language is under-resourced. Deep learning techniques utilised in NMT (Jian Liu et al. 2018) and pre-trained multilingual LLMs (Jian Liu et al. 2019) have also helped in this task.

Another important task within IE is Relation Extraction (RE), whose goal is to predict the semantic relationship between two entities, if any. The best results on RE are obtained by fine-tuning LLMs, which are supplied with a classification head. One of the most pressing problems in RE is the scarcity of manually annotated examples in real-world applications, particularly when there is a domain and language shift. In recent years, new methods have emerged that only require a few-shot or zero-shot examples. Prompt-based learning, e. g., uses task and label verbalisations that

can be designed manually or learned automatically (Schick and Schütze 2021) as an alternative to fine-tuning. In these methods, the inputs are augmented with prompts and the LM objective is used in learning and inference. This paradigm shift has allowed IE tasks to be framed as a QA problem (Sulem et al. 2022) or as a constrained text generation problem (S. Li et al. 2021) using prompts, questions or templates.

4.6 Natural Language Generation and Summarisation

Natural Language Generation (NLG) has become one of the most important and challenging tasks in NLP (Gehrmann et al. 2021). NLG automatically generates understandable texts, typically using a non-linguistic or textual representation of information as input (Reiter and Dale 1997; Gatt and Krahmer 2018; Junyi Li et al. 2021a). Applications that generate new texts from existing text include MT from one language to another (see Section 4.4), fusion and summarisation, simplification, text correction, paraphrase generation, question generation, etc. With the recent resurgence of deep learning, new ways to solve text generation tasks based on different neural architectures have arisen (Junyi Li et al. 2021b). One advantage of these neural models is that they enable end-to-end learning of semantic mappings from input to output in text generation. Existing datasets for most supervised text generation tasks are small (except MT). Therefore, researchers have proposed various methods to solve text generation tasks based on LLMs. Transformer models such as T5 (Raffel et al. 2020) and BART (M. Lewis et al. 2020) or a single Transformer decoder block such as GPT (Brown et al. 2020) are currently standard architectures for generating high quality text.

Due to the rapid growth of information generated daily online (Gambhir and Gupta 2017), there is a growing need for automatic summarisation techniques that produce short texts from one or more sources efficiently and precisely. Several extractive approaches have been developed for automatic summary generation that implement a number of machine learning and optimisation techniques (J. Xu and Durrett 2019). Abstractive methods are more complex as they require NLU capabilities. Abstractive summarisation produces an abstract with words and phrases that are based on concepts that occur in the source document (Du et al. 2021). Both approaches can now be modeled using Transformers (Yang Liu and Lapata 2019).

4.7 Human-Computer Interaction

The demand for technologies that enable users to interact with machines at any time utilising text and speech has grown, motivating the use of dialogue systems. Such systems allow the user to converse with computers using natural language and include Siri, Google Assistant, Amazon Alexa, and ChatGPT, among others. Dialogue sys-

tems can be divided into three groups: task-oriented systems, conversational agents (also known as chatbots) and interactive QA systems.

The distinguishing features of task-oriented dialogue systems are that they are designed to perform a concrete task in a specific domain and that their dialogue flow is defined and structured beforehand. For example, such systems are used to book a table at a restaurant, call someone or check the weather forecast. The classical implementation of this type of system follows a pipeline architecture based on three modules: the NLU module, the dialogue manager and the NLG module. While classical dialogue systems trained and evaluated these modules separately, more recent systems rely on end-to-end trainable architectures based on neural networks (Bordes et al. 2017; Hosseini-Asl et al. 2020).

Conversational agents enable engaging open-domain conversations, often by emulating the personality of a human (S. Zhang et al. 2018). The Alexa prize,[12] for instance, focused on building agents that could hold a human in conversation as long as possible. These kinds of agents are typically trained in conversations mined from social media using end-to-end neural architectures (Roller et al. 2021).

Interactive QA systems try to respond to user questions by extracting answers from either documents (Rajpurkar et al. 2018) or knowledge bases (T. Yu et al. 2018). In order to be able to have meaningful interactions, interactive QA systems have a simple dialogue management procedure taking previous questions and answers into account (Choi et al. 2018). The core technology is commonly based on LLMs (Qiu et al. 2020) where some mechanism is included to add context representation (Vakulenko et al. 2021).

5 Language Technology beyond Language

Knowledge about our surrounding world is required to properly understand natural language utterances (Bender and Koller 2020). That knowledge is known as world knowledge and many authors argue that it is a key ingredient to achieve human-level NLU (Storks et al. 2019). One of the ways to acquire this knowledge is to explore the visual world together with the textual world (Elu et al. 2021). CNNs have been the standard architecture for generating representations for images (LeCun and Bengio 1995) during the last decade. Recently, self-attention-based Transformer models (Vaswani et al. 2017) have emerged as an alternative architecture, leading to exciting progress on a number of vision tasks (Khan et al. 2021). Compared to previous approaches, Transformers allow multiple modalities to be processed (e. g., images, videos, text and speech) using similar processing blocks and demonstrate excellent scalability properties. Encoder-decoder models in particular have been gaining traction recently due to their versatility on solving different generative tasks (Junnan Li et al. 2022; Xi Chen et al. 2022).

[12] https://developer.amazon.com/alexaprize

Regarding downstream tasks, caption generation is a typical visio-linguistic task, where a textual description of an image must be generated. The first approaches to solve this problem combined CNNs with RNNs in an encoder-decoder architecture (Vinyals et al. 2015). Further improvements were achieved when attention was included (K. Xu et al. 2015) and some researchers have proposed utilising object-based attention instead of spatial attention (Anderson et al. 2018). Although it is not currently clear which attention mechanism is better, the quality of the text generated by these models is high as measured by metrics such as BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005)

Visual generation, in contrast to caption generation, requires an image to be generated from a textual description. One of this task's most significant challenges is to develop automatic metrics to evaluate the quality of the generated images and their coherence with the input text. The first effective approaches were based on Generative Adversarial Networks (Goodfellow et al. 2014) and Variational Autoencoders (Kingma and Welling 2013). Cho et al. (2020) demonstrate that multimodal Transformers can also generate impressive images from textual input. Nevertheless, novel advancements in diffusion models (Sohl-Dickstein et al. 2015; Ho et al. 2020) have defined the current state-of-the-art in image generation (Ramesh et al. 2022). These models learn to iteratively reconstruct noisy images and, recently, their size and computational cost has been reduced as diffusion can be now applied in a reduced latent space instead of an image's pixel space (Rombach et al. 2022).

Another typical task is Visual Question Answering (VQA), where given an image and a question about the contents of that image, the right textual answer must be found. There are many VQA datasets in the literature (Antol et al. 2015; Johnson et al. 2017). Some demand leveraging external knowledge to infer an answer and, thus, they are known as knowledge-based VQA tasks (P. Wang et al. 2017a,b; Marino et al. 2019). These VQA tasks demand skills to understand the content of an image and how it is referred to in the textual question, as well as reasoning capabilities to infer the correct answer. Multimodal Transformers, such as OFA (P. Wang et al. 2022) and PaLI (Xi Chen et al. 2022), define the state-of-the-art in several of these tasks.

Visual Referring Expressions are one of the multimodal tasks that may be considered an extension of a text-only NLP task, i. e., referring expressions (Krahmer and Deemter 2012) in NLG systems. Its objective is to ground a natural language expression to objects in a visual input. There are several approaches to solve this task (Golland et al. 2010; Kazemzadeh et al. 2014). The most recent ones use attention mechanisms to merge both modalities (L. Yu et al. 2018) or are based on multimodal Transformers (Ding et al. 2022).

A natural extension of textual entailment, Visual Entailment is an inference task for predicting whether an image semantically entails a text. Vu et al. (2018) initially proposed a visually-grounded version of the textual entailment task, where an image is augmented to include a textual premise and hypothesis. However, Xie et al. (2019) propose visual entailment, where the premise is an image and the hypothesis is textual. As an alternative to entailment, there are other grounding tasks that classify whether an image and its caption match (Suhr et al. 2018; F. Liu et al. 2022) or

tasks that measure the similarity between sentences with visual cues, such as vSTS (Lopez de Lacalle et al. 2020).

Multimodal MT (MMT) seeks to translate natural language sentences that describe visual content in a source language into a target language by taking the visual content as an additional input to the source language sentences (Elliott et al. 2017; Barrault et al. 2018). Different approaches have been proposed to handle MMT, although attention models that associate textual and visual elements with multimodal attention mechanisms are the most common (Huang et al. 2016; Calixto et al. 2017).

6 Conclusions

Language tools and resources have increased and improved since the end of the last century, a process further catalysed by the advent of deep learning and LLMs over the past decade. Indeed, we find ourselves today in the midst of a significant paradigm shift in LT and language-centric AI. This revolution has brought noteworthy advances to the field along with the promise of substantial breakthroughs in the coming years. However, this transformative technology poses problems, from a research advancement, environmental, and ethical perspective. Furthermore, it has also laid bare the acute digital inequality that exists between languages. In fact, as emphasised in this chapter, many sophisticated NLP systems are unintentionally exacerbating this imbalance due to their reliance on vast quantities of data derived mostly from English-language sources. Other languages lag far behind English in terms of digital presence and even the latter would benefit from greater support. Moreover, the striking asymmetry between official and non-official European languages with respect to available digital resources is concerning. The unfortunate truth is that DLE in Europe is failing to keep pace with the newfound and rapidly evolving changes in LT. One need look no further than what is happening today across the diverse topography of state-of-the-art LT and language-centric AI for confirmation of the current linguistic unevenness. The paradox at the heart of LT's recent advances is evident in almost every LT discipline. Our ability to reproduce ever better synthetic voices has improved sharply for well-resourced languages, but dependence on large volumes of high-quality recordings effectively undermines attempts to do the same for low-resource languages. Multilingual NMT systems return demonstrably improved results for low- and zero-resource language pairs, but insufficient model capacity continues to haunt transfer learning because large multilingual datasets are required, forcing researchers to rely on English as the best resourced language.

Nonetheless, we believe this time of technological transition represents an opportunity to achieve full DLE in Europe. There are ample reasons for optimism. Recent research in the field has considered the implementation of cross-lingual transfer learning and multilingual language models for low-resource languages, an example of how the state-of-the-art in LT could benefit from better digital support for low-resource languages.

Forecasting the future of LT and language-centric AI is a challenge. Just a few years ago, nobody would have predicted the recent breakthroughs that have resulted in systems able to deal with unseen tasks or maintaining natural conversations. It is, however, safe to predict that even more advances will be achieved in all LT research areas and domains in the near future. Despite claims of human parity in many LT tasks, *Natural Language Understanding is still an open research problem* far from being solved since all current approaches have severe limitations. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual LLMs and self-supervised systems opens up the way to leverage LT for less developed languages. However, the development of these new LT systems would not be possible without sufficient resources (experts, data, HPC facilities, etc.) as well as the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application. Focusing on state-of-the-art results exclusively with the help of leaderboards without encouraging deeper understanding of the mechanisms by which they are achieved can generate misleading conclusions, and direct resources away from efforts that would facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication, to create transparent digital language equality in Europe in all aspects of society, from government to business to citizen.

References

Adelani, David, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek (2022). "Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 773–800. https://aclanthology.org/2022.wmt-1.72.

Agerri, Rodrigo, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa (2021). *Deliverable D1.2 Report on the State of the Art in Language Technology and Language-centric AI*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-state-of-the-art.pdf.

Aharoni, Roee, Melvin Johnson, and Orhan Firat (2019). "Massively Multilingual Neural Machine Translation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3874–3884. DOI: 10.18653/v1/N19-1388. https://aclanthology.org/N19-1388.

Ahmed, Nur and Muntasir Wahed (2020). "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research". In: *CoRR* abs/2010.15581. https://arxiv.org/abs/2010.15581.

Ahn, David (2006). "The stages of event extraction". In: *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Sydney, Australia: Association for Computational Linguistics, pp. 1–8. https://aclanthology.org/W06-0901.

Akhbardeh, Farhad, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri (2021). "Findings of the 2021 Conference on Machine Translation (WMT21)". In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 1–88. https://aclanthology.org/2021.wmt-1.1.

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang (2018). "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, pp. 6077–6086. DOI: 10.1109/CVPR.2018.00636. http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Anderson%5C_Bottom-Up%5C_and%5C_Top-Down%5C_CVPR%5C_2018%5C_paper.html.

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh (2015). "VQA: Visual Question Answering". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, pp. 2425–2433. DOI: 10.1109/ICCV.2015.279. https://doi.org/10.1109/ICCV.2015.279.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey (2019). "The missing ingredient in zero-shot neural machine translation". In: *arXiv preprint arXiv:1903.07091*. https://arxiv.org/abs/1903.07091.

Artetxe, Mikel and Holger Schwenk (2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *Transactions of the Association for Computational Linguistics* 7, pp. 597–610. DOI: 10.1162/tacl_a_00288. https://aclanthology.org/Q19-1038.

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. https://aclanthology.org/W05-0909.

Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank (2018). "Findings of the Third Shared Task on Multimodal Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 304–323. DOI: 10.18653/v1/W18-6402. https://aclanthology.org/W18-6402.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada, pp. 610–623.

Bender, Emily M. and Alexander Koller (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. https://aclanthology.org/2020.acl-main.463.

Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258 [cs.LG]. https://arxiv.org/abs/2108.07258.

Bordes, Antoine, Y-Lan Boureau, and Jason Weston (2017). "Learning End-to-End Goal-Oriented Dialog". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon,*

France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net. https://openreview.net/forum?id=S1Bb3D5gg.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Calixto, Iacer, Qun Liu, and Nick Campbell (2017). "Doubly-Attentive Decoder for Multi-modal Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1913–1924. DOI: 10.18653/v1/P17-1175. https://aclanthology.org/P17-1175.

Chen, Xi, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut (2022). "Pali: A jointly-scaled multilingual language-image model". In: *arXiv preprint arXiv:2209.06794*.

Chen, Xie, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li (2021). "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset". In: *ICASSP*. IEEE, pp. 5904–5908.

Chen, Yubo, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao (2015). "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 167–176. DOI: 10.3115/v1/P15-1017. https://aclanthology.org/P15-1017.

Cho, Jaemin, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi (2020). "X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8785–8805. DOI: 10.18653/v1/2020.emnlp-main.707. https://aclanthology.org/2020.emnlp-main.707.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer (2018). "QuAC: Question Answering in Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2174–2184. DOI: 10.18653/v1/D18-1241. https://aclanthology.org/D18-1241.

Chomsky, Noam (1957). *Syntactic structures.* The Hague: Mouton.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research* 12, pp. 2493–2537.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. https://aclanthology.org/2020.acl-main.747.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis,

Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Ding, Henghui, Chang Liu, Suchen Wang, and Xudong Jiang (2022). "VLT: Vision-Language Transformer and Query Generation for Referring Segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dione, Cheikh M Bamba, Alla Lo, Elhadji Mamadou Nguer, and Sileye Ba (2022). "Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof<-> French". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6654–6661.

Doddapaneni, Sumanth, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra (2021). "A primer on pretrained multilingual language models". In: *arXiv preprint arXiv:2107.00676*. https://arxiv.org/abs/2107.00676.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (2021). "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus". In: *arXiv preprint arXiv:2104.08758*.

Du, Zhengxiao, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang (2021). "All NLP Tasks Are Generation Tasks: A General Pretraining Framework". In: *arXiv preprint arXiv:2103.10360*. https://arxiv.org/abs/2103.10360.

Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia (2017). "Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 215–233. DOI: 10.18653/v1/W17-4718. https://aclanthology.org/W17-4718.

Elu, Aitzol, Gorka Azkune, Oier Lopez de Lacalle, Ignacio Arganda-Carreras, Aitor Soroa, and Eneko Agirre (2021). "Inferring spatial relations from textual descriptions of images". In: *Pattern Recognition* 113, p. 107847.

Emezue, Chris C and Bonaventure FP Dossou (2022). "MMTAfrica: Multilingual machine translation for African languages". In: *arXiv preprint arXiv:2204.04306*.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin (2021). "Beyond english-centric multilingual machine translation". In: *Journal of Machine Learning Research* 22.107, pp. 1–48.

Fan, Yuchen, Yao Qian, Feng-Long Xie, and Frank K Soong (2014). "TTS synthesis with bidirectional LSTM based recurrent neural networks". In: *Fifteenth annual conference of the international speech communication association*.

Gales, Mark and Steve Young (2008). *The application of hidden Markov models in speech recognition*. Now Publishers Inc.

Gambhir, Mahak and Vishal Gupta (2017). "Recent automatic text summarization techniques: a survey". In: *Artificial Intelligence Review* 47.1, pp. 1–66.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer (2018). "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/W18-2501. https://aclanthology.org/W18-2501.

Gatt, Albert and Emiel Krahmer (2018). "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation". In: *Journal of Artificial Intelligence Research* 61, pp. 65–170.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder,

Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou (2021). "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics". In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, pp. 96–120. DOI: 10.18653/v1/2021.gem-1.10. https://aclanthology.org/2021.gem-1.10.

Golland, Dave, Percy Liang, and Dan Klein (2010). "A Game-Theoretic Approach to Generating Spatial Descriptions". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 410–419. https://aclanthology.org/D10-1040.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, pp. 2672–2680. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang (2020). "Conformer: Convolution-augmented Transformer for Speech Recognition". In: *Interspeech*, pp. 5036–5040.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851.

Hosseini-Asl, Ehsan, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher (2020). "A simple language model for task-oriented dialogue". In: *Advances in Neural Information Processing Systems* 33, pp. 20179–20191.

Huang, Po-Yao, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer (2016). "Attention-based Multimodal Neural Machine Translation". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pp. 639–645. DOI: 10.18653/v1/W16-2360. https://aclanthology.org/W16-2360.

Hujon, Aiusha V, Thoudam Doren Singh, and Khwairakpam Amitab (2023). "Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings". In: *Procedia Computer Science* 218, pp. 1–8.

Izacard, Gautier and Edouard Grave (2021). "Distilling Knowledge from Reader to Retriever for Question Answering". In: *International Conference on Learning Representations*. https://openreview.net/forum?id=NTEz-6wysdb.

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick (2017). "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 1988–1997. DOI: 10.1109/CVPR.2017.215. https://doi.org/10.1109/CVPR.2017.215.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6282–6293. https://aclanthology.org/2020.acl-main.560.

Juang, Biing-Hwang and Lawrence R Rabiner (2005). "Automatic speech recognition–a brief history of the technology development". In: *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1, p. 67.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Pro-*

cessing (EMNLP). Online: Association for Computational Linguistics, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. https://aclanthology.org/2020.emnlp-main.550.

Kazemzadeh, Sahar, Vicente Ordonez, Mark Matten, and Tamara Berg (2014). "ReferItGame: Referring to Objects in Photographs of Natural Scenes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 787–798. DOI: 10.3115/v1/D14-1086. https://aclanthology.org/D14-1086.

Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah (2021). *Transformers in Vision: A Survey*. arXiv: 2101.01169 [cs.CV]. https://arxiv.org/abs/2101.01169.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Koehn, Philipp, Franz J. Och, and Daniel Marcu (2003). "Statistical Phrase-Based Translation". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133. https://aclanthology.org/N03-1017.

Krahmer, Emiel and Kees van Deemter (2012). "Computational Generation of Referring Expressions: A Survey". In: *Computational Linguistics* 38.1, pp. 173–218. DOI: 10.1162/COLI_a_00088. https://aclanthology.org/J12-1006.

LeCun, Yann and Yoshua Bengio (1995). "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. https://aclanthology.org/2020.acl-main.703.

Lewis, Patrick, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel (2021). *PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them*. arXiv: 2102.07033 [cs.CL].

Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi (2022). "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International Conference on Machine Learning*. PMLR, pp. 12888–12900.

Li, Junyi, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen (2021a). "TextBox: A Unified, Modularized, and Extensible Framework for Text Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 30–39. DOI: 10.18653/v1/2021.acl-demo.4. https://aclanthology.org/2021.acl-demo.4.

Li, Junyi, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen (2021b). "Pretrained Language Model for Text Generation: A Survey". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Zhi-Hua Zhou. Survey Track. International Joint Conferences on Artificial Intelligence Organization, pp. 4492–4499. DOI: 10.24963/ijcai.2021/612. https://doi.org/10.24963/ijcai.2021/612.

Li, Sha, Heng Ji, and Jiawei Han (2021). "Document-Level Event Argument Extraction by Conditional Generation". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 894–908. https://aclanthology.org/2021.naacl-main.69.

Liu, Fangyu, Guy Emerson, and Nigel Collier (2022). "Visual spatial reasoning". In: *arXiv preprint arXiv:2205.00363*.

Liu, Jian, Yubo Chen, Kang Liu, and Jun Zhao (2018). "Event Detection via Gated Multilingual Attention Mechanism". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial*

Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 4865–4872. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16371.

Liu, Jian, Yubo Chen, Kang Liu, and Jun Zhao (2019). "Neural Cross-Lingual Event Detection with Minimal Parallel Resources". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 738–748. DOI: 10.18653/v1/D19-1068. https://aclanthology.org/D19-1068.

Liu, Jing, Rupak Vignesh Swaminathan, Sree Hari Krishnan Parthasarathi, Chunchuan Lyu, Athanasios Mouchtaris, and Siegfried Kunzmann (2021). "Exploiting Large-scale Teacher-Student Training for On-device Acoustic Models". In: *Proc. International Conference on Text, Speech and Dialogue (TSD)*.

Liu, Kang, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao (2020). "Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges". In: *AI Open* 1, pp. 22–39. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2021.02.004. https://www.sciencedirect.com/science/article/pii/S266665102100005X.

Liu, Yang and Mirella Lapata (2019). "Text Summarization with Pretrained Encoders". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. DOI: 10.18653/v1/D19-1387. https://aclanthology.org/D19-1387.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer (2020). "Multilingual Denoising Pre-training for Neural Machine Translation". In: *Transactions of the Association for Computational Linguistics* 8, pp. 726–742. DOI: 10.1162/tacl_a_00343. https://aclanthology.org/2020.tacl-1.47.

Lopez de Lacalle, Oier, Ander Salaberria, Aitor Soroa, Gorka Azkune, and Eneko Agirre (2020). "Evaluating Multimodal Representations on Visual Semantic Textual Similarity". In: *Proceedings of the Twenty-third European Conference on Artificial Intelligence, ECAI 2020, June 8-12, 2020, Santiago Compostela, Spain*.

Marino, Kenneth, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi (2019). "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, pp. 3195–3204. DOI: 10.1109/CVPR.2019.00331. http://openaccess.thecvf.com/content%5C_CVPR%5C_2019/html/Marino%5C_OK-VQA%5C_A%5C_Visual%5C_Question%5C_Answering%5C_Benchmark%5C_Requiring%5C_External%5C_Knowledge%5C_CVPR%5C_2019%5C_paper.html.

Mikolov, Tomás, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, pp. 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Miller, George A. (1992). "WordNet: A Lexical Database for English". In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. https://aclanthology.org/H92-1116.

Min, Zeping (2023). "Attention Link: An Efficient Attention-Based Low Resource Machine Translation Architecture". In: *arXiv preprint arXiv:2302.00340*.

Nagrani, Arsha, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman (2020). "Voxsrc 2020: The second voxceleb speaker recognition challenge". In: *arXiv preprint arXiv:2012.06867*. https://arxiv.org/abs/2012.06867.

Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman (2017). "VoxCeleb: A Large-Scale Speaker Identification Dataset". In: *Interspeech*, pp. 2616–2620.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen (2021). "Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 80–90. DOI: 10.18653/v1/2021.eacl-demos.10. https://aclanthology.org/2021.eacl-demos.10.

Nguyen, Thien Huu and Ralph Grishman (2016). "Modeling Skip-Grams for Event Detection with Convolutional Neural Networks". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 886–891. DOI: 10.18653/v1/D16-1085. https://aclanthology.org/D16-1085.

Nguyen, Toan Q. and David Chiang (2017). "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 296–301. https://aclanthology.org/I17-2050.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016). "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499*. https://arxiv.org/abs/1609.03499.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. https://aclanthology.org/P02-1040.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Proceedings of CoNLL*, pp. 1–40. https://www.aclweb.org/anthology/W12-4501.

Qian, Yao, Yuchen Fan, Wenping Hu, and Frank K Soong (2014). "On the training aspects of deep neural network (DNN) for parametric TTS synthesis". In: *ICASSP*. IEEE, pp. 3829–3833.

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang (2020). "Pre-trained models for natural language processing: A survey". In: *Science China Technological Sciences* 63.10, pp. 1872–1897.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever (2022). "Robust speech recognition via large-scale weak supervision". In: *arXiv preprint arXiv:-2212.04356*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of Machine Learning Research* 21.1, pp. 5485–5551.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (2018). "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 784–789. https://aclanthology.org/P18-2124.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125*.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Reiter, Ehud and Robert Dale (1997). "Building applied natural language generation systems". In: *Natural Language Engineering* 3.1, pp. 57–87.

Ribeiro, Marco Tulio, Carlos Guestrin, and Sameer Singh (2019). "Are Red Roses Red? Evaluating Consistency of Question-Answering Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6174–6184. DOI: 10.18653/v1/P19-1621. https://aclanthology.org/P19-1621.

Robertson, Stephen and Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Found. Trends Inf. Retr.* 3.4, pp. 333–389. ISSN: 1554-0669. DOI: 10.1561/1500000019. https://doi.org/10.1561/1500000019.

Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston (2021). "Recipes for Building an Open-Domain Chatbot". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 300–325. DOI: 10.18653/v1/2021.eacl-main.24. https://aclanthology.org/2021.eacl-main.24.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer (2022). "High-resolution image synthesis with latent diffusion models". In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

Safari, Pooyan, Miquel India, and Javier Hernando (2020). "Self-attention encoding and pooling for speaker recognition". In: *Interspeech*, pp. 941–945.

Saito, Yuki, Shinnosuke Takamichi, and Hiroshi Saruwatari (2017). "Statistical parametric speech synthesis incorporating generative adversarial networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.1, pp. 84–96.

Schick, Timo and Hinrich Schütze (2021). "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 2339–2352. DOI: 10.18653/v1/2021.naacl-main.185. https://aclanthology.org/2021.naacl-main.185.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan (2021). "CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6490–6500. DOI: 10.18653/v1/2021.acl-long.507. https://aclanthology.org/2021.acl-long.507.

Siddhant, Aditya, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia (2022). "Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning". In: arXiv: 2201.03110 [cs.CL].

Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli (2015). "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR, pp. 2256–2265.

Storks, Shane, Qiaozi Gao, and Joyce Y Chai (2019). "Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches". In: *arXiv preprint arXiv:1904.01172*, pp. 1–60.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. https://aclanthology.org/P19-1355.

Suhr, Alane, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi (2018). "A corpus for reasoning about natural language grounded in photographs". In: *arXiv preprint arXiv:1811.00491*.

Sulem, Elior, Jamaal Hay, and Dan Roth (2022). "Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1075–1085. https://aclanthology.org/2022.naacl-main.79.

Sun, Zewei, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li (2022). "Rethinking document-level neural machine translation". In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3537–3548.

Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423. eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. https://doi.org/10.1093/mind/LIX.236.433.

Vakulenko, Svitlana, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha (2021). "Question rewriting for conversational question answering". In: *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 355–363.

Vaswani, Ashish, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit (2018). "Tensor2Tensor for Neural Machine Translation". In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Boston, MA: Association for Machine Translation in the Americas, pp. 193–199. https://aclanthology.org/W18-1819.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015). "Show and tell: A neural image caption generator". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 3156–3164. DOI: 10.1109/CVPR.2015.7298935. https://doi.org/10.1109/CVPR.2015.7298935.

Vu, Hoa Trong, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt (2018). "Grounded Textual Entailment". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2354–2368. https://aclanthology.org/C18-1199.

Wang, Peng, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel (2017a). "Explicit Knowledge-based Reasoning for Visual Question Answering". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. Ed. by Carles Sierra. ijcai.org, pp. 1290–1296. DOI: 10.24963/ijcai.2017/179. https://doi.org/10.24963/ijcai.2017/179.

Wang, Peng, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel (2017b). "Fvqa: Fact-based visual question answering". In: *IEEE transactions on pattern analysis and machine intelligence* 40.10, pp. 2413–2427.

Wang, Peng, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang (2022). "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework". In: *International Conference on Machine Learning*. PMLR, pp. 23318–23340.

Wang, Rui, Junyi Ao, Long Zhou, Shujie Liu, Zhihua Wei, Tom Ko, Qing Li, and Yu Zhang (2022). "Multi-view self-attention based transformer for speaker recognition". In: *ICASSP*. IEEE, pp. 6732–6736.

Weaver, Warren (1955). "Translation". In: *Machine translation of languages* 14.15-23, p. 10.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2021). "Finetuned Language Models Are Zero-Shot Learners". In: *arXiv preprint arXiv:2109.01652*. arXiv: 2109.01652 [cs.CL]. https://arxiv.org/abs/2109.01652.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144*. https://arxiv.org/abs/1609.08144.

Wu, Zhizheng and Simon King (2016). "Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.7, pp. 1255–1265.

Xie, Ning, Farley Lai, Derek Doran, and Asim Kadav (2019). "Visual entailment: A novel task for fine-grained image understanding". In: *arXiv preprint arXiv:1901.06706*. https://arxiv.org/abs/1901.06706.

Xu, Jiacheng and Greg Durrett (2019). "Neural Extractive Text Summarization with Syntactic Compression". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3292–3303. DOI: 10.18653/v1/D19-1324. https://aclanthology.org/D19-1324.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 2048–2057. http://proceedings.mlr.press/v37/xuc15.html.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41.

Yang, Sen, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li (2019). "Exploring Pre-trained Language Models for Event Extraction and Generation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5284–5294. DOI: 10.18653/v1/P19-1522. https://aclanthology.org/P19-1522.

You, Weiqiu, Simeng Sun, and Mohit Iyyer (2020). "Hard-Coded Gaussian Attention for Neural Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7689–7700. DOI: 10.18653/v1/2020.acl-main.687. https://aclanthology.org/2020.acl-main.687.

Yu, Licheng, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg (2018). "MAttNet: Modular Attention Network for Referring Expression Comprehension". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, pp. 1307–1315. DOI: 10.1109/CVPR.2018.00142. http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Yu%5C_MAttNet%5C_Modular%5C_Attention%5C_CVPR%5C_2018%5C_paper.html.

Yu, Tao, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev (2018). "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3911–3921. DOI: 10.18653/v1/D18-1425. https://aclanthology.org/D18-1425.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich (2020). "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault. Association for Computational Linguistics, pp. 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148.

Zhang, Hanyi, Longbiao Wang, Yunchun Zhang, Meng Liu, Kong Aik Lee, and Jianguo Wei (2020). "Adversarial Separation Network for Speaker Recognition." In: *Interspeech*, pp. 951–955.

Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston (2018). "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2204–2213. DOI: 10.18653/v1/P18-1205. https://aclanthology.org/P18-1205.

Zhang, Ziqiang, Yan Song, Jian-shu Zhang, Ian McLoughlin, and Li-Rong Dai (2020). "Semi-Supervised End-to-End ASR via Teacher-Student Learning with Conditional Posterior Distribution". In: *Interspeech*, pp. 3580–3584.

Zhu, Jinhua, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu (2020). "Incorporating BERT into Neural Machine Translation". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=Hyl7ygStwB.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer Learning for Low-Resource Neural Machine Translation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1568–1575. DOI: 10.18653/v1/D16-1163. https://aclanthology.org/D16-1163.

Chapter 3
Digital Language Equality: Definition, Metric, Dashboard

Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way

Abstract This chapter presents the concept of Digital Language Equality (DLE) that was at the heart of the European Language Equality (ELE) initiative, and describes the DLE Metric, which includes technological factors (TFs) and contextual factors (CFs): the former concern the availability of Language Resources and Technologies (LRTs) for the languages of Europe, based on the data included in the European Language Grid (ELG) catalogue, while the latter reflect the broader socio-economic contexts and ecosystems of the languages, as these determine the potential for LRT development. The chapter discusses related work, presents the DLE definition and describes how it was implemented through the DLE Metric, explaining how the TFs and CFs were quantified. The resulting scores of the DLE Metric for Europe's languages can be visualised and compared through the interactive DLE dashboard, to monitor the progress towards DLE in Europe.[1]

1 Introduction and Background

The META-NET White Paper Series (Rehm and Uszkoreit 2012) showed the clear imbalance in terms of technology support for 31 European languages as of 2012 (see Chapter 1). Beyond the official European and national languages, more than 60 regional and minority languages (RMLs) are protected by the European Charter for Regional or Minority Languages and the Charter of Fundamental Rights of

Federico Gaspari · Owen Gallagher · Andy Way
Dublin City University, ADAPT Centre, Ireland, federico.gaspari@adaptcentre.ie, owen.gallagher@adaptcentre.ie, andy.way@adaptcentre.ie

Annika Grützner-Zahn · Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, annika.gruetzner-zahn@dfki.de, georg.rehm@dfki.de

Maria Giagkou · Stelios Piperidis
R. C. "Athena", Greece, mgiagkou@athenarc.gr, spip@athenarc.gr

[1] This chapter is based on Gaspari et al. (2021, 2022a,b), Giagkou et al. (2022), and Grützner-Zahn and Rehm (2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_3

the EU. Against this background, the EU-funded project European Language Equality (ELE) has addressed the issue of Digital Language Equality (DLE) in Europe, with the intention of tackling the imbalances across Europe's languages, that have widened even further in the meantime, as explained in Chapter 4. ELE's contribution to advancing DLE in Europe hinges on a systematically developed and inclusive all-encompassing strategic research, innovation and implementation agenda (SRIA) and a related roadmap to drive forward much needed efforts in this direction (see Chapter 45). The present chapter describes the notion of DLE and the associated metric that are at the heart of these plans, and presents the DLE dashboard that visualises the digital support of each European language, so as to monitor the overall progress towards DLE in Europe, also in a comparative fashion across languages.

Despite the persisting imbalances, Europe has come a long way in recognising and promoting languages as fundamental rights of its people and essential components of its unique combined cultural heritage, and this awareness is reflected in research and policy advancements of the last two decades. Krauwer (2003) represented one of the earliest calls for action towards the development of Language Resources and Technologies (LRTs), in particular for under-resourced languages. In the following years, several projects and initiatives contributed to the progress of Europe's languages in terms of technological and digital support; some of the main efforts in this area that laid the foundation for subsequent substantial progress were, e. g., Euromatrix (Eisele et al. 2008), iTranslate4.eu (Yvon and Hansen 2010), FLaReNet (Soria et al. 2012) and CLARIN (Hinrichs and Krauwer 2014). Additionally, META-NET, an EU Network of Excellence forging the Multilingual Europe Technology Alliance, was established and a group of projects (T4ME, CESAR, METANET4U, META-NORD) promoted and supported the development of Language Technologies (LTs) for all European languages (Rehm and Uszkoreit 2012, 2013; Rehm et al. 2016). The EU project CRACKER (Cracking the Language Barrier, 2015-2017) continued the work of META-NET, concentrating on additional strategy development and community building (Rehm et al. 2020). The most recent EU-funded projects continuing efforts in this area were European Language Grid (ELG, Rehm 2023b) and European Language Equality (ELE, Rehm et al. 2022), which collaborated closely, leading to the development of the DLE Metric and the DLE dashboard presented in this chapter.

2 Related Work

While our work on DLE focused specifically on the languages of Europe, it is located in a broader context of related recent efforts with a wider remit, which are briefly reviewed here to pinpoint issues of interest for the subsequent presentation of the definition of DLE, its metric and the dashboard. Joshi et al. (2020) investigate the relation between the languages of the world and the resources available for them as well as their coverage in Natural Language Processing (NLP) conferences, providing evidence for the severe disparity that exists across languages in terms of technological support and attention paid by academic, scientific and corporate play-

ers. In a similar vein, Blasi et al. (2022, p. 5486) argue that the substantial progress brought about by the generally improved performance of NLP methods "has been restricted to a minuscule subset of the world's approx. 6,500 languages", and present a framework for gauging the global utility of LTs in relation to demand, based on the analysis of a sample of over 60,000 papers published at major NLP conferences. This study also shows convincing evidence for the striking inequality in the development of LTs across the world's languages. While this severe disparity is partly in favour of a few, mostly European, languages, on the whole, the vast majority of the languages spoken in Europe are at a disadvantage.

Simons et al. (2022) develop an automated method to evaluate the level of technological support for languages across the world. Scraping the names of the supported languages from the websites of over 140 tools selected to represent a good level of technological support, they propose an explainable model for quantifying and monitoring digital language support on a global scale. Khanuja et al. (2022) propose an approach to evaluate NLP technologies across the three dimensions of inclusivity, equity and accessibility as a way to quantify the diversity of the users they can serve, with a particular focus on equity as a largely neglected issue. Their proposal consists of addressing existing gaps in LRT provision in relation to societal wealth inequality. Khanuja et al. (2022) lament in particular the very limited diversity of current NLP systems for Indian languages, and to remedy this unsatisfactory situation they demonstrate the value of region-specific choices when building models and creating datasets, also proposing an innovative approach to optimise resource allocation for fine-tuning. They also discuss the steps that can be taken to reduce the biases in LRTs for Indian languages and call upon the community to consider their evaluation paradigm in the interest of enriching the linguistic diversity of NLP applications.

Acknowledging that LTs are becoming increasingly ubiquitous, Faisal et al. (2022) look into the efforts to expand the language coverage of NLP applications. Since a key factor determining the quality of the latest NLP systems is data availability, they study the geographical representativeness of language datasets to assess the extent to which they match the needs of the members of the respective language communities, with a thorough analysis of the striking inequalities. Bromham et al. (2021) examine the effects of a range of demographic and socio-economic aspects on the use and status of the languages of the world, and conclude that language diversity is under threat across the globe, including in industrialised and economically advanced regions. This study finds that half of the languages under investigation faced serious risks of extinction, potentially within a generation, if not imminently. This is certainly an extremely sombre situation to face up to, which calls for a large-scale mobilisation of all possible efforts by all interested parties to avoid such a daunting prospect, particularly in Europe, where multilingualism is recognised as an important part of diversity. Establishing a working definition of DLE, devising a metric to measure the situation of each European language with respect to DLE and implementing an interactive dashboard to monitor progress in this direction are vital elements of this large-scale endeavour.

3 Digital Language Equality: Key Principles and Definition

The DLE Metric and the DLE dashboard can be used to measure, visualise and compare the position of Europe's languages with respect to DLE on the basis of up-to-date and carefully chosen quantitative indicators. In this context, language *equality* does not mean *sameness* on all counts, regardless of the respective environments of the languages; in fact, the different historical developments and current situations of the very diverse languages under consideration are duly taken into account, along with their specific features, different needs and realities of their communities, e. g., in terms of number of speakers, ranges of use, etc., which vary significantly. It would be naive and unrealistic in practice to disregard these facts, and to set out to erase the differences that exist between languages, which are vital reflections of the relevant communities of speakers and key components of Europe's shared cultural heritage. This is also a core value of multilingualism in Europe, where all languages are regarded as inherent components of the cultural and social fabric that connects European citizens in their diversity.

In addition, the notion of DLE stays well clear of any judgement of the political, social and cultural status or value of the languages, insofar as they collectively contribute to a multilingual Europe that should be supported and promoted. Alongside the fundamental concept of *equality*, we also recognise the importance of the notion of *equity*, meaning that for some European languages, and for some of their needs, a targeted effort is necessary to advance the cause of equality. For example, the availability of, and access to, certain resources and services (e. g., to revitalise a language, or to promote education through that language) may be very important for some of Europe's languages, but by and large these are not pressing issues, for instance, for most official national languages. With this in mind, the definition of DLE and the implementation of the DLE Metric discussed below are intended to accurately capture the needs and expectations of the various European languages, and especially the shortfalls with respect to being adequately served in terms of resources, tools and technological services in the digital age, so as to support the large-scale efforts to achieve DLE, also through data analytics and visualisation in the DLE dashboard.

The definition of DLE drew inspiration, among others, from the META-NET White Paper Series (Rehm and Uszkoreit 2012) and from the BLARK concept (Basic Language Resource Kit, Krauwer 2003), which have been instrumental in assessing the level of technological support for specific languages, and in particular in identifying those that lag behind in the digital age and in encouraging the targeted interventions required to fill the gaps in LT support. These starting points were further elaborated by the ELE consortium in collaboration with its vast networks of contacts and partnerships, also in light of the latest developments in LRTs and in language-centric AI techniques and of the evolution of the relevant institutional, academic, industrial and business landscape that has grown and diversified considerably in the last two decades, as discussed in other chapters of this book. Following a systematic and inclusive consultation effort in the ELE consortium, the following consensus was achieved (Gaspari et al. 2021, p. 4).

Digital Language Equality (DLE) is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age.

This definition was applied to 89 European languages in the project: all 24 official EU languages, 11 additional official national languages and 54 RMLs. This definition, in turn, provided the conceptual basis to design and implement a metric to enable the quantification of the level of technological support of each European language with descriptive, diagnostic and predictive value to promote DLE in practice. This approach allows for comparisons across languages, tracking their progress towards the ultimate collective goal of DLE in Europe, as well as the prioritisation of interventions to meet any needs, especially to fill identified gaps, focusing on realistic and feasible targets, as part of the implementation of the all-encompassing SRIA and related roadmap devised by ELE to drive the advancement towards DLE, as described in detail in Chapter 45.

4 Implementing the Digital Language Equality Metric

Based on the definition of DLE, we describe the associated metric as follows (Gaspari et al. 2021, p. 4):

The Digital Language Equality (DLE) Metric is a measure that reflects the digital readiness of a language and its contribution to the state of technology-enabled multilingualism, tracking its progress towards the goal of DLE.

The DLE Metric is computed for each European language on the basis of a range of quantifiers, grouped into technological factors (TFs, that correspond to the available resources, tools and services, Gaspari et al. 2022a) and situational contextual factors (CFs, that reflect the broad socio-economic ecosystem of each language, which determines the potential for technology and resource development, Grützner-Zahn and Rehm 2022).

The setup and formulation of the metric are modular and flexible, i. e., they consist of well-defined separate and independent, but tightly integrated quantifiers. In particular, the TFs were devised so as to be compatible with the metadata schema adopted by the European Language Grid cloud platform[2] (Labropoulou et al. 2020; Piperidis et al. 2023). The ELG cloud platform bundles together datasets, corpora, functional software, repositories and applications to benefit European society, industry and academia and administration, and provides a convenient single access point to LRTs for Europe's languages (Rehm 2023a).

[2] https://www.european-language-grid.eu

In addition, the definition of DLE and its associated metric have been designed to be transparent and intuitive for linguists, LT experts and developers, language activists, advocates of language rights, industrial players, policy-makers and European citizens at large, to encourage the widest possible uptake and buy-in to the cause of DLE across Europe. In establishing the DLE definition and its associated metric, an effort was made for them to be founded on solid, widely agreed principles, but also striking a balance between a methodologically sound and theoretically convincing approach, and a transparent formulation. The rationale behind this approach was that the DLE definition and its metric should be easily understood and able to inform future language and LT-related policies at the local, regional, national and European levels in order to guide and prioritise future efforts in the creation, development and improvement of LRTs according to the SRIA and roadmap (see Chapter 45), with the ultimate goal of achieving DLE in Europe by 2030.

Through data analytics and visualisation methods in the DLE dashboard (see Section 7), European languages facing similar challenges in terms of LT provision can be grouped together, and requirements can be formulated to support them in remedying the existing gaps and advancing towards full DLE. A crucial feature of the DLE Metric is its dynamic nature, i. e., the fact that its scores can be updated and monitored over time, at regular intervals or whenever one wishes to check the progress or the status of one or more European languages. This is why the DLE Metric is a valuable tool to achieve DLE for all European languages, and a key element of the sustainable evidence-based SRIA and of the roadmap guiding future interventions promoting LTs and language-centric AI across Europe.

5 Technological Factors

In order to objectively quantify the level of technological support for each of Europe's languages, a number of TFs were considered. The following description presents their main categories, illustrating the breadth and diversity of the LRTs that they capture through the ELG catalogue (Rehm 2023a; Piperidis et al. 2023; Labropoulou et al. 2020). In that regard, we assume that the ELG catalogue, with its more than 13,000 LRTs at the time of writing, provides a representative picture of the state of play of technology support of Europe's languages.

The first category of TFs is based on the availability of LRs, i. e., corpora, datasets or collections of text documents, text segments, audio transcripts, audio and video recordings, etc., monolingual or bi-/multilingual, raw or annotated. This category also encompasses language models and computational grammars and resources organised on the basis of lexical or conceptual entries (lexical items, terms, concepts, etc.) with their supplementary information (e. g., grammatical, semantic, statistical information, etc.), such as lexica, gazetteers, ontologies, term lists, thesauri, etc.

The resulting technological DLE score for each European language is a reflection of the LRTs available in the ELG catalogue for that language. While the number of available LRs is an essential aspect of a language's digital readiness, the specific

types and features of these LRs are equally important, insofar as they indicate how well a language is supported in the different LT areas. To capture such aspects in the DLE Metric, in addition to raw counts of available LRs, the following LR features have also been taken into account and attributed specific weights in the scoring mechanism (see Table 1, p. 66, in the Appendix):

- resource type
- resource subclass
- linguality type
- media type covered or supported
- annotation type (where relevant)
- domain covered (where relevant)
- conditions of use

The second category of TFs is based on the availability of tools and services offered via the web or running in the cloud, but also downloadable tools, source code, etc. This category encompasses, for example, NLP tools (morphological analysers, part-of-speech taggers, lemmatisers, parsers, etc.); authoring tools (e. g. spelling, grammar and style checkers); services for information retrieval, extraction, and mining, text and speech analytics, machine translation, natural language understanding and generation, speech technologies, conversational systems, etc. The features of tools and services that are considered and assigned weights in the scoring system of the DLE Metric (see Table 2, p. 67), are as follows:

- language (in)dependent
- type of input processed
- type of output provided
- type of function
- domain covered (where relevant)
- conditions of use

5.1 Weights and Scores

The weights given to the feature values of the LRTs quantify their contribution to the DLE score with regard to the relevant TFs. The scoring system (see Tables 1 and 2) is based on the assumption that for any language some features of LRTs contribute more effectively to achieving DLE than others. Higher weights are assigned to feature values related to 1. more complex LRTs, e. g., tools that process or support more than one modality, 2. more expensive and labour-intensive datasets or tools, e. g., in terms of the effort required to build them, 3. more open or freely available datasets and tools, and 4. additional envisaged applications that could be supported.

One guiding consideration in developing the DLE Metric, and especially in assigning the weights of the features and their values for the TFs, is to make the fewest possible assumptions about the (preferred or supposedly ideal) use-cases and actual

application scenarios that may be most relevant to users. These can vary widely for all languages on the basis of a number of factors impossible to establish a priori. We therefore refrained from predetermining particular preferred end-uses when implementing the full specification of the DLE Metric, which otherwise would risk it being unsuitable for some end-users and applications. Here we briefly review some of the key features of the TFs, focusing on those that can have several values.

For instance, a feature of LRs that can receive several values is that of *Annotation Type*, where applicable. In the implementation of the DLE Metric, we assign a constant very small fixed weight, also based on the fact that some LRs can possess several annotation types in combination. A similar consideration applies to the *Domain* feature (again, where relevant), which has many possible values both for LRs and for tools and services: in these cases, the weights assigned to *Domain* values are fixed and relatively small, again considering that multiple domains can be combined in a single LR, tool or service. In addition to *Domain*, another feature that appears both in LRs and tools and services is *Conditions of use*: the weights proposed for this feature of the TFs are identical for the corresponding values of *Conditions of use* across datasets and tools and services. In the case of (much) more restrictive licensing terms, lower weights are assigned than to liberal use conditions, so they contribute (much) less to the partial technological DLE score for the LRT in question, and therefore to the overall technological DLE score for the specific language.

5.2 Configuration of the Technological Factors

Before coming up with the final implementation of the weighting and scoring system for the TFs (see Tables 1 and 2), we experimented with a range of different setups. We used the contents of the ELG catalogue as of early 2022, which at that time contained about 11,500 records, out of which about 75% were datasets and resources (corpora, lexical resources, models, grammars) and the rest were tools and services. These records contained multiple levels of metadata granularity. The ELG repository had been populated with LRTs following extensive efforts by a wide range of language experts and reflected the input of this community of experts, mobilised in ELE, to ensure comprehensive coverage, which is why we considered the ELG catalogue representative with regard to the existence of LRTs for Europe's languages, so it was used as the empirical basis for the computation of the technological DLE scores.

The ELG catalogue includes metadata for LRs and LTs. In ELG, each resource and tool/service has several features and associated values, based on the schemes presented in Tables 1 and 2. Each feature was initially assigned a tentative weight to calculate preliminary technological DLE scores of each language, comparing the resulting scores of a number of alternative preliminary setups. During this fine-tuning of the weights, we considered especially where each language stood in relation to the others and how their relative positioning changed as a result of assigning different weights to the various feature values. This was an efficient and effective method to

gradually refine the setup of the TFs and propose the implementation of the weights in the scoring mechanism that was eventually adopted (see Tables 1 and 2).

The experiments showed that the global picture of the technological DLE scores for the languages of Europe tended not to change dramatically as the weights assigned to the feature values were manipulated. We experimented both with very moderate and narrow ranges of weights, and with more extreme and differentiated weighting schemes. Since, ultimately, any changes were applied across the board to all LRTs included in the ELG catalogue for all languages, any resulting changes propagated proportionally to the entire set of languages, thus making any dramatic changes rather unlikely, unless one deliberately rewarded (i. e., gamed) features known to disproportionately affect one or more particular languages. It is clear that this would have been a biased and unfair manipulation of the DLE Metric, and was therefore avoided, as we wanted the relevant scores to be a fair, and bias-free, representation of the status of all European languages with respect to DLE.

These preliminary experiments carried out in early 2022 to finalise the setup of the TFs for the DLE Metric demonstrated that the overall distribution of the languages tended to be relatively stable. This was due partly to the sheer amount of features and possible feature values that make up the TFs. As a result, even if one changed the weights, with the exception of minor and local fluctuations, three main phenomena were generally observed while testing the DLE Metric and its TF scores.

1. The overall positioning of the languages remained largely stable, with a handful of languages standing out with the highest technological DLE scores (English leading by far, typically over German, Spanish and French, with the second language having roughly half the technological DLE score of English), the many minimally supported languages still displaying extremely low technological DLE scores, and a large group of similarly supported languages in the middle.
2. Clusters of languages with similar LT support according to intuition and expert opinion remained ranked closely together, regardless of the adjustments made to specific weights for individual features and their values.
3. Even when two similarly supported languages changed relative positions (i. e., language A overtook language B in terms of technological DLE score) as a result of adjusting the weights assigned to specific features and their values, their absolute technological DLE scores still remained very close, and the changes in ranking tended not to affect other neighbouring languages on either side in a noticeable manner.

During the preliminary testing that eventually led to the final setup of the TFs in the DLE Metric presented in Tables 1 and 2, we performed focused checks on pairs or small sets of languages spoken by comparable communities and used in nearby areas or similar circumstances, and whose relative status in terms of LT support is well known to the experts. These focused checks involved, e. g., Basque and Galician, Irish with respect to Welsh, and the dozen local languages of Italy (also with respect to Italian itself), etc. Overall, the general stability and consistency demonstrated by the technological DLE scores across different setups of weight assignments for the various features and their possible values for TFs provided evidence of its validity

as an effective tool to guide developments and track progress towards full DLE for all of Europe's languages. In essence, the setup eventually selected (Tables 1 and 2) ensures that the DLE Metric optimally captures the real situation of all of Europe's languages in the digital age, tracking the progress towards DLE.

5.3 Computing the Technological Scores

Based on the above, the steps to calculate the technological DLE score which is part of the DLE Metric are as follows:

1. Each LRT in the ELG catalogue obtains a score ($Score_{LRT}$), which is equal to the sum of the weights of its relevant features (see Tables 1 and 2 for the weights and associated values). Specifically for features *Annotation Type* and *Domain*, instead of simply adding the respective weight, the weight is multiplied by the number of unique feature values the LR in question has (see Section 5.1).
   Example: Suppose an LRT in the ELG catalogue (LRT1) has the following features: corpus, annotated, monolingual, with three different annotation types (morphology, syntax, semantics), with text as media type, covering one domain (e. g., finance), with condition of use *research use allowed*. Then, using the weights as specified in Table 1, LRT1 is assigned the following score:

$$Score_{LRT1} = 5 + 1 + 2.5 + (3 * 0.25) + 1 + (1 * 0.3) + 3.5 = 14.05$$

2. To compute the technological DLE score for language X ($TechDLE_{LangX}$) we sum up the $Score_{LRT}$ of all LRTs that support language X (LRT1, LRT2, …LRTN), i. e.,

$$TechDLE_{LangX} = \sum_{i=1}^{N} Score_{LRTi}$$

Similarly, any tool or service included in the ELG catalogue receives a partial score with the same procedure, on the basis of the weights presented in Table 2. As the ELG catalogue organically grows over time, the resulting technological DLE scores are constantly updated for all European languages. These scores can be visualised through the DLE dashboard (see Section 7), providing an up-to-date and consistent (i. e., comparable) measurement of the level of LT support and provision that each language of Europe has available, also showing where the status is not ideal or not at the level one might expect.

5.4 Technological DLE Scores of Europe's Languages

Figure 1 shows the technological DLE scores for all of Europe's languages as of late February 2023, obtained on the basis of the final weighting and scoring mechanism described in the previous sections.

Not surprisingly, based on the TFs of the DLE Metric, at the time of writing in early 2023, English is still by far the most well-resourced language of Europe, leading the way over German and Spanish, that follow with very similar technological DLE scores, which are roughly half that of English. French has a marginally lower score, which places it in fourth position. Italian, Finnish and Portuguese follow at some distance, and it is interesting to note that the next cluster of languages that are spoken by sizeable communities in Europe (e. g., Polish, Dutch, Swedish), still in the top ten of the overall list of languages, have a technological DLE score that is roughly six times lower than that of English: a stark reminder based on evidence provided by the ELG catalogue and measured through the DLE Metric of the persisting imbalances in the overall digital support of Europe's languages, showing that urgent decisive action is needed to achieve DLE (Chapter 4 provides a more detailed cross-language comparison).

5.5 Open Issues and Challenges

The technological DLE scores based on the TFs do not take into account the size of the LRs or the quality of the LRTs included in ELG. While these are important features, there exist a large variety of size units for LRs, and the way of measuring data size is not standardised, especially for new types of LRs such as language models. Regarding the quality of tools and services in particular, while some information on the Technology Readiness Level[3] scale is available in ELG, the large number of null values does not make it easy to take this aspect into account for consistency reasons. These are shortcomings that can be revisited in subsequent efforts, with a view to overcoming these limitations and further improving the overall accuracy and granularity of the technological DLE scores going forward.

As far as datasets are concerned, in particular, there could be benefits in setting a minimum size criterion to include LRs such as corpora or grammars in the computation of the technological DLE score, e. g., to avoid using very small resources that cannot be realistically applied in actual technology development scenarios. However, it is difficult to establish arbitrarily what this minimum size threshold should be, also in recognition of the specifics of the languages of Europe. As a result, the decision was made not to set any minimum size requirement for LRs. The thinking behind this choice was that relatively small datasets are common in less-resourced languages, for particular domains, etc., and there is the possibility to merge small datasets to create bigger ones that would, in fact, be useful, for instance in domain

[3] https://en.wikipedia.org/wiki/Technology_readiness_level

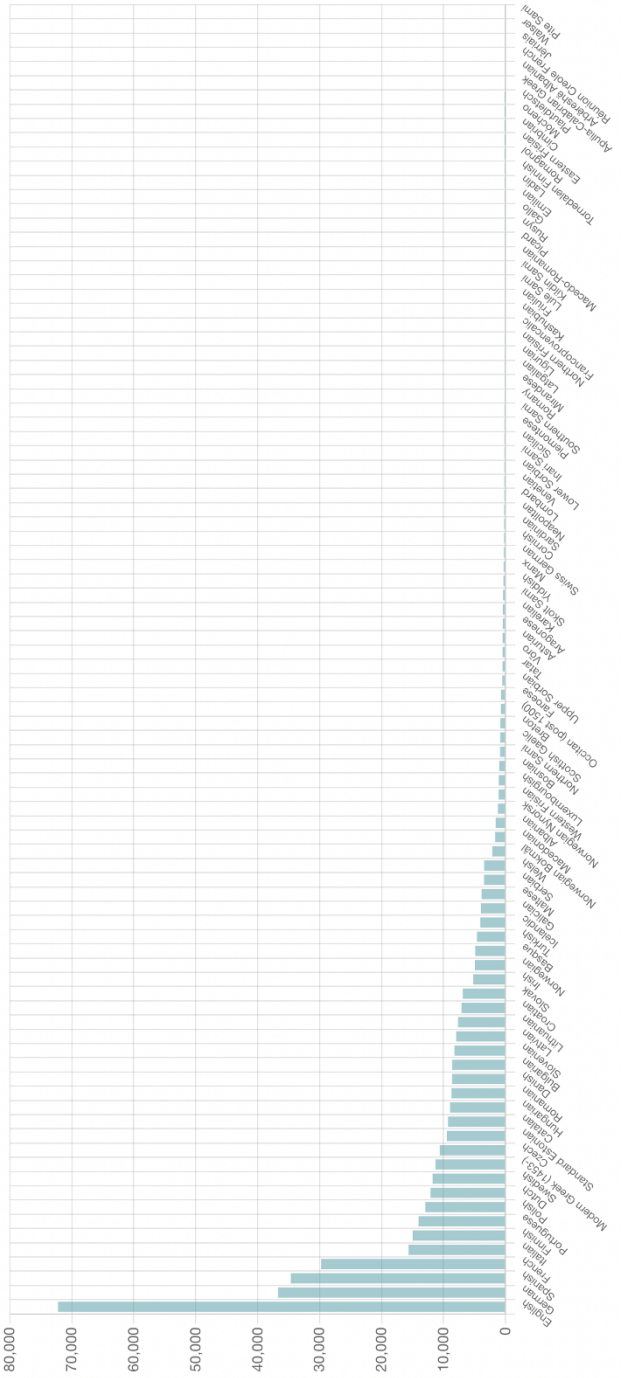

Fig. 1 Technological Digital Language Equality scores as of late February 2023

adaptation for MT, to mention but one example. More broadly, by proposing the DLE Metric we intend to foster a culture of valuing all and any LRTs, especially for less-resourced languages, judiciously balancing the importance given to the size, quantity, diversity and quality of the LRTs, being mindful that several of Europe's languages are in dire need of support.

6 Contextual Factors

While the technological scores based on the TFs represent the technological support of a language, they do not reflect the overall socio-political environment of a language. There are other factors that influence how a language thrives in the digital age, such as political will, funding, being the object of research projects, economic interest, etc. The importance of creating a picture that reflects this environment of a language community was recently also considered by other researchers. Several data-driven studies analyse the relationship between the technical support of a language and non-technological factors (see Section 2).

Related approaches attempt to measure the influence of non-technological factors on the development of LRTs considering often only individual factors in the realm of economy (usually the Gross Domestic Product, GDP), research (e. g., number of publications in specific conferences) and the size of the language community. In the DLE Metric, the Contextual Factors (CFs) are defined as the "general conditions and situations of the broader context" of a language community (Gaspari et al. 2021, p. 7). This definition includes factors from all areas of life assuming that those have an influence on the development and use of LRTs.

Economy Factors in this area reflect the general and the LRT-specific part of the economy. The overall welfare of the language community and the size of the potential market are important factors for companies to invest in the development of LRTs for a language.

Education The language and digital literacy level of a language community influences the use of a language online and on digital devices. Additionally, to be able to develop LRTs, researchers with technical but also linguistic skills of the respective languages are needed.

Funding Investment in research and innovation in the area of LT is necessary for basic and applied research on which technology development is based.

Industry Companies, both well-established and startups, are important drivers of the development and distribution of LT applications, tools and services.

Law The legal framework can hinder progress or steer developments in certain directions.

Media The creation and distribution of news, newspapers, magazines, films, etc. in a language constitutes, on the one hand, a possible large dataset for the development of LRTs, and on the other hand, demonstrates the willingness to make content accessible to the language community.

Online The online representation of a language community indicates that active community members are willing and determined to use the language in the digital world. Additionally, the availability of online data in the respective language gives researchers or developers the opportunity to create LRs.

Policy Strategic plans and agendas at local, regional and national levels indicate the political will to support a topic and the direction in which policy-makers intend to lead society in the future.

Public Administration Public authorities represent the state to its citizens. The inclusion and support of languages spoken in the country or region by public authorities enables participation and utilisation within the society.

Research & Development & Innovation Innovations depend on basic and applied research and on the development of products that are ready for the market. This requires a minimum of research positions in relevant institutions and supporting infrastructure.

Society The social attitude towards a language has a great influence on how much investment, effort and time are put into the preservation of a language by the language community and by the state.

Technology The technological infrastructure reflects the possibility for a language community to access and take a part in the digital world.

6.1 Computing the Contextual Scores

6.1.1 Data Sources and Collection

Initially, 72 potential contextual factors were identified through the collection of factors considered relevant in publications such as, among others, the STOA study (STOA 2018), the META-NET White Paper Series (Rehm and Uszkoreit 2012) and EFNIL's European Language Monitor (ELM);[4] we also consulted with the 52 ELE project partners. The 72 tentative CFs were clustered into 12 areas (see above) representing different aspects of a language's context (Gaspari et al. 2021).

To be measurable, each factor had to be quantified with an indicator, which depended on the existence and accessibility of corresponding data. First, different data sources were collected including, among others, EUROSTAT,[5] ELM, Ethnologue[6] and various reports and articles. Second, possible indicators for each factor were considered and matched with the available data. GDP, for example, was considered to be a suitable indicator for the factor "economic size".

Eventually, 27 of the 72 initial factors had to be excluded due to missing data. This affected especially factors from the areas "research & development & innovation", "society" and "policy". Data about policies is essentially too broad and reflects rather

[4] http://www.efnil.org/projects/elm

[5] https://ec.europa.eu/eurostat

[6] https://www.ethnologue.com

coarsely whether policies exist or not. For instance, the factor "presence of local, regional or national strategic plans, agendas, committees working on the language, LT, NLP, etc." was quantified on data indicating whether a national agenda with regard to AI and LTs exists. Considering also local and regional plans and the existence and maybe also number and size of committees would require much more detailed data. The factors excluded from the class "research & development & innovation" covered mainly figures about the LT research environment, while broader numbers about the research situation of the whole country were indeed available. Tables 4-15 in the Appendix show all factors from the preliminary definition (Gaspari et al. 2021, 2022b), their class and the indicator they were quantified with. Overall, 46 factors were quantified with at least one appropriate indicator, and some with two indicators representing different perspectives like total numbers and numbers per capita.

The data was collected in late 2021. Many sources provided their data as spreadsheets, while some data was published as HTML documents. The data for 15 indicators had to be collected manually from reports and articles. We attempt to update the contextual factors on an annual basis. Preliminary tests indicate that updating the contextual DLE scores for all EU languages takes up to two weeks of work by one member of staff who is familiar with the structure and nature of the CFs.

6.1.2 Data Processing

The collected CF data was very heterogeneous: it had different formats, was based on country or language community level, included differing languages or countries and consisted of different data types. Data preparation took several steps, including data format standardisation, harmonising language names based on Glottolog (Hammarström et al. 2021) and data merging. Some sources provided plain text from which a score had to be manually determined. Features mentioned in the text, e. g., regarding the existence of a national LT policy, were quantified with a number and this number was assigned to countries or language communities. If the text included more than one feature, the numbers were added up, e. g., if a country published several policies covering the topic AI and LTs. Table 3 (p. 68) shows a list of the indicators transformed from plain text.

The DLE Metric processes data on a per-language basis. Thus, data collected on the *country* level had to be converted to the *language* level. In total, the factors were quantified with three different types of data, namely absolute numbers, proportional numbers, and scores. Total numbers were split proportionally, using the percentage of speakers of the language per country. The percentages were calculated through population size and number of speakers. Due to some gaps and old records, experts from the ELE consortium were asked to provide missing or more up-to-date and reliable data. The figures for Alsatian, Faroese, Gallo, Icelandic, Macedonian and the Saami languages were corrected accordingly.

Languages often taught as a second language (English, German, French, Spanish) were only included in the mapping if the language had an official status in the country. For example, the figures for English consist of the figures of the UK, Ireland and

Malta (in other European countries, English does not have official status). If the language was an official national language in at least one country, only language communities with more than one percent were included to simplify the mapping. Total numbers per capita of a language community, proportional numbers, and scores were applied to the language communities without adjustment.

If a language was spoken in more than one country, total numbers were added up, while proportional numbers, scores and total numbers per capita were calculated through the average; the different sizes of the language communities were partly taken into account, hence, the data values of bigger language communities were weighted double for the calculation of the average. However, a more complex inclusion of the size of the language community would result in more fine-grained figures, which would probably affect the contextual DLE scores to some extent.

6.1.3 Calculation of the Contextual Digital Language Equality Score

The data referring to each language community was converted into contextual DLE scores, which indicate the extent to which a language has a context that supports the possibility of evolving digitally or not. Without the political will, funding, innovation and economic interest in the respective region, the probability of achieving DLE is low. Given the underlying complexity, in order for the contextual scores to be easily conceptualised and comparable across languages, a relative score between 0 and 1 was assigned to each language, with 0 representing a context with no potential for the development of LT, and 1 representing the best potential. To keep this part of the DLE Metric as transparent as possible, we decided to base the calculation on an average of the factors. Therefore, the intermediate goal was to calculate a score between 0 and 1 for each factor. The language with the lowest value for the respective factor was attributed 0, while the language with the highest value received 1. The following steps were conducted to calculate the contextual DLE score for each European language:

1. Calculation of the range: highest value – lowest value;
2. $\frac{(value-minimum)*100}{range}$ = Percentage weighting of a language within the range;
3. The result is a relative value: to obtain a score between 0-1 the result is divided by 100;
4. Apply steps 1-3 for all languages and factors;
5. Calculate the average of all factors per language;
6. Weighting of the scores with the three chosen factors of a. number of speakers, b. scores based on the language status, and c. whether the language is an official EU language or not.

The three weighting factors were considered to be particularly relevant for the context to develop LRTs due to the influence of the number of speakers on the investment by large companies and its official status in the EU on the amount of funding. The weighting included two steps: 1. calculating the average of the overall scores, the scores for the number of speakers and the legal status and 2. adding 0.07 to the

score for each official EU language. The second step was separated from the average calculation, because the indicator consisted of two values, 1 if it is an official EU language and 0 if it is not. The average calculation would result in an excessively strong boost for all official EU languages. Hence, with the data for the contextual factors available at the end of 2021, English already had a score of around 0.7-0.8 without the boost. Smaller values for EU languages would have penalised English, which would not have represented reality.

We created five different versions of the possible configurations of the CFs to conduct a thorough comparative evaluation. The factors were classified based on a number of overall properties, i. e., if a data point can be updated automatically or if the data is considered high quality (see Tables 4-15). Data quality was chosen to avoid bias in the overall result caused by extreme maximum and minimum values. For example, for the quantification of the factor "number of podcasts", several platforms were found which could have provided numbers of podcasts in different European languages, but because of different target audiences, the values were highly skewed to the languages spoken by those target audiences. Factors which were quantified with data reflecting no big differences between languages were also excluded by the quality criterion, e. g., the literacy level of all countries varied between 98 and 99 percent, i. e., hardly at all. To be able to update the metric on a regular basis without much manual effort after the end of the ELE project, the possibility of collecting the data fully automatically was picked as the other main criterion.

Based on these criteria, the following CF configurations were examined:

1. Factors with available data: 46 factors
2. Factors that can be updated automatically: 34 factors
3. Factors with good or high data quality: 26 factors
4. Factors that can be updated automatically and that also have good or high data quality: 21 factors
5. A set of manually curated factors using four criteria: automatically updatable, good/high data quality, a maximum of two factors per class, balance between data types: 12 factors (Table 16 shows the factors included in this configuration)

Including fewer factors in the metric increased the risk of omitting an important factor. On the other hand, including fewer factors also reduced the risk of distorting the metric with more data.

6.2 Experts Consultation

Considering that appropriate baselines do not exist, we validated the five different results through the consultation of experts. Individual contextual scores can be interpreted by comparing them to the scores of other languages.

The panel consisted of ELE consortium partners. We selected the members based on their expertise and experience in the areas of LT, Computational Linguistics and Linguistics. Moreover, the experts represented different European countries and

were very familiar with the background of their countries and languages spoken there. We reached out to 37 of the 52 ELE partner organisations. They received the results of the five configurations of the metric and were asked to provide assessments regarding the languages they knew, to explain how they would have expected the results to be, and to indicate the most appropriate configuration.

In total, 18 partners provided assessments. The feedback consisted of overall ratings of the five configurations as well as detailed comments regarding individual languages. As a consequence, most answers related to official EU languages. RMLs for which feedback was received are spoken in the UK, Spain, Italy and the Nordic countries. We received feedback on 56 of the 89 languages.

In general, using all factors was evaluated as risky due to the possible distortion of results caused by data of bad quality. The results of configuration 1 were considered unexpected, with high scores for languages such as Emilian, Gallo and Franco-Provencial, probably caused by distorted data. The second configuration was criticised, too, except for positive comments on the automatic nature of the metric. The results were less distorted but evaluated as worse compared to configurations 3-5. The results of configurations 4 and 5 were similar. Focusing on quality data improved the results significantly. With fewer factors, configuration 5 provided similar results as configuration 4. Configuration 5 was assessed positively regarding the transparency of fewer factors and the possibility to balance the classes.

Overall, the results of the fifth configuration were assessed to represent the context of the language communities in the most adequate way, while there is still room for improvement for a few languages. Table 17 (p. 73) provides more details.

Several suggestions for improvements were made. Since only pan-European data sources were taken into account for reasons of consistency and comparability, one recommendation concerned extending the data through relevant national and regional sources. One expert pointed out that the context of European languages spoken in countries outside of Europe was excluded, and these missing statistics on the development of LRTs would greatly impact the overall scores, e. g., Portuguese in Brazil. Another suggestion referred to missing factors, such as the inclusion of the vitality status of a language being particularly important for RMLs, or the integration of a factor representing competition of a national language with English as the other official national language which often still dominates daily life, e. g., in Ireland, and prevents more widespread use of the other national language in these areas. Another idea was to replace the official EU status as a weighting factor with the country's membership in the European Economic Area (EEA), since these countries also have access to European research funds.

Suggestions were also made regarding the presentation of the results. Language communities having particularly complex political backgrounds are most likely to be misrepresented by a simple calculation based on country-specific data, and should be highlighted and presented with the limits of solely data-driven work for such cases. It was also suggested that languages without a writing system should be emphasised as special cases for the development of LRTs.

Some feedback expressed reservations about the whole approach. A few reviewers pointed out that a single methodology should not be used to take into account

the different complex contexts and realities of Europe's language communities. For example, languages like Maltese, Irish and the other Celtic languages, which scored better than expected according to our experts, are of note here. The relative prosperity of the United Kingdom, even though it is no longer an EU Member State, seems to boost the RMLs spoken in the UK, although in reality these RMLs are strongly dominated by English. The same applies to Ireland, which has a strong economy, a large ICT sector and significant investments in (English) AI and LT research and development, but a very low level of support for Irish LT.

Another point of criticism was the inclusion of data not applied on a per capita basis. As a result, despite having relatively good support, some small language communities were unable to achieve a high score. The size of the language community has an impact on the economic interest, investment, number of researchers, etc. for the language, but for small language communities that have already invested a lot in their language and infrastructure, some of the scores obtained may appear too low compared to the expectations of the experts.

These criticisms can be debated at length, especially in the interest of finding effective solutions to the identified issues, but are very difficult to avoid altogether with such a quantitative approach as the one that is required to define and measure the CFs as part of the DLE Metric.

These first stable results for the CF calculation were improved based on a more fine-grained data mapping from country to language community level and the feedback of the experts. The aggregation of data points from different countries for languages spoken in several countries, e. g., French, was based on the average with a boost for the data points collected from the countries in which the language has an official national status. This process was replaced by the calculation of a weighted average based on the number of speakers of the language communities which reflects the distribution of the language communities better and prevents distortion through too small or too big language communities. In addition, the boost for EU Member States was changed to a boost for countries in the EEA, the vitality status was added as a penalty for declining languages, and those competing with English as the other dominant official national language were also penalised. The results of this adaptation decreased the number of languages that eventually achieved an excessively high contextual score.

6.3 Contextual DLE Scores of Europe's Languages

In all examined configurations, the top third is dominated by the official EU languages, while the RMLs are part of the long tail to the right. Official national languages which are not official EU languages are ranked between the official EU languages and the RMLs. Figure 2 shows the final results after the adaptation.

As expected, English has the best context for the development of LRTs by far. It is followed by German and French. Italian and Spanish are shown in positions 4 and 5. The position of Spanish *after* Italian is caused by the inclusion of data from

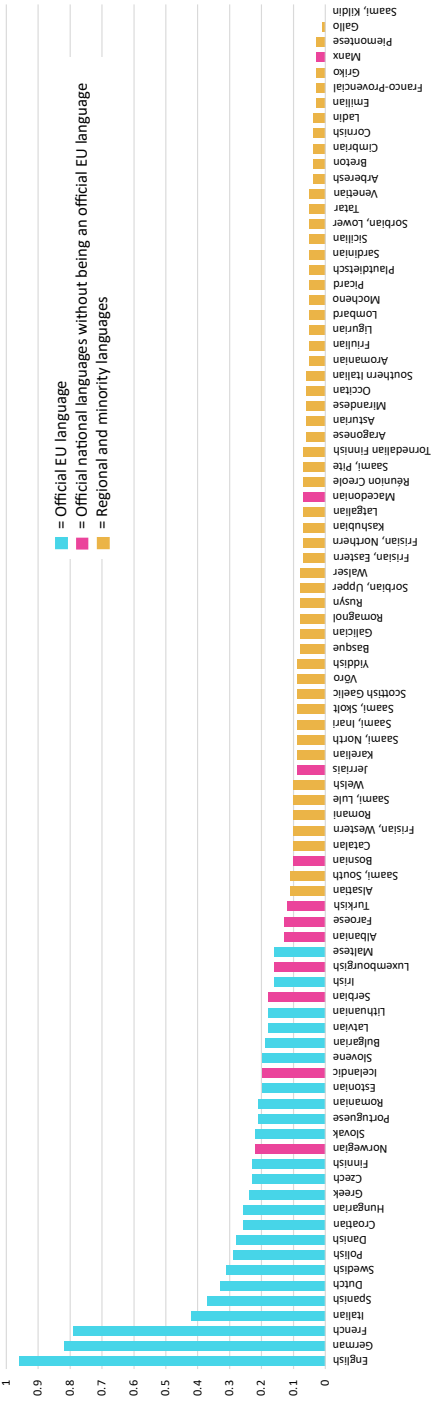

Fig. 2 Contextual Digital Language Equality scores as of late February 2023

European countries only. If data had been included from countries outside of Europe, Spanish, Portuguese, French and English would have had much higher scores. After the five leading languages, variations between the different configurations can be seen. Swedish, Dutch, Danish, Polish, Croatian, Hungarian and Greek are ranked in the upper half of the official EU languages. The official EU languages with the lowest scores are Latvian, Lithuanian, Bulgarian, Romanian, Maltese and Irish which joined this group after the last adjustment.

Among the group of official national languages which are not official EU languages, Norwegian, Icelandic and Serbian are the top performers, achieving contextual DLE scores in line with the middle- and lower-scoring official EU languages, while Manx[7] is presented as a downward outlier. Languages such as Norwegian, Luxembourgish, Faroese and Icelandic achieve better scores than Albanian, Turkish, Macedonian and Bosnian.

The RMLs are led by languages spoken in the more Northern countries like some Saami languages, Western Frisian and Welsh or languages spoken by quite big language communities like Catalan. A total of 23 RMLs achieve contextual DLE scores equal to or lower than 0.05 in the final results, while 30 of the languages obtain scores between 0.06 and 0.1. Kildin Saami and Griko are the languages with the lowest scores.

6.4 Open Issues and Challenges

The contextual DLE scores calculated have some limitations (see Section 6.2). First, expanding the dataset to include regional or national sources would result in 1. a higher number of factors, 2. improved data quality, as the gaps in individual indicators may be filled, 3. quantification of more factors with more than one indicator, to reflect different perspectives, and 4. a more complex mapping to language communities based on regional data resulting in a significant impact on RMLs.

Second, the data cleaning procedure can be improved. One possibility would be to replace outliers with values outside twice the standard deviation by the respective maximum or minimum values of the data series. Data gaps could be filled using data from previous years and skewed data could be corrected using a square root transformation. These processing steps could decrease the impact of distorted data.

An improvement of the mapping from country level to language level could represent regional or urban-rural divides more accurately, especially for larger countries. In particular, the missing mapping of proportional data, scores and total numbers per capita has a major impact on the resulting contextual DLE scores. Here, regional data could help calculate the average deviation of individual regions or language commu-

[7] Manx and Jerriais have been assigned to the group of national languages without being an official EU language, as both languages are recognised as official languages of Jersey and the Isle of Man. Neither island is part of the United Kingdom, but crown dependencies. Therefore, the two languages can be considered both official national languages or RMLs.

nities from other proportional data and to transfer this deviation to proportional data only found on the national level, and similarly for the total figures per capita.

Romaine (2017, p. 49) stresses the importance of an "on-going monitoring of individual communities" for a reliable evaluation of the situation regarding language diversity, which was taken into account with the inclusion of the criterion of automatic updatability of the factors. One problem concerns the eventual interdependencies of the values: the scores of *all* languages may change if new values for some language communities are added, even if the situation of another language community itself has not changed. A temporal dimension could be added to mitigate this.

7 Digital Language Equality Dashboard

In order to provide a precise and easy-to-use tool for presenting and monitoring the TFs and CFs that contribute to the DLE Metric, we designed and implemented a web-based dashboard as part of the European Language Grid.[8] It is available at:

https://live.european-language-grid.eu/catalogue/dashboard

The dashboard shows the contents of the ELG database as interactive visuals dynamically created by user queries, thus providing constantly up-to-date and consistent (i. e., comparable) measurements of the level of LT support and provision across all of Europe's languages (Figure 3). The dashboard provides the figures, statistics and graphs, as appropriate, for:

- the TFs and CFs of the DLE Metric, calculated according to the detailed technical description presented above;
- LRTs hosted in the ELG catalogue, which constitute the source/base data for the TFs that are at the basis of the technological DLE score.

Architecturally, the DLE dashboard consists of two layers: the database of the ELG catalogue and the frontend. The ELG database contents are indexed and saved in JSON. Each user query retrieves the respective results from JSON and exposes them to the front end. While the TFs are calculated dynamically (see Section 5.3) and they reflect the status of the ELG catalogue's database at the time of accessing the dashboard, in the current implementation the CFs are calculated offline, stored in a separate file and exposed to the respective tab of the dashboard's frontend.

[8] https://www.european-language-grid.eu

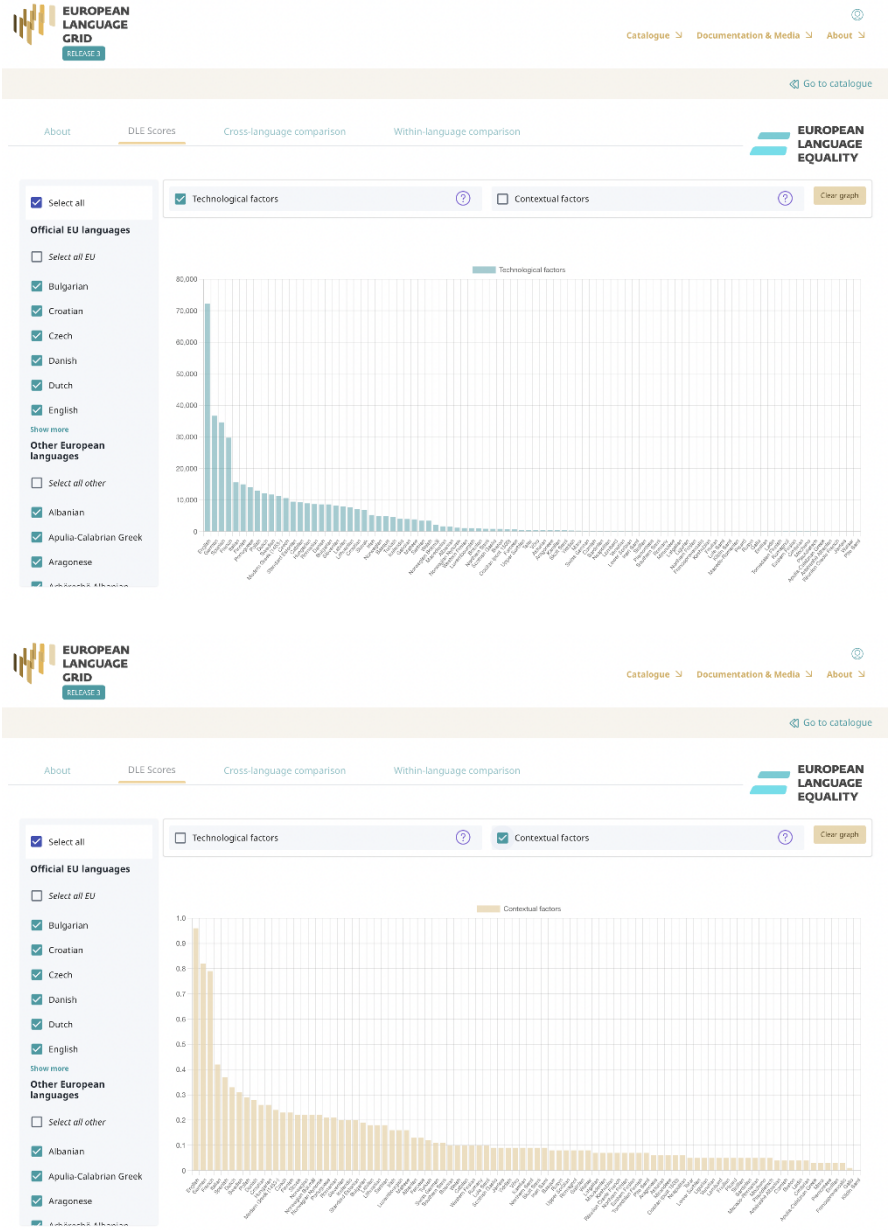

Fig. 3 DLE dashboard showing the technological (top) and contextual DLE scores (bottom)

8 Conclusions and Future Work

This chapter has introduced the definition of DLE adopted in ELE and has described the DLE Metric, explaining the roles and setups of the complementary TFs and CFs and how the scores are computed. By providing an empirically-grounded and realistic quantification of the level of technological support of the languages of Europe, the DLE Metric is intended to contribute to future efforts to level up the digital support of all of Europe's languages, most notably with the implementation of the evidence-based SRIA and roadmap that will drive future efforts in equipping all European languages with the LRTs needed to achieve full DLE (see Chapter 45). The DLE Metric provides a transparent means to track and monitor the actual progress in this direction, as the technological and contextual DLE scores can be visualised through the DLE dashboard.

The overview of the TFs and CFs is accompanied by discussions of the scoring and weighting mechanisms adopted for the computation of the technological and contextual DLE scores, following extensive testing and expert consultations comparing alternative setups. The chapter explains the overall design of the features and their values with the scores and weighting mechanisms that contribute to the DLE Metric scores, based on data included in the ELG catalogue and the factors eventually selected to represent the specific ecosystems of the languages and their communities. As a result of this, the notion of DLE and its associated metric introduced in this chapter represent valuable tools on which to base future efforts to measure and improve the readiness of Europe's languages for the digital age, also taking into account the situational contexts in which the various languages are used via the CFs.

Thanks to the descriptive, diagnostic and predictive value of the DLE Metric, the community now has a solid and verifiable means of pursuing and evaluating much-needed developments in the interest of all languages of Europe and their speakers. The DLE Metric is relevant to a wide range of stakeholders at local, regional, national and European levels who are committed to preventing the extinction of European languages under threat and who are interested in promoting their prosperity for the future. Such stakeholders include decision- and policy-makers, industry leaders, researchers, developers, and citizens across Europe who will drive forward future developments in the fields of LT and language-centric AI in the interest of DLE.

References

Blasi, Damian, Antonios Anastasopoulos, and Graham Neubig (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5486–5505. DOI: 10.18653/v1/2022.acl-long.376. https://aclanthology.org/2022.acl-long.376.

Bromham, Lindell, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua (2021). "Global predictors of language endangerment

and the future of linguistic diversity". In: *Nature Ecology & Evolution* 6, pp. 163–173. https://doi.org/10.1038/s41559-021-01604-y.

Eisele, Andreas, Christian Federmann, Hans Uszkoreit, Herve Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann, and Yu Chen (2008). "Hybrid machine translation architectures within and beyond the EuroMatrix project". In: *Proceedings of the 12th Annual conference of the European Association for Machine Translation, EAMT 2008, Hamburg, Germany, September 22-23, 2008*. Ed. by John Hutchins, Walther Hahn, and Bente Maegaard. European Association for Machine Translation, pp. 27–34. https://aclanthology.org/2008.eamt-1.6/.

Faisal, Fahim, Yinkai Wang, and Antonios Anastasopoulos (2022). "Dataset Geography: Mapping Language Data to Language Users". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3381–3411. DOI: 10.18653/v1/2022.acl-long.239. https://aclanthology.org/2022.acl-long.239.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way (2022a). "Introducing the Digital Language Equality Metric: Technological Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 1–12. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf.

Gaspari, Federico, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way (2022b). *Deliverable D1.3 Digital Language Equality (full specification)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166 ELE. https://european-language-equality.eu/reports/DLE-definition.pdf.

Gaspari, Federico, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou (2021). *Deliverable D1.1 Digital Language Equality (preliminary definition)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/DLE-preliminary-definition.pdf.

Giagkou, Maria, Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Athanasia Kolovou, and Leon Voukoutis (2022). *Deliverable D1.37 Database and Dashboard*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/DLE-dashboard.pdf.

Grützner-Zahn, Annika and Georg Rehm (2022). "Introducing the Digital Language Equality Metric: Contextual Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 13–26. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank (2021). *Glottolog 4.5*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.5772642.

Hinrichs, Erhard and Steven Krauwer (2014). "The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 1525–1531. http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. https://aclanthology.org/2020.acl-main.560.

Khanuja, Simran, Sebastian Ruder, and Partha Talukdar (2022). *Evaluating Inclusivity, Equity, and Accessibility of NLP Technology: A Case Study for Indian Languages*. DOI: 10.48550/ARXIV.2205.12676. https://arxiv.org/abs/2205.12676.

Krauwer, Steven (2003). "The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap". In: *Proceedings of the International Workshop Speech and Computer (SPECOM 2003)*. Moscow, Russia.

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva (2020). "Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3421–3430. https://www.aclweb.org/anthology/2020.lrec-1.420/.

Piperidis, Stelios, Penny Labropoulou, Dimitris Galanis, Miltos Deligiannis, and Georg Rehm (2023). "The European Language Grid Platform: Basic Concepts". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cham: Springer, pp. 13–36. DOI: 10.1007/978-3-031-17258-8_2. https://doi.org/10.1007/978-3-031-17258-8_2.

Rehm, Georg, ed. (2023a). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg (2023b). "European Language Grid: Introduction". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 1–10.

Rehm, Georg, Federico Gaspari, German Rigau, Maria Giagkou, Stelios Piperidis, Annika Grützner-Zahn, Natalia Resende, Jan Hajic, and Andy Way (2022). "The European Language Equality Project: Enabling digital language equality for all European languages by 2030". In: *The Role of National Language Institutions in the Digital Age – Contributions to the EFNIL Conference 2021 in Cavtat*. Ed. by Željko Jozić and Sabine Kirchmeier. Budapest, Hungary: Nyelvtudományi Kutatóközpont, Hungarian Research Centre for Linguistics, pp. 17–47.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg etc.: Springer. http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

Rehm, Georg, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvalds-

son, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiș, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite (2016). "The Strategic Impact of META-NET on the Regional, National and International Level". In: *Language Resources and Evaluation* 50.2, pp. 351–374. DOI: 10.1007/s10579-015-9333-4. http://link.springer.com/article/10.1007/s10579-015-9333-4.

Romaine, Suzanne (2017). "Language Endangerment and Language Death". In: *The Routledge Handbook of Ecolinguistics*. Abingdon, Oxfordshire: Routledge, pp. 40–55. DOI: 10.4324/9781315687391.ch3. https://www.routledgehandbooks.com/doi/10.4324/9781315687391.ch3.

Simons, Gary F., Abbey L. Thomas, and Chad K. White (2022). *Assessing Digital Language Support on a Global Scale*. DOI: 10.48550/ARXIV.2209.13515. https://arxiv.org/abs/2209.13515.

Soria, Claudia, Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, and Nicoletta Calzolari (2012). "The FLaReNet Strategic Language Resource Agenda". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), pp. 1379–1386. http://www.lrec-conf.org/proceedings/lrec2012/summaries/777.html.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. https://data.europa.eu/doi/10.2861/136527.

Yvon, François and Viggo Hansen (2010). "iTranslate4.eu: Internet translators for all European languages". In: *Proceedings of the 14th Annual conference of the European Association for Machine Translation, EAMT 2010, Saint Raphaël, France, May 27-28, 2010*. European Association for Machine Translation. https://aclanthology.org/2010.eamt-1.41/.

Appendix

Feature	Value	Weight
Resource Type	Corpus	5
	Lexical conceptual resource	1.5
	Language description	3.5
Subclass	Raw corpus	0.1
	Annotated corpus	2.5
	Computational lexicon	2
	Morphological lexicon	3
	Terminological resource	3.5
	Wordnet	4
	Framenet	4
	Model	5
	Each of the others (there are 15 more)	0.5
Linguality Type	Multilingual	5
	Bilingual	2
	Monolingual	1
Media Type	Text	1
	Image	3
	Video	5
	Audio	2.5
	Numerical text	1.75
Annotation Type	*Each of these – can be combined in a single LR*	0.25
Domain	*Each of these – can be combined in a single LR*	0.3
Conditions of Use	Other specific restrictions	0.5
	Commercial uses not allowed	1
	No conditions	5
	Derivatives not allowed	1.5
	Redistribution not allowed	2
	Research use allowed	3.5

Table 1 Weights assigned to the technological factors of the DLE Metric for language resources

Feature	Value	Weight
Language Independent	False	5
	True	1
Input Type	Input text	2
	Input audio	5
	Input image	7.5
	Input video	10
	Input numerical text	2.5
Output Type	Output text	2
	Output audio	5
	Output video	10
	Output image	7.5
	Output numerical text	2.5
Function Type	Text processing	3
	Speech processing	10
	Information extraction and information retrieval	7.5
	Translation technologies	12
	Human-computer interaction	15
	Natural language generation	20
	Support operation	1
	Image/video processing	13
	Other	1
	Unspecified	1
Domain	*Each of these – can be combined in a single tool*	0.5
Conditions of Use	Unspecified	0
	Other specific restrictions	0.5
	No conditions	5
	Commercial uses not allowed	1
	Derivatives not allowed	1.5
	Redistribution not allowed	2
	Research use allowed	3.5

Table 2 Weights assigned to the technological factors of the DLE Metric for tools and services

Factor	Merging of the Scores	Conversion from Text to Scores
Public funding available for LTs	Adding up scores for each country	1 for regional funding 1 for national funding 1 for intranational funding 1 for ESIF 1 for EUREKA 1 for EUROSTAT
Legal status and legal protection	Adding up scores per language	10 for statutory national language 10 for de facto national working language 2 for statutory provincial language 2 for statutory provincial working language 1 for recognised language
Publicly available media outcomes	Adding up two scores: one score for language transfer practices for cinema works screened and one for television works broadcast	2 for dub 1.5 for voice over 1.5 for sub and dub 1 for sub
	Adding up scores + division by the number of answers	Broadcast in original language: 5 for mostly/always, 2.5 for sometimes Broadcast with dubbing: 4 for mostly/always, 2 for sometimes Broadcast in original language with voice-over: 3 for mostly/always, 1.5 for sometimes Dual-channel sound: 2 for mostly/always, 1 for sometimes Broadcast with subtitles: 1 for mostly/always, 0.5 for sometimes
Presence of local, regional or national strategic plans	One of the scores per country	1 for no plan/strategy 2 for a plan without mentioning LT 3 for a plan mentioning LT 4 for a plan mentioning LT and minority and regional languages
Political activity	Adding up scores per country	1 score for each document 1 score for each document mentioning LT 2 for each document exclusively about LT 1 for a document covering a specific language 2 for each document published 2020/2021 1 for each document published 2019/2018

Table 3 Contextual factors: Conversion from plain text into scores

ECONOMY	
Factor	**Indicator**
Size of the economy	Annual GDP
	GDP per capita* **
Size of the LT/NLP market	LT market in million Euro
Size of the language service, translating or interpreting market	Number of organisations from the industry in the ELG catalogue* **
Size of the IT/ICT sector	Perc. of the ICT sector in the GDP* **
	ICT service exports in balance of payment* **
Investment instruments into AI/LT	GDE on R&D in relevant areas*
Regional/national LT market	No indicator found
Average socio-economic status	Annual net earnings, 1.0 FTE worker* **
	Life expectancy at age 60**

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 4 Contextual factors: Proposed factors for class "Economy"

EDUCATION	
Factor	**Indicator**
Higher Education Institutions operating in the language	No indicator found
Higher education in the language	No indicator found
Academic positions in relevant areas	Head count of R&D personnel
Academic programmes in relevant areas	No indicator found
Literacy level	Literacy rate*
Students in language/LT/NLP curricula	Total no. of students in relevant areas* **
Equity in education	Proportional tertiary educ. attainment* **
Inclusion in education	Percentage of foreigners attaining tertiary education* **

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 5 Contextual factors: Proposed factors for class "Education"

FUNDING	
Factor	**Indicator**
Funding available for LT research projects	No. of projects funded in relevant areas*
	Score from the national funding programmes
Venture capital available	Venture capital amounts in Euro
Public funding for interoperable platforms	Number of platforms**

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 6 Contextual factors: Proposed factors for class "Funding"

INDUSTRY	
Factor	**Indicator**
Companies developing LTs	No. of enterprises in the ICT area* **
Start-ups per year	Percentage of "Enterprise births"**
Start-ups in LT/AI	Number of AI start ups* **

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 7 Contextual factors: Proposed factors for class "Industry"

LAW	
Factor	**Indicator**
Copyright legislation and regulations	No indicator found
Legal status and legal protection	Scores out of the legal status* **

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 8 Contextual factors: Proposed factors for class "Law"

MEDIA	
Factor	**Indicator**
Subtitled or dubbed visual media	Scores out of language transfer practices* Scores out of answers about broadcast practices
Transcribed podcasts	Number of entries in the CBA*

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 9 Contextual factors: Proposed factors for class "Media"

ONLINE	
Factor	**Indicator**
Digital libraries	Percentage of contribution to Europeana
Impact of language barriers on e-commerce	Percentage of population buying cross-border**
Digital literacy	No indicator found
Wikipedia pages	Number of articles in Wikipedia* **
Websites exclusively in the language	No indicator found
Websites in the language (not exclusively)	Perc. of websites in the languages* **
Web pages	No indicator
Ranking of websites delivering content	12 selected websites supporting the languages
Labels and lemmas in knowledge bases	Number of lexemes in Wikipedia* **
Language support gaps	Language matrix of supported features*
Impact on E-commerce websites	T-Index*

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 10 Contextual factors: Proposed factors for class "Online"

POLICY	
Factor	**Indicator**
Presence of strategic plans, agendas, etc.	Scores out of a list of the published national AI strategies Scores from questionnaire about strategies
Promotion of the LR ecosystem	No indicator found
Consideration of bodies for the LR citation	No indicator found
Promotion of cooperation	No indicator found
Public and community support for resource production best practices	No indicator found
Policies regarding BLARKs	No indicator found
Political activity	Scores out of the list of documents

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 11 Contextual factors: Proposed factors for class "Policy"

PUBLIC ADMINISTRATION	
Factor	**Indicator**
Languages of public institutions	No. of constitutions written in the language
Available public services in the language	Percentage of a maximum score about digital public services**
	Score for digital public services**

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 12 Contextual factors: Proposed factors for class "Public administration"

RESEARCH & DEVELOPMENT & INNOVATION	
Factor	**Indicator**
Innovation capacity	Innovation Index* **
Research groups in LT	Number of research organisations
Research groups/companies predominantly working on the respective language	No indicator found
Research staff involved in LT	No indicator found
Suitably qualified Research staff in LT	No indicator found
Capacity for talent retention in LT	No indicator found
State of play of NLP/AI	No indicator found
Scientists working in LT/on the language	Number of researchers in relevant areas*
Researchers whose work benefits from LRs and LTs	No indicator found
Overall research support staff	Head count of research support staff* **
Scientific associations or general scientific and technology ecosystem	No indicator found
Papers about LT and or the language	Number of papers about LT**
	Number of papers about the language* **

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 13 Contextual factors: Proposed factors for class "Research & Development & Innovation"

SOCIETY	
Factor	**Indicator**
Importance of the language	No indicator found
Fully proficient (literate) speakers	Number of L1 speakers*
Digital skills	Perc. of individuals with basic digital skills* **
Size of language community	Total number of speakers* **
Population not speaking the official language(s)	No indicator found
Official or recognized languages	Total no. of languages with official status*
	Number of bordering languages
Community languages	Number of community languages*
Time resources of the language community	No indicator found
Society stakeholders for the language	No indicator found
Speakers' attitudes towards the language	Total number of participants wanting to acquire the language
Involvement of indigenous peoples	No indicator found
Sensitivity to barriers	No indicator found
Usage of social media or networks	Total number of social media users* **
	Percentage of social media users* **

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 14 Contextual factors: Proposed factors for class "Society"

TECHNOLOGY	
Factor	**Indicator**
Open-source technologies of LTs	No indicator found
Access to computer, smartphone etc.	Perc. of households with a computer* **
Digital connectivity and internet access	Perc. of households with broadband* **

*Indicator marked * is automatically updateable – Indicator marked ** provides good quality data*

Table 15 Contextual factors: Proposed factors for class "Technology"

Class	Factor
Economy	Size of economy
	Size of the ICT sector
Education	Students in LT/language
	Inclusion in education
Industry	Companies developing LTs
Law	Legal status and legal protection
Online	Wikipedia pages
R & D & I	Innovation capacity
	Number of papers
Society	Size of language community
	Usage of social media
Technology	Digital connectivity, internet access

Table 16 Contextual factors included in the final configuration (configuration 5)

Appropriate	Ranked too high	Ranked too low	Contrary Opinion
English	Irish	Norwegian	French
Dutch	Italian	Spanish	German
Danish	Swedish	Portuguese	Saami, Northern
Polish	Hungarian	Czech	Latvian
Greek	Croatian	Romanian	
Finnish	Maltese	Bulgarian	
Estonian	Faroese	Icelandic	
Slovene	Scottish Gaelic	Emilian	
Slovak	Cornish	Sicilian	
Lithuanian	Manx		
Serbian	Saami, Southern		
Basque	Saami, Pite		
Catalan	Saami, Lule		
Galician	Saami, Skolt		
Asturian	Saami, Inari		
Aragonese	Sardinian		
Welsh	Romagnol		
Griko			
Lombard			
Ligurian			
Venetian			
Southern Italian			
Friulian			
Piemontese			
Ladin			
25	17	9	4

Table 17 Contextual factors: Assessment of the languages in the final configuration (configuration 5) by the panel of experts

Chapter 4
European Language Technology in 2022/2023

Maria Giagkou, Teresa Lynn, Jane Dunne, Stelios Piperidis, and Georg Rehm

Abstract This chapter presents the results of an extensive empirical investigation of the digital readiness of European languages, and provides a snapshot of the support they are offered through technology as of 2022. The degree of digital readiness was assessed on the basis of the availability of language resources and technologies for each language under investigation and a cross-language comparison was performed. As a complementary approach, the perspectives and opinions of LT users, developers and the regular citizen were acquired in order to fully understand the EU's LT landscape. Both the objective empirical findings and the voice of the community clearly indicate that there is an extreme imbalance across languages when it comes to the individual levels of technological support. Although the LT field as a whole has demonstrated remarkable progress during the last decade, this progress is not equally evidenced across all languages, posing, more acutely than ever before, a threat of digital extinction for many of Europe's lesser supported languages.[1]

1 Introduction

More than ten years ago, the study "Europe's Languages in the Digital Age" concluded that most European languages are under threat in the digital age. The study, prepared by more than 200 experts and documented in 32 volumes of the META-NET White Paper Series (Rehm and Uszkoreit 2012), assessed Language Technology (LT) support for each language in four different areas: automatic translation,

Maria Giagkou · Stelios Piperidis
R. C. "Athena", Greece, mgiagkou@athenarc.gr, spip@athenarc.gr

Teresa Lynn · Jane Dunne
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie, jane.dunne@adaptcentre.ie

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

[1] This chapter includes findings from Way et al. (2022) and makes use of the general sections written by the ELE consortium for the language reports (Giagkou et al. 2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_4

speech interaction, text analysis and the availability of language resources (LRs). The results were alarming: most of the 32 European languages investigated were evaluated as severely under-resourced and some almost completely neglected.

During the last ten years since the publication of the META-NET White Papers, the LT field as a whole has seen remarkable progress. In particular, the advent of data-driven approaches such as deep learning and neural networks, together with the considerable increase in the number and quality of LRs for a number of languages, have yielded previously unforeseeable results. However, is this remarkable progress equally evidenced across all languages, or is the gap between "big" and "small" languages documented in 2012 still present in 2022/2023?

The question of whether languages can be considered digitally equal has become increasingly relevant in recent years, with a growing number of studies attempting to quantify digital readiness and compare languages in this respect. Methods have varied, with some assessing the level of technology support based on mentions of a language at NLP publication venues or language resource catalogues (e. g., Blasi et al. 2022; Joshi et al. 2020; Ranathunga and Silva 2022) or on websites describing LT tools and services (e. g., Simons et al. 2022). However, the overall conclusion is always the same; from a technological perspective, there is a striking imbalance across languages in terms of support, and it is clear that not all languages benefit equally and fairly from the overall progress in LT advances.

In the ELE project, we took an empirical approach to quantifying digital readiness of a language and providing an evidence-based grounding on which languages can be compared. We started by applying the Digital Language Equality (DLE) Metric (see Chapter 3) to examine both the current state of technology support and the potential for short- and mid-term development of LT (Section 2). We continued with a quantitative investigation of the various perspectives and dimensions of current technological support, as this is reflected in the Language Resources and Technologies (LRTs) collection of the European Language Grid (ELG, Rehm 2023). The results of this empirical assessment were then supplemented by surveys and consultations with a broad representation of LT developers and LT users and consumers, who provided feedback and insight as to their experiences with LTs for EU languages (Section 3). Furthermore and most importantly, we focused on a large number of European languages and provided updates of the META-NET White Papers in the form of the ELE Language Reports (Giagkou et al. 2022), condensed versions of which are presented in Chapters 5–37. It is only through such a holistic examination that a clear picture of the current status and future prospects of DLE can be gained.

2 How Do Europe's Languages Compare?

In this section, we first describe our source of evidence and methodology (Section 2.1), followed by a presentation of our findings (Section 2.2).

2.1 Source of Evidence and Methodology

To compare the level of technology support across languages, we considered the language technology tools and resources in the catalogue of the European Language Grid (Rehm 2023; Piperidis et al. 2023; Labropoulou et al. 2020). The comparative evaluation was performed on various dimensions.

- The current state of technology support, as indicated by the availability of tools and services[2] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (NLG, e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing
  - Human-computer interaction (HCI, e. g., tools for conversational systems)
- The potential for short- and mid-term development of LTs, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Language models
  - Lexical resources (incl. dictionaries, wordnets, ontologies, etc.)

We measured the LT support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently, each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. *Weak or no support:* the language is present (as content, input or output language) in <3% of the ELG resources of the same type

[2] Tools tagged as "language independent" without mentioning any specific language are *not* taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

2. *Fragmentary support:* the language is present in ≥3% and <10% of the ELG resources of the same type
3. *Moderate support:* the language is present in ≥10% and <30% of the ELG resources of the same type
4. *Good support:* the language is present in ≥30% of the ELG resources of the same type

The thresholds for defining the four bands (i. e., 3%, 10% and 30%) were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters were then used to define the bands per application area and resource type. The overall level of support for a language was calculated based on the average coverage of all dimensions investigated.

The ELG platform harvests several major LR/LT repositories[3] and, on top of that, more than 6,000 additional LRTs were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions. At the time of investigation, the ELG catalogue comprised more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages, both official and regional as well as minority ones.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain categories of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information provided on corpus size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records are expected to improve over time.

For the purposes of a high-level comparison, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change as it reflects a snapshot of the available resources at the time of investigation.

That said, we consider the current status of the ELG catalogue and the higher-level findings below representative with regard to the current existence of LT resources for Europe's languages.

[3] At the time, ELG harvested ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and the datasets section of Hugging Face (Labropoulou et al. 2023).

2.2 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG catalogue. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at the national or regional level in at least one European country and other minority and lesser spoken languages,[4] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold of the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises these findings.

Looking into particular dimensions of data availability, it is evident that an abundance of training data for developing LTs is available only for a few languages with high commercial interest. For many (the majority of) European languages, this is not the case and only corpora which are minuscule in comparison to English are available. When investigating the current availability of some of the data types mentioned in the previous paragraph, as represented in the resources hosted in ELG in January 2023,[5] it is apparent that even the best-supported languages in this dimension, Spanish and English, are still only moderately covered (Figure 2). With respect to multimodal data, all languages with the exception of English are weakly covered, with some, e. g., Maltese and Luxembourgish, severely underrepresented (Figure 3).

Although the data gaps per language are different, some data types are particularly sparse across many languages. These include: large language models, both monolingual and multilingual; multimodal data, especially speech in conversational settings (dialogues) from speakers of different ages, genders and linguistic/dialectal backgrounds, but also video corpora for sign languages; domain-specific data (e. g., medical, legal or media among many others of interest); data for language use on

[4] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser and Yiddish. The scores for all of these languages are very low, placing all of them in the *weak or no support* group.

[5] The DLE dashboard enables more fine-grained comparisons. It dynamically visualises the contents of the ELG catalogue and offers an up-to-date snapshot of the current availability of LRTs (see Chapter 3): https://live.european-language-grid.eu/catalogue/dashboard.

Table 1 State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

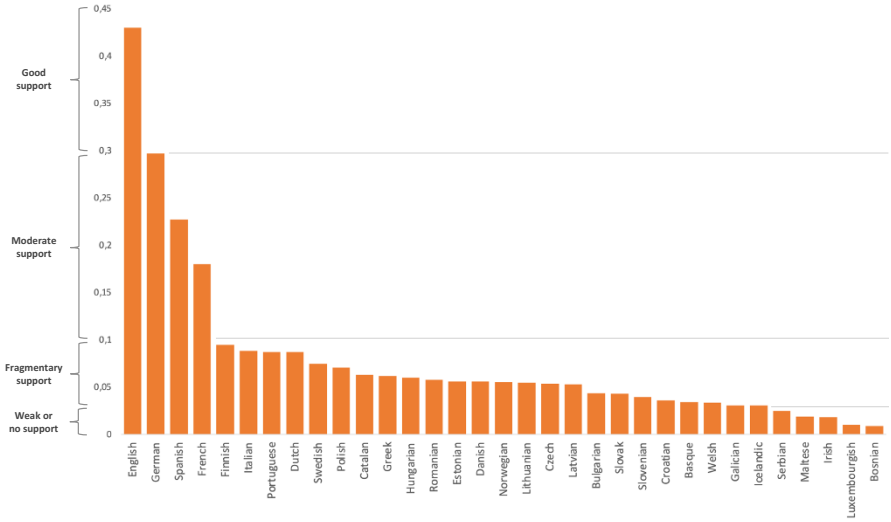

Fig. 1 Overall state of technology support for selected European languages (2022)

social media; semantic resources (e. g., semantic annotations and knowledge bases); data for language pathologies; benchmarks, i. e., well-designed gold-standard corpora for evaluating LT systems or fine-tuning language models.

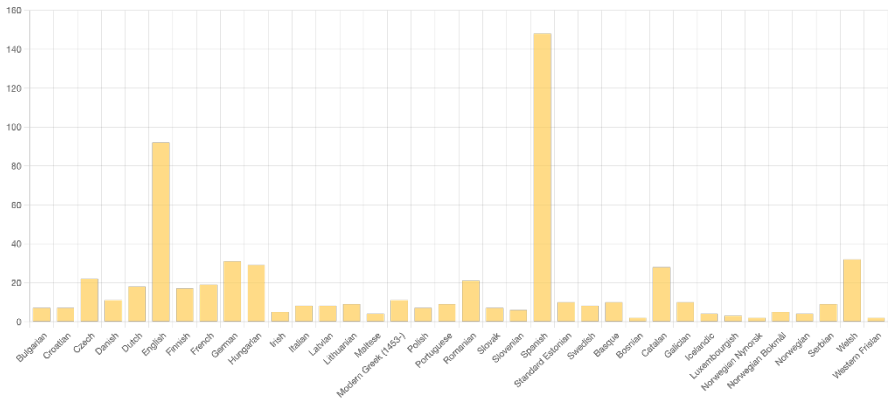

Fig. 2 Number of language models available in the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones (as of January 2023)

Similarly to data, the identified gaps for technologies are very diverse across languages. While overall LTs for English are numerous and at the state-of-the-art level, a number of very small minoritised languages lack even basic tools such as spell checkers. In the worst case, they are not even supported by operating systems. Nevertheless, there seems to be a generalised consensus that, when it comes to languages

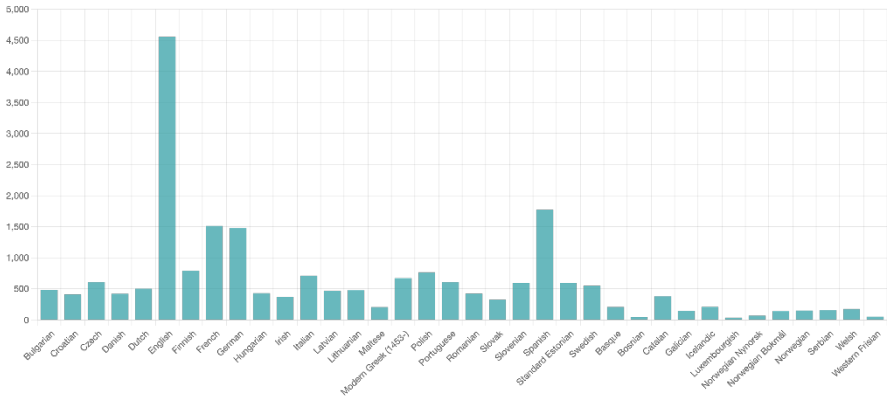

Fig. 3 Number of multimodal datasets (i. e., media type: audio, video or image) available in the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones (as of January 2023)

for which at least a minimum level of technological support has been achieved, the technologies most urgently needed include: discourse processing, bias detection and anonymisation, conversational systems and question-answering in the wider context of HCI, NLG (with summarisation mentioned frequently) and Natural Language Understanding (NLU), e. g., even English and German are currently supported by less than 100 HCI or NLG systems on ELG, while some languages like Bosnian and Norwegian Nynorsk are not supported at all (Figures 4 and 5).

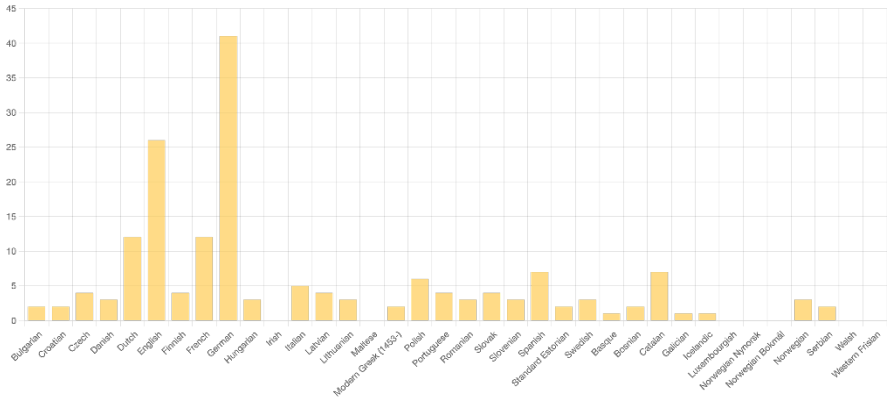

Fig. 4 Number of Human-Computer Interaction systems described in the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones (as of January 2023)

The results of this analysis are only informative of the relative positioning of languages, but not of the technological progress achieved by a specific language. The

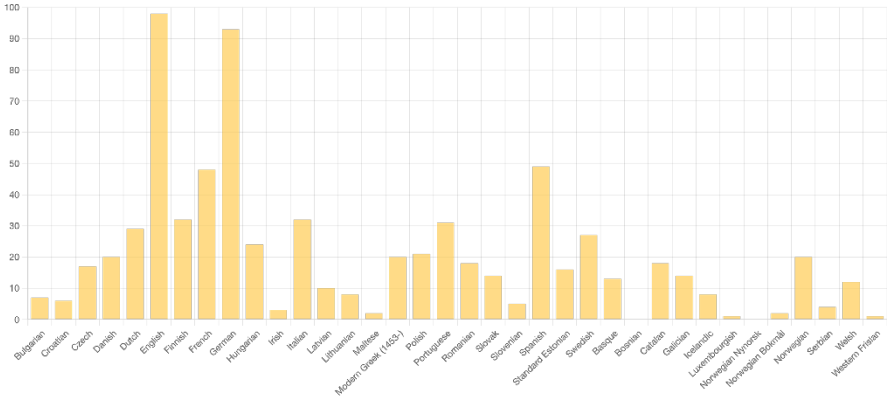

Fig. 5 Number of Natural Language Generation systems described in the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones (as of January 2023)

LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of LRTs. It is at the same time undebatable that the technology requirements for a language to be considered digitally supported by today's standards have changed significantly in the last ten years (e. g., the prevalent use of virtual assistants, chatbots, improved text analytics capabilities, etc.). Nevertheless, the imbalance in distribution across languages which was documented in the META-NET White Papers in 2012 still exists, and the huge distance between the best supported languages and the minimally supported ones was still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, or at least reduced, in order to move towards DLE and avert the risks of digital language extinction.

It should be noted that this analysis does not include a fifth level, *excellent support*, for the grouping of languages, in addition to the four levels described in Section 2.1. Currently, no European language, not even English, is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language. Although recently there have been many breakthroughs in AI, Computer Vision, Machine Learning and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself on request, and be effected as required on the fly and at scale. A language can only be considered excellently supported by technology if and when the goal of Deep Natural Language Understanding has been reached.

3 The Voice of the Community

The findings in Section 2 are extremely valuable in terms of highlighting the status quo across Europe with respect to LT support. However, facts and figures alone cannot paint the full picture. The perspectives and opinions of LT users, developers and the average citizen were also required in order to fully understand the EU's LT landscape. As a project from the community for the community, the ELE consortium wanted to ensure that as many voices as possible were heard and taken as input for the ELE strategic agenda and roadmap.

A broad spectrum of stakeholders was consulted to achieve this wider insight into the levels of LT support across European languages (also see Chapter 38, p. 229 ff.). We distinguish between three main stakeholder groups: *LT developers* (industry and research), *LT users* (commercial and academic users) and *EU citizens*, i. e., the general public who use and consume LTs in everyday personal and professional settings, often without even realising it. Each group is diverse, some including many subgroups, representing a variety of sectors and domains. For the latter, we looked at the interesting subdivisions of commercial and academic users as well as EU citizens. The first two groups are represented in the ELE consortium with several networks, initiatives and associations, representing the views of their constituencies, highlighting their wishes, demands and needs towards full DLE in Europe.

Further insight was gained from a number of online surveys and expert interviews targeting LT developers, users and consumers. The surveys investigated language coverage, evaluated the current situation of LT in Europe and encouraged participants to share their predictions and visions for the future. In this section, we look, in particular, at the evaluation of the current situation to see how these opinions compare to the empirical results presented in Section 2 and also in Chapter 39 (p. 245 ff.).

3.1 Developers of Language Technologies

European LT developers are a diverse group of stakeholders, comprising *academic* and *industrial entities* in the field of LT. Beyond research, they develop pre-commercial prototypes, algorithms, applications and systems. An initial grouping is, thus, *LT industry* and *LT research* (also see Rehm et al. 2023, 2020). This section focuses on their view about the situation as of 2022, while Section 3 in Chapter 38 presents their forward-looking predictions going towards 2030.

In addition to the horizontal grouping into research and industry, a vertical categorisation can be performed with regard to the multi- and interdisciplinary nature of LT. LT is in the intersection of Linguistics and Computational Linguistics, Computer Science and AI, while at the same time encompassing methods and findings from Cognitive Science and Psychology, Mathematics, Statistics, Philosophy and other fields. As a result, the ELE stakeholder group of LT developers were identified not only within the strict limits of *LT* per se, but also in the neighbouring disciplines of *AI* and *Digital Humanities/Social Science and Humanities* (DH/SSH).

Europe has a long-standing research, development and innovation tradition in LT with over 800 centres performing excellent, highly visible and internationally recognised research on all European and many non-European languages. In terms of companies, the European LT industry was estimated to comprise 435 companies (LT-Innovate 2016) or 473 LT vendors in the EU26 plus Iceland and Norway in 2017 (Vasiljevs et al. 2019). In January 2023, the ELG catalogue comprised more than 800 commercial entities including integrators and a certain number of user companies.

In order to disseminate the survey widely, we mobilised existing European networks, associations, initiatives and projects. Some of the well-established and long-standing pan-European LT networks were represented in the ELE consortium and they constituted the core ELE LT developers stakeholders groups (i. e., CLAIRE, CLARIN, LT-Innovate, META-NET and ELG). The ELE partners that represented these initiatives not only contributed their views to the project but also facilitated access to and elicitation of the views of their constituency and members. In particular, they coordinated the distribution of the survey to their members, conducted interviews and focused consultation meetings, where needed and appropriate, and consolidated their feedback (Thönnissen 2022; Eskevich and Jong 2022; Rufener and Wacker 2022; Hajič et al. 2022; Hegele et al. 2022).

The survey encompassed 45 questions in total. A respondent was presented with 32 (minimum) to 45 (maximum) questions, including "if other" questions. In all, 35 questions were mandatory and 27 were closed questions (single or multiple choice). The survey was structured into four main parts: Part A. Respondents' profiling, Part B. Language coverage, Part C. Evaluation of current situation, and Part D. Predictions and visions for the future (see also Chapter 38, p. 229 ff., and Chapter 39, p. 245 ff.). For assessing the current situation from the perspective of LT developers, we focus on the findings based on responses to Parts B and C of the survey.

The LT developers survey was filled in by 321 different respondents who represent 223 different organisations (Way et al. 2022). 73% of the organisations are research or academic institutions and 22% are private companies. In 5% of responses the "Other" value was indicated as the type of organisation and this has been further specified as freelancer/private practitioner or currently unemployed, government agency, not-for-profit organisation, etc. Of note here is the response to the question "*What languages does your organisation conduct research in and/or for what languages do you offer services, software, resources, models etc.?*". Figure 6 shows the languages supported by survey respondents' organisations. All official EU languages are covered as well as other state official, regional and/or co-official European languages. The five most frequently mentioned languages are, yet again, English, German, Spanish, French and Italian.

In order to evaluate the current situation and to further grasp the main challenges and obstacles the European LT community faces, the survey participants were asked to indicate their level of agreement with a set of potential obstacles (Figure 7). As part of a free text question, respondents were also given the opportunity to elaborate on the obstacles and challenges indicated in the questions and/or add any other obstacle/challenge not previously listed.

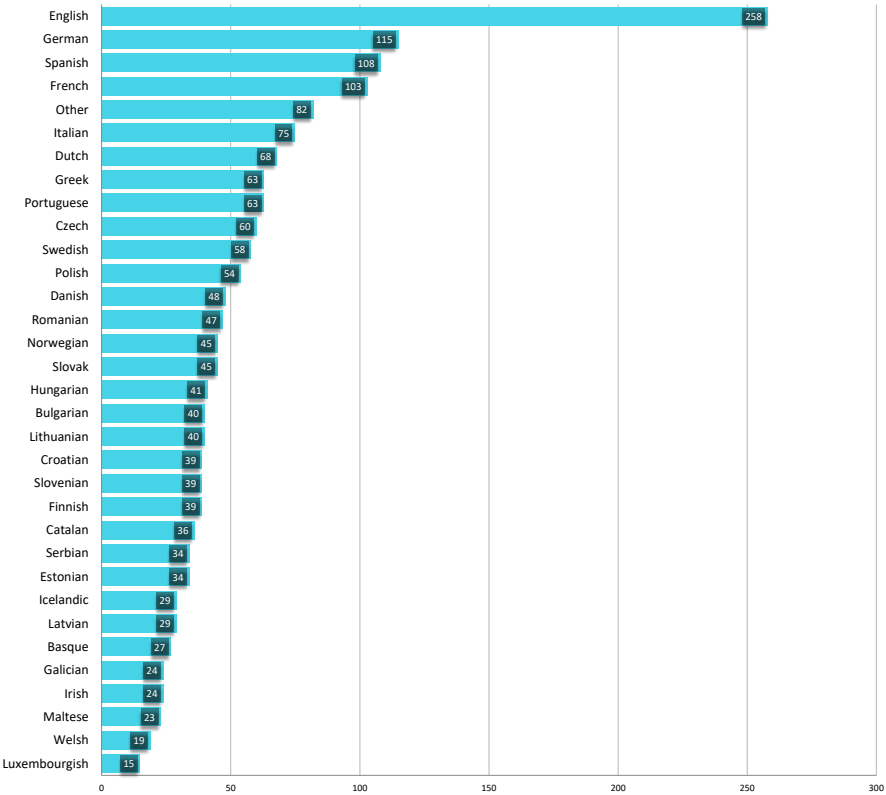

Fig. 6 LT developers survey – languages supported by the respondents' organisations in their research and development activities

With respect to questions about the status quo of the languages, most of the participants agreed or strongly agreed that the importance of multilinguality in the European landscape does not always receive adequate recognition, and the smaller languages appear not to be attractive enough for industry and investors (74% agreed or strongly agreed on this point). This was backed up by comments relating to how industrial players can find a commercial interest in pre-competitive investments for "larger" languages, while this will rarely be the case for "smaller" ones. It was suggested that in that situation, the role of additional investors for the development of LTs for "smaller" languages should be played by bodies either at national or EU level. Moreover, it was noted that it is very often the case that small languages can rely on public funding only, which however is considered insufficient. For this reason, it was argued that public investments for small languages are necessary on a larger scale to really make them available to the wider community. It was also observed that the cost of developing LTs for a language is usually constant, regardless of the number of speakers of that language. Furthermore, for languages with larger numbers of speakers, it can often be easier to collect LRs: for instance, the larger

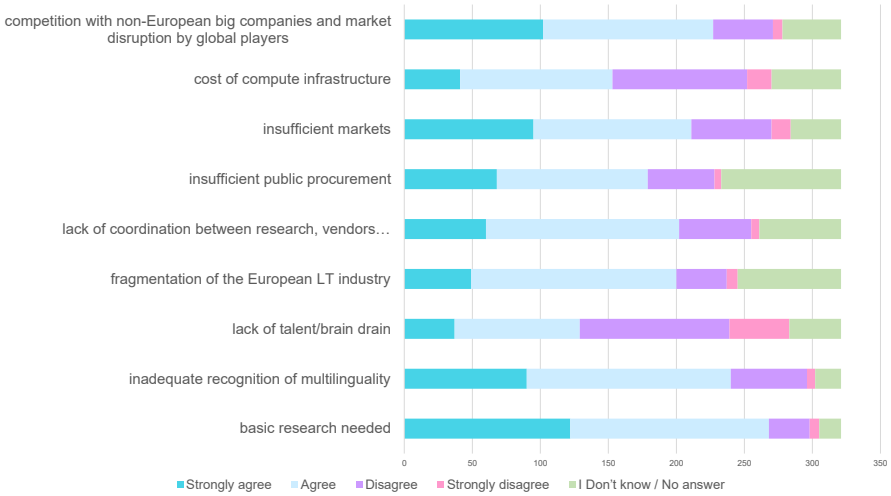

Fig. 7 LT developers survey – challenges the European LT community currently faces, according to LT developers

the number of speakers, the more online content is produced, which in turn can be collected and provide the raw language data necessary for the development of LRTs.

It was reported that this situation was even worse for non-standard languages: local dialects, non-standard written language on social media platforms, non-standard language for speech recognition, and non-standard language as used by migrants or citizens with a migration background. There is hardly ever funding available for creating LRs for non-standard varieties. There is equally little incentive for researchers to publish their work on small languages, resulting in the dominance of the English language in scientific literature.

3.2 Users of Language Technologies

Commercial users were those respondents representing companies in the sector of Information and Communication Technologies (ICTs) and eCommerce (e. g., Megabyte Ltd, A Capela group, Telecats), energy (e. g., Shell, Menai Science Park Ltd) and business services (e. g., Spencer Stuart, Inuits, Projectus grupa). They also included respondents from the following groups: self-employed language professionals (e. g., translators); professionals working on different economic sectors (e. g., banking, health); independent professionals/consultants; professionals working in public administration; media and publishing professionals.

Academic users included researchers, data scientists, university professors, language teachers, lecturers, and Master's and PhD students. Some non-governmental organisations (NGOs) were also represented in the survey, such as Federal Lezghin

National and Cultural Autonomy, and representatives of public administration, such as National Youth Service (Ministry of Education, Children and Youth, Luxembourg), Hungarian National Research, Development and Innovation Office and the Government of the Balearic Islands. In addition, Wikipedia partners collected responses from representatives of the various Wikipedia projects, such as Wikimedia Community User Group Malta, Wikimedia Hungary, Wikimedia UK, and Wikimedia Community Ireland, to name a few. The full list of stakeholders of the LT users and consumers survey is presented in Way et al. (2022).

Six well-known European initiatives disseminated the survey within their networks and produced one report each, based on their respective constituencies. These include the European Federation of National Institutions for Language (EFNIL, Kirchmeier 2022), the European Language Equality Network (ELEN, Hicks 2022), the European Civil Society Platform for Multilingualism (ECSPM, Gísladóttir 2022), the New European Media initiative (NEM, Hrasnica 2022), the Association of European Research Libraries (LIBER, Blake 2022) and Wikipedia (Heuschkel 2022).

The survey obtained a total of 246 responses. The results show that contributions came from a diverse range of economic sectors and professional activities, but most of the respondents worked in the education and research sector with 130 responses (53%) out of 246, that is, most respondents were researchers, university professors, assistant professors, lecturers or held other academic positions. The survey was also filled out by representatives of NGOs, large enterprises, SMEs, government departments and independent contractors and consultants in diverse economic sectors. The 15 (6%) respondents who selected the option "other" represented non-governmental bodies, non-profit organisations, public sector organisations, social organisations and independent government departments.

Of relevance to assessing the current situation, we note here the responses to the question "*In general terms, how do you evaluate the performance of the tools you use for the official European language(s) you work with*". Responses were captured through a 4-point Likert scale (where 1 indicated *very poor support*, 2 *poor support*, 3 *good support* and 4 *excellent support*). The list of LTs evaluated can be seen in Way et al. (2022). Figure 8 shows the average score for each of the European languages evaluated. The results show striking differences in technological support between European languages. Unsurprisingly, English is very well supported with a mean score of 3.4, while the group formed by German, French and Spanish follows with a mean score between 2.4 and 2.5. All other European languages were considered to have either *poor support* (mean scores ranging from 1 to 1.3), *very poor support* or *no support* at all with scores below 1.

3.3 European Citizens as Consumers of Language Technologies

In addition to the consultation with stakeholders that represent communities of users and consumers, a survey targeting European citizens was carried out to make sure that their voices also play a decisive role in the pursuit of full DLE in Europe. This

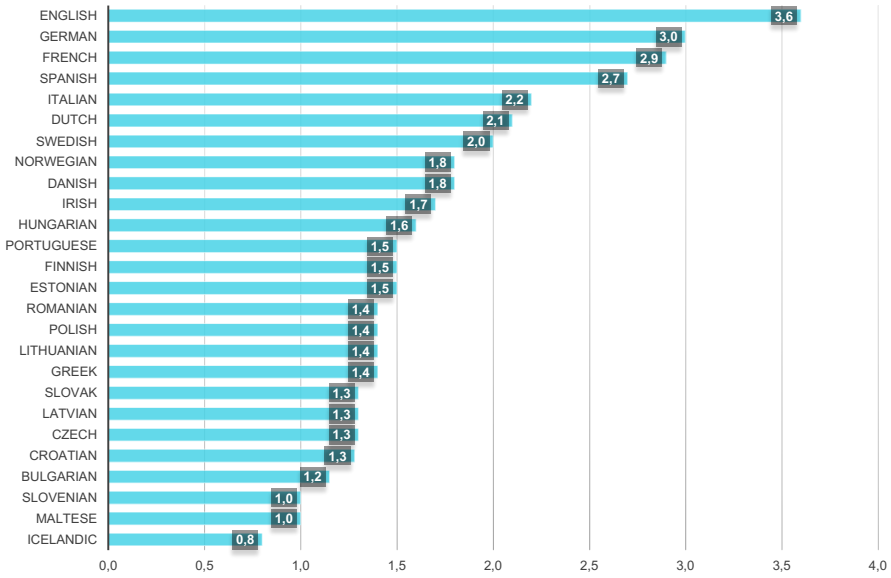

Fig. 8 LT users survey – level of technological support: average scores for the European language(s) that respondents work with

consultation with a larger and more diverse cohort of consumers allowed us to obtain a more accurate picture of the current scenario in terms of LT support across European languages and have a more representative basis for a technological and scientific forecasting on how LTs can be deployed and applied in Europe by 2030.

The citizens' survey was launched in January 2022 and closed on 01 May 2022. It was made available in 35 languages and disseminated across 28 countries.[6] For each country we created a standalone survey so that respondents only saw the version in the language of the country in which they were based. For countries with more than one official language, we created a standalone version of the survey in each language spoken in the country, e. g., four surveys were set up in Spain (in Spanish, Catalan, Galician and Basque). This approach allowed us to specifically target regions where we were more likely to find communities of respondents that were speakers of that language. More details on this survey and the community consultation methodology are presented in Chapter 38 (p. 229 ff.).

In total, 21,108 complete responses were collected. However, as the collection of survey responses through commercial online services is known to present some known issues that can render results unreliable (Lawlor et al. 2021), closer inspection revealed a number of flags indicating unreliable responses. These responses were filtered from the dataset, and as such, a final 20,586 responses were analysed.

[6] While ELE investigated about 90 European languages, we only produced translated versions for those languages for which native speaker post-editing was available. The 35 languages covered by the survey represent the support offered through the ELE consortium members.

Respondents provided profiling questions and were asked to list all of the languages they speak. Of particular interest in our examination of the current situation is the response to question 6 "*Please rate all the types of software applications, apps, tools or devices you use for your language(s)*".

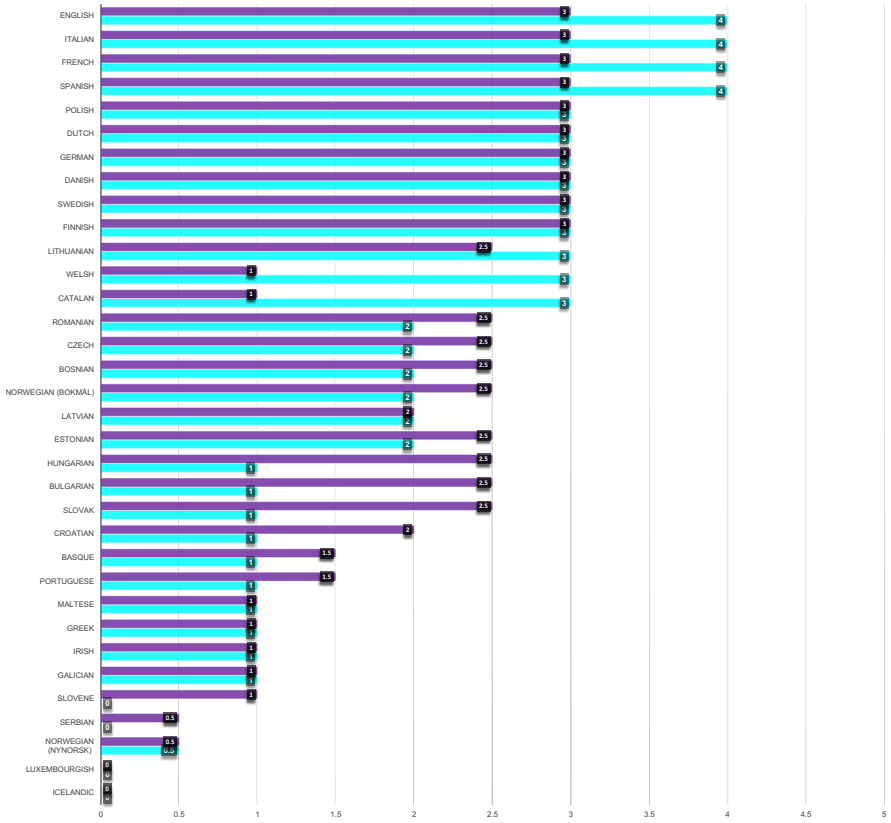

Fig. 9 EU citizens survey – responses to question 6: *Please rate all the types of software applications, apps, tools or devices you use for your language(s). Tools you do not use for your language(s) do not need to be rated.* Note that purple indicates the median and blue the mode.

The list of eight tools presented was: Search apps (e. g., Google, Bing); personal assistant apps (e. g., Siri, Alexa); proofreading apps (e. g., spelling and grammar checkers, autocorrect); translation apps (e. g., Google Translate, DeepL); automatic subtitling (e. g., news report, YouTube); language learning apps (e. g., Babbel, Rosetta Stone); chatbots (e. g., for customer support) and screen readers. The aim of this question was to understand the perception of the average EU citizen and LT user of the quality of the tools that they use for each language they speak.

The ratings were based on a 5-point Likert scale, i. e., respondents had the option of rating 1-star (*poor*) through to 5-stars (*excellent*) for each of the eight tools

presented, and for each language they had selected in the previous question. In the interest of space, Figure 9 presents only the languages for which language reports were produced (see Chapters 5–37) and only shows responses from the perspective of each language, as opposed to each tool. Due to the large size of the dataset and the varying proportion of responses for each language, the figures presented here are based on the calculation of the median score (purple) and the mode (blue). Tools that were not available or used by a respondent did not receive a score. In these instances, the tool was assigned a rating of zero, as a penalty for lesser-used tools across all languages. This explains the low scores for languages such as Serbian, Luxembourgish and Icelandic, which either have very few available or low-rated existing LTs.

To some degree, the results reflect the trend presented for the technological DLE scores of the relevant languages (see Chapter 3) in terms of the quantification of the technological factors of the DLE Metric. The difference between the median score for English and the next well-resourced languages is not as stark, however. This could be explained by the fact that the ratings of the tools are bound to an upper limit of five and as a result, the scores are "flatter" and closer to each other. On the other hand, we can see that the mode score reveals that tools for English, French, Spanish and Italian received more frequent higher ratings. Nevertheless, the results provide a clear insight into the average European user's perception of the quality of LT support for their languages.

4 Conclusions

We examined around 90 European languages with the goal of creating a snapshot of their digital readiness in 2022. We made use of the inventory of LRTs in the European Language Grid and assessed the technological readiness of each language based on the availability of LRTs. From this, we carried out a cross-language comparison on this empirical basis, as well as an analysis of feedback from developers and users of LTs across Europe, including input from over 20,000 EU citizens.

The status as analysed in 2022 is very clear: there is an extreme imbalance across languages when it comes to the individual levels of technological support. While the META-NET White Paper Series reported a similar imbalance ten years ago, what is surprising is the little comparative change seen across the board since then. The same trend of acute digital inequality continues, and worse still, the gap between English and the rest of the EU languages is getting wider. Even though some of the widely spoken languages in Europe and beyond (Spanish, German and French) have demonstrated considerable progress and are among the top performers, their distance from English is intolerable. Moreover, a striking asymmetry is evidenced between official and non-official EU or EEA languages.

Our results reiterate that digital language *inequality* poses a direct threat to Europe's linguistic and cultural diversity. Europe has become or is about to become a continent where *digital diglossia* is the de facto context for many EU citizens, with the exception of English native speakers. When going about their online lives, EU

citizens too often find it more efficient or even absolutely necessary to rely on other, more widely supported languages (predominantly English) for certain services and information because this gives them greater access to high-quality and reliable content to a broader audience, and allows them to use more advanced technologies. This is true particularly for the younger generations, thus increasing the generational language gap and bringing lesser-resourced languages ever closer to digital extinction.

References

Blake, Oliver (2022). *Deliverable D2.10 Report from LIBER*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LIBER.pdf.

Blasi, Damian, Antonios Anastasopoulos, and Graham Neubig (2022). "Systematic Inequalities in Language Technology Performance across the World's Languages". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5486–5505. DOI: 10.18653/v1/2022.acl-long.376. https://aclanthology.org/2022.acl-long.376.

Eskevich, Maria and Franciska de Jong (2022). *Deliverable D2.3 Report from CLARIN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLARIN.pdf.

Giagkou, Maria, Stelios Piperidis, Georg Rehm, and Jane Dunne, eds. (2022). *ELE Language Report Series*. Project deliverables; European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/deliverables/.

Gísladóttir, Guðrún (2022). *Deliverable D2.7 Report from ECSPM*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ECSPM.pdf.

Hajič, Jan, Tea Vojtěchová, and Maria Giagkou (2022). *Deliverable D2.5 Report from META-NET*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-META-NET.pdf.

Hegele, Stefanie, Katrin Marheinecke, and Georg Rehm (2022). *Deliverable D2.6 Report from ELG*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELG.pdf.

Heuschkel, Maria (2022). *Deliverable D2.12 Report from Wikipedia*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-Wikipedia.pdf.

Hicks, Davyth (2022). *Deliverable D2.9 Report from ELEN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELEN.pdf.

Hrasnica, Halid (2022). *Deliverable D2.11 Report from NEM*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-NEM.pdf.

LT-Innovate (2016). *The LT-Innovate Innovation Agenda*. http://www.lt-innovate.org/sites/default/files/2904-LTi_Innovation_Agenda.pdf.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. https://aclanthology.org/2020.acl-main.560.

Kirchmeier, Sabine (2022). *Deliverable D2.8 Report from EFNIL*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-EFNIL.pdf.

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva (2020). "Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3421–3430. https://www.aclweb.org/anthology/2020.lrec-1.420/.

Labropoulou, Penny, Stelios Piperidis, Miltos Deligiannis, Leon Voukoutis, Maria Giagkou, Ondřej Košarko, Jan Hajič, and Georg Rehm (2023). "Interoperable Metadata Bridges to the wider Language Technology Ecosystem". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 107–127.

Lawlor, Jennifer, Carl Thomas, Andrew T Guhin, Kendra Kenyon, Matthew D Lerner, UCAS Consortium, and Amy Drahota (2021). "Suspicious and fraudulent online survey participation: Introducing the REAL framework". In: *Methodological Innovations* 14.3. DOI: 10.1177/20597991211050467. https://doi.org/10.1177/20597991211050467.

Piperidis, Stelios, Penny Labropoulou, Dimitris Galanis, Miltos Deligiannis, and Georg Rehm (2023). "The European Language Grid Platform: Basic Concepts". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cham: Springer, pp. 13–36. DOI: 10.1007/978-3-031-17258-8_2. https://doi.org/10.1007/978-3-031-17258-8_2.

Ranathunga, Surangika and Nisansa de Silva (2022). "Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, pp. 823–848. https://aclanthology.org/2022.aacl-main.62.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Katrin Marheinecke, Rémi Calizzano, and Penny Labropoulou (2023). "Language Technology Companies, Research Organisations and Projects". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 171–185.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rufener, Andrew and Philippe Wacker (2022). *Deliverable D2.4 Report from LT-innovate*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LTInnovate.pdf.

Simons, Gary F., Abbey L. L. Thomas, and Chad K. K. White (2022). "Assessing Digital Language Support on a Global Scale". In: *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4299–4305. https://aclanthology.org/2022.coling-1.379.

Thönnissen, Marlies (2022). *Deliverable D2.2 Report from CLAIRE*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLAIRE.pdf.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI: 10.2759/142151. https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Way, Andy, Georg Rehm, Jane Dunne, Jan Hajič, Teresa Lynn, Maria Giagkou, Natalia Resende, Tereza Vojtěchová, Stelios Piperidis, Andrejs Vasiljevs, Aivars Berzins, Gerhard Backfried, Marcin Skowron, Jose Manuel Gomez-Perez, Andres Garcia-Silva, Martin Kaltenböck, and Artem Revenko (2022). *Deliverable D2.17 Report on all external consultations and surveys*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/external-consultations.pdf.

Chapter 5
Language Report Basque

Kepa Sarasola, Itziar Aldabe, Arantza Diaz de Ilarraza, Ainara Estarrona, Aritz Farwell, Inma Hernáez, and Eva Navas

Abstract Since 1968 Basque has been immersed in a process of revitalisation that has faced formidable obstacles. Nonetheless, significant progress has been made in numerous areas. The Language Technology community widely accepts the standardised language and constructs efficacious LT tools. After thirty years of collaborative work, research has resulted in state-of-the-art technology and robust, broad-coverage NLP for Basque. However, a dramatic difference remains between Basque and other European languages in terms of both the maturity of research and the state of readiness with respect to language technology solutions.

1 The Basque Language

Basque is spoken by 28.4% (751,500) of Basques in a territory that spans part of northern Spain and southern France. Of these, 93.2% reside on the Spanish side and the remaining 6.8% in the French region. The Basque Autonomous Community in Spain has established Basque as a co-official language. The Chartered Community of Navarre grants co-official status to Basque only in northern Navarre. Basque has no official status in the French Basque Country. The same is true for the European Union, which limited the status of official European languages to state languages.

As a non-Indo-European language isolate, Basque grammar differs considerably from surrounding languages, though it has borrowed up to 40% of vocabulary from Romance languages and uses the Latin script. The five main spoken dialects are noticeably distinct from one another and it was not until 1968 that the Royal Academy of the Basque Language unified Basque. Since then, it has been immersed in a process of revitalisation that has faced formidable obstacles. Nevertheless, significant progress in numerous areas has fostered the necessary sociolinguistic conditions for the successful development and dissemination of LT. This positive course of events,

Kepa Sarasola · Itziar Aldabe · Arantza Díaz de Ilarraza · Ainara Estarrona · Aritz Farwell · Inma Hernáez · Eva Navas
University of the Basque Country, Spain, kepa.sarasola@ehu.eus, itziar.aldabe@ehu.eus, a.diazdeilarraza@ehu.eus, ainara.estarrona@ehu.eus, aritz.farwell@ehu.eus, inma.hernaez@ehu.eus, eva.navas@ehu.eus

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_5

bolstered by years of collaborative work, has resulted in state-of-the-art technology and robust, broad-coverage NLP for Basque (Hernáez et al. 2012). Still, a dramatic difference remains between Basque and other European languages in terms of research maturity and readiness with respect to solutions (Sarasola et al. 2022).

Data collected by the Basque Institute of Statistics (EUSTAT), shows that 85% of people aged 15+ in the Basque Autonomous Community (1,603,000 individuals) used the internet between June and September 2021. According to the PuntuEUS Observatory, which measures the presence of Basque on the internet, there are currently 12,470 websites with the Basque language code (.eus) as the top-level domain. In 2020, the percentage of websites with content in Basque was 84.4%.

2 Technologies and Resources for Basque

The LT support of Basque is reflected in the European Language Grid (ELG). Half of the resources are corpora, while the rest includes resources, grammars and models. Basque language models in ELG may be divided into monolingual and multilingual. Among the former is BERTeus, a Basque language model pre-trained on crawled newspaper articles and the Basque Wikipedia. The latter include IXAmBERT, a multilingual pre-trained language model for English, Spanish and Basque.

Most Basque monolingual corpora are annotated at some linguistic level. The largest, the ETC corpus and the Lexical Observatory Corpus, contain 48-355 million words. The EPEC corpus contains 300,000 words of standard written text, manually tagged at different grammatical levels. Bi- or multilingual corpora, the majority of Basque corpora in ELG, are composed of comparable or parallel data. HAC, a cross-lingual corpus for Basque, Spanish, French and English and the Basque-Spanish EiTB corpus of aligned comparable sentences contain 629,916 and 564,625 translation units, respectively. In comparison to text corpora, resources that include other modalities are relatively few. However, several databases for ASR, TTS and speech-to-speech translation (S2ST) have been built over the last decade. Large public datasets for high quality speech synthesis for Basque are not available for commercial use, but smaller datasets developed at the UPV/EHU are on hand for research. S2ST, a new research area that requires bilingual data, has made inroads with respect to Basque: there is a bilingual Basque-Spanish dataset containing over eight years of Basque parliamentary sessions.

Lexical resources outweigh conceptual ones, followed by dictionaries, thesauri, terminological resources, ontologies and wordnets. The Egungo Euskararen Hiztegia (Contemporary Basque Dictionary) and the Orotariko Euskal Hiztegia (General Basque Dictionary) count among the most important dictionaries. Additionally, there are euLex and the Euskararen Datubase Lexikala and three variants of WordNet (EusWordNet, Multilingual Central Repository 3.0, SLI Galnet).

Basque tools and services in ELG span a range of applications, but none are listed for information extraction and retrieval, language generation and summarisation or human-computer interaction. Instead, most may be classified as spellcheckers or fall

under text analysis, speech processing, and translation technologies. There are three spellcheckers of note, while pipelines for sentence segmentation, tokenisation, PoS tagging, lemmatisation and dependency parsing may be constructed with UDPipe, ixaKat or IXA-pipes. Other types of linguistic processing are also available, ranging from word sense disambiguation and lexical similarity to RST parsers.

There are two major TTS engines that read texts with high quality synthetic voices in Basque or Spanish. Just one Basque company offers a speech recognition service. Google's Cloud Speech-to-Text is available for Basque, but only in default and command and search models. There are no additional enhanced models as there are for English, French or Spanish, no option for using Google's Cloud TTS, and Amazon does not include Basque in their TTS or ASR services. Besides Google Translate, there are four locally developed neural systems that provide high quality translation between specific language pairs.

Although most basic LT tools are available, a significant gap remains between Basque and other languages in terms of data. This difference is also observed in speech resources and domain-specific data. If we wish to fine-tune models for better performance, domain-specific corpora are required. These examples underline the endemic digital inequality that exists in LT, although one bright spot for languages with few resources, such as Basque, is that pre-trained mono- and multilingual models have proven quite useful in NLP tasks, even when based on far smaller corpora. As a final note, it is worth mentioning most Basque resources have been produced by research groups at the University of the Basque Country and other public entities. Regrettably, resources produced by companies involved in publicly funded projects are not always open-sourced and greater pressure must be applied to ensure they are.

3 Recommendations and Next Steps

While Basque's digital condition may not be endangered, it does remain vulnerable. More work must be done to deepen its integration into social network applications, expand its use in business and employment services, and extend its reach into entertainment products. Moreover, there are significant gaps in the availability of language data and tools that must be addressed so that research may be improved and better commercial applications developed. The more obvious lacunae include a lack of sufficient multimodal corpora, public datasets, and advanced language models. While it is true that pre-trained mono- and multilingual models are employed to great effect in a variety of NLP tasks, a dearth of domain-specific data in Basque continues to hinder the ability to fine-tune models. This is an area that not only requires attention with respect to Basque, but also underscores the chasm in LT between the most utilised languages, such as English, and those with far fewer digital resources. It is as understandable as it is troublesome that a high percentage of Basque speakers meet with obstacles in their online lives, too often finding it easier or necessary to rely on other, more widely available languages for determined services and informa-

tion. This prima facie case of linguistic inequality, not limited to Basque, does not bode well for the future of Europe's cultural heritage.

Fortunately, a remedy may yet be found if action is taken now. Basque's digital health would benefit from bolder and nimbler LT strategies at the European, national and regional levels. In this context, the Spanish Government has approved the New Language Economy PERTE with the purpose of reinforcing the value of official languages in the digital transformation process. Out of a €1.1 billion budget, at least €30 million will be earmarked for supporting projects in co-official languages. Similarly, the Basque Government has launched GAITU, an action plan that aims to integrate Basque into LT between 2021-2024. Finally, the opportunity to take a role in the CLARIN infrastructure would also result in the creation and maintenance of resources. These types of actions should guarantee that data and resources will be made publicly accessible whenever possible because the amount of available data will determine the quality of prospective applications. Licences that provide fewer restrictions on content creation should be more widespread so that greater amounts of linguistic data may be collected. Infrastructures and trained personnel are required to manage the influx of data and curate it for research and development. At one level, taking these steps will help ensure that LT continues to adapt to Basque's digital needs and keep pace with advances at the global level. At another, such a strategy would impart greater visibility to LT and reinforce its vital role in enabling Basque to thrive in today's rapidly evolving socio-digital space.

References

Hernáez, Inmaculada, Eva Navas, Igor Odriozola, Kepa Sarasola, Arantza Diaz de Ilarraza, Igor Leturia, Araceli Diaz de Lezana, Beñat Oihartzabal, and Jasone Salaberria (2012). *Euskara Aro Digitalean – The Basque Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/basque.

Sarasola, Kepa, Itziar Aldabe, Arantza Diaz de Ilarraza, Ainara Estarrona, Aritz Farwell, Inma Hernaez, and Eva Navas (2022). *Deliverable D1.4 Report on the Basque Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-basque.pdf.

Chapter 6
Language Report Bosnian

Tarik Ćušić

Abstract It is objective to state that there are no language technologies for the Bosnian language or initiatives for the digitalisation of the Bosnian language. Therefore, it is necessary to take initial steps towards technological support for the Bosnian language, in order to prevent its digital extinction. In Bosnia and Herzegovina, no programmes aimed at the research and development of language technology products have been initiated. The Bosnian language is present in the digital sphere more or less as much as it is included in foreign, multilingual tools and resources, which are mostly related to Machine Translation (Google Translate and others).

1 The Bosnian Language

The Bosnian language belongs to the West-South Slavic subgroup of the Slavic branch of the great Indo-European linguistic family. Bosnian has about 2.5 million native speakers in Europe. It is the official language in Bosnia and Herzegovina, along with Croatian and Serbian, where it is spoken by 1.87 million people, or 53% of the population. Bosnian is the native language of Bosniaks in Bosnia and Herzegovina, but also of members of other ethnic groups. Outside of Bosnia and Herzegovina, Bosnian is one of the official languages in Montenegro. Bosnian is also an officially recognised minority language in Croatia, Serbia, North Macedonia and Kosovo. In Western Europe and North America, Bosnian is used by about 150,000 people, and by 100,000 to 200,000 people in Turkey.

There is no single language law in Bosnia and Herzegovina that regulates the issue of official language use. However, Bosnian (along with Croatian and Serbian) is listed as one of the official languages in laws and regulations on primary education, secondary education and higher education.

Two writing systems are used in the Bosnian language: Latin and Cyrillic. Both Latin and Cyrillic have 30 letters each; Latin has 27 monographs and three digraphs

Tarik Ćušić
University of Sarajevo, Bosnia and Herzegovina, tarik.cusic@izj.unsa.ba

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_6

(dž, lj, nj), and Cyrillic has 30 monographs. In the past, the Bosnian language was also recorded with Glagolitic, Bosnian Cyrillic (Bosančica) and Arebica.

According to the morphological classification, the Bosnian language belongs to the group of synthetic languages of the inflectional type: it has a larger number of inflections, i. e., different grammatical forms of words; it is characterised by the frequent merging of different morphemes, by a multitude of changes within individual forms and at the boundaries of morphemes, etc.

The Bosnian language belongs to the group of languages marked by the syntactic structure of SVO: Subject–Verb–Object, e. g., *Mahir sluša rok* [Mahir listens to rock.]. There are three types of word order in the Bosnian language: basic word order (grammatical-semantic), actualised word order (contextually conditioned) and obligatory word order (prosodically conditioned) (Jahić et al. 2000, p. 465-473).

In January 2021, 3.27 million people lived in Bosnia and Herzegovina (49.2% of them in urban areas): the total number of mobile connections was 3.73 million, which is 113.9% of the total population; there were 2.32 million internet users (71% of the population) and 1.8 million active social media users (55% of the population).[1]

There are more than 25,000 .ba domains registered.[2] The languages of websites under the .ba domain are mostly Bosnian, Croatian and Serbian, while some websites, due to their character and purpose, are bilingual: Bosnian – English, Croatian – English, Serbian – English and the like.

2 Technologies and Resources for Bosnian

Very few resources (i. e., corpora, language models or lexica) are available for Bosnian to date. In fact, Bosnian lacks a reference monolingual corpus that would be a valuable asset for both linguistic research and LT development. With regard to bi- or multilingual corpora, although they are rare, Bosnian is included as part of some corpora. Examples are the SETimes corpus, a parallel corpus in ten languages with its Bosnian part consisting of 2.2 million words, and the Oslo Corpus of Bosnian Texts, a 1.5 million words corpus consisting of different genres of texts published in the 1990s. The Bosnian part of the CC-100 corpus comprises 14 million tokens (Conneau et al. 2020).

In a relatively recent project aiming at compiling Web corpora of Bosnian (bsWaC) (Ljubešić and Klubička 2014), 8,388 seed URLs for Bosnian were obtained via the Google Search API queried with bigrams of mid-frequency terms obtained from corpora built with focused crawls of newspaper sites. Each TLD was crawled for 21 days with 16 cores used for document processing. The web corpus of the Bosnian language comprises 722 million tokens (Ljubešić and Klubička 2016).

[1] https://datareportal.com

[2] https://www.domaintools.com

With respect to available language technologies, Bosnian is supported in a number of machine translation systems, mainly commercial ones, like Apptek, Tradukka and iTranslate. Google Translate also supports Bosnian.

CroNER is a tool for recognising and classifying named entities in natural language texts in Croatian. CroNER recognises nine different classes of named entities. Although developed for Croatian, CroNER can successfully be applied to texts in closely related languages such as the Bosnian language.

A relatively recent (2017) mobile application for *The orthography of the Bosnian language* (Halilović 1996) can be used to learn the spelling of the Bosnian language and certain grammar rules. The mobile application allows you to search words or book chapters that contain this "orthography". This medium also allows for more flexibility than a book: You can consult "orthography" almost always, on the tram, in a cafe, during a walk. The aim was to bring the book closer to the younger generation and to promote the use of technology in education.

The Language Institute of the University of Sarajevo has developed a digital platform for the Bosnian language, e-bosanski.[3] Its goal is to offer language material about Bosnian in an online format. The material currently available is the Bosnian Dictionary of Accent Variations – Sound (Online) and Converter of Alphabets.

The Dictionary of Accent Doublets is a dictionary entry in the Bosnian Accent Manual (with a sound accent book) by a group of authors: Jasmin Hodžić, Aida Kršo and Haris Ćatović.[4] The corpus of audio recordings is designed to acquire competencies in accentuation, especially for practising general mutual accent differences in individual accents, regardless of the realised examples in everyday speech or in the Bosnian accent norm. It contains over 1,000 accent doublets selected from over 7,000 examples that make up the already excerpted material for a future study on the sources of Bosnian accentuation. Practically, this means that sound recordings for different accent variations of the same words are hosted on this platform. The Sounded Dictionary of Names is a separate part of the dictionary appendix of the future study of the Prosodem variant of personal names by the author Jasmin Hodžić. 111 names with accent variations are currently provided, i. e., recordings of different accent variations of the same names. The platform also encompasses the Accent Reader[5] and Accent Exercises.[6] The Accent Reader provides material from a hundred accented and sounded literary texts. The texts are related to everyday Bosnian life and tradition. Videos with the pronunciation of all vowels under different accents in the Bosnian language are available, including short-descending, short-ascending, and long-descending and long-ascending accents.

The platform additionally provides a Converter of Alphabets, i. e., a converter from the Latin alphabet to Glagolitic, Bosnian Cyrillic (Bosančica) and Arebica.

The Language Institute of the University of Sarajevo plans to create a large historical online dictionary of the Bosnian language that will include language material

[3] https://www.e-bosanski.ba

[4] https://www.e-bosanski.ba/rad/

[5] https://www.youtube.com/playlist?list=PL230XGW7TwJoq3ZNvg7IF7VpcsieCLW-n

[6] https://www.youtube.com/playlist?list=PL230XGW7TwJo2MgihumhTIX52_QxFBQrT

from the Middle Ages (inscriptions and charters), aljamiado texts, texts from oral literature and so-called Krajina letters. The online dictionary will provide word search functionalities, retrieving the context of the word (sentence, verse, document) from the original work.

3 Recommendations and Next Steps

As is evident from the analysis above, there are no large monolingual corpora that are representative of the modern use of the Bosnian language, or for the development of large language models (Ćušić 2022). Therefore, it is necessary to start from scratch. Current data is not sufficient in either the general or specific domains. At the national level, the Council of Ministers of Bosnia and Herzegovina is a public body that could pass the necessary acts to support the development of LT for the Bosnian language, but it is unlikely that this will happen, because language is a sensitive issue in Bosnia.

References

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proc. of the 58th Annual Meeting of the Assoc. for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. ACL, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

Ćušić, Tarik (2022). *Deliverable D1.36 Report on the Bosnian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-bosnian.pdf.

Halilović, Senahid (1996). *Pravopis bosanskoga jezika*. Preporod.

Jahić, Dževad, Senahid Halilović, and Ismail Palić (2000). *Gramatika bosanskoga jezika*. Dom štampe.

Ljubešić, Nikola and Filip Klubička (2014). "{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian". In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 29–35.

Ljubešić, Nikola and Filip Klubička (2016). *Bosnian web corpus bsWaC 1.1*. Jožef Stefan Institute, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1062.

Chapter 7
Language Report Bulgarian

Svetla Koeva

Abstract This chapter reports on the current status of technology support for Bulgarian and highlights certain gaps. The analysis is based on the services and resources available in the European Language Grid in early 2022. While the LT field as a whole has significantly progressed in the last ten years, we conclude that there is still a yawning technological gap between English and Bulgarian, and even between German, French, Italian, Spanish and Bulgarian. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality for Bulgarian.

1 The Bulgarian Language

Bulgarian is the official language of the Republic of Bulgaria. It is spoken by over eight million native speakers. According to an assessment by the National Statistical Institute for the 2021 census, the population of Bulgaria is about 6,500,000. A report by the World Bank states that about 1.7 million Bulgarians lived abroad in 2020.

The official alphabet is Cyrillic. Bulgarian was the first Slavic language to have its own writing system, which dates from the 9th century. Bulgarian belongs to the family of South Slavic languages and forms part of the Balkan linguistic union. Bulgarian exhibits a number of specific characteristics that contribute to the richness of the language but can also be a challenge for natural language processing (NLP), e. g., a rather flexible word order, combined with the lack of morphological distinction for nominal cases and regular subject omission.

The Bulgarian constitution states that Bulgarian is the official language in the Republic of Bulgaria. All education and teaching provided as part of the current state curriculum, from preschool to university, is in Bulgarian. The Institute for Bulgarian Language of the Bulgarian Academy of Sciences is the official institution that monitors changes in the Bulgarian language, determines literary norms and reflects these changes in both orthography and grammar.

Svetla Koeva
Institute for Bulgarian Language Prof. Lyubomir Andreychin, BAS, Bulgaria, svetla@dcl.bas.bg

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_7

According to W3Techs, Bulgarian accounts for just 0.1% of the language content on the web (as of November 2021). Bulgarian internet users in 2020 increased by 31% in comparison to 2007 and already 46% of the total population uses the internet. Bulgarian Wikipedia, as an important source of data for NLP, has a considerably smaller size than the biggest Wikipedias.

Bulgaria's membership in the EU, together with the ideas of unity and diversity, and globalisation while preserving national identity, provides a real opportunity for the equal use of Bulgarian together with the other major European languages.

2 Technologies and Resources for Bulgarian

Language Technology (LT) provides solutions for the following main application areas: Text Analysis; Speech Processing; Machine Translation; Information Extraction and Information Retrieval; Natural Language Generation; and Human-Computer Interaction. This study on LT for Bulgarian is based mainly on the European Language Grid as of February 2022 (Koeva and Stefanova 2022).

Technological developments in recent years have enabled the processing of huge amounts of language data, and allowed the application of complex models and algorithms, which will lead to significant progress (including for Bulgarian). Bulgarian is present in several monolingual and multilingual corpora. Some of the multilingual corpora are sentence-aligned, which allows for cross-lingual research. However, large multilingual corpora are usually created automatically from the internet (often from Wikipedia). Annotated corpora with manually validated or manually assigned linguistic information are smaller in number and volume. There are very few examples of multimodal corpora. Among the multilingual annotated corpora where Bulgarian is present, there are two relatively large collections: Universal Dependencies treebank v2.8.1, and the annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1), both freely available. There is an expanding collection of datasets and models for Bulgarian (at Hugging Face).

Bulgarian is relatively well-resourced when it comes to dictionaries and thesauri. Most dictionaries have been developed at the Institute for Bulgarian Language, but due to copyright restrictions, some of them only offer single user queries or access for research purposes only. Parts of the Bulgarian WordNet are available for download, extended with semantic classes, new semantic relations and semantic frames.

There are several NLP libraries providing sets of linguistic annotations for Bulgarian (tokenisation, sentence splitting, paragraph detection, lemmatisation, named entity recognition, dependency parsing, etc.). A number of libraries provide deep learning techniques and knowledge graphs, and report good levels of accuracy and speed (e. g., Spark NLP). Recently, two NLP pipelines (including a tokeniser, a sentence splitter, a tagger, a lemmatiser and a dependency parser) have become available: UDpipe and NLP-Cube, trained for languages with UD Treebanks, including Bulgarian.

Generally, LTs for Bulgarian still dominate text analysis while multimodal input data (such as simultaneous text, images, audio and video) is rarely processed.

The quality of speech technology for Bulgarian is not yet satisfactory. There are still no accessible and reliable speech-to-text systems for Bulgarian, especially working in real time. Excluding the automatic translation offered by multinational enterprises, there are other available MT systems from and into Bulgarian with different types of access. The assessment of the quality of existing MT services, the number of language pairs, and the coverage of thematic domains still determines MT technologies for Bulgarian as underdeveloped. Recently, there have been serious advances in research on information extraction for Bulgarian: event extraction, sentiment analysis, fake news detection, fact-checking.

There is no dedicated funding or infrastructure for Bulgarian LTs. Many of the achievements and advancements in the development of language data and tools for Bulgarian have been the result of short-term funded projects and PhD theses.

A number of Bulgarian LT companies are very successful, for example, Ontotext, operating in the field of semantic technologies with its product GraphDB.

When we compare the two studies – Blagoeva et al. (2012) and Koeva and Stefanova (2022) – we can see that there is a development in LRs and LTs for Bulgarian, but this is also true for other European languages. Furthermore, in a comparative analysis in 2012, Bulgarian was ranked 15th in terms of technological support, while it is ranked 21st in 2022. Nowadays, technological progress is rapid, and we should consider language models such as GPT-3 and its successors for Bulgarian and the other European languages, which will necessitate significant investments.

3 Recommendations and Next Steps

Many commonly used AI technologies are still not available for Bulgarian (Human-Computer Interaction, multimodal processing, etc.), while for others, if technology has made advances, there are no available applications (summarisation, question answering, etc.). Progress is typically made abroad and Bulgarian is part of some multilingual systems for MT and speech analysis. There is already a need for open real time MT services from and to Bulgarian combining text and speech, taking into account context, communicative purposes and different environments. Thus, speech and text technologies for Bulgarian have to be combined with technologies for other modalities: real time image and video processing working simultaneously in multilingual environments. Natural language understanding and generation of Bulgarian have to become part of multilingual and multimodal processing.

Digital Bulgarian needs large-scale, long-term support, harmonised with the support for all European languages. The sporadic funding of various tasks and particular languages should be replaced by common goals and objectives for all European languages, which if provided with the necessary funding will lead to vast improvements. Efforts cannot be focused only on Bulgarian or on any single language, because multilingual and multimodal resources and technologies are currently needed.

A BLARK-like (Basic Language Resources Kit) minimum set of LRs and LTs for all European languages should be developed and maintained, taking into account that the minimum requirements change rapidly. In 2022, this set should contain large integrated models for as many applications as possible: real-time, multimodal, cross-domain and multilingual LRs and LTs; and a variety of domain-specific datasets.

Convenient and well-regulated access to data is essential for the development of new products, applications and services. To achieve a significant advance over the current situation, an increase of available (open and copyright-free) data for Bulgarian and other European languages is needed, as is an improvement in the legal conditions for (re)using data at the European level.

There is a need for dedicated education and training programmes in the field of LT and AI, as it has proven difficult to source researchers, linguists or engineers with the right combination of skills (e. g., Bulgarian language, computer science, linguistics).

To avoid the reduplication of efforts and to promote data-sharing, it is needed to strengthen and reinforce the European hubs and repositories, such as ELG, intended for ready-to-use datasets, models, tools and services. This will increase the overall language support and ensure the sustainability of LT solutions.

To conclude, although a number of technologies and resources for Bulgarian exist, there are far fewer technologies and resources for Bulgarian than for English as well as for some other European languages. Our vision is high-quality LT for all European languages that supports political and economic unity through cultural diversity.

References

Blagoeva, Diana, Svetla Koeva, and Vladko Murdarov (2012). *Българският език в дигиталната епоха – The Bulgarian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/bulgarian.

Koeva, Svetla and Valentina Stefanova (2022). *Deliverable D1.5 Report on the Bulgarian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-bulgarian.pdf.

Chapter 8
Language Report Catalan

Maite Melero, Blanca Calvo, Mar Rodríguez, and Marta Villegas

Abstract Despite its vulnerable position as a minoritised language, the presence of Catalan in the digital sphere is relatively strong, thanks to an active online community with a high technological profile. Technological support for Catalan is slowly growing, following the recent advances in AI and increased awareness of the value of language data and technologies among public and private bodies. However, more effort is needed to promote the creation of open-source solutions and resources so as to lower the investment barrier for companies to build technology for Catalan.

1 The Catalan Language

Catalan is a Romance language, closely related to Occitan, spoken in four European states – Andorra, Spain, France and Italy – where it shares space with three big languages (Spanish, French and Italian). Andorra is the only state where Catalan is the national language. In Spain, it is mainly spoken in Catalonia, Valencia, and the Balearic Islands, where it is official together with Spanish. In Valencia, the traditional denomination of the language is Valencian. Catalan is also spoken in Alghero (Sardinia) and in the south of France. The total number of habitual speakers of Catalan is estimated to be about 4.5 million. Despite its vulnerable position as a *minoritised language*, the presence of Catalan in the digital sphere is relatively strong. A good example is the Catalan Wikipedia, which ranks 20th globally in terms of number of articles. The use of Catalan in websites that offer their services in Catalonia (and the rest of the Catalan-speaking territories) has been steadily growing from an estimate of 38.75% in 2002 to the current estimate of 66.03%. The digital presence of the language is uneven across sectors. Only 30.3% of the 480 most popular brands in Catalonia have their website translated into Catalan, but close to 100% of universities, NGOs and culture-related Catalan organisations have their website in Catalan. In contrast, few public organisations at the Spanish level, and none at the European

Maite Melero · Blanca Calvo · Mar Rodríguez · Marta Villegas
Barcelona Supercomputing Center, Spain, maite.melero@bsc.es, blanca.calvo@bsc.es, mar.rodriguez@bsc.es, marta.villegas@bsc.es

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_8

level, offer a Catalan version of their website. As for social media and streaming platforms, popular sites such as Instagram, Netflix, Spotify, HBO, Linkedin or TikTok do not offer localised Catalan versions. In spite of this, Catalan web users are considerably active online: Catalan is the 10th EU language (and 19th in the world) in terms of number of tweets, 9th of the EU (and 17th in the world) in terms of number of users who tweet in this language and 5th of the world in number of tweets per user. In the last ten years, grassroots social-media initiatives have emerged, such as Valençúbers, Canal Malaia or Creators.tv. These efforts have given visibility to more than 500 Catalan content creators on various channels, such as YouTube, Instagram, Twitter, TikTok or Twitch and have generated millions of views.

The presence of Catalan in technological products is slowly growing but very unevenly. Large technology corporations tend to consider Spain as a single language market, and consequently rarely include Catalan in innovative and popular AI applications, such as voice assistants, although they do include it among the languages offered by some of their cloud services (e. g., Google Translate and Google Cloud STT and TTS, Amazon Lex and other AWS services, etc.).

2 Technologies and Resources for Catalan

From the mid-nineties, machine translation between Catalan and Spanish began to be used intensively by press editors aiming at producing bilingual publications. Among the products developed during those years, FreeLing, a text analysis tool, and the AnCora corpus still stand out (Moreno et al. 2012). The Corpus textual informatitzat de la llengua catalana (CTILC), manually annotated with lemma and morphological information, was collected by the Institute for Catalan Studies (IEC) during the same period, while the Academy of the Valencian Language collected the Corpus Informatitzat del Valencià and the Corpus Toponímic Valencià, reflecting the Valencian subvariant. Another relevant institution created during the first years of digitisation is TERMCAT, a public entity entrusted with the creation of terminological resources and the standardisation of neologisms.

Current LTs rely heavily on the use of large language models trained on large corpora (Melero et al. 2022). The recent CaText is the largest web corpus in Catalan with acceptable quality, while AnCora remains the largest and more complete annotated corpus. There is a noticeable lack of specialised annotated corpora in Catalan for a variety of domains, genres and tasks, both for fine-tuning and evaluation purposes. Luckily this trend is starting to turn, and a series of datasets annotated for text classification, question answering, summarisation, textual similarity, and named entity recognition, among others, are being created in the framework of the AINA project. AINA has also released monolingual language models trained on CaText. One of the most popular and widely used LTs is machine translation (MT). To train MT models, bilingual parallel data is needed. Most of the largest bilingual corpora are between Catalan and Spanish, although many are not publicly available. Both the OPUS initiative and the Paracrawl project offer multilingual models also containing Catalan

texts. Several online platforms offer translation services for Catalan, like Google Translate and MS Bing, although one of the best rated ones, DeepL, does not yet include Catalan. In addition, some open-source initiatives have built downloadable translation models that include Catalan, such as rule-based Apertium (to and from most Romance languages) and neural-MT Softcatalà (to and from some European languages). More work is needed in MT involving Catalan to improve existing models and add more languages, like Chinese, Russian or Arabic. This would have major impacts, e. g., on e-commerce, the integration of migrants, and the international diffusion of Catalan audiovisual productions. Speech recognition and synthesis are trained on audio datasets. The Mozilla Common Voice project has been very successful among the Catalan community, having produced over 1,300 hours of recorded speech. Another important resource is ParlamentParla, an open-source speech corpus consisting of around 611 hours of parliamentary speeches. Aside from that, we find smaller transcribed audio corpora for specific purposes (e. g., prosody, clinical, social and geographical variation). Local companies, like Verbio, offer customised solutions involving STT and TTS technologies in Catalan, such as automatic subtitling for Catalan television. Catotron is a recent open-source TTS tool for Catalan developed by CollectivaT using deep learning models trained on an audio corpus provided by Catalan television. Catalan Sign Language (LSC) is used by more than 25,000 people in Catalonia. There is an ongoing project to collect an LSC corpus carried out by the IEC and the Pompeu Fabra University. The current amount of data is still insufficient to develop translators and other technology related to LSC, thus more efforts should be devoted to this sensible area.

In Catalonia, the AI strategy (Catalonia.ai) is led by the Department of Digital Policies, which has recently approved the AINA project to promote the development of technological applications in Catalan, in collaboration with the Barcelona Supercomputing Center. AINA has already started to produce concrete results (see above). There is a sizeable number of research groups focused on NLP or speech technologies in universities and research centres across Catalonia and Valencia. There is also a vibrant ecosystem of small and medium local enterprises providing language services and developing intelligent applications, some of them offering Catalan, although less often than desired due to the initial investment barrier. Among the relevant stakeholders it is worth mentioning the role played by Softcatalà since the beginning of digitisation. Softcatalà is a non-profit association whose basic aim is to promote the use of Catalan in computer science, the internet and new technologies. Since their origins, in 1998, they have contributed to open-source software localisation and have developed free tools, such as spell-checkers, translation models, synonym dictionaries and multilingual dictionaries.

3 Recommendations and Next Steps

The recent advances in AI-powered LTs have resulted in an increased awareness of the Catalan society and political bodies, of the importance of LT and language data.

However, public administrations still host very large volumes of non-confidential data, that is suitable for developing cutting-edge technology but remains unexploited and inaccessible. We feel that due to this increased awareness, this is beginning to change. We expect that the European directives on the reuse of public information will soon be fully implemented in the Catalan administration, and open access to language data, which is recognised as essential for the development of new applications and services in Catalan, will become standard. Given the particularities of the Catalan market, supporting open-source solutions would decrease dependence on large corporations for developing cutting-edge technology for Catalan. Moreover, having access to open-source solutions and resources will allow small and medium-sized companies (and potentially also large ones) to develop applications in Catalan without having to face the initial investment barrier. A significant way of stimulating the market and driving the demand of technology in Catalan is to increase the innovation capacity of Catalan public services by incorporating technological solutions that include Catalan. This will eventually benefit the citizens down the line as well. Finally, the creation of an independent Centre of Excellence dedicated to Catalan LTs would be a way of 1. increasing visibility and sustainability of infrastructures and resources, both existing ones and those soon-to-be-created by current projects, 2. offering more educational and training LT programmes in Catalonia to increase the number of trained experts, 3. facilitating technology transfer between academia and industry, 4. boosting a growing economic sector, while guaranteeing the position of Catalan in the digital challenge.

References

Melero, Maite, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas (2022). *Deliverable D1.6 Report on the Catalan Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-catalan.pdf.

Moreno, Asunción, Núria Bel, Eva Revilla, Emília Garcia, and Sisco Vallverdú (2012). *La llengua catalana a l'era digital – The Catalan Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/catalan.

Chapter 9
Language Report Croatian

Marko Tadić

Abstract This chapter presents a summary of the *Language Report on Croatian* (Tadić 2022) on general features of the language and the level of technological support it receives since the previous report (Tadić et al. 2012). The chapter includes information about the typological and structural features of Croatian, its status and usage in the digital sphere and its support through Language Technologies.

1 The Croatian Language

The Croatian language belongs to the West-South Slavic subgroup of the Balto-Slavic branch of the Indo-European linguistic family. It is the native language of over 5 million speakers. Croatian consists of the dialects and standard national language of the Croats, and is the official language of just under 4 million people in Croatia. Along with Bosnian and Serbian, it is one of the three official languages in Bosnia and Herzegovina, where it is spoken by about 400,000 people. Croatian is also spoken by national minorities in Croatia as well as by autochthonous Croatian minorities in Serbia, Montenegro, Slovenia, Hungary, Austria, Slovakia and Italy. Croatian is also used abroad. The largest Croatian diaspora is located in Germany, followed by the US, Canada and Australia. In 2013 Croatian became the 24th official EU language. According to the 2011 census, 90.42% of the country's inhabitants are ethnic Croats, with Croatian the native language for 95.6%. Croatian is the main language used and taught in schools. The literacy ratio in Croatia is 99.2%. Croatian was written with three scripts (Glagolitic, Cyrillic, Latin), and the Latin script became dominant in the 16th century. It was standardised after 1835, when the Croatian Latin alphabet settled on its modern-day form.

The phoneme inventory of the Croatian standard language consists of 6 vowels and 25 consonants. Croatian differentiates ten parts of speech, five of which inflect (nouns, adjectives, numbers (partially), pronouns, verbs) and four do not (prepositions, conjunctions, particles, exclamations), while some adverbs inflect only in com-

Marko Tadić
University of Zagreb, Croatia, marko.tadic@ffzg.hr

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_9

parison. The grammatical categories that inflect for the majority of declinable words are gender (3 values), number (2 values), and case (7 cases). Definiteness is marked on adjectives and animacy in the accusative singular form of masculine nouns and adjectives. Verbs use categories of manner (5 values), person (3 values), number (2 values), voice (2 values) and tense (7 values). The verbs *biti* ('to be') and *htjeti* ('will') are auxiliary. Verbs also have an elaborate aspectual system (imperfective and perfective with additional subvalues such as inchoativity, iterativity, partitivity etc.) and they could also be reflexive. Adjectives and adverbs can inflect for comparison (3 values). Croatian is characterised by an SVO syntactic pattern and relatively free word order. Double-negation is required. The agreement of components in gender, number and case is typical.

The Croatian Web Archive catalogues and stores web resources: portals, websites of institutions, associations, events, scientific projects, books, journals, etc. from 1998. The Croatian Wikipedia has 211,970 articles (31 May 2022) and is ranked 47th. Croatian is prevalently used on major social media. Croatian appears in Google Translate (since 2008) and Bing Translator as a source or target language. Most social media offer translations of posts in/from Croatian, while popular open-source software as well as systems and interfaces by Apple, Google and Microsoft are localised.

2 Technologies and Resources for Croatian

In the last decade, the development of Croatian LT advanced primarily because Croatia joined the EU in 2013. The position of the 24th official EU language resulted in the inclusion of Croatian in large multilingual NLP campaigns, and it started to be researched by non-Croatian NLP experts, too. Although in some areas a number of fundamental resources are not yet available for Croatian, progress has been made in LR collection, text analytics, language models, computer assisted language learning and machine translation (MT), but speech processing is still seriously underdeveloped. A number of EU and nationally funded projects were running mostly in academic institutions. Fundamental tools for lemmatisation, PoS tagging, NER and syntactic analysis have been provided, but there are no robust and reliable industrial systems. In the area of NLU, there is a newer version of Croatian Wordnet (v2.1) and in 2016, a layer of semantic roles was added to the Croatian Dependency Treebank thus providing basic LRs for semantic processing at lexical and clausal levels.

After the release of the Croatian National Corpus v3 in 2013, there were significant advances in large corpus collection, e. g., hrWac v2.1, ParlaMint-HR 2.1, MARCELL Croatian Legislative Corpus, etc. A number of smaller specialised corpora appeared: for processing social media, for sentiment analysis, for investigation of speech disorders, or language learning.

Available bilingual data are either stand-alone parallel corpora, e. g., hrenWaC 2.0, bi-texts in the domain of tourism, or the MARCELL Croatian-English Parallel Corpus of Legislative Texts, or the results of data collection campaigns, e. g.,

ParaCrawl, Bible translations, and parallel corpora collected from public institutions. Croatian became a language of interest in multilingual initiatives and shared tasks: Universal Dependencies (UD), C4Corpus, Deltacorpus, EU Patents, EU EAC TM, JRC DGT TMs, ParlaMint corpora and Comparable Wikipedias of South Slavic Languages, OSCAR, SETimes, TED talks, OPUS, W2C, and WikiMatrix.

The largest freely available lexical resources are inflectional lexicons: Croatian Morphological Lexicon (HML) and hrLEX v1.3. There is only one general language dictionary freely available for online search to its fullest extent: Hrvatski jezični portal accessing the lexicographical base of a publisher. Access to other lexica is limited through a proprietary app by another publisher. Other larger lexica are specialised like the Croatian Old Dictionaries Portal, Dictionary of Neologisms, or dictionaries compiled by the Institute of Croatian Language and Linguistics covering spelling, phrasemes, valencies, collocations and Croatian Terminology Portal. Special types of lexica are Croatian Derivative Lexicon, CroDeriv and DerivBase.HR, that represent the first steps of processing at the level of derivative morphology and both are connected with the Universal Derivations.

The development of NooJ grammar models accelerated because it was present in teaching at undergraduate and postgraduate levels of linguistics and information sciences. Recently, after the introduction of language model approaches, a similar model was built for Croatian but usually in combination with other languages, such as CroSloEngualBERT, BERTić or ELMo embeddings models.

The best pipeline for processing Croatian is developed within the UD initiative (UDPipe) and it found its way into the GATE, Weblicht and ELG platforms. The UD data served also to produce the Croatian segment in UDify. Apart from the Croatian Language Processing Pipeline developed in 2013, there is the CLASSLA fork for the Stanford Stanza pipeline for processing South Slavic languages. Also, at the lexical and event semantics level, two popular online services feature Croatian, among other languages: Wikifier and Event Registry. In Babelnet, Croatian is well represented and ranked 41st with almost 3 million synsets.

Support for Croatian as a source and target language in MT systems was provided as early as in MT@EC, followed by CEF AT and eTranslation. The introduction of the NMT paradigm increased the translation quality, as shown in the CEF project EU Council Presidency Translator, developed for the Croatian EU Presidency in 2020. The system outperformed Google Translate in $\text{hr} \rightarrow \text{en} \rightarrow \text{hr}$ directions by several BLEU points and in 2020 it translated more than 60 million tokens.

From 2015 to 2016 within the ESF-funded project HR4EU, a Portal for Learning Croatian as a Foreign Language was produced.

Despite some attempts at the Universities of Zagreb and Rijeka, speech technology is the most underdeveloped area for Croatian; no free industry-level applications exist. The commercial players have started to offer speech modules for Croatian. The Collins Multilingual databases WordBank and PhraseBank have included Croatian since 2016, while the GlobalPhone Croatian Pronunciation Dictionary has been available since 2013. TalkBank is the final offering in this limited set of speech data for Croatian. Support for Android devices is provided at the level of the operating system, but it does not exist in the iOS environment.

A nationally funded programme for LT ran from 2007 to 2012. It disseminated LT research from the Faculty of Humanities and Social Sciences, University of Zagreb to a number of other institutions in Croatia. The Croatian LT Society has a mission to unofficially coordinate LT activities in Croatia. The dominant role regarding further development of Croatian LT in the last decade was played by the EU through its FP7, ICT-PSP, H2020 and CEF programmes, funding the involvement of several Croatian research teams where expertise persists to this day, but R&D rarely involves industry. Croatia joining CLARIN ERIC provided additional impetus. At the national level, several projects were funded through the Croatian Research Council.

3 Recommendations and Next Steps

There is a lot of currently inaccessible data that could make an impact on the future of Croatian LT and are still not recognised as language data, e. g., texts produced by public administrations, aligned audio and subtitles archived by the national broadcaster, and the Croatian Scientific Journals Portal with open access. The long-term plan is to secure the presence of Croatian NLP modules in the major NLP platforms such as spaCy, FreeLing, NLP Cube, TextRazor, Cloud Natural Language, Apache Open NLP, etc., in order to secure the wider usage of LT for Croatian and its digital language equality with other languages.

References

Tadić, Marko (2022). *Deliverable D1.7 Report on the Croatian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-croatian.pdf.

Tadić, Marko, Dunja Brozović-Rončević, and Amir Kapetanović (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/croatian.

Chapter 10
Language Report Czech

Jaroslava Hlaváčová

Abstract This chapter provides basic data about Language Technology for the Czech language. After a brief introduction with general facts about the language (history, linguistic features, writing system, dialects), we touch upon Czech in the digital sphere. The main achievements in the field of NLP are presented: important datasets (corpora, treebanks, lexicons etc.) and tools (morphological analyzers, taggers, automatic translators, voice recognisers and generators, keyword extracters etc).

1 The Czech Language

Czech, one of the West Slavonic languages, has about 10 million speakers, most live in the Czech Republic (Czechia). In other parts of the world, there are about 200,000 speakers. Czech is the official language in Czechia, and since May 2004 it has been one of the administrative languages of the EU. It is used in administrative, judicial and other official proceedings (see Bojar et al. 2012, for more details).

The Czech language has several varieties, especially in its spoken form. Literary (standard) Czech is a prestige variety, which is taught in schools and strongly preferred in official texts and the media. In everyday communication, most people prefer other varieties of Czech. The most widespread one is common Czech, based on the Central Bohemia dialects. In Moravia and Silesia, dialects such as Hanak, Lach, and Czecho-Moravian are still used actively. Common Czech and these dialects differ from the literary variety, especially in morphology, and to a lesser extent in terms of the lexicon and pronunciation. Other differences are marginal.

In writing, initially, the medieval Latin alphabet was used and for sounds not present in Latin, digraphs were used. In the early 15th century, the religious reformer Jan Hus replaced the digraphs with single letters with diacritics ("háček" for the palatal/palatalised consonants – ť, ď, ň, ř, š, ť, ž; "čárka" and for long vowels – á, é, í, ó, ú, ý). The only digraph surviving in modern Czech is ch. Long u might have a ring ů, coming from the chain of changes ó → uo → ů.

Jaroslava Hlaváčová
Charles University, Czech Republic, hlavacova@ufal.mff.cuni.cz

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_10

The Czech Republic has .cz as the top-level internet domain. It came into effect in January 1993 after the split of the former Czechoslovakia, which had the domain .cs. As of 21 October 2021, 1,412,102 websites with the top-level domain .cz were registered. There were 9.66 million internet users in Czechia in January 2022.[1] This number increased by 120,000 (+1.3%) between 2021 and 2022. Internet penetration in Czechia stood at 90.0% in January 2022. There were 8.05 million social media users in Czechia in January 2022 (about 75% of the total population). The number increased by 660,000 (+8.9%) between 2021 and 2022.

2 Technologies and Resources for Czech

There are several groups in Czech universities working on all areas of NLP (Hlavacova 2022). They are especially Charles University in Prague, University of West Bohemia, Czech Technical University, Technical University of Liberec, Masaryk University in Brno, Brno University of Technology and Palacký University in Olomouc. Apart from academia, many companies develop LT, usually (but not always) with a narrower focus. The LINDAT/CLARIAH-CZ Research Infrastructure for Language Technologies brings together all the achievements in one place which makes them easily accessible to the wide public.

The main sources of contemporary Czech data are the corpora of the series SYN (Hnátková et al. 2014). SYN2000, SYN2005, SYN2010, SYN2015 and SYN2020 are balanced (representative) corpora of written Czech, morphologically annotated, around 100 million tokens each. SYN2006PUB, SYN2009PUB and SYN2013PUB are corpora of contemporary Czech newspapers and magazines sized 300 MW, 700 MW and 935 MW, respectively. All of the SYN corpora are joined together into a single corpus, the last version being SYN v10 (Křen et al. 2021), the corpus of contemporary written (printed) Czech. It contains 5.9 GW.

The Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0) is a richly annotated and genre-diversified resource (Hajič et al. 2020a). It consolidates the existing PDT-corpora of Czech data, annotated using the standard PDT scheme.

Bilingual data is represented mainly by Czech-English corpora. The 4th release of the Czech-English corpus CzEng 1.0 (Bojar et al. 2011) contains 15 million parallel sentences from seven different types of sources automatically annotated at the surface and deep layers of syntactic representation.

Universal Dependencies is a project that seeks to develop cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on (universal) Stanford dependencies, Google universal part-of-speech tags, and the Interset interlingua for morphosyntactic tagsets.

[1] https://datareportal.com/reports/digital-2022-czechia

The main tools for NLP are UDPipe (Straka 2020), a trainable pipeline for segmentation, tokenisation, POS tagging, morphological analysis, lemmatisation and dependency parsing of raw texts, and MorphoDiTa: Morphological Dictionary and Tagger (Straka and Straková 2015). It performs morphological analysis, morphological generation, tagging and tokenisation and is distributed as a standalone tool or as a library, along with trained linguistic models.

The best-performing tool for Czech-English translation is the deep-learning system CUBBITT (Popel et al. 2021). In a context-aware blind evaluation by human judges, CUBBITT significantly outperformed professional-agency English-to-Czech news translation in preserving text meaning (translation adequacy). While human translation is still rated as more fluent, CUBBITT is shown to be substantially more fluent than previous state-of-the-art systems. Most participants of a Translation Turing test struggle to distinguish CUBBITT translations from human translations.

The work on speech recognition and indexing for digitised oral history archives MALACH (Holocaust survivors' testimony, archive of the Institute for the Study of Totalitarian Regimes)[2] continues and new tools are being developed.

The Alquist Dialogue System[3] is the social bot developed by a team of students from the Czech Technical University in Prague. Alquist is an advanced Conversational AI bot carrying out entertaining and engaging conversations with humans on popular topics such as movies, sports, news, etc. In 2017 and 2018, it gained second place in the Alexa Prize contests in a competition with over 100 academic teams.

The basic lexicon is MorfFlex (Hajič et al. 2020b), the morphological dictionary of Czech, with full inflectional information for every word form, encoded in a positional tag. Wordforms are organised into paradigms according to their formal morphological behaviour. They are identified by a unique lemma.

3 Recommendations and Next Steps

The National Artificial Intelligence Strategy (2019-2030) of the Czech Republic was released in 2019 by the Ministry of Industry and Trade. It mentions NLP among the disciplines related to human-machine interaction, i. e., as one of the prominent fields to be supported. At the same time, AICzechia,[4] a national initiative for cooperation between Czech stakeholders in the field of AI, was established. In terms of NLP applications, it wants to target traditional areas such as defence/security, media and government, but also new domains such as social networks, smart homes and business support. It will maintain and expand activities in international organisations in the field (META-NET, CLARIN ERIC, LT Innovate, BDVA, ISCA, ACL, IEEE, ELRA and LDC). These documents indicate that AI, including NLP, will continue to be supported.

[2] https://ufal.mff.cuni.cz/malach/en

[3] http://alquistai.com

[4] https://www.aiczechia.cz

References

Bojar, Ondřej, Silvie Cinková, Jan Hajič, Barbora Hladká, Vladislav Kuboň, Jiří Mírovský, Jarmila Panevová, Nino Peterek, Johanka Spoustová, and Zdeněk Žabokrtský (2012). *Čeština v digitálním věku – The Czech Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/czech.

Bojar, Ondřej, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna (2011). *Czech-English Parallel Corpus 1.0 (CzEng 1.0)*. LINDAT/CLARIAH-CZ, Charles University. ÚFAL MFF UK. http://hdl.handle.net/11234/1-1458.

Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský (2020a). *Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0)*. LINDAT/CLARIAH-CZ, Charles University. Czech Republic. http://hdl.handle.net/11234/1-3185.

Hajič, Jan, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková (2020b). *MorfFlex CZ 2.0*. LINDAT/CLARIAH-CZ, Charles University. http://hdl.handle.net/11234/1-3186.

Hlavacova, Jaroslava (2022). *Deliverable D1.8 Report on the Czech Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-czech.pdf.

Hnátková, Milena, Michal Křen, Pavel Procházka, and Hana Skoumalová (2014). "The SYN-series corpora of written Czech". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. Reykjavík, Island, pp. 160–164.

Křen, Michal, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal (2021). *SYN v9: large corpus of written Czech*. LINDAT/CLARIAH-CZ, Charles University. http://hdl.handle.net/11234/1-4635.

Popel, Martin, Markéta Tomková, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský (2021). *CUBBITT Translation Models (en-cs) (v1.0)*. LINDAT/CLARIAH-CZ, Charles University. http://hdl.handle.net/11234/1-3733.

Straka, Milan (2020). *UDPipe 2*. Prague, Czech Republic: ÚFAL MFF UK.

Straka, Milan and Jana Straková (2015). *MorphoDiTa*. ÚFAL MFF UK.

Chapter 11
Language Report Danish

Bolette Sandford Pedersen, Sussi Olsen, and Lina Henriksen

Abstract This chapter summarises the current level of language technologies (LT) and resources for Danish (Pedersen et al. 2022). Even if Danish LTs are now being used in many areas of society, their *quality* still needs to be improved in order to make them more useful and inclusive for the majority of the population. To this end, the development of large, high-quality language resources and data sets still proves to be a bottleneck. We report, however, on an increased awareness of sharing and reusing language resources and data sets across public institutions, academia and industry. New, large governmental initiatives within the area of AI and LT have been initiated which support this development.

1 The Danish Language

Danish is a Mainland Scandinavian language and the official language of Denmark, which has about 5.831 million inhabitants. Danish phonology distinguishes itself from that of several of its neighbouring languages by exhibiting a very large number of vowels and by having glottal stop as a meaning differentiating feature (e. g., '!hund' ('dog') vs 'hun' ('she')). Furthermore, phonetical reductions are very common, a fact which complicates Danish speech technology since word boundaries become very hard to identify, to give just one example.

In the written language, the fact that compounds are spelled as one word (as in other Germanic languages) complicates the development of language tools, and furthermore, compounds are generated dynamically and so only partially accounted for in dictionaries. The very extensive use of particles with semi-lexicalised meanings poses a challenge to LT systems. The constructions often occur discontinuously in spoken and written Danish, a fact which tends to require large amounts of language data in order to be well represented in the corresponding language models.

Bolette Sandford Pedersen · Sussi Olsen · Lina Henriksen
University of Copenhagen, Denmark, bspedersen@hum.ku.dk, saolsen@hum.ku.dk, linah@hum.ku.dk

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_11

The influence of the English language on Danish language users is increasing. Loan words and fixed phrases do not influence the language system as such, but the syntax is influenced in some cases. For instance, some Danish verbs change their valency pattern because of the influence from English, as is the case for 'at gro' ('to grow') which is now beginning to occur as transitive, as in 'kan man gro trøfler i Danmark?' ('can you grow truffles in Denmark?'). In addition, the placement of adverbials tends to be increasingly influenced by English.

2 Technologies and Resources for Danish

In recent years a number of repositories for Danish LT have been established. The following overview is primarily based on these, including the Danish CLARIN platform, the repository of The Danish Agency for Digitisation, sprogteknologi.dk, the Danlp list and the DaCy repository (for references, see Pedersen et al. 2022, 2012).

Large Danish text corpora have typically been collected by institutions that develop dictionaries. These host very large balanced corpora today, but due to intellectual property rights they are not entirely open source and ready to use for industry. For research and non-commercial purposes, the DK-CLARIN Reference Corpus of General Danish (45 million words) has been available for a decade at the CLARIN-DK repository. Recently, the Danish GigaWord initiative has been launched, a freely available billion-word corpus of Danish texts assembled by a group of researchers.

In recent years several statistical and neural language models for Danish have been processed and are based primarily on the above mentioned corpora. Schneidermann et al. (2020) report on six different neural models with different correlations with a hand-crafted similarity data set. Recently, also a number of contextualised, pre-trained models have been developed for Danish. The Scandival benchmark evaluates these and other models. Overall, recent models enable improved language processing for Danish with, e. g., a better grasp of the variation of word meaning in running text. Here, diversity in the training data is becoming more relevant since it can result in biases with respect to gender, ethnicity, regional origin etc.

Parallel text corpora are primarily used to build statistical models for machine translation. These models are highly dependent on really large amounts of text data within all domains. The number of parallel corpora including Danish has increased somewhat over the last few years; especially corpora where the other language is English. In recent years the EU initiative European Language Resource Coordination (ELRC) has helped increase awareness of the value of parallel corpora, in collaboration with three nationally located anchor points. Large public speech corpora are generally in short supply for Danish, a fact which complicates the development of speech technologies. However, a few such resources exist at a medium scale, i. e., the Danish NST ASR Database at the Norwegian Språkbanken, compiled originally by the company Nordisk Sprogteknologi, DanPASS, and the Danish Parliament Speech Corpora. The production of a large, transcribed and time-encoded speech corpus is foreseen as part of the government's new AI initiative, launched in 2022.

The Danish Universal Dependencies Treebank (UD-DDT), which has annotations for dependency structure, part-of-speech and named entities, constitutes a basic resource in terms of syntax. The STO lexicon also contains syntactic information such as valency information. Lexical semantic resources of various kinds are also available for Danish. The Danish wordnet, DanNet, is the largest semantic resource with around 70,000 concepts. More specific resources are framenets and Danish sentiment lexicons, various lists of person names, addresses, place names, and some dialect dictionaries. A joint computational lexical project, Central WordRegister for Danish (COR), combines several of these resources in one joint resource.

Danish preprocessing tools such as lemmatisers, part-of-speech taggers, named entity recognisers, and parsers have existed for Danish for several years and are continuously upgraded, partly based on the above-mentioned resources. Even if there is still room for improvement, these tools generally achieve high accuracy and are integrated today into most advanced systems.

Integrated LTs can count on services such as speech, machine translation, and abstracting systems. Dictus ApS and Omilon are examples of Danish companies that deliver dictation solutions to citizens and organisations such as the Danish Parliament, the healthcare system, schools, Danish TV-stations and many more. Speech technology is also used in some chatbots and virtual assistants, and examples of services working for Danish are Siri and Google Assistant. Their performance, however, still leaves room for improvement. Open-source packages for developing speech recognition for Danish are generally scarce. An example is the open-source Python package DanSpeech (now Alvenir) from the Technical University of Denmark.

Currently, most public institutions outsource their translation tasks to private companies, and this trend is rising. Machine translation is applied in almost all fields of translation, and the quality is improving. Recent benchmarking reports for Danish-English and English-Danish show acceptable BLEU scores over 0.70 (depending on the domain) for Google Translate, eTranslation, and DeepL. However, translation quality decreases dramatically when Danish is used in combination with other languages. Other services include technologies such as anonymisation, sentiment analysis, automatic abstracting, summarisation, fake news detectors etc. of which only a few currently exist off-the-shelf for Danish. Areas such as opinion mining and sentiment analysis are growth areas since many companies and institutions in Denmark feel an increasing need to monitor opinions and sentiments on the web.

3 Recommendations and Next Steps

Several factors play a role in how fast and how well a language community like the Danish one adapts to new technological advances. Even if Denmark is one of the most digitised countries in the world, its relatively small size – both as a language community and commercial market, together with our high proficiency in English – seem to have delayed the investments and developments in Danish language processing and LT. The specific characteristics of Danish may also play a role. However,

we see renewed interest in LT at all levels of Danish society. New stakeholders are emerging day by day together with the increasing tendency of introducing language-centric AI in nearly all aspects of society; recent tentative counts indicate that more than 70 Danish SMEs have entered the LT scene. With this development comes more focus and better understanding of the challenges of language processing and of why a continuous upgrade of Danish language resources is indispensable. This increased acknowledgement and tendency towards sharing resources across fields is seen in academia, industry and public administration and will definitely boost LT for Danish in the coming years. New governmental investments in AI and LT are supporting industry and research in this development. All this being acknowledged, the need for continuous coordinated efforts based in public institutions, industrial settings as well as research still remains. One precondition for supporting this effort is to ensure that sufficient highly qualified staff are educated in NLP. It is recommended that NLP study programmes are sufficiently supported and prioritised at higher educational and ministry levels. Governmental focus on industry, research, and education is indispensable if we are to ensure that Danish stays on track to being a digitally fully functional language, also in future language-centric AI solutions.

References

Pedersen, Bolette Sandford, Sussi Olsen, and Lina Henriksen (2022). *Deliverable D1.9 Report on the Danish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-danish.pdf.

Pedersen, Bolette Sandford, Jürgen Wedekind, Steen Bøhm-Andersen, Peter Juel Henrichsen, Sanne Hoffensetz-Andresen, Sabine Kirchmeier-Andersen, Jens Otto Kjærum, Louise Bie Larsen, Bente Maegaard, Sanni Nimb, Jens-Erik Rasmussen, Peter Revsbech, and Hanne Erdman Thomsen (2012). *Det danske sprog i den digitale tidsalder – The Danish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/danish.

Schneidermann, Nina, Rasmus Stig Hvingelby, and Bolette Sandford Pedersen (2020). "Towards a Gold Standard for Evaluating Danish Word Embeddings". In: *12th International Conference on Language Resources and Evaluation, Conference Proceedings (LREC 2020)*. Ed. by Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, pp. 4754–4763.

Chapter 12
Language Report Dutch

Frieda Steurs, Vincent Vandeghinste, and Walter Daelemans

Abstract This chapter provides a new state of affairs (Steurs et al. 2022) with regard to language technology (LT) for Dutch (after Odijk 2012). LT for Dutch is highly developed, and the Netherlands and Flanders have a strong and cooperative LT community. A lot of digital data is freely available through CLARIN and the Dutch Language Institute (INT). However, data and software have to be updated continuously, and there is a need for a new overarching programme to support research initiatives.

1 The Dutch Language

Dutch is a West-Germanic language spoken by about 25 million people as a first language in the Netherlands and Belgium and by 5 million people as a second language (Steurs 2021). It is a close relative to both German and English and shares with German the survival of two to three grammatical genders, as well as the use of modal particles, final-obstruent devoicing, and similar word order. The vocabulary is mostly Germanic and incorporates slightly more Romance loans than German but far fewer than English. Some characteristics are challenging for computational processing, such as a relatively free word order with differences between main and subordinate clauses, and productive compounding. Separable verb prefixes can occur far from the verb and the meaning of a separable verb is often non-compositional. Written Dutch is a monocentric standardised language, with lexical and pronunciation variety between the Netherlands and Flanders. In contrast to its written uniformity, Dutch lacks a unique prestige dialect and has a large dialectal continuum consisting of 28 main dialects. Dutch is used by 1.3% of all websites and is the 12th most used language in terms of number of websites. The Dutch one is the sixth-largest Wikipedia edition. Dutch is used often in social media, which leads to new linguistic trends and sublanguages, for which corpora are required to allow investigation.

Frieda Steurs · Vincent Vandeghinste
Dutch Language Inst., The Netherlands, frieda.steurs@ivdnt.org, vincent.vandeghinste@ivdnt.org

Walter Daelemans
University of Antwerp, Belgium, walter.daelemans@uantwerpen.be

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_12

2 Technologies and Resources for Dutch

The Dutch Language Institute (INT) keeps a detailed list of tools and resources for Dutch at K-Dutch,[1] many of these are downloadable.[2] The Language Machines website also makes plenty of LT tools available as webservices.[3]

SoNaR (Oostdijk et al. 2013) is a reference corpus containing different text genres. Parallel data is available through the Dutch Parallel Corpus (Paulussen et al. 2013) and through OPUS. The Spoken Dutch Corpus (Oostdijk et al. 2002) contains 900 hours (9 million words) of speech and is manually transcribed and linguistically annotated. A new large up-to-date corpus for spoken Dutch containing more recent language and more variants is in high demand. Notwithstanding the popularity of social media, it is difficult to collect and share such data due to restrictions in the EU's GDPR, and only a limited part of SoNaR contains this register.

GiGaNT is a computational lexicon with a historical and a modern component. Open Dutch WordNet (Postma et al. 2016) is a freely available Dutch lexical semantic database. A more contemporary version with better coverage would be desirable.

Hugging Face, a hub for pre-trained language models (LMs), lists 112 pre-trained LMs for Dutch, some of these can perform generation. Word2vec embeddings are available from Tulkens et al. (2016). Nevertheless, there is still demand for very large-scale LMs for Dutch, and for LMs on certain domains and registers.

In terms of text analysis, Frog (van den Bosch et al. 2007) provides lemmas, morphology, PoS tagging, named entities, chunking, and dependency information. Alpino (van Noord 2006) provides deep linguistic dependency parsing. Pattern (De Smedt and Daelemans 2012) and LeTs (Van de Kauter et al. 2013) are multilingual tools for text analysis, including Dutch. SpaCy, Stanza, Weblicht and UDPipe contain Dutch models. Dutch NER is available in OpenNLP.

Text-to-speech and speech recognition (ASR) are commercially available, often in two language variants, and also for research purposes (both variants).

Dutch is present in most commercial online translation services, which provide a limited amount of translation for free. eTranslation from the European Commission provides unlimited translation, including from and to Dutch.

There is currently no joint Flanders-Netherlands overarching programme for the further development of tools and resources for Dutch. The LT community in the Netherlands and Flanders would be very much in favour of setting up a follow-up programme to the STEVIN programme (Spyns and D'Halleweyn 2013), a joint programme to provide the essentials for Dutch language technology (2004-2011).

The Nederlandse AI Coalitie (Dutch AI Coalition) lists the use-case Nederlandse AI voor het Nederlands (Dutch AI for Dutch). The Nederlandstalige Spraak Coalitie (Dutch Speech Coalition) stimulates development of speech technology for Dutch. NOTaS (Dutch Organisation for Language and Speech Technology) joins the various players in the field in ensuring that the Dutch LT industry leads the way in technologi-

[1] https://kdutch.ivdnt.org

[2] https://taalmaterialen.ivdnt.org

[3] https://webservices.cls.ru.nl

cal developments. The Flanders AI programme supports research in NLP, especially on Conversational Agents for Dutch. Computational Linguistics in the Netherlands (CLIN), a yearly conference, is a meeting point for LT researchers in the Netherlands and Flanders. The CLIN Journal provides an international forum for open access publication of high-quality scholarly articles in all areas of LT, with special attention on Dutch. Belgium NLP Meetup is a meeting group for anyone interested in Natural Language Processing. CLARIN is a European research infrastructure in which the Netherlands and Belgium participate. The Dutch portal pages CLAPOP list resources created in CLARIN NL and CLARIAH NL. The CLARIN portal at INT provides access to CLARIN tools from the Netherlands and Flanders. In both regions, CLARIN is part of CLARIAH, in which it joins forces with DARIAH, an infrastructure for the arts and humanities.

3 Recommendations and Next Steps

Dutch is not in a bad shape digitally. Plenty of data sets and tools are available, and the uptake of Dutch in major NLP applications seems assured. Many of the open tools rely on open data sets, often created in the STEVIN programme. Both the Dutch language and NLP technology have changed in the meantime, thus making a new effort at least of the size of the STEVIN programme necessary. It is important to allow tools to learn from recent language use. It is paramount that a new programme is set up in which researchers from the Netherlands and Flanders, and perhaps also beyond, cooperate in the design and construction of corpora that document recent language, be it in written, spoken, or microblog form.

LT is already embedded in our everyday lives, and we may be using it without realising, when checking for spelling errors, using search engines or calling the bank to perform a transaction. It is an important ingredient of applications that cut across various domains. In the health domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in educational settings and applications, for instance for educational content mining, for automatic assessment of free text answers, for feedback to learners and teachers, or for evaluation of pronunciation in a foreign language. In the legal domain, LT proves an indispensable component for the search, classification and codification of huge legal databases to legal question answering and prediction of court decisions. If Dutch wants to remain a part of this strong LT-driven society, we need new investments in research projects.

References

De Smedt, Tom and Walter Daelemans (2012). “Pattern for Python”. In: *Journal of Machine Learning Research* 13, pp. 2031–2035.

Odijk, Jan (2012). *Het Nederlands in het Digitale Tijdperk – The Dutch Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/dutch.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat, and R. Harald Baayen (2002). "Experiences from the Spoken Dutch Corpus Project". In: *Proceedings of LREC 2002*. European Language Resources Association (ELRA).

Oostdijk, Nelleke, Martin Reynaert, and Ineke Hoste Véroniqueand Schuurman (2013). "The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch". In: *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*. Ed. by Peter Spyns and Jan Odijk. Berlin, Heidelberg: Springer, pp. 219–247. DOI: 10.1007/978-3-642-30910-6_13.

Paulussen, Hans, Lieve Macken, Willy Vandeweghe, and Piet Desmet (2013). "Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French". In: *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*. Ed. by Peter Spyns and Jan Odijk. Berlin, Heidelberg: Springer, pp. 185–199. DOI: 10.1007/978-3-642-30910-6_11.

Postma, Marten, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen (2016). "Open Dutch WordNet". In: *Proc. of the 8th Global Wordnet Conference*. Bucharest, Romania.

Spyns, Peter and Elisabeth D'Halleweyn (2013). "The STEVIN Programme: Result of 5 Years Cross-border HLT for Dutch Policy Preparation". In: *Essential Speech and Language Technology for Dutch*. Ed. by Spyns P. and Odijk J. Berlin, Heidelberg: Springer, pp. 21–39.

Steurs, Frieda (2021). "Nederlands een grote taal? Een kewstie van meten". In: *Neerlandica Wratislaviensia*, pp. 17–29.

Steurs, Frieda, Vincent Vandeghinste, and Walter Daelemans (2022). *Deliverable D1.10 Report on the Dutch Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-dutch.pdf.

Tulkens, Stéphan, Chris Emmery, and Walter Daelemans (2016). "Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource". In: *Proceedings of LREC 2016*. European Language Resources Association (ELRA).

Van de Kauter, Marjan, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste (2013). "LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit". In: *Computational Linguistics in the Netherlands Journal* 3, pp. 103–120. https://clinjournal.org/clinj/article/view/28.

van den Bosch, Antal, Gertjan Busser, Sander Canisius, and Walter Daelemans (2007). "An efficient memory-based morphosyntactic tagger and parser for Dutch". In: *Computational linguistics in the Netherlands*. Ed. by P. Dirix, I. Schuurman, V. Vandeghinste, and F. van Eynde. LOT, pp. 191–206.

van Noord, Gertjan (2006). "At Last Parsing Is Now Operational". In: *TALN Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*. Leuven, Belgique: ATALA, pp. 20–42. https://aclanthology.org/2006.jeptalnrecital-invite.2.

Chapter 13
Language Report English

Diana Maynard, Joanna Wright, Mark A. Greenwood, and Kalina Bontcheva

Abstract This chapter focuses on the status of the English language, primarily acting as a benchmark for the level of technological support that other European languages could receive (see Maynard et al. 2022; Ananiadou et al. 2012). While it is rather unlikely that any other European language will ever reach this level, due to the continuing development of support for English, and thus serves as a moving goalpost, nevertheless it provides a good criterion for relative assessment. While the inequalities in the amount of technological support available for English compared with other European languages may act as a deterrent for working on the latter, nevertheless it serves as a useful mechanism for applying cross-lingual transfer methods in order to build language models and generate labelled data for lower resource languages.

1 The English Language

English is a truly international language, due in no small part to the worldwide influence of the British Empire since the 17th century, and later to the influence of the United States. It has become the primary language of international discourse and is the lingua franca in many professional contexts, as well as in a number of regions with diverse native languages. English is the most spoken language in the world, with an estimated 1.36 billion total speakers.English is also the most widely taught foreign language in the world. There are almost three times as many people who speak English as a second language compared to native speakers, with a total of 360 million first language speakers and around one billion second language speakers.

English is an Indo-European language and shares a number of features of other Germanic languages. It uses the Latin alphabet with a left-to-right writing system, and has the ISO-639-1 code (*en*). It is classed as a pluricentric language, meaning that it has no single standard codified form but rather several interacting ones, typically set by or corresponding to different countries (e. g., US vs. British English).

Diana Maynard · Joanna Wright · Mark A. Greenwood · Kalina Bontcheva
University of Sheffield, United Kingdom, d.maynard@sheffield.ac.uk, j.wright@sheffield.ac.uk, m.greenwood@sheffield.ac.uk, k.bontcheva@sheffield.ac.uk

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_13

English is the most commonly used language online, representing about 60.4% of the top 10 million websites.[1] As of 31 March 2020, the internet was estimated to have around 1.186 billion English speaking users (25.9% of all internet users around the world).[2] In terms of internet penetration, out of the 1.531 billion English speakers estimated for 2021 according to Internet World Stats, 77.5% of them are internet users. The number of English-speaking users has enjoyed a relatively modest growth rate of 742.9% in the last 20 years, compared with Arabic at 9,348%.

2 Technologies and Resources for English

While there has been an increasing interest in developing data and tools for multilingual language processing in the last 20 years, as witnessed by the topics of long-standing shared tasks such as CONLL, nevertheless English continues to be overwhelmingly dominant in every aspect of language processing. This is partially as a result of the dominance of the use and status of English in the digital sphere and as an international language, but also a circular problem related to the availability of existing low-level language processing tools and training data, which provide an easy starting point for further development.

Thousands of corpora are freely available for English. The majority of these are covered by a Creative Commons licence, although they may come with restrictions (e. g., attribution or no commercial use). Some are covered by shared task participation agreements, implying that they are freely available at least to task participants. A number of corpora are released under licences controlled by ELRA and thus only available to ELRA members. The LDC grows by around 30 to 35 new corpora each year, and while these do not all include English, it does mean that new resources with contemporary language use appear with reasonable regularity.

Hundreds of monolingual lexical/conceptual resources are available, most of which are domain-specific, including a few ontologies. It is likely that a huge number of freely available additional resources are available beyond those listed in the main language resource catalogues such as ELRA and LDC. The same is true for bilingual resources that include English. Additionally, a number of multimodal resources exist (where text is one of the forms), mostly concerned with pronunciation.

English is very well-served generally by spelling and grammar-checking tools. Most operating systems have built-in spell-checking tools, for example, aspell and hunspell on Linux. Most programming languages have at least one spell-checking library. Similarly, there are many summarization systems available as open source or commercially, including HuggingFace Transformers. Text-to-speech (TTS) systems are also well supported with a number of open source and commercial models.

There are several major infrastructures or toolkits for language processing available, including GATE, Stanford CoreNLP, Stanford Stanza, NLTK, spaCy, Hugging-

[1] https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/

[2] https://www.internetworldstats.com/stats7.htm

Face Transformers, and OpenNLP, which all contain a variety of processing tools which can be used individually or as a collection. All of these support at least tokenisation, sentence splitting, PoS tagging, and named entity extraction. Some support many more tools such as sentiment analysis, or have specific support for domains such as medicine. Overall, there are thousands of models available, especially for text summarization, translation, TTS and various kinds of classification.

For low-level processing tasks, such as tokenisation, sentence splitting and PoS tagging, there are a few standalone tools and services contained in the ELG platform, but many more are provided as part of standard APIs. In general, tools for tokenisation and sentence splitting for European languages are more or less language-independent. POS tagging is also a reasonably well-solved problem for English.

In terms of Information Extraction, there are dozens of NER systems for English, of which roughly half are domain-specific, with domains/genres including biomedical, Twitter, dendrochronology, environment, chemistry and politics. This is also an area which has seen many ML models released. Tools which fall broadly into the Information Retrieval (IR) category cover a wide range of tasks, including question answering. A number of these are cross-lingual. Many systems enable search in a specified language but can return results in other languages, including English. There are a number of commercial IR engines available, both for generic and specialised tasks. Concerning Machine Translation, there are hundreds of tools, of which a large number contain English as either input or output. The most common pairing (regardless of direction) is English/German.

In terms of LT providers, we have identified 53 major industrial organisations in the UK, including players such as BBC News Labs, the JISC, and Oxford University Press, and 246 research groups or organisations based at 94 different universities. These research groups are split between various faculties and departments, comprising mostly Computer Science and Language departments, but also others such as Medicine, Architecture, Life Sciences and Education, Creative Industries, and Maths. In Ireland there are also extensive LT industry bodies and research centres (e. g., Apple, Accenture, Google, SoapBox Labs, AYLIEN, and CeADAR), whose primary focus is on supporting the English-speaking rather than Irish-speaking population.

3 Recommendations and Next Steps

English is extremely well supported by LT, which is unsurprising given its status in the digital world. Almost every tool and infrastructure or toolkit is first developed to handle English before being applied to other languages. Similarly, an enormous amount of data is available for English. These two factors have a circular effect: due to the amount of data available, training and testing new tools is much easier for English than other languages, and this leads to new models, tools, and resources being developed. The frequency with which English is used for online communication also provides a wealth of data from which to create new corpora, and the availability of a wide range of tools also makes it easier to annotate these with linguistic informa-

tion. As tools improve, the accuracy and usefulness of pre-annotated corpora also improve, thereby making further tool development easier.

On the one hand, this is an excellent situation for those working on English data, and given the widespread use of English in the digital world, the usefulness of new tools is clear. On the other hand, this can be a double-edged sword for the development of LTs and LRs for other languages. The availability of data, tools and resources for English has fed the enormous success of neural models for developing LT applications, but the lack of data for other languages means that such deep learning models trained on English are not directly applicable. Recently, however, advances have been made in the development of cross-lingual transfer learning in order to build NLP models for a low-resource target language by leveraging labelled data from languages such as English with a high level of resources, or via a staged process whereby training data from English feeds the development of languages with moderate resources, which may have greater similarity to low-resource languages and can feed a further transfer process. Additionally, multilingual transfer settings enable training data in multiple source languages to be leveraged to further boost performance of low-resource languages. On the negative side, almost all languages are inevitably playing "catch-up" compared with English, and as can be seen from our survey, the differences in LTs and LRs available for European languages are striking. It is hard even to grasp a sense of how much is available for English, since resources are so disparate, and the figures reported in the collections of ELG, ELRA and other repositories are only the tip of the iceberg.

References

Ananiadou, Sophia, John McNaught, and Paul Thompson (2012). *The English Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/english.

Maynard, Diana, Joanna Wright, Mark A. Greenwood, and Kalina Bontcheva (2022). *Deliverable D1.11 Report on the English Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-english.pdf.

Chapter 14
Language Report Estonian

Kadri Muischnek

Abstract This chapter gives a brief overview of Estonian LT tools and resources (Muischnek 2022; Liin et al. 2012). The Estonian language has only around one million speakers and so the market for Estonian LT products is also a small one. In general, the current situation of Estonian LT is acceptable for a small language, but far from perfect. The main force driving the development of Estonian LT has been the public sector and so the resources and tools developed by publicly funded projects are mainly open source. Nonetheless, during the last decade, the private sector has also engaged in creating tools and solutions for Estonian.

1 The Estonian Language

Differently from most languages spoken in Europe, Estonian is not an Indo-European language, but belongs to the Balto-Finnic group of the Finno-Ugric languages. Typologically, Estonian represents a transitional form from an agglutinating to a fusional language. The characteristic features of Estonian include the accent on the first syllable, a high frequency of vowels as opposed to consonants, three different lengths of vowels and consonants, the lack of grammatical gender and articles, and a basic vocabulary different from that of the Indo-European languages.

Estonian has a rich morphological system: nominals inflect for case and number, and verbs for person, number, tense, mood and voice. Compounding is relatively free and productive in Estonian and derivation is another productive device for forming new lexical items. The word order of Estonian is rather free and mostly governed by information structure. The most important rule is V2: the verb occupies the second position in the clause (Erelt 2003).

Estonian is the official language of the Republic of Estonia and it is used in all spheres of life although there are some concerns regarding the use of Estonian in science and higher education. It is written using a supplemented Latin alphabet; in addition to ASCII characters, it also includes the letters Ä, Ö, Ü, Õ, Š and Ž.

Kadri Muischnek
University of Tartu, Estonia, kadri.muischnek@ut.ee

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_14

The Estonian population has good access to the internet and digital services: 92% of households have an internet connection and many services are available online.[1]

2 Technologies and Resources for Estonian

Large monolingual Estonian corpora are collected regularly; the most recent one, Estonian National Corpus 2021, contains ca. 2.4 billion tokens.[2] Estonian is included in the multilingual resources of the EU languages and we have at least a minimum necessary amount of audio resources.

Annotated data is expensive to create and, thus, scarce. Notable examples are the Estonian UD treebanks[3] and the Embeddia dataset for hate speech detection[4].

Lexical-conceptual resources are mostly lexicons, machine-readable dictionaries and terminological databases. An important resource is the Estonian Wordnet.[5]

There is only one full-coverage computational grammar for Estonian: Constraint Grammar.[6] Several large language models trained exclusively on Estonian data[7] [8] [9] [10] and also multilingual ones[11] [12] have been created.

Text analysis tools cover sentence segmentation, tokenisation, morphological analysis and dependency parsing for the standard written language. As soon as the text deviates from the standard, the quality of the analysis decreases. The most basic tool for analyzing morphologically complex Estonian is a morphological analyzer.[13] For parsing Estonian one can use CG[14] or several dependency parsing models[15] [16] [17] trained on the Estonian UD treebanks. The EstNLTK Python library[18] contains a pipeline starting from tokenisation and ending with syntactic analysis and informa-

[1] https://andmed.stat.ee/en/stat/majandus
[2] https://doi.org/10.15155/3-00-0000-0000-0000-08D17L
[3] https://universaldependencies.org
[4] http://embeddia.eu/outputs/
[5] https://www.cl.ut.ee/ressursid/teksaurus/
[6] https://github.com/EstSyntax/EstCG
[7] https://huggingface.co/tartuNLP/EstBERT
[8] https://huggingface.co/EMBEDDIA/est-roberta
[9] https://www.clarin.si/repository/xmlui/handle/11356/1277
[10] https://huggingface.co/tartuNLP/gpt-4-est-large
[11] https://huggingface.co/xlm-roberta-base
[12] https://huggingface.co/EMBEDDIA/finest-bert
[13] https://github.com/Filosoft/vabamorf
[14] https://github.com/EstSyntax/EstCG
[15] https://stanfordnlp.github.io/stanza
[16] https://github.com/EstSyntax/EstSpaCy
[17] https://lindat.mff.cuni.cz/services/udpipe/
[18] https://github.com/estnltk/estnltk

tion extraction (NER etc). The TEXTA Toolkit[19] provides resources for text analytics and enables document classification, terminology extraction and topic detection.

The TalTech's speech recognition system[20] provides speech recognition and other services, e. g., automated subtitling.[21] There are also several models for speech synthesis,[22] including a neural network-based one.[23]

Estonian is featured in Google Translate, Microsoft Translator and the EU's translation tool eTranslation. However, independent MT services are important for the government sector, so the central translation platform project was initiated.

In terms of Information Extraction, there are several NER models, as part of EstNLTK[24] or on top of BERT[25] and also resources for time expression extraction.[26]

Existing virtual assistant solutions (Alexa, Siri, etc.) provide little value for Estonian as they do not understand the language. On the other hand, simple "Estonian-speaking" chatbots are widely used on the websites of companies and institutions to provide help for common problems.

The need for LT support has been acknowledged by Estonian government agencies and policy-makers. Since 2006 there has been a series of National Programmes for Language Technology, with the current one in force until the year 2027.[27]

A new national AI strategy[28] (2022–23) has been published recently. The Estonian Language Development Plan[29] states the development of LT as a priority. The national research infrastructures relating to LT in Estonia are the Center of Estonian Language Resources[30] and the Competence Center for Natural Language Processing.[31] Estonia is a member of CLARIN, ELRC, and ELG.

3 Recommendations and Next Steps

In terms of gaps, Estonian lacks both annotated data and tools for certain tasks and, as annotating data is a time- and workforce-consuming process, it can be seen as an obstacle. Furthermore, Estonian lacks parallel Estonian – non-English data as a result of direct translation between these language pairs. Bigger and/or special multimodal

[19] https://github.com/texta-tk/texta

[20] https://tekstiks.ee

[21] https://github.com/alumae/kiirkirjutaja

[22] http://www.eki.ee/heli/index.php

[23] https://neurokone.ee

[24] https://github.com/estnltk/estnltk

[25] https://github.com/TartuNLP/bert-ner-service

[26] https://github.com/soras/Ajavt

[27] https://www.hm.ee/sites/default/files/documents/2022-10/estonian_language_technology_2018-2027.pdf

[28] https://e-estonia.com/wp-content/uploads/factsheet-ai-strategy-feb2023.pdf

[29] https://www.hm.ee/en/ministry/ministry/strategic-planning-2021-2035

[30] https://www.keeleressursid.ee/en/

[31] https://portaal.eki.ee

corpora are needed, e. g., containing children's or senior's speech, accented speech etc. We also need more audio data for natural and noisy communication situations: spontaneous conversations, spontaneous meetings etc. Estonian lacks annotated resources containing non-normative language varieties, such as the written language variants used on social media or specialised languages used by professionals (healthcare, legal sphere etc.). Computational semantics for Estonian is under-resourced; we lack resources and tools for semantic role labeling, coreference resolution, relation extraction and event extraction, also for polarity detection. There is a need for text simplification, summarisation and paraphrasing tools and resources. In the field of discourse modelling and pragmatics, good and useful theoretical application-oriented research has been carried out, but that has yet to be put into practice. There is also a growing popularity of Digital Humanities and a need to process older written variants of Estonian.

Despite various national and international programmes, initiatives and strategies, there is still a lack of continuity in funding as research funding in Estonia is entirely project-based, which is not sufficient to address the gaps in research and technology support for the Estonian language.

References

Erelt, Mati (2003). *Estonian Language*. Linguistica Uralica Supplementary Series. Tallinn: Estonian Academy Publishers.

Liin, Kristra, Kadri Muischnek, Kaili Müürisep, and Kadri Vider (2012). *Eesti keel digiajastul – The Estonian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/estonian.

Muischnek, Kadri (2022). *Deliverable D1.12 Report on the Estonian Language*. Reports on European Language Equality (ELE) | Coordinator: Prof. Dr. Andy Way, Co-Coordinator: Prof. Dr. Georg Rehm, received funding from the European Union (EU project no. LC-01641480 – 101018166). https://european-language-equality.eu/reports/language-report-estonian.pdf.

Chapter 15
Language Report Finnish

Krister Lindén and Wilhelmina Dyster

Abstract During the last ten years, digitalisation has changed the way we interact with the world creating an increasing demand for language-based AI services. In the field of language technology, the Finnish language is still only moderately equipped with products, technologies and resources. The situation has improved in recent years, but still support for automated translation leaves room for ample improvement, as the general support for spoken language is modest in industry applications although some recent research results are encouraging. We take stock of the existing resources for Finnish and try to identify some remaining gaps.

1 The Finnish Language

Finnish is the native language of about 4.9 million people in Finland and the second language of 0.5 million Finns (see Koskenniemi et al. 2012; Lindén and Dyster 2022). Finnish is spoken in several European countries as well as the United States and Australia. Finnish is an official language in the European Union. The Finnish constitutional law and language law define Finnish and Swedish as the national languages of Finland. Moreover, Finnish is an official minority language in Sweden.

The Finnish literary language has a relatively short history. It has been used in religious literature and the church since the 16th century. Laws have been written in Finnish since the 18th century. Until the 19th century, Swedish was used in administration, education and literature, when the foundation of contemporary Finnish was laid and Finnish became a sovereign language in all societal activity.

Dialects are divided into two categories: the Western and the Eastern dialects. The difference is mostly in the pronunciation and word forms (*meijän*, *männä* in the East, *meirän*, *mennä* in the West) and partly in the vocabulary (*vasta* in the East, *vihta* in the West). The differences are clear, and speakers from different areas can be identified by their intonation. However, the differences are minor enough to allow speakers of different dialects to understand each other.

Krister Lindén · Wilhelmina Dyster
University of Helsinki, Finland, krister.linden@helsinki.fi, wilhelmina.dyster@helsinki.fi

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_15

Finnish is used widely and actively on the internet and social media. Almost all Finnish households (96%) have access to the internet. Traficom, the Finnish Transport and Communications Agency, reported in November 2020 that the total number of registered FI-domains had reached 500,000.

2 Technologies and Resources for Finnish

The development of Finnish language data and tools has progressed steadily over the past 30 years. Since 1995, the Language Bank of Finland[1] and since 2015 CLARIN and FIN-CLARIN have offered a wide variety of text and speech corpora and tools. Today, a large number of fundamental tools and datasets are available for Finnish. Below we present some relevant resources in the different domains of LT.[2]

There are several large monolingual corpora with contemporary textual and spoken language as well as some multilingual corpora. Overall, general domain data seems to be prevalent, e. g., data collected from discussion forums or using web crawls. In addition, news texts, legislative texts and parliamentary speech are well-represented domains. The Language Bank of Finland has the expertise to handle sensitive data, but for example, health domain corpora are still scarce.

The Institute for the Languages of Finland has comprehensive collections of lexical corpora. The Helsinki Term Bank for the Arts and Sciences (HTB)[3] is a multidisciplinary project aiming to gather a permanent terminological database for all fields of research in Finland. The Comprehensive Grammar of Finnish[4] was published in 2004 by the Finnish Literature Society. FinBERT[5] is a version of Google's deep transfer learning model for Finnish, developed by the TurkuNLP Group. FinBERT is pre-trained with 1 million steps on over 3 billion tokens of Finnish text drawn from news, online discussion, and web crawls. Important software packages are: 1. The Helsinki Finite-State Transducer (HFST)[6] can be used to implement morphological analysers. 2. The Turku Neural Parser Pipeline developed by TurkuNLP[7] is an open source dependency parsing pipeline. 3. The Aalto University Automatic Speech Recognition System (Aalto-ASR)[8] provides functionalities for ASR from audio files and for automatic forced alignment of text and speech. 4. OPUS-MT[9] focuses on the development of free resources and tools for machine translation, with

[1] https://kielipankki.fi

[2] META-SHARE Finland contains additional resources, see https://metashare.csc.fi.

[3] https://tieteentermipankki.fi

[4] https://kaino.kotus.fi/visk/

[5] https://github.com/TurkuNLP/FinBERT

[6] https://hfst.github.io

[7] http://turkunlp.org/Turku-neural-parser-pipeline/

[8] http://urn.fi/urn:nbn:fi:lb-2021082323

[9] https://github.com/Helsinki-NLP/Opus-MT

currently over 1,000 pre-trained neural MT models. 5. Finto AI[10] is a service for automated subject indexing, which can be used to suggest subjects for texts in Finnish, Swedish and English. 6. Wavelet-based embedding models for speech synthesis for Finnish have been developed at the University of Helsinki.

The Language Bank of Finland supports academic research and provides some support for the industrial use of academic resources which are also available for commercial use. CSC (IT Center for Science) is tasked with providing one of the three EuroHPC supercomputers, LUMI. The whole system is designed with AI, machine learning and data analytics in mind. LUMI's first pilot phase was concluded by the end of 2021, and LUMI will reach its full capacity in 2022.[11]

Generally, the Finnish market is extremely active in the AI field. According to the State of AI in Finland report by FAIA (2020), "there are over 1,250 companies that use different AI applications, of which roughly 750 have developed their own technology." A rapidly growing startup ecosystem boosts AI/LT development.

3 Recommendations and Next Steps

In November 2019, VAKE (currently the Climate Fund) published a report (Jauhiainen et al. 2019), specifying the next phase of the language-centric AI development programme and identifying topics in need of intervention. In November 2020, Finland launched an updated national AI strategy. The AI 4.0 Programme promotes the use of AI and other digital technologies in companies, with a special focus on SMEs. In the first interim report,[12] published in April 2021, the programme presented a vision for the future of the Finnish manufacturing industry, stating that by 2030 the Finnish manufacturing industry will be clean, efficient and digital. As stated in the report, seamless collaboration between high-speed telecommunications networks, cloud computing and AI are central to the digital transformation.

According to the VAKE report, we need the availability and accessibility of components for processing speech with open licences to create prototypes or develop methods into full-scale production versions in the hands of companies. To this end, collaboration between stakeholders is needed: an ecosystem with a forum or a platform where different-level actors can come together to exchange experiences and seek new projects and collaboration opportunities.

Currently, 1. there are some multi-modal resources, but still no advanced discourse processing tools for Finnish; 2. several research projects are working on advanced information retrieval (IR) and data mining for Finnish; 3. the legal situation has become clearer with the General Data Protection Regulation (GDPR), but we are still waiting for Finland to fully implement the Digital Single Market Directive (DSM); 4. we have some specific corpora of high quality, but the commercial sector

[10] https://ai.finto.fi

[11] https://www.lumi-supercomputer.eu

[12] http://urn.fi/URN:ISBN:978-952-327-643-7

in Finland still needs large, up-to-date resources for product development targeted at everyday users and technologies to collect specialised data sets; 5. work on semantics has still not led to significant applications, but this is explored in the context of advanced research projects on IR and extraction; 6. in speech technology, the recent biggest leaps forward have been made using neural network technology. This has also led to some improvements for the commercial sector offering speech-based services, but speech and video corpora are no longer considered hard to collect with the advent of mobile phones and teleconferencing.

Speech corpora and especially resources for spontaneous speech recognition and various genres of speech synthesis are currently being developed. The need for extensive and varied text materials can to some extent be rectified for research purposes through corpus collections of publicly produced language material when properly considering GDPR and the DSM directive. This will enable the creation of language models. However, we still need a variety of specialised data sets for domain-specific purposes to adapt open-source or proprietary software components. Developing dedicated components from scratch requires giga-scale data sets which may be difficult to compile for small language communities and in specialised domains. This points to a need for a general-purpose language-centric AI which can leverage cross-language and cross-domain resources and benefit from adaptation to local language varieties and specialised domains with small or medium-sized data sets.

References

Jauhiainen, Tommi, Mietta Lennes, and Terhi Marttila (2019). *Suomenkielisen tekoälyn kehittämisohjelma – esiselvitys*. http://hdl.handle.net/10138/319478.

Koskenniemi, Kimmo, Krister Lindén, Lauri Carlson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirka Hyvärinen, Imre Bartis, Pirkko Nuolijärvi, and Aino Piehl (2012). *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/finnish.

Lindén, Krister and Wilhelmina Dyster (2022). *Deliverable D1.13 Report on the Finnish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-finnish.pdf.

Chapter 16
Language Report French

Gilles Adda, Ioana Vasilescu, and François Yvon

Abstract This chapter presents a survey of the current state of technologies for the automatic processing of the French language. It is based on a thorough analysis of existing tools and resources for French, and also provides an accurate presentation of the domain and its main stakeholders (Adda et al. 2022). The chapter documents the presence of French on the internet and describes in broad terms the existing technologies for the French language. It also spells out general conclusions and formulates recommendations for progress towards deep language understanding for French.

1 The French Language

French is typologically a Romance language, closely related to other languages whose origin is Latin (e. g., Italian, Spanish, Portuguese, Romanian). French inherited Gaulish features from the Celtic dialects spoken by ethnic groups that previously populated the territory conquered by the Romans, and was later influenced by Germanic dialects as a consequence of the invasions that marked the fall of the Roman Empire. Modern French uses the Latin alphabet and has retained many Latin linguistic features. For instance, French is a nominative-accusative and article-based language (SVO) that greatly simplified the nominal and verbal declensions. French developed a large vocalic system including 12 oral and 4 nasal vowels.

With 128 million "native and real speakers" worldwide and an estimate of close to 300 million speakers overall (Collectif 2019), French appears only as the 16th most spoken native language, but as the 6th most spoken language in the world, after English, Chinese Mandarin, Spanish, Hindi and Russian. French is an official language in close to 30 countries, most notably in Europe (France: 65m speakers, Belgium: 7m speakers, Switzerland: 3m speakers, and Luxembourg), Africa, Canada and Haiti. In Europe, it is estimated that 129 million people speak French making it the 3rd most spoken second language, after English and German. French-speaking

Gilles Adda · Ioana Vasilescu · François Yvon
Université Paris-Saclay, CNRS, LISN, France, gilles.adda@limsi.fr, ioana.vasilescu@limsi.fr, francois.yvon@limsi.fr

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_16

countries together constitute *La Francophonie*, with the *Organisation Internationale de la Francophonie* coordinating policies between 88 associated states and entities.

Collectif (2019) notes that in 2018 French occupies the fourth place on the internet behind English, Chinese and Spanish, with a comfortable lead over the next set of languages. Pimienta (2022) observes that although French remains in fourth place on the internet in 2022, the gap to the following languages has considerably narrowed. The presence of French on the internet derives from its role as an international language: French is an official language of the EU and one of the three working languages of the European Commission. French is also a working language at the Organisation for Economic Co-operation and Development, and at the United Nations. French is also one of the three official languages of the European Patent Office and one of the four working languages of the African Union.

2 Technologies and Resources for French

Looking at the technology landscape for French, most state-of-the-art tools and applications rely almost exclusively on generic machine learning technologies, a major change with respect to the previous survey (Mariani et al. 2012): the most important ingredients for system building are data and, to a lesser extent, compute resources. We will, therefore, focus on the most critical language resources and give a general overview of the various technologies derived from them.

Large-scale, general purpose lexica for French associating lemmas or word forms to morpho-syntactic information are widely available. There is no official French National Corpus that would contain a representative subset of the language, balanced across periods, genres and domains, as may exist for other languages. However, sizable corpora (up to billions of tokens) of mixed genres are accessible and searchable.

The CommonCrawl project aggregates Web data that is orders of magnitude larger than these resources; and it is updated on a regular basis. Using French subsets of CommonCrawl, it has been possible to train large language models (LMs): FlauBERT uses a corpus of 12B running words, while CamemBERT uses the 22B words OSCAR. Other large LMs for French are available for research and commercial use; they help to boost the state-of-the-art for multiple NLP tasks.

Large-scale annotated (segmented in sentences, speakers and turns, transcribed) speech databases, containing thousands of hours of recordings are available for several genres. Such resources have enabled advanced technologies for French (transcription, synthesis, NLU). However, the collection of large sets of recordings remains a pressing issue to widen the applicability of these technologies, an objective addressed by Mozilla's Common Voice[1] or the Voice Lab project.[2]

Basic NLP tools were already well covered in 2012 and they have benefited from improvements in machine learning. Open source industrial strength tokenizers, lem-

[1] https://commonvoice.mozilla.org/fr

[2] http://www.levoicelab.org

matizers and POS taggers for French are available. We note, however, that no recent systematic performance comparisons exist for these tasks; most of these tools process "generic" French and too little exists for specific sublanguages.

Having moved to fully neural, the availability of Machine Translation systems for French mostly depends on the availability of parallel corpora. Good resources exist for French, especially when matched with an English translation.

As for most social science and humanities domains, the digital revolution has created new avenues for language analysis. Such methodological changes are also happening for French and impact all linguistic domains, with the creation of corpora, tools and methods. Regarding corpora, both written and spoken varieties of French are well covered, although for historical reasons written sources are more common.

Owing to its role as an international language and the comparatively large size and advanced development of French-speaking markets, French is relatively well covered by international LT services: French-English has been one of the earliest translation pairs on the Web, and French versions of Siri, Amazon Echo and Google Home have been available for years. The development of LTs for French far exceeds the activity observed in France or other French-speaking countries.

Institutional support to LTs is mostly operated by the ANR (the French National Research Agency), albeit with a lack of continuous funding; large variability in funding over the years is not favourable to planning. The French research community is nonetheless active, with a dozen significant academic clusters all over France, as well as Belgium, Canada and Switzerland, covering the full spectrum of NLP. This research has greatly benefited from the development of the Jean Zay platform, an open high-performance computing infrastructure tailored to AI applications.

3 Recommendations and Next Steps

Many open-domain French corpora are the result of uncoordinated initiatives and consequently only partially cover the needs of domain-specific applications. This state of affairs results in 1. a lack of visibility of tools and data that are only known to restricted sub-communities, and 2. a waste of resources, as existing datasets are underused, or even duplicated, when other pressing needs remain unsatisfied. A first recommendation is thus to institutionalise clearer policies for the archiving of LRs for French, when they are produced by public research projects.

A second recommendation, aimed to increase the diversity and size of existing corpora, is to open the large datasets produced by public administration and institutions (e. g., in health, culture, media, justice or education) which are hard to access. Policies are needed to amplify the actions of the European CEF/ELRC programme to incentivize the development of open repositories with clear access rules.

Applications that involve social network data (e. g., opinion mining, fake news and hate speech detection) require specific actions, as they are often associated with delicate legal issues (related to proprietary rights or personal information) that limit their dissemination and exploitation. To reduce the dependency on current data poli-

cies of content holders, a third recommendation would be to secure access to sensitive data for research purposes and to facilitate the dissemination of publicly produced databases and models (e. g., using privacy-preserving techniques).

Recommendation four is the definition of a strategic roadmap for identifying, building, curating, annotating and securing resources for language varieties or domains that are critical for research, industry or for the administration in each French-speaking country, based on a precise analysis of the gaps in the existing datasets (some were alluded to above). This roadmap should also identify cases where resources can be transferred from English through MT.

Recommendation five aims to ensure, through recurrent funding, that evaluation campaigns specifically targeting French for a large number of applications are organized on a regular basis and widely advertised, so that systems are evaluated under real world conditions, so as to document their biases, defects and harmful impacts.

The final recommendation is to increase the support for research on themes such as fair and explainable deep learning for large language models, deep language analysis algorithms and technologies, multimodal resources for the study of language acquisition through interactions and grounding, and the study of pathological language processing. This multidisciplinary research should involve all relevant communities.

References

Adda, Gilles, Annelies Braffort, Ioana Vasilescu, and François Yvon (2022). *Deliverable D1.14 Report on the French Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-french.pdf.

Collectif (2019). *La langue française dans le monde*. OIF/Gallimard.

Mariani, Joseph, Patrick Paroubek, Gil Francopoulo, Aurélien Max, François Yvon, and Pierre Zweigenbaum (2012). *La langue française à l' Ère du numérique – The French Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/french.

Pimienta, Daniel (2022). "La place du français sur Internet". In: *La langue française dans le monde 2022*. OIF/Gallimard, pp. 26–27. https://www.francophonie.org/sites/default/files/2022-03/Synthese_La_langue_francaise_dans_le_monde_2022.pdf.

Chapter 17
Language Report Galician

José Manuel Ramírez Sánchez, Laura Docío Fernández, and Carmen García-Mateo

Abstract This chapter reports on the current state of Language Technology (LT) for Galician. The main conclusion is that there are a limited number of resources, products, and technologies for the Galician language with text-based technologies and services being more mature than those based on speech processing. We start with general facts about Galician, followed by a high-level qualitative description of the LT situation for Galician, and conclude with recommendations for bridging the gap between Galician LT with Spanish and the other co-official languages of Spain.

1 The Galician Language

Galician is part of the Romance family of languages, closely related to Portuguese, and it is one of the co-official languages of Spain. The linguistic rights of Galician speakers are guaranteed and regulated under the Linguistic Normalisation Act, especially those related to administration, education, and media. Galician has about 1,926,000 speakers. There are still large Galician-speaking communities outside Spain (mainly in Europe and America). Their total size is unknown due to the variety and complexity of these communities.

The online presence of Galician is limited, with less than 0.1% of websites using it.[1] Nevertheless, some initiatives try to increase the presence of Galician on the web (PuntoGal[2] and Galipedia[3] are good examples). The official survey *Enquisa estrutural a fogares. Coñecemento e uso do galego* shows a generally low internet penetration and use by European standards, but between the ages of 15 and 44, the numbers are very similar to other European regions.[4]

José Manuel Ramírez Sánchez · Laura Docío Fernández · Carmen García-Mateo
Univ. of Vigo, Spain, jmramirez@gts.uvigo.es, ldocio@gts.uvigo.es, carmen.garcia@uvigo.es

[1] https://w3techs.com/technologies/details/cl-gl-

[2] https://dominio.gal

[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] http://www.ige.gal/estatico/html/gl/OperacionsEstruturais/PDF/Resumo_resultados_EEF_Galego_2018.pdf

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_17

A substantial amount of digital content in the Galician language is generated by public institutions of the Autonomous Community of Galicia. In the last few years, the number of products and services developed has increased considerably, aimed at incorporating Galician into the digital society. The web portal of the Real Academia Galega and the Xunta de Galicia translator are noteworthy examples. Although some large enterprises (Microsoft, Apple, Google, Meta) offer a few products with support for Galician, many others do not (TikTok, Twitch, Adobe). However, there is a total lack of support for Galician in the virtual assistants market, where none of the popular solutions allows interaction via Galician.

2 Technologies and Resources for Galician

The 2012 META-NET White Paper on Galician (García-Mateo and Arza 2012) was moderately optimistic about the state of LT support for the language. Ten years later, the LT status for Galician has changed a bit (Sánchez and García-Mateo 2022). In our analysis, we noticed an increase in the resources and corpora created between 2018-2021 (67.7% of those indexed). However, tools and services developed in the same period have not increased to the same degree (37.3% of those indexed). There is a significant imbalance in the distribution of resources and corpora by technologies. Data in text format are the most common (more than 90%), whereas corpora for other technologies are very few (5% are multi-modal, and almost 2% are audio only).

Most of the resources come from three origins: non-Galician universities and research centres, Galician public institutions, and non-Galician private companies or public institutions. It is important to note that most of the resources, services, and tools created by non-Galician entities tend to belong to multilingual projects or products that include Galician as one of several languages. However, most of the resources, services, and tools created by Galician entities tend to focus on Galician, offering high-quality results.

Regarding the accessibility and use of resources for Galician, most of them have been developed by open source projects, research centres, or universities under GNU/GPL licences. Around 20% of the indexed items are not available for commercial purposes, and more than 10% of resources are under a proprietary licence.

The situation of Galician in terms of data and resources is optimistic for most of the technologies that process and use text. However, regarding multimedia data, there is an enormous gap. In that sense, speech processing technologies seem less mature than technologies based on text processing.

For Galician, key results regarding technologies and resources include:

- There are large reference text databases in modern and historical Galician with a balanced mix of various domains (economics, technology, or the legal field) (Piñeiro 2019; García-Mateo et al. 2014).
- There are some databases annotated with syntactic, semantic, or discursive information. However, the number and size of these resources decrease as more complex linguistic and semantic information is needed.

- Parallel databases with millions of tokens exist between Galician and other languages such as Spanish, Portuguese, and English (OPUS[5] is a good example). These databases have been used to develop machine translation systems in production and education environments for Portuguese or Spanish.
- A relevant model to highlight is Bertinho (Vilares Calvo et al. 2021), a monolingual BERT model for Galician. Bertinho implements state-of-the-art technology, and it is possible to use it in many NLP tasks. However, its developers state that Bertinho does not reach the size or performance of other monolingual versions, such as BETO for Spanish.
- Available multimedia resources are relatively limited, with little domain variability and usually recordings of readings. The acoustic quality is excellent though.
- Another gap is related to human-computer interaction, where the necessary tools and resources to put together chatbots, virtual assistants, and similar systems are poor or outdated.

Spain has national plans for both Artificial Intelligence (AI, Gobierno_de_España 2020a) and LT (specifically for NLP, Gobierno_de_España 2020b). These plans focus more on the potential, opportunities, and needs of Spanish LT, putting less emphasis on co-official languages such as Galician. Two national associations bring together the community of researchers on issues related to LT: *Sociedad Española de Procesamiento del Lenguaje Natural* with a focus on NLP, and the *Red Temática en Tecnologías del Habla* with its focus on speech processing.

The Autonomous Community of Galicia has its own strategy for AI.[6] This document describes the current environment of AI in Galicia and provides a roadmap for public investments and developments until 2030. There is also an initiative called Proxecto Nós, a regional LT plan for Galician focused on digital challenges promoted by the Galician regional government. Furthermore, there are many more projects related to LT in the Galician university environment, both from a linguistic and technological point of view. Another interesting fact is that from the number of companies in the Galician ICT industrial environment that use AI, only 21% are focused on cognitive assistants and just 12% on NLP. The Galician LT industry is very small, but a very active environment of spin-offs and public programmes exists dedicated to transferring knowledge from universities to the market.

3 Recommendations and Next Steps

The main goal of LT for Galician is to reach the level of other co-official languages of Spain, such as Catalan or Basque. In this sense, increasing the use of LT in Galician public services and institutions could be a necessary line of action to support and stimulate research and development of new resources and better tools. Galician

[5] https://opus.nlpl.eu

[6] https://amtega.xunta.gal/sites/w_amtega/files/20210608_estrategia_ia_gl.pdf

institutions are the producers of high-quality resources and tools for Galician. However, there is a lack of standardisation and dissemination of these products. An office that centralises and standardises all the LT resources and tools created for Galician could be a significant contribution to unifying all efforts.

Support for open source solutions (data and software) would be a good long-term strategy for small-market languages. These solutions allow the development and research of new technologies without having to face an initial investment barrier. Furthermore, an open-source policy encourages the creation of strong communities and guarantees some technological sovereignty from the interests of global markets and multinational corporations.

References

García-Mateo, Carmen and Montserrat Arza (2012). *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/galician.

García-Mateo, Carmen, Antonio Cardenal López, Xosé Luis Regueira, Elisa Fernández Rei, Marta Martinez, Roberto Seara, Rocío Varela, and Noemí Basanta (2014). "CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2653–2657.

Gobierno_de_España (2020a). *Estrategia Nacional de Inteligencia Artificial 2020*. https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2020/201202_np_ENIAv.pdf.

Gobierno_de_España (2020b). *Estrategia Procesamiento del Lenguaje Natural 2020*. https://drive.google.com/file/d/1eXlFdRNTmOx4sm3FQ439Z8zaeNqEFGiK/view.

Piñeiro, Centro Ramón (2019). *Corpus de Referencia do Galego Actual (CORGA) [3.2]*. http://corpus.cirp.gal/corga/.

Sánchez, José Manuel Ramírez and Carmen García-Mateo (2022). *Deliverable D1.15 Report on the Galician Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-galician.pdf.

Vilares Calvo, David, Marcos García González, and Carlos Gómez Rodríguez (2021). "Bertinho: Galician BERT representations". In: *Procesamiento del lenguaje natural*, pp. 13–26.

Chapter 18
Language Report German

Stefanie Hegele, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, and Georg Rehm

Abstract German is the second most widely spoken language in the EU. The last decade has seen strongly perceptible language change, trending towards the simplification of the grammatical system, a rapidly growing number of anglicisms, a decreasing prevalence of dialects, and an increase in socio-political debates on matters such as language policies and gender-neutral language. Many technologies and resources for German are available, which is also due to numerous well-established research institutions and a thriving Language Technology (LT) and Artificial Intelligence (AI) industry. In order to withstand in the digital sphere, it is important that incentives for research, digital education and also concrete opportunities for marketing and deploying LT applications are put at the forefront of future AI strategies.

1 The German Language

With more than 150 million native and non-native speakers (Eberhard et al. 2021), German is the second most widely spoken language in the European Union. Germany, Austria and Switzerland form the DACH region, which is not only home to the three (codified) standard varieties of the German language, but also boasts a wealth of regiolects and dialects. Perceptible language change in German has been omnipresent for decades, leaving the language community to decide what becomes the norm. According to three reports on the state of the German language,[1] published in the years 2013-2021 by the Union of the German Academies of Sciences and Humanities, changes lean heavily towards the simplification of the grammatical system.

Stefanie Hegele · Katrin Marheinecke · Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany,
stefanie.hegele@dfki.de, katrin.marheinecke@dfki.de, georg.rehm@dfki.de

Barbara Heinisch · Dagmar Gromann
University of Vienna, Austria, barbara.heinisch@univie.ac.at, dagmar.gromann@univie.ac.at

Antonia Popp · Annette Rios · Martin Volk
University of Zurich, Switzerland, popp@cl.uzh.ch, rios@cl.uzh.ch, volk@cl.uzh.ch

[1] https://www.akademienunion.de/publikationen/sammelbaende

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_18

There has also been a huge expansion in vocabulary. Over the last decades, many Anglicisms have been introduced into the language, that either replace existing German words or fill vocabulary gaps. Dialects have been more and more displaced.

German uses grammatical gender. However, nouns that refer to the social gender are often biased towards the male form. Proponents of a gender-inclusive language advocate that German needs a grammar that explicitly includes women and non-binary people, making all people feel equally addressed.

Public debates about language policy positions are becoming more frequent and also more heated. They attract a great deal of media attention in Germany. The New Right tries to use the topic of language in a targeted manner and to instrumentalise it in terms of national identity (Lobin 2021).

There are a number of non-governmental, publicly funded organisations that promote the study of German and encourage international cultural exchange, such as the Goethe Institute, the Society for the German Language, or the Institute for the German Language.[2]

Regarding language education, the PISA study has continued to confirm the strong correlation between socio-economic background and educational success.[3] Fears that the increased use of social media and emojis would worsen young people's writing skills cannot be confirmed. Instead, the emergence of new written forms should be noted (Beißwenger and Pappert 2020; Storrer 2014).

German is currently the second most studied foreign language in the EU, but is also gaining importance in Africa and Asia.

German has a widespread online presence and the fourth largest Wikipedia. Internet use continues to rise. According to the European Statistical Office (Eurostat), in both Germany and Austria, there are more than 85% of regular internet users and close to 70% of people with basic or above basic digital skills.

2 Technologies and Resources for German

German has many linguistic characteristics and particularities such as relatively free word order and fairly long nested sentences (Eroms et al. 2003) that pose challenges for Natural Language Processing tasks. Nevertheless, German is well supported by Language Technology (LT) applications and resources compared to most other European languages. A number of large-scale resources and state-of-the-art technologies have been produced for Standard German. However, dialect-specific resources currently account for only a small percentage.

There exist a large number of German corpora of different sizes, ranging from a few hundred sentences up to millions. The sources are most often newspaper texts or texts collected from the web and social media. Various terminological resources, lexica, dictionaries or word lists have also been developed for German. Annotations

[2] https://www.goethe.de, https://gfds.de, https://www.ids-mannheim.de

[3] https://www.bmbf.de/bmbf/shareddocs/pressemitteilungen/de/pisa-2018-deutschland-stabil-ueber-oecd-durchschnitt.html

cover a large spectrum of syntactic, semantic, and discourse structure markup. The most frequent corpus domains include health, news, politics and social media. Currently, there are only a few language models publicly available for German.

In addition, there are numerous free multilingual resources available online for German, e. g., the LEO dictionary. Other widely used MT systems are DeepL and Google Translate which cover the translation from German into dozens of languages. EUROPEANA functions like a multimedia portal and digital library with content from different sources.[4] By the end of 2015, Germany, Austria and Switzerland had contributed around 16% to the more than 24 million objects.

Hundreds of tools, both open source and commercial, that work either exclusively for German or multiple languages including German have been developed. The vast majority process text input. Even though speech technology has already been successfully integrated into many everyday applications, from spoken dialogue systems and voice-based interfaces to mobile phones and car navigation systems, audio is only supported by a small fraction of tools, and image and video by even less.

Research over the last decade and the deployment and integration of LT components to end-to-end processing pipelines has successfully led to the design of high-quality software with many tools supporting more than one function. The most frequent tasks supported by the current collection of German tools include text and data analytics, information extraction, named entity recognition, information retrieval and speech recognition. Tools developed by universities and research centres are typically available for all users free of charge.

The research community in Germany, Austria and Switzerland has been growing rapidly over the last decade. Numerous universities offer study programmes focused on Language Technology, NLP, Computational Linguistics and closely related disciplines. Recent breakthroughs in AI have not only led to cutting-edge technology developed by big companies, but have also inspired numerous startups and SMEs in the field. Current funding programmes, even though mostly targeted towards AI, have also helped to improve research in the field in general, and also have supported a number of research projects working on German in particular. While overall AI strategies vary in the German-speaking regions, the situation for LT/NLP research and development in Germany is, all aspects considered, rather good. The German government aims to invest about 3 billion Euros until 2025 to implement the strategy, including the creation of new AI centres, new funding programmes, new professorships, new international collaborations (e. g., with France) and a new national roadmap for AI standardisation.

3 Recommendations and Next Steps

The scope of resources and range of tools are still limited when compared to English, and they are not yet good or ample enough to develop the kind of technologies re-

[4] https://www.europeana.eu/de

quired to support a truly multilingual knowledge society. High quality data sets and large language models represent a major step forward in AI. Our empirical results show that German is still partially lagging behind in this area (Hegele et al. 2022; Burchardt et al. 2012). There are also gaps in the areas of speech and text processing. In addition, existing technologies do not cover the many different varieties of regional languages and dialects that exist in Germany, Austria and Switzerland. Furthermore, many resources are not available due to copyright reasons, confidentiality, (national) security reasons etc.

While German is among the three best supported European languages (next to Spanish and French), the gap towards English is indeed significant. Without a substantial and timely intervention by the European Union, for many European languages this gap will continue to increase, endangering their digital existence.

References

Beißwenger, Michael and Steffen Pappert (2020). "Sprachverfall durch Emojis? Eine pragmalinguistische Perspektive auf den Beitrag von Bildzeichen zur digitalen Kommunikationskultur". In: *Aptum. Zeitschrift für Sprachkritik und Sprachkultur* 16, pp. 32–50.

Burchardt, Aljoscha, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk (2012). *Die Deutsche Sprache im digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/german.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (2021). *Ethnologue: Languages of the World*. Dallas, Texas. http://www.ethnologue.com.

Eroms, Hans-Werner, Gerhard Stickel, and Gisela Zifonun (2003). *Schriften des Instituts für Deutsche Sprache*.

Hegele, Stefanie, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, and Georg Rehm (2022). *Deliverable D1.16 Report on the German Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-german.pdf.

Lobin, Henning (2021). *Sprachkampf – Wie die Neue Rechte die deutsche Sprache instrumentalisiert*. Berlin: Dudenverlag.

Storrer, Angelika (2014). "Sprachverfall durch internetbasierte Kommunikation?" In: *Sprachverfall?* Ed. by Albrecht Plewnia and Andreas Witt. Berlin: De Gruyter, pp. 171–196.

Chapter 19
Language Report Greek

Maria Gavriilidou, Maria Giagkou, Dora Loizidou, and Stelios Piperidis

Abstract Technological support for Greek, one of Europe's lesser spoken languages, has progressed in the past decade, while LRTs have both increased in volume and improved in quality and coverage. Despite this progress, when compared to the 'big languages', Greek is obviously disadvantaged. Prominent among the challenges is the fact that LT is not included in the language policies or AI strategies of Greece and Cyprus, i. e., the significance of language-centric AI is still not officially recognised. Lack of continuity in research and development funding is an additional factor hampering progress. A Europe-wide coordinated initiative focused on overcoming the differences in language technology readiness for European languages coupled with national targeted actions is considered necessary.

1 The Greek Language

Greek is the official language of Greece, one of the two official languages of Cyprus and, since 1981, one of the official languages of the European Union. It is spoken as a mother tongue by about 95% of the 10.7 million inhabitants of Greece, by around 840,000 Greek Cypriots, and approximately 5 million people of Greek origin worldwide.

Greek is a heavily inflectional language, and has an extensive set of derivational affixes. As regards syntax, it presents a free word order, the neutral order being Verb-Subject-Object or Subject-Verb-Object. The Greek writing system has been the Greek alphabet for most of its history. The Modern Greek alphabet consists of 24 letters. The official orthography of Modern Greek is the simplified *monotonic* (single stress) system, which utilises only stress mark and diaeresis.

Maria Gavriilidou · Maria Giagkou · Stelios Piperidis
R.C. "Athena", Greece, maria@athenarc.gr, mgiagkou@athenarc.gr, spip@athenarc.gr

Dora Loizidou
University of Cyprus, Cyprus, loizidou.dora@ucy.ac.cy

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_19

2 Technologies and Resources for Greek

In the last decade, language resources have both increased in volume and improved in quality and variety (Gavriilidou et al. 2022, 2012). Resources and basic NLP tools are provided by academia, research centres and private companies as outputs of various endeavours (research projects conducted by academic institutions, funded by EU or national funds, commercial projects or self-funded) and made available under various licensing conditions (freely distributed, only for research etc.).

Contemporary written language is represented in three main general domain monolingual text corpora: the Hellenic National Corpus developed by ILSP, the corpora of the Centre for the Greek Language, and the Corpus of Greek texts of the University of Athens. Nonetheless, the size of available corpora does not suffice for valid synchronic linguistic research, and cannot guarantee the development of language models. Multiple bi-/multilingual text corpora which include Greek, developed mostly automatically by leveraging web crawling techniques, have been extensively used for the development and training of MT systems.

Multimodal resources have been developed sporadically, with most systematic efforts concentrated on sign language corpora and lexica. Recent attempts to construct multimodal language resources for speech pathology applications are also noteworthy. With regards to lexical resources, the presence of Greek in various international bi-/multilingual resources (e. g., IATE, WordNet, ConceptNet etc.) is encouraging. Finally, Greek features in some multilingual and/or monolingual language models; recently, three BERT models have been developed for Greek.

Existing basic NLP tools have been improved by adopting deep-learning methodologies and neural networks. The existing pipelines include tools for various types of annotation, i. e., sentence splitting, tokenization, POS tagging, lemmatisation, chunking, and dependency parsing. All pipelines are available for use through the ELG and CLARIN:EL infrastructures. Tools for more advanced tasks such as monolingual information extraction, event detection and named entity recognition have also improved over the last few years, by being trained on new datasets and applied to a variety of domains. Other applications, such as anonymisation, natural language generation and sentiment analysis can be found at different levels of robustness and completeness. Concerning multilingual text processing, MT systems such as eTranslation, Google Translate and DeepL, have significantly improved their coverage of Greek, while a number of MT systems have also been developed by smaller companies in Greece and other EU Member States, and by academic and research organisations. Speech processing has seen important progress: dictation systems for Greek with domain-specific implementations and high-calibre speech synthesis technologies have been made available by commercial providers. Several Greek-speaking digital assistants are also currently available.

Most available LRTs described above are relevant only for Standard Modern Greek. Dialectal varieties of Greek, such as Cypriot Greek, used mainly in oral speech and in specific written speech types (e. g., in poetry and literature), are not equally supported by technology. As Cypriot Greek is distinguished from Standard Modern Greek on several linguistic levels of analysis, it is often the case that exist-

ing LTs trained on Standard Modern Greek data fail to appropriately process Cypriot Greek. At the same time, LRTs developed specifically for Cypriot Greek are sparse. These are mainly general-use lexical resources (dictionaries, glossaries, wordlists). In order to protect this dialectal variety of Modern Greek, as well as the heritage and culture of its speakers, LT research should specifically treat Cypriot Greek.

Public research and academic organisations in Greece and Cyprus play a major role in developing LT, mainly through their participation in national and EU-funded projects in the fields of LT and AI, despite the fact that in the last ten years, there has been no funding programme specifically supporting LT in Greece. Participation in large-scale infrastructures, initiatives and projects, such as CLARIN:EL, ELRC and ELG, has boosted not only R&D in Greek LT, but it has also facilitated sharing and reuse of LRTs. As far as the LT industry is concerned, Greek is part of the portfolios of several multinational commercial providers, while it is also supported by a small but active LT industry in Greece and Cyprus, consisting mainly of SMEs and providing various LT-related services, indicatively: AI, LT (event detection, basic NLP, lexical resources and terminologies), MT and Localisation, Speech Processing (mainly recognition), and Data Science/Big Data Analytics.

3 Recommendations and Next Steps

Despite the progress of Greek LT during the past decade, when comparing Greek to the 'big languages', the abysmal difference in terms of quantity, size and quality of LRTs is evident. Efforts in the coming years should be concentrated on the further development of large-scale monolingual corpora that can be used for training large language models. Semantically annotated datasets, semantic lexica and knowledge bases, and datasets that can be used for anonymisation, simplification, summarisation, text levelling and question answering systems should also be prioritised. Speech and multimodal data are scarcely available, limiting the potential for the development of conversational agents, among others. Greek is dramatically deprived particularly when it comes to conversational data or speech in informal settings that is generated by speakers of different ages, genders and linguistic/dialectal backgrounds. The transition to ubiquitous human-computer interaction in Greek, supported by state-of-the-art research results in NLU and NLG is, unfortunately, still far away. Further challenges posing impediments to the development of LT for Greek include: 1. Scarcity of data: as Greece and Cyprus are small countries, the production of digital language data is limited; 2. Lack of experience in the use of LT: the deployment of digital tools and methods in many disciplines, including life sciences and humanities, has only recently been introduced. Researchers/professionals in these domains need still to be convinced about its benefits; 3. Issues related to IPR or GDPR render resource owners hesitant about sharing their datasets. Non-explicit, unclear terms of use and distribution restrict sharing, use and repurposing of digital texts and language processing tools. The majority of resources pose restrictions on

the types of uses they allow, thus discouraging prospective users, hampering new research and development and leading to repetition in resource creation.

One of the main reasons for the disadvantaged position of Greek is that LT is not included in the language policy of Greece and Cyprus, i. e., the significance of language-centric AI has not been recognised yet. While sporadic efforts, self-funded or partially supported within IT or AI programmes, have yielded results, they are not adequate to boost Greek LT to a state-of-the-art level, nor to help Greece keep pace with developments worldwide. Lack of continuity in R&D funding has been experienced for many years, with short-term projects alternating with periods of drought. While it is important that infrastructural initiatives for LT have been thriving in Greece, their future funding is not secured and their sustainability may be at stake.

A strategy for keeping Greek up to pace with LT developments and ensuring Greek thrives in the digital sphere should foresee: 1. maintenance, extension and sustainability of LT-related infrastructures; 2. national and/or European coordinated actions for ensuring access to open high-performance compute infrastructure; 3. coordinated actions for the development of large-scale LRs ready to power large language models; 4. targeted actions to fill the observed gaps in speech and multimodal data; 5. measures ensuring that the importance of LT and language-centric AI is recognised and included in national policies and strategies; 6. coordinated actions to further enhance digital literacy in the research communities and society as a whole; 7. coordinated actions to promote the culture of data sharing, including open source software, involving all stakeholders, the public sector, research and industry.

References

Gavriilidou, Maria, Maria Giagkou, Dora Loizidou, and Stelios Piperidis (2022). *Deliverable D1.17 Report on the Greek Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-greek.pdf.

Gavriilidou, Maria, Maria Koutsombogera, Anastasios Patrikakos, and Stelios Piperidis (2012). *Η Ελληνικη Γλωσσα στην Ψηφιακη Εποχη – The Greek Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/greek.

Chapter 20
Language Report Hungarian

Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, and Tamás Váradi

Abstract The revolutionary expansion of language technologies (LT) in the last decade and the emergence of neural networks has heavily impacted LT. This is reflected in the development of Hungarian NLP as well, as numerous high-quality LMs, tools and datasets have been created. However, new, huge datasets are still needed to train LMs. Due to being a lesser resourced Uralic language with a smaller number of speakers, Hungarian LT has to face challenges often different from those of large Indo-European languages like English. Here we present a snapshot of this important period in the development of Hungarian LT, with special attention to language resources, and we outline some of the possible next steps.

1 The Hungarian Language

Hungarian, spoken by 13-14 million people globally, is the official language of Hungary and a few Hungarian-majority regions and municipalities in Serbia and Slovenia. 9.8 million speakers live in Hungary and a further 2.5 million speakers use Hungarian as a recognised minority language in neighbouring countries that once belonged to Hungary. An additional 1 million Hungarian speakers live scattered around the globe. There are slight differences across these language variants.

Hungarian belongs to the Finno-Ugric branch of the Uralic language family (Simon et al. 2012). Its linguistic relatives include Finnish and Estonian, with a total number of speakers below 7 million combined. This has implications for Hungarian Language Technology (LT), which cannot draw much support from the technological development of its Uralic relatives. Developers of Hungarian LT face problems such as the extensive case system and agglutination in the language; as nominals inflect for number, case, and person, and verbs inflect for person, number, tense, and mood both in definite and indefinite conjugation paradigms. The Hungarian case system – with around 20 cases (Thomason 2005) – is particularly complex compared

Kinga Jelencsik-Mátyus · Enikő Héja · Zsófia Varga · Tamás Váradi
Research Centre for Linguistics, Hungary, jelencsik-matyus.kinga@nytud.hu, heja.eniko@nytud.hu, varga.zsofia@nytud.hu, varadi.tamas@nytud.hu

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_20

to Indo-European languages. The Hungarian language is written using an extended version of the Latin script, the 44-letter Hungarian alphabet.

Most of the Hungarian-specific LT resources are developed either in Hungary or as part of large, multilingual Pan-European initiatives. The language variant these resources represent is almost exclusively standard Hungarian. Even in the case of corpora, most of the material that creators include comes from within Hungary, with only some exceptions (e. g., Hungarian National Corpus 2).

2 Technologies and Resources for Hungarian

In recent years, the number of application areas of Hungarian LT has greatly increased, and several good quality Hungarian language models, tools, corpora and lexical resources have been created. Huge developments can be seen in the field of AI as well. Below we give a snapshot of Hungarian NLP in this period of swift changes, with a special emphasis on language resources (Jelencsik-Mátyus et al. 2022).

Most monolingual corpora available for Hungarian were not built specifically for LT, however, there is huge improvement in this area. Nowadays, monolingual corpora for Hungarian not only include collections of curated data (see the Hungarian National Corpus 2.0), but also datasets compiled by web crawling (e. g., Webcorpus 2.0). New resources are now built with higher levels of annotation, and often with the purpose to serve as test and training data. For example, HuLu (Hungarian Language Understanding Evaluation Benchmark Kit) can be used primarily for the evaluation and analysis of natural language understanding (NLU) systems, and it aims to be the Hungarian version of the GLUE and SuperGLUE benchmarks. At the same time, multilingual textual data containing Hungarian are abundant with almost 250 datasets, as Hungarian is often included in large EU and non-EU projects alongside dozens of other languages. While multilingual corpora vary across being comparable or parallel, general or domain-specific, there are very few domain-specific monolingual Hungarian corpora, especially from the legal domain. However, datasets an order of magnitude larger are needed to build effective language models. Several corpora to support building LMs are now under construction.

The number of multimodal corpora for Hungarian is quite low, with the most common form being an audio dataset backed with transcripts. Importantly, there are no publicly available domain-specific multimodal datasets of considerable size in Hungarian, so R&D projects need to compile their own resources to train and evaluate speech processing systems.

As BERT has become a standard in NLP, a number of BERT models have been trained for Hungarian (see HuBERT, HILBERT, emBERT). Besides BERT, models with other architectures are being adapted to Hungarian; a couple of experimental models were developed by the HILANCO consortium.

Solutions for the most common tasks in text analysis are available in state-of-the-art NLP tools and pipelines for Hungarian (see UDPipe, HuSpaCy, e-magyar and Magyarlánc). To cover higher levels of text analysis, industrial stakeholders de-

veloped some cutting-edge text analysis toolkits, e. g., Neticle's media monitoring system. However, rapidly expanding demands pose an ever-growing number of challenges for Hungarian LT developers.

There are numerous multilingual speech processing tools covering Hungarian, but only a few Hungarian-specific applications are available. As the DNN approach has become prominent both in TTS and ASR research and development, although there are some high-quality applications for Hungarian, new challenges have been identified. There is a lack of computational and speech resources, i. e., competitive GPU-grids and high-variability natural speech recordings, that hinder the development of TTS and ASR solutions. As for commercial applications, see, for instance, Clementine's Clemvoice that provides services including speech processing, or SpeechTex specialising in TTS for the legal domain.

Neural machine translation (NMT) has become the leading paradigm for MT at large, and for Hungarian as well. A state-of-the-art NMT system is implemented by the Hungarian Research Centre for Linguistics. To carry out high-performance NMT, however, having high quality parallel language data both from general and specific domains is essential. The Hungarian provider Globalese does this by enabling human translators to train the company's NMT engines based on their own parallel data.

Although there are some commercial solutions covering Hungarian (e. g., IntelliDockers engines or SAS), we are not aware of any summarisation tool developed for Hungarian but, as a first step towards such a tool, initial extractive and abstractive summarisation tools were built based on Hungarian-specific Transformer models. A GPT-2 model (with news and poem generators) was also built for Hungarian.

Chatbots and simple task-based systems are increasingly used, but systems that can carry out more open-ended conversations in Hungarian are not yet available.

In the last years, several solutions have been created for information retrieval. Recently, vector space models have been trained with a searchable online interface. Text classification, tag recommendation, topic modelling and sentiment analysis tools have been built to support Hungarian health services and the press.

Following the growth of AI in several fields, numerous national programmes and umbrella organisations were founded recently. The two most prominent organisations in Hungary are the Artificial Intelligence National Laboratory and the Artificial Intelligence Coalition. Their goals include facilitating cooperation and communication between research centres, universities, and industrial AI developers; and, eventually, strengthening the position of Hungarian AI internationally.

3 Recommendations and Next Steps

The emergence of neural technologies has massively reshaped how language data is used in a uniform way in most subfields of NLP. As we have seen in examples ranging from speech processing to summarisation and machine translation, although plenty of monolingual and multilingual corpora were compiled in the past years, there is an ever-growing need for novel datasets for fine-tuning, testing and bench-

marking. Due to their importance, the automatic generation of such resources should be considered as well.

Thanks to the efforts made over the last decade, there are now multiple toolchains performing good-quality linguistic analysis. At the same time, more intricate tasks are still left to be covered, e. g., processing solutions for social media texts should also be expanded. Human-computer interaction is a field that appears to be of utmost importance, but complex conversational agents are not yet available for Hungarian.

There is still ample room for strengthening cooperation between R&D and industry, and their links with the public sector market (e. g., public administration). The future of R&D of Hungarian LT and AI is primarily dependent on various funding agents, as the LT-connected market in itself is currently unable to provide sufficient financial background. Finally, due to the complexity of LT-related knowledge, the need for good-quality and well-organised LT education should be addressed in the long run.

References

Jelencsik-Mátyus, Kinga, Enikő Héja, Zsófia Varga, Tamás Váradi, László János Laki, and Győző Yang Zijian (2022). *Deliverable D1.18 Report on the Hungarian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-hungarian.pdf.

Simon, Eszter, Piroska Lendvai, Géza Németh, Gábor Olaszy, and Klára Vicsi (2012). *A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/hungarian.

Thomason, Sarah G. (2005). "Typological and theoretical aspects of Hungarian in contact with other languages". In: *Hungarian Language Contact outside Hungary*. Amsterdam, Philadelphia: John Benjamins, pp. 11–28.

Chapter 21
Language Report Icelandic

Eiríkur Rögnvaldsson

Abstract In 2019, the Icelandic Government launched a three-year Language Technology Programme for Icelandic (LTPI). Within this programme, a number of language resources and tools have been built from scratch and several pre-existing resources and tools have been enhanced and improved. This programme is now finished and the situation for Icelandic with respect to language technology has improved considerably. In spite of this, Icelandic still remains a low-resourced language compared to most official European languages.

1 The Icelandic Language

Icelandic is a North Germanic language with its roots in Old Norse. It is the only official language of Iceland apart from Icelandic Sign Language. Even though it is only spoken by around 350,000 people in Iceland and by several tens of thousands of Icelanders living abroad, it is not considered endangered according to UNESCO's Language Vitality Scales[1] or EGIDS.[2] The language community is very homogeneous, and dialectal variation is negligible.

Icelandic is a morphologically rich language; nouns, pronouns, adjectives and verbs are inflected for several grammatical features. The language is fusional, such that a single ending usually stands for more than one morphological category. Typologically, Icelandic is an SVO (subject-verb-object) language with a strong V2 rule that requires the verb to appear in the second (or first) position of the sentence. However, because of the rich inflectional system, word order is relatively free.

The Icelandic alphabet is based on the Latin alphabet with a number of additions, especially vowel symbols with an acute accent, *á é í ó ú ý Á É Í Ó Ú Ý*, and the vowel symbols *æ Æ* and *ö Ö* which are also used in a number of other languages. Furthermore, Icelandic employs two more eccentric symbols: *ð Ð* (eth, not to be

Eiríkur Rögnvaldsson
Árni Magnússon Institute for Icelandic Studies, Iceland, eirikur@hi.is

[1] https://ich.unesco.org/doc/src/00120-EN.pdf

[2] https://www.ethnosproject.org/expanded-graded-intergenerational-disruption-scale/

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_21

confused with "d with a stroke", *đ*) which is also used in Faroese, and *þ Þ* (thorn) which is not used in any other language.

Iceland has the highest percentage of internet users in Europe. In 2020, 98% of Icelandic households had internet access.[3] In the same year, 68,344 websites had .is as the top level domain.[4] Icelandic is sufficiently represented on the internet, with a number of media websites and an Icelandic Wikipedia, for instance, but most people also frequently visit news sites in English, access various types of information in English, etc. Even though Icelandic is the main language used on social media, English is also prominent.

2 Technologies and Resources for Icelandic

The Icelandic Government launched the Language Technology Programme for Icelandic (LTPI) in September 2019. The self-owned foundation Almannarómur[5] was entrusted with the role of conducting the programme. Almannarómur, in turn, commissioned the SÍM Consortium,[6] comprising members from academia, NGOs and the private sector, to carry out the research and development work in this project. Researchers, developers and LT users are well represented in the Consortium.

Most of the existing resources and tools for Icelandic are direct or indirect outputs of this programme. Almost all of these resources and tools are stored in the CLARIN-IS repository.[7] They can be downloaded for free, most of them under standard open licences, and used in any kind of application.

The Icelandic Gigaword Corpus (IGC) is a monolingual corpus comprising almost 2.7 billion tokens of different genres. Most of the texts are from 2001-2022. A few parsed corpora exist, most of them having been automatically parsed. Greynir-Corpus contains 10 million sentences from news sources which have been parsed into full constituency trees. The Icelandic Contemporary Corpus is a constituency parsed corpus built by using an Icelandic model of the Berkeley Neural Parser and containing 30 million clauses from the IGC. A number of small specialised corpora have also been developed.

There exist a number of bilingual English-Icelandic corpora. Most of them are domain-specific corpora from ELRC and are not aligned. However, a few general purpose aligned corpora exist, the most important being ParIce with 5.3 million translation units. Much larger bilingual corpora are needed, especially between Icelandic and English but also between Icelandic and other languages such as Polish.

A few audio corpora exist. The most important one is Talrómur which consists of 122,417 short audio clips of eight different speakers reading short sentences, amount-

[3] https://www.statista.com/statistics/185663/internet-usage-at-home-european-countries/

[4] https://www.isnic.is/is/tolur

[5] https://almannaromur.is/en

[6] https://icelandic-lt.gitlab.io

[7] https://repository.clarin.is

ing to 12,780 minutes in total. A large crowdsourcing project, Samrómur, is now ongoing. In May 2022, a total of 2.85 million sentences from 28,000 speakers had been recorded, 247,800 minutes in all. No video corpora have been built for Icelandic.

The Database of Modern Icelandic Inflection (DMII) is supposed to contain the inflectional paradigms of the whole vocabulary of Icelandic. The current version has a vocabulary of about 305,000 lemmas, and 6.2 million inflectional forms. The DMII Core is an extract of DMII and contains the core vocabulary of Modern Icelandic, around 58,000 entries. The monolingual Dictionary of Contemporary Icelandic has 56,000 entries and is constantly being updated. Sound files with recordings of all the headwords in the dictionary are also available.

The company Miðeind, a member of SÍM, has been developing a translation system between English and Icelandic using neural networks. Although still under development, it already gives very promising results. The pilot version is offered as a web-based service.[8] Miðeind is also developing AI models and some of them are already available, such as GreynirTranslate (mBART25 NMT), general domain IS-EN and EN-IS translation models based on a multilingual BART model.

There exist a number of tools for analysing Icelandic text. Among them are two packages that each include various tools. IceNLP is a package which contains a tokeniser, part-of-speech tagger, lemmatiser, and shallow parser. Greynir is a more recent package that can parse text into constituency trees, find lemmas, inflect noun phrases, assign part-of-speech tags and more.

A number of tools for speech processing are currently being developed within the LTPI, among them a new speech recogniser and a speech synthesiser, but these are not yet publicly available although prototypes have been publicly demonstrated.

Embla is the first voice assistant app for the Icelandic language, available both for iOS and Android. It combines a speech recogniser, a speech synthesiser and the Greynir tool which it uses to search for answers to questions that the user poses. Greynir extracts information from Icelandic text which allows natural language querying of that information and facilitates natural language understanding.

In the national AI strategy from April 2021, the importance of developing LT resources and tools for Icelandic is explicitly mentioned.[9] In the policy statement of the new Government that took office in November 2021,[10] it is explicitly stated that the strategic R&D LT programme will be prolonged throughout the current election period, until 2025.

3 Recommendations and Next Steps

Ten years ago, the status of Icelandic LT was rather poor (Rögnvaldsson et al. 2012), but the LTPI has revolutionised the situation (Rögnvaldsson 2022). The forming of

[8] https://velthyding.is

[9] https://www.stjornarradid.is/gogn/rit-og-skyrslur/stakt-rit/2021/04/29/Stefna-Islands-um-gervigreind/

[10] https://www.stjornarradid.is/library/05-Rikisstjorn/Agreement2021.pdf

the SÍM Consortium has led to a very fruitful cooperation among all stakeholders. Researchers who used to work individually on small projects now work together on implementing projects on a much bigger scale. The number of researchers and students involved in LT has multiplied and new startup companies have emerged.

The LTPI has delivered high-quality applications that hopefully contribute to the digital vitality of Icelandic. But even so, Icelandic still lacks a number of important resources now that the LTPI is finished. Among them are spoken language corpora; parallel corpora (Icelandic and other languages than English, such as Polish and the Scandinavian languages); corpora for different purposes (sentiment analysis, question answering, summarisation); annotated multimodal corpora; and term lists. Furthermore, Icelandic lacks tools for sentiment analysis, summarisation, question answering, natural language understanding and generation, dialogue management, disambiguation, text and speech translation, automatic subtitling, advanced speech synthesis (intonation, empathy) and specialised grammar checking.

In order for these resources and tools to be developed, the continuation of the LTPI must be secured. It is also of vital importance that Icelandic is compatible with products of the large international IT companies. A delegation of LT specialists led by the President of Iceland and the Minister of Culture recently visited Amazon, Apple, META and Microsoft in order to convince them to include Icelandic in their products, offering them access to all deliverables of the LTPI. A large-scale European cooperation would also be a welcome assistance in preparing Icelandic for the future.

References

Rögnvaldsson, Eiríkur (2022). *Deliverable D1.19 Report on the Icelandic Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-icelandic.pdf.

Rögnvaldsson, Eiríkur, Kristín M. Jóhannsdóttir, Sigrún Helgadóttir, and Steinþór Steingrímsson (2012). *Íslensk tunga á stafrænni öld – The Icelandic Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/icelandic.

Chapter 22
Language Report Irish

Teresa Lynn

Abstract Language technology (LT) underpins many applications that enable our digitally enhanced lives (virtual assistants, search engines, translation tools, spell-checkers, language learning tools etc.). However, these advances do not benefit all Irish citizens equally. Due to a lack of sufficient LTs for Irish, Irish speakers regularly need to revert to using English. Such a language shift plays a major role in the risk of digital extinction, i. e., an eventual decline in language use due to lack of technological support. This chapter highlights work carried out on Irish LT, and the gaps and challenges that still need to be addressed (Lynn 2022).

1 The Irish Language

Irish is the first official and national language of the Republic of Ireland, with English as the second official language. Irish Sign Language has had official legal recognition since 2017. Figures from the 2016 census report that 39.8% (1.7 million) of the population can speak Irish, while only 1.5% (roughly 73,000) speak Irish on a daily basis outside the education system. Irish is also recognised as a minority language in Northern Ireland and has been an official language of the European Union since 2007 (and full working language of the EU since 2022).

Irish has three main dialects. However, there is no spoken standard variety, which has implications for speech technology development. The written form was standardised in 1958 with the publication of An Caighdeán Oifigiúil (The Official Standard). Irish has rich morphology and a verb-subject-object (VSO) word order, which can pose challenges for applications such as alignment tools and machine translation (MT) when paired with English (SVO). Its inflectional nature (suffixation, initial mutation, etc.) leads to sparsity in Irish datasets, which impacts data-driven LT.

There are dispersed 'Gaeltacht' regions across Ireland where Irish is spoken daily as a first language. However, English is becoming increasingly used in these regions, partially due to its monopolising digital presence. Outside Gaeltacht regions, Irish

Teresa Lynn
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_22

is also spoken in some urban areas. Irish is a compulsory core subject at primary and secondary level, and the number of Irish-medium pre-schools, primary and secondary schools is growing in both the Republic of Ireland and Northern Ireland.

The Official Languages Act (2003) has the objective of ensuring the improved provision of public services through the Irish language. In addition, the 20 Year Strategy for the Irish Language (2010-2030) and the accompanying Action Plan for the Irish Language (2018-2022) recognise the State's commitment to the language's revival. The National AI Strategy for Ireland focuses on English language-based AI. The recently published Digital Plan for Irish outlines urgent needs in LT.

Mainstream media produces much valuable audio and text-based Irish content. Irish language content is only found across roughly 1,500 (0.5%) of Ireland-based .ie domains, with low numbers of businesses localising their websites to Irish. The use of Irish in social media is prevalent among users across the main platforms. However, there is still minimal support for Irish. Google Translate and Bing Microsoft Translator still prove unreliable within particular domain settings, and much controversy has arisen around the frequent misuse of unverified automated translations. Facebook does not yet provide the option to translate Irish language posts. Google Search and Gmail interfaces were localised by volunteer translators.

2 Technologies and Resources for Irish

This summary is based on the European Language Grid (ELG). Some progress has been made in text analytics, MT, and speech technologies, mainly thanks to data collection and corpus creation from short term academic projects, funded by EU-projects and national funds, or self-funded. However, it should be noted that the ELG figures for Irish resources are inflated in some cases due to 1. the inclusion of version updates of some multilingual datasets like Universal Dependencies and ParaCrawl, 2. large multilingual datasets of which only a small proportion represents Irish, and 3. Irish web-crawled data made available through overlapping projects.

Irish is still very much a low-resourced language, with few changes in terms of LT support since Judge et al. (2012). The lack of data resources, skill-sets and dedicated funding has left a gap for many fundamental technologies. While there are extensive LT industry bodies and research centres in Ireland, little attention has been given to Irish LT. Irish-language related projects are mostly funded through The Department of the Gaeltacht's Irish Language Support Schemes and Foras na Gaeilge.

The two largest monolingual Irish corpora (New Corpus for Ireland – Irish, NCII, and Gaois Corpus of Contemporary Irish) are both restricted in terms of access due to copyright. To address this, the development of the open-source National Corpus of Ireland is underway, where resources such as word-frequency and n-gram lists, as well as language models will be made available.

Some NLP-task specific corpora have been produced as part of PhD research (e. g., POS-tagged corpora, treebanks, MWE-tagged corpora, spoken corpora). There is a considerable lack of Irish monolingual corpora for specific domains (e. g., legal,

medical, education etc.). The Irish Wikipedia (An Vicipéid) dataset was used in the development of Multilingual BERT and the Irish gaBERT language model.

The availability of bilingual texts for the purposes of English-Irish MT increased largely due to Ireland's involvement in the European Language Resource Coordination (ELRC) project, and other EU funded initiatives. The majority of this data is available to download from Ireland's National Relay Station: eStór under the EU Open Data Directive. As such, both statistical and neural English-Irish MT engines have been built at DCU through PhD research. Irish is included amongst the languages supported by the European Commission's eTranslation platform. Google, Bing, and the IRIS MT system, are all free general-purpose Irish MT systems.

An XFST Finite State suite of tools includes an Irish tokeniser, lemmatiser, morphological analyser, POS-tagger, a constraint grammar and a chunking tool. Dependency parsing models are available through UDPipe and Stanza. There is only one open-source spell-checker (GaelSpell) and grammar checker (An Gramadóir).

Steady progress has been made in speech synthesis for the three main dialects. Applications have been developed to make these voices available to the public (e. g., in accessibility aids and computer assisted language learning, CALL). Live recordings and crowdsourced recordings of predominantly native speakers using the online facility Míle Glór are being collected and processed for the development of the first ASR system. The Mozilla Common Voice project has also collected a small dataset of Irish speech (both native and non-native speakers) through crowdsourcing efforts.

Irish is relatively well-resourced when it comes to electronic dictionaries, terminology databases, thesauri, gazetteers and glossaries. Most dictionary developments (funded by Foras na Gaeilge) due to copyright restrictions, only offer single user queries or data access for research purposes only. The National Morphology Database and accompanying computational grammar library (Gramadán) are open-source. The National Terminology Database is used by the general public, students, freelance translators and translators at EU institutions. The Pota Focal site hosts a dictionary, glossary, verb valency dictionary and thesaurus, the latter of which is powered by Líonra Séimeantach na Gaeilge (LSG), an Irish Wordnet.

In terms of Natural Language Processing, the GaelTech project (2017-2023) at DCU focuses on POS-tagging, syntactic parsing, language modelling and the processing of user-generated content, code-switching and multiword expressions.

3 Recommendations and Next Steps

Many commonly used and necessary technologies are still not available for Irish: relatively little progress in ASR, and no research or system development for Automatic Subtitling, Information Retrieval and Extraction, Natural Language Generation, Semantic Role Labelling, Named Entity Recognition, Sentiment Analysis, Question-Answering, Virtual Agents, Adaptive Learning or Anonymisation. The following highlights some strategies to address this. 1. *Change of focus* A shift in focus (away from the development of dictionaries and terminologies for language learning or

translation) is required to recognise LT as an equally important axis for continued language use. 2. *Untapped Potential* Language data is broadly unknown and undervalued amongst Irish citizens and across the public sector. If collected and applied appropriately, this data could make a huge impact on the future of Irish LT. For example: development of ASR and automatic subtitling systems through data from the archives of the national broadcasters (RTÉ, TG4); a named entity recogniser using the national placenames, biographies databases, and the Database of Irish-language Surnames; CALL systems using language learning corpora. 3. *Need for Dedicated LT Programmes* Due to the lack of dedicated education and training programmes in this field, it has proven difficult to source researchers, linguists or engineers with the right combination of skills (e. g., Irish language, computer science, linguistics) in previous LT projects. 4. *Long-term strategy* There is a clear need for: a strategy for safeguarding Irish in a digital age; support for dedicated LT education and training; investments in data collection and annotation; development of production-ready LT tools. 5. *Open-source culture* Many high quality resources available for Irish are under copyright protection, rendering them unusable for general purpose. Where possible all data and tools developed for Irish should be open-source, ensuring that access is widened to others that have the skills or resources to develop them further. 6. *Corporate Social Responsibility* While Ireland is a major European hub for technological innovation in AI and NLP industries, this investment only serves the English-speaking population of Ireland. As part of a corporate social responsibility policy, support for Irish language requires much more serious consideration.

References

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha (2012). *An Ghaeilge sa Ré Dhigiteach – The Irish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/irish.

Lynn, Teresa (2022). *Deliverable D1.20 Report on the Irish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-irish.pdf.

Chapter 23
Language Report Italian

Bernardo Magnini, Alberto Lavelli, and Manuela Speranza

Abstract In the last few years, three important factors have influenced the Italian Language Technology (LT) community: 1. in 2015, the foundation of the Associazione Italiana di Linguistica Computazionale (Italian Association for Computational Linguistics, AILC); 2. the organisation of CLiC-it, the annual Italian Conference on Computational Linguistics; 3. the organisation of the EVALITA (Evaluation of NLP and Speech Tools for Italian) evaluation campaigns. This situation is producing a widespread expansion of interest in LT for Italian in academia and industry.

1 The Italian Language

Italian is an official language in Italy (where other languages are co-official within certain regions), San Marino and the Vatican City State and it is one of the official languages in Switzerland. It has official minority status in Slovenia and Croatia and formerly had official status in Albania, Malta, Monaco, Montenegro and Greece. It used to be an official language in the former colonial areas of Italian East Africa and Italian North Africa. It is among the minority languages of Bosnia and Herzegovina and Romania, although it is not protected in these countries. Italian is also spoken by very large immigrant and expatriate communities in the Americas and Australia. Italian is a major European language, being one of the official languages of the European Union, the Organisation for Security and Co-operation in Europe and one of the working languages of the Council of Europe.

Italian is the native language of around 15% of the EU population (European Commission 2012), thus the second most widely spoken language after German (Keating 2020), and has 61.8 million first language speakers according to WorldInfo.[1] Around 56 million native speakers of Italian reside in Italy; it has been estimated that another more than 200,000 first language speakers of Italian reside in Switzerland, Belgium, France, Germany, and the United Kingdom, and smaller groups of speakers reside

Bernardo Magnini · Alberto Lavelli · Manuela Speranza
Fondazione Bruno Kessler, Italy, magnini@fbk.eu, lavelli@fbk.eu, manspera@fbk.eu

[1] https://www.worlddata.info/languages/italian.php

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_23

in Croatia, Luxembourg, Malta, Romania and Slovenia. Italian is in fourteenth place in the ranking of the most used languages on the internet, as W3Techs estimates it to be used by 0.7% of the top 10 million websites.[2]

Italian belongs to the Indo-European language family of the Romance languages. Its writing system is close to being a phonemic orthography and almost all native Italian words end with vowels. Italian grammar is typical of the grammar of Romance languages in general. Cases exist for pronouns but not for nouns and there are two genders (masculine and feminine). Nouns, adjectives, and articles inflect for gender and number. Subject pronouns are usually dropped, their presence implied by verbal inflections. There are numerous contractions of prepositions with subsequent articles and numerous productive suffixes (e. g., for diminutive and augmentative). Many native speakers of Italian residing in Italy are native bilingual speakers of Italian and one of the Italian dialects (which may differ significantly from Italian).[3]

The Digital Report, a survey conducted in 2020 by "We Are Social" in collaboration with Hootsuite, with the aim of collecting data on the use of the internet and social platforms both at the global and the local level (Starri 2021), reports that in Italy over 1 million people connected to the internet for the first time in 2020 (a 2.2% increase), for a total of over 50 million internet users.

2 Technologies and Resources for Italian

A considerable part of the publicly available language resources for Italian have been produced in the EVALITA evaluation campaigns.[4] In the context of EVALITA, 62 tasks (with the availability of corresponding annotated data) have been organised in total. These tasks range from lemmatisation to sentiment analysis, covering both written texts and speech tools.

The last LT funding programme in Italy dates back to 1999–2001. Since then, there has been no specific programme, nor is one foreseen in the near future. Italian, as one of the bigger EU languages, is better equipped than other languages, but further research is needed before truly effective language technology solutions will be ready for everyday use, as well as to not lag behind the much better resourced English language. There is no national research infrastructure dedicated to LTs in Italy. Corpora (both annotated and unannotated, benchmarks, tools for several tasks) for the Italian language are, however, available either through websites of single research institutions, or through shared infrastructures at the European level, including the CLARIN repository and the European Language Grid.

Despite the lack of national funding programmes, the Italian community is rather active at the international level. Italy hosted EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics) in Trento and

[2] https://w3techs.com/technologies/overview/content_language

[3] https://www.istat.it/it/files/2017/12/Report_Uso-italiano_dialetti_altrelingue_2015.pdf

[4] https://www.evalita.it

ACL 2019 (57th Annual Meeting of the Association for Computational Linguistics) in Florence. Italian researchers have been chairing several LT conferences, including various editions of the Language Resources and Evaluation Conference (LREC), ACL 2021 (Programme Chair) and ACL 2022 (General Chair).

In the last few years, a series of initiatives have been taking place in the Italian NLP community. In 2007, the first edition of EVALITA (Evaluation of NLP and Speech Tools for Italian) was held. The general objective of EVALITA is to promote the development of language and speech technologies for the Italian language, providing a shared framework where different systems and approaches can be evaluated in a consistent manner. As a side-effect of the evaluation campaign, both training and test data are available to the scientific community as benchmarks for future improvements (Magnini et al. 2022). The first EVALITA edition was followed by six additional successful editions, the last in 2020.

Following the strong interest raised by EVALITA, the Associazione Italiana di Linguistica Computazionale[5] (Italian Association for Computational Linguistics, AILC) was founded in 2015, with the goal of establishing common ground for the Italian LT community.

A second relevant initiative on LT in Italy is CLiC-it, the annual Italian Conference on Computational Linguistics.[6] The first edition of CLiC-it was held in Pisa in 2014. CLiC-it has become the most important forum for computational linguistics in Italy, and has obtained the important goal of stimulating the production of high-quality research and resources for the Italian language.

Another relevant initiative concerning Italian LTs is the work carried out by the European Language Resource Coordination (ELRC).[7] One of the aims of the Italian ELRC is to mobilise public sector bodies to share their high-quality translated data. Additionally, many of the EVALITA resources and technologies (Patti et al. 2023) have been made available through the European Language Grid (Rehm 2023).[8]

Finally, it is worth mentioning the Lectures on Computational Linguistics, an AILC initiative targeting students (both Master's and PhD) and aiming at providing core competence in the LT field.[9]

The Italian academic LT community is relatively well distributed over the whole Italian territory, both in university departments in human sciences (e. g., linguistics, digital humanities, cognitive sciences) and in departments in computer science. In addition, there are departments of the National Research Council (CNR) and local research institutions, which are very active in the field of computational linguistics and NLP. As for industrial providers, in Italy there are more than one hundred companies that can be considered active developers in the LT field.

More details about technologies and resources for Italian can be found in Magnini et al. (2022) and the META-NET White Paper on Italian (Calzolari et al. 2012).

[5] https://www.ai-lc.it

[6] https://www.ai-lc.it/en/conferences/clic-it/

[7] https://lr-coordination.eu

[8] https://www.european-language-grid.eu

[9] https://www.ai-lc.it/lectures/

3 Recommendations and Next Steps

Given this rather favourable context (new neural approaches and strong community initiatives), we are seeing a widespread expansion of interest in Language Technology for Italian, both in academia and in industry. At the same time, the LT bar is continuously moving upwards, which requires adequate efforts and investments. These are particularly needed in areas such as for instance dialogue systems, where Italian is still lacking sufficient language resources, and in application domains such as biomedicine, where progress is still limited.

References

Calzolari, Nicoletta, Bernardo Magnini, Claudia Soria, and Manuela Speranza (2012). *La Lingua Italiana nell'Era Digitale – The Italian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/italian.

European Commission (2012). *Europeans and their Languages*. https://europa.eu/eurobarometer/surveys/detail/1049.

Keating, Dave (2020). "Despite Brexit, English Remains The EU's Most Spoken Language By Far". In: *Forbes*. https://www.forbes.com/sites/davekeating/2020/02/06/despite-brexit-english-remains-the-eus-most-spoken-language-by-far/.

Magnini, Bernardo, Alberto Lavelli, and Manuela Speranza (2022). *Deliverable D1.21 Report on the Italian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-italian.pdf.

Patti, Viviana, Valerio Basile, Andrea Bolioli, Alessio Bosca, Cristina Bosco, Michael Fell, and Rossella Varvara (2023). "Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 295–300.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Starri, Matteo (2021). *Digital 2021 – I dati italiani*. https://wearesocial.com/it/blog/2021/02/digital-2021-i-dati-italiani/.

Chapter 24
Language Report Latvian

Inguna Skadiņa, Ilze Auziņa, Baiba Valkovska, and Normunds Grūzītis

Abstract Ten years ago, when META-NET conducted a study on Language Technology support for Europe's languages, Latvian was assessed as a language with little or no support (Skadiņa et al. 2012). During the last decade, progress has been made in the development of language resources and tools for Latvian, particularly with respect to advanced datasets and language models, machine translation solutions, speech technologies, and technologies for natural language understanding and human-computer interaction. This chapter provides a summary of the current state of the Latvian language, the only official language of Latvia, in the digital environment and highlights the most important activities in the language technology field.

1 The Latvian Language

Latvian is the official language of the Republic of Latvia. There are about 1.5 million native speakers, 1.38 million of which live in Latvia. By the end of 2017, Latvian was the mother tongue of 60.8% of the country's resident population. Latvian is spoken as a second language by around 0.5 million people of other ethnicities. Latvian has three dialects: the Central, Livonic, and High Latvian dialect.

The Latvian language uses the phono-morphological basis of orthography. Latvian punctuation is based on the grammatical punctuation principle. Latvian orthography almost fully corresponds to the pronunciation. The present-day Latvian orthography basis is the Latin script. The Latvian standard alphabet consists of 33 letters, including letters with diacritical marks.

Standard Latvian has 26 consonant phonemes, 12 vowels (six short and six long), and 10 diphthongs. Vowel length is phonemic and plays an important role in distinguishing the lexical and grammatical meaning of words. Most Latvian words are stressed on the first syllable. Syllables with long vowels, diphthongs, and diphthongical combinations of vowel and sonorant in the centre are subject to certain intonation

Inguna Skadiņa · Ilze Auziņa · Baiba Valkovska · Normunds Grūzītis
University of Latvia, Latvia, inguna.skadina@tilde.com, ilze.auzina@lumii.lv, baiba.valkovska@lumii.lv, normunds.gruzitis@lumii.lv

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_24

patterns. In a few areas, three patterns of tone or intonation are distinguished: level (also drawling, even) tone, falling tone, and broken tone.

From a language typology perspective, Latvian has a classic Indo-European (Baltic) system. However, for regional and historical reasons, Latvian grammar also displays some features more similar to those found in Finno-Ugric languages (Kalnaca and Lokmane 2021). Latvian is a fusional, mainly suffixing language with a rich system of forms and word formation. A distinction is made between inflected and non-inflected word classes. Nouns inflect for number and case, adjectives inflect for case, number, gender and definiteness, and verbs may inflect for tense, mood, voice and person (Nau 1998). Word order is relatively free, i. e., pragmatically governed, but the basic word order is subject verb object (SVO).

2 Technologies and Resources for Latvian

Research and development activities in Latvia are being supported through different EU and national finance instruments and are usually organised around short-term projects. The lack of a dedicated LT programme, however, leads to fragmentation of research and development activities and complicates the development of larger resources and long-term cooperation between institutions. Progress and key achievements are regularly reported through the Baltic HLT conferences and other events (Skadiņa 2019; Skadiņa et al. 2022, provide recent overviews).

Most open-access monolingual text corpora for Latvian are listed on Korpuss.lv (Saulīte et al. 2022). Modern Latvian is primarily represented through the Balanced Corpus of Modern Latvian (LVK2018, Dargis et al. 2020). A balanced subset of LVK2018 includes several annotation layers: named entities, co-references, Universal Dependencies (UD), FrameNet and PropBank annotations, as well as Abstract Meaning Representation (AMR) (Gruzitis et al. 2018). Many parallel corpora are openly accessible from OPUS, ELG and ELRC-SHARE. Bilingual and multilingual corpora are also stored on Korpuss.lv and the Tilde Data Library.[1] Domain-specific parallel corpora for the development of domain-specific MT engines are lacking.

The first Latvian speech corpus was created in 2012/2013. It contains 100 hours of transcribed speech. However, access is limited, and currently the only open-access Latvian speech corpora are very small. Multimodal corpora are still not available for Latvian, although the development of a sign language corpus is ongoing in the State Research Programme “Letonika”.

Tezaurs.lv is the largest open lexical dataset and online dictionary for Latvian (Spektors et al. 2016). It is regularly updated, and currently contains more than 380k single- and multi-word entries, compiled from 300+ sources. A Latvian WordNet is being created as an extension to Tezaurs.lv. Different lexicons (mostly bilingual) are available from the Letonika.lv portal, including dictionaries for widely used language pairs, as well as dictionaries of the languages of the Baltic countries.

[1] https://tilde.com/products-and-services/data-library

Various text analysis tools such as tokenisers and sentence splitters, morphological analysers and taggers, spelling and grammar checkers, syntactic and semantic parsers, named entity recognisers, and text classifiers are available for Latvian. Open-source components are integrated into a Latvian NLP pipeline as a service.[2]

Regarding natural language understanding and generation, experiments with the interlingual UD, FrameNet, AMR, BERT, GPT, etc. models for Latvian demonstrate the potential of combining machine learning and knowledge-based approaches.

With respect to machine translation (MT), the situation has changed a lot since 2012. Besides MT solutions provided by global companies, the company Tilde provides customised MT solutions for complex, highly inflected languages, particularly smaller European languages. MT systems developed by Tilde have been recognised among the best systems for four consecutive years (2017-2020) at WMT international news translation shared tasks (Pinnis et al. 2019). These results allowed Tilde together with partners to develop the EU Council Presidency Translator which has been used already in eight countries (Pinnis et al. 2020).

Several speech recognition and synthesis systems have been developed for Latvian by Tilde, the national news agency LETA, and the University of Latvia. Several virtual assistants can communicate in Latvian, e. g., Hugo.lv (Skadins et al. 2020) lists more than 10 virtual assistants for different public services.

Latvia is a member of CLARIN (Skadiņa et al. 2020) and focuses on Latvian and Latgalian resources and tools. CLARIN-LV participates in the CLARIN Knowledge Center for Systems and Frameworks for Morphologically Rich Languages.

3 Recommendations and Next Steps

Today, Latvian has a rather stable position in the digital world. However, the situation could change dramatically, if efforts and investments in LT are not increased in R&D and language policy. Strong national and European support is necessary for further Latvian research and development activities, including dedicated long-term LT programmes, that provide equal support for both research and industrial activities. To narrow the digital divide, there is pressing urgency for novel techniques that would bring less-resourced languages to a level comparable to the state-of-the-art results for resource-rich languages. Moreover, close synchronisation between national and international activities is necessary.

References

Dargis, Roberts, Kristine Levane-Petrova, and Ilmars Poikans (2020). “Lessons Learned from Creating a Balanced Corpus from Online Data”. In: *Human Language Technologies – The Baltic Perspective*. Vol. 328. IOS Press, pp. 127–134. DOI: 10.3233/FAIA200614.

[2] http://nlp.ailab.lv

Gruzitis, Normunds, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens (2018). "Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 4506–4513.

Kalnaca, Andra and Ilze Lokmane (2021). *Latvian Grammar*. University of Latvia.

Nau, Nicole (1998). *Latvian*. Vol. 217. Lincom Europa.

Pinnis, Mārcis, Toms Bergmanis, Kristīne Metuzāle, Valters Šics, Artūrs Vasiļevskis, and Andrejs Vasiļjevs (2020). "A Tale of Eight Countries or the EU Council Presidency Translator in Retrospect". In: *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2)*. Association for Machine Translation in the Americas, pp. 525–537.

Pinnis, Mārcis, Rihards Krislauks, and Matiss Rikters (2019). "Tilde's Machine Translation Systems for WMT 2019". In: *Proceedings of the 4th Conference on Machine Translation (Volume 2)*. Florence, Italy: ACL, pp. 327–334. DOI: 10.18653/v1/W19-5335.

Saulīte, Baiba, Roberts Darģis, Normunds Grūzītis, Ilze Auziņa, Kristīne Levāne-Petrova, Lauma Pretkalniņa, Laura Rituma, Pēteris Paikens, Artūrs Znotiņš, Laine Strankale, Kristīne Pokratniece, Ilmārs Poikāns, Guntis Bārzdiņš, Inguna Skadiņa, Anda Baklāne, and Valdis Saulespurēns (2022). "Latvian National Corpora Collection – Korpuss.lv". In: *Proceedings of the 13th LREC Conference*.

Skadiņa, Inguna (2019). "Some Highlights of Human Language Technology in Baltic Countries". In: *Databases and Information Systems*. Vol. 315. IOS Press, pp. 18–30. DOI: 10.3233/978-1-61499-941-6-18.

Skadiņa, Inguna, Ilze Auziņa, Normunds Grūzītis, and Arturs Znotiņš (2020). "Clarin in Latvia: From the preparatory phase to the construction phase and operation". In: *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries*, pp. 342–350.

Skadiņa, Inguna, Ilze Auziņa, Baiba Valkovska, and Normunds Grūzītis (2022). *Deliverable D1.22 Report on the Latvian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-latvian.pdf.

Skadiņa, Inguna, Andrejs Veisbergs, Andrejs Vasiļjevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte (2012). *Latviešu valoda digitālajā laikmetā – The Latvian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/latvian.

Skadins, Raivis, Marcis Pinnis, Arturs Vasilevskis, Andrejs Vasiljevs, Valters Sics, Roberts Rozis, and Andis Lagzdins (2020). "Language Technology Platform for Public Administration". In: *Human Language Technologies – The Baltic Perspective*. Ed. by Utka Andrius, Vaicenoniene Jurgita, Kovalevskaite Jolantai, and Kalinauskaite Danguole. Vol. 328. FAIA. IOS Press, pp. 182–190.

Spektors, Andrejs, Ilze Auzina, Roberts Dargis, Normunds Gruzitis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, and Baiba Saulite (2016). "Tezaurs.lv: the largest open lexical database for Latvian". In: *Proceedings of LREC 2016*.

Chapter 25
Language Report Lithuanian

Anželika Gaidienė and Aurelija Tamulionienė

Abstract Significant progress has been made in adapting the Lithuanian language to the digital environment. A number of digital language resources and basic language analysis tools, as well as complex online language services and the Lithuanian language ontology have been developed, while a number of computer programs and tools have been localised. Computer applications relevant to society are being Lithuanianised, and the standardisation of computer terms is being carried out. Lithuanian researchers actively participate in the cooperation and mobility activities of international associations, and a core of Lithuanian specialists working in the field of IT application, and developing innovative work in this field, has been formed. Lithuania also strives for all citizens to have full access to digital solutions, which adds importance to the policy of adapting them for those living with disabilities.

1 The Lithuanian Language

Lithuanian is a Baltic language from the Indo-European family. Lithuanian and Latvian are the two surviving Baltic languages. Since 2004, Lithuanian has been one of the official languages of the European Union. Lithuanian is the state language of the Republic of Lithuania and is enshrined in the Constitution as such. The use of Lithuanian in public life is regulated by the State Law on the Lithuanian Language (1995). According to data from 2012, there were about 3.6 million Lithuanian speakers. In terms of number of speakers, Lithuanian ranks 144th in the world.

Lithuanian is the most conservative of the Indo-European living languages, and it has best preserved many of its archaic features. From a typological point of view, Lithuanian has many unique features, including abundant forms of variation, the synthesis of tonal and dynamic stress, and the diverse order of words reflecting the complex syntactic level of discourse communication. Standard Lithuanian was formed at the beginning of the 20th century on the basis of one of the Aukštaitian dialects.

Anželika Gaidienė · Aurelija Tamulionienė
Institute of the Lithuanian Language, Lithuania, anzelika.gaidiene@lki.lt, aurelija.tamulioniene@lki.lt

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_25

It is characterised by a great variety of regional variants, the two main dialects are Aukštaitian and Samogitian (Vaišnienė and Zabarskaitė 2012). The Lithuanian alphabet was formed in the 16th to 20th centuries on the basis of the Latin alphabet, to which nasal vowels (ą, ę, į, ų) and letters with diacritics (č, š, ž, ė, ū) were added. The current Lithuanian language alphabet has 32 letters: 12 vowels, 20 consonants, and 3 letter combinations (ch, dz, dž). The grammatical structure is of a flexural type; the vocabulary is the most variable level of the language. Some words disappear and are replaced by new ones. In the current Lithuanian language, there is a pronounced abundance of terms in various fields. The vocabulary of the Lithuanian language consists of old words inherited from the Proto-Indo-European language, borrowings, and new words based on inherited words and borrowings.

According to 2021 data, in the 16 to 74 age group, almost 87% of the Lithuanian population uses the internet; this figure is as high as 100% in the 16 to 24 age group, and 55.2% in the 65 to 74 age group. In 2021, about 225,000 .lt domains were registered, of which more than 2,000 contain distinctive Lithuanian letters (ė, ž, etc.). In addition, Lithuania remains among the leaders in fibre-optic internet service. In Lithuania, the coverage of the fibre-optic network is 46.8%.

2 Technologies and Resources for Lithuanian

The level and advancement of language technologies in Lithuania can first and foremost be appraised by the degree of achievement of the goals rooted in the 2014 – 2020 guidelines (State Commission of the Lithuanian Language 2014) for the expansion of the Lithuanian language in information technologies. Notably, those goals have been achieved with a great deal of success, yet some follow-up actions are needed, depending on the progress of the rapidly shifting language technologies on the global market and amidst society (Gaidienė and Tamulionienė 2022).

Lithuania continues to create and develop general resources needed for the purposes of building language technologies and devising their applications. There are a number of monolingual and bilingual corpora. The largest corpus of the Lithuanian language is the Corpus of the Contemporary Lithuanian Language. There are also several morphologically and syntactically annotated corpora (Morphologically Annotated Corpus, MATAS; Syntactically Annotated Treebank, ALKSNIS). There are also a number of parallel corpora (e. g., the LILA corpus). Most corpora are open access. Nonetheless, considering the demand for language data, it needs to be said that corpus data has to be augmented and new corpora (especially multilingual parallel data) should be developed to reflect as many areas of language use as possible.

Lithuania continues to develop digital dictionaries and databases. Users have free online access to the latest Dictionary of the Standard Lithuanian Language as well as other dictionaries, such as Dictionary of the Modern Lithuanian Language, Dictionary of the Lithuanian Language and the ongoing Database of Lithuanian Neologisms. Other resources such as the Dictionary of Synonyms, the Dictionary of Antonyms, other various bilingual dictionaries etc. are also freely accessible online.

Despite this abundance of digital dictionaries, considering the demands of language technologies and of the public, the dictionaries of synonyms, antonyms, and phraseology need to be updated, and the dictionaries of pronunciation and combinability (among others) digitalised.

Semantic networks and ontologies in Lithuania are few in number. There is the General Ontology of the Lithuanian Language, the open-access ontology of Lithuanian medical terms Snomed CT, and the service E-terms (Ontologies). There are several Lithuanian wordnets that can be developed further, e. g., LitWorNet. However, the available ontologies and wordnets are inadequate and need to be expanded.

The ALPMAVIS machine translation system is freely available. The company Tilde offers MT systems based on the latest neural networks for free. Continued development of MT systems would require more bilingual parallel corpora as well as specialised text data to ensure better quality translation output.

The available Lithuanian Speech-to-text Transcription Service covers different domains: administrative, legal, medical, and standard colloquial. There are also services where speech recognition technology is used to voice-control computers, such as Browser (browsing voice control), Controller (computer voice control), and so on. Some of the services available in Lithuania feature speech synthesis technology, including Pronouncer, the Lithuanian Speech Synthesiser for the Blind, and so on. The Lithuanian language needs more annotated speech databases, which calls for concerted efforts to build speech databases for different fields, dialects, age groups, and sound environments (among others), and to make them available to the public.

The various projects that have been implemented in Lithuania have produced key open-source tools for the basic analysis of digital texts in the Lithuanian language, such as a segmenter, a lemmatiser, a morphological analyser, a part of speech tagger, and so on (State Commission of the Lithuanian Language 2020). In terms of generating natural language, Lithuania is only making its first steps in this area.

3 Recommendations and Next Steps

Since 2012, significant progress has been made in developing various digital language resources and tools/services in Lithuania. Although Lithuanian is a language with a small number of speakers, it is progressing rapidly in the area of LT. As for digital resources and tools/services, there are still areas requiring further advances. Though a number of Lithuanian language resources are already available, considering the demands of LTs and of the public, new monolingual dictionaries and bilingual dictionaries as well as various lexicons still need to be developed or updated. Ontologies, wordnets, corpora have to be enlarged and expanded; multilingual parallel corpora required for MT need to be developed, etc. Concerning terminology, additional and updated compendia or terms are needed; the structure and technological solutions of the databases of terms vary, making it more difficult to utilise data for other technological solutions; there is also a shortage of open terminologi-

cal data. Lithuanian is in need of digital grammars, annotated speech databases and other resources that would accelerate the progress of language technologies.

Lithuania requires an increase in digital language resources – corpora with texts and recordings, the development of LTs and the creation of public services based on them – so that no group of society or region can feel the digital divide and foreign languages can be integrated more easily into Lithuanian society. The Guidelines for the Development of the Lithuanian Language in the Digital Environment and the Progress of Language Technologies for 2021–2027 map out the essential tasks or challenges of Lithuanian LTs, what should be done in Lithuania in the near future, and in which directions to work: 1. To increase the competence of specialists working in the field of language technologies and to improve the ability of society as a whole to use the opportunities provided by language technologies. 2. To accumulate and enrich open, reliable, high-quality, reusable digital language resources and other digital language data sets. 3. To develop the language technology infrastructure, the application of language technologies in the public sector and public services, to create and improve publicly available information technology solutions and tools.

References

Gaidienė, Anželika and Aurelija Tamulionienė (2022). *Deliverable D1.23 Report on the Lithuanian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-lithuanian.pdf.

State Commission of the Lithuanian Language (2014). *The Guidelines for the Development of the Lithuanian Language Language Technologies for 2014–2020*. http://www.vlkk.lt/kalbos-politika/lietuviu-kalbos-pletros-informacinese-technologijose-gaires/lietuviu-kalbos-pletros-informacinese-technologijose-2014-2020-m-gaires.

State Commission of the Lithuanian Language (2020). *The Guidelines for the Development of the Lithuanian Language in the Digital Environment and the Progress of Language Technologies for 2021–2027*. https://www.e-tar.lt/portal/lt/legalAct/71152ab00eee11ebb74de75171d26d52.

Vaišnienė, Daiva and Jolanta Zabarskaitė (2012). *Lietuvių kalba skaitmeniniame amžiuje – The Lithuanian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/lithuanian.

Chapter 26
Language Report Luxembourgish

Dimitra Anastasiou

Abstract The Grand Duchy of Luxembourg is a small and multilingual country. The national language is Luxembourgish, and the legislative language is French. French, German and Luxembourgish are the three administrative and judicial languages. There are about 650,000 inhabitants and the majority of Luxembourgers speak four languages. As of March 2021, there were 59,000 Wikipedia articles written in Luxembourgish. Luxembourgish is very under-resourced when it comes to data resources and tools. This chapter provides a brief overview of the current level of support that Luxembourgish receives through technology (Anastasiou 2022).

1 The Luxembourgish Language

Luxembourg is a very small, but highly multilingual country. At various times, it was part of different European empires. Today, Luxembourg is the third European capital, along with Brussels and Strasbourg. It has the honour of hosting many of the EU's important institutions, including the Publications Office of the EU, the Directorate-General for Translation, and the Translation Centre for the Bodies of the EU.

As for the population of Luxembourg, the Statistics portal of the Grand Duchy of Luxembourg (STATEC) published a demographic atlas in 2019. According to this atlas, between 1981 and 2018, the Luxembourgish population increased by about 65%, from 364,597 to 602,005. There are 12 officially declared towns and 102 municipalities. Luxembourg City has the highest percentage of foreigners with 70.8%.

The languages spoken vary depending on the social situation or region. The regions with the highest density of Luxembourgish speakers include the north (85%) and the east (81%) of the country. According to STATEC, three out of four residents work in a multilingual environment and 25% of the population has to speak four or more languages at work. French is the most spoken language at work (78%), followed by English (51%) and Luxembourgish (48%). Luxembourgish is the most widely spoken language at home (53%), followed by French (32%) and Portuguese

Dimitra Anastasiou
Luxembourg Institute of Science and Technology, Luxembourg, dimitra.anastasiou@list.lu

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_26

(19%). It should be noted that Luxembourgish is not an official language of the EU. The Luxembourgish language is a Moselle-Franconian dialect, which was historically the mainly spoken language up to the 19th century in Luxembourg. On 24 February 1984, a law was enacted which made Luxembourgish an officially recognised language. In September 2018, the law was amended to add German sign language as an official language of Luxembourg. According to the provisions of the Languages Law of 1984, French, German or Luxembourgish may be used in administrative and judicial matters. Citizens can interact with the administration in any of these three languages, and officials must attempt 'as far as possible' to respond in the language used by the applicant. Legislative documents are written in French and an important consequence of this on the judicial level is that only the text in French is deemed authentic for all levels of public administration.

In terms of vocabulary, Luxembourgish has a substantial number of loan words from French and German, but its morpho-syntax follows Germanic patterns (Gilles and Trouvain 2013). With the exception of the alveolo-palatal fricatives and the approximant [w], the consonant inventory of Luxembourgish is quite similar to Standard German. In addition, Luxembourgish has a set of eight diphthongs, which is considerably larger than for Standard German which has just one (Gilles and Trouvain 2013). Gilles (2019) examines the complex language situation of Luxembourg. There is an officially recognised system with regards to the orthography of Luxembourgish, called "OLO" (ofizjel lezebuurjer ortografi); it can be found at the Zenter fir d'Lëtzebuerger Sprooch (ZLS)/Centre for the Luxembourgish Language.[1]

2 Technologies and Resources for Luxembourgish

Luxembourgish-specific tools include a grapheme-to-phoneme conversion for Luxembourgish based on 30,000 manually phonetically transcribed words, two spell-checkers, a PoS-tagger (including a tokenizer and lemmatizer), and sentence splitter (Sirajzade and Schommer 2019), and a mobile application called Schnëssen.[2] This crowdsourcing app collects data on the present-day language situation of Luxembourgish; users can participate in a large set of audio recordings tasks and in sociolinguistic surveys. A recently published tool, LëtzRead,[3] is a free browser extension to integrate Luxembourgish-learning just by browsing the web (displaying certain words in Luxembourgish). Moreover, The library spaCy for advanced NLP has been trained for Luxembourgish, and the text-to-speech (TTS) tool MaryTTS has also been extended to support Luxembourgish. Luxembourgish data resources are mainly monolingual corpora, but there is also a Luxembourgish COVID glossary as well as an orthography trainer. The biggest text corpus in Luxembourgish contains 170 million words from a wide range of genres (Parliamentary debates, lit-

[1] https://portal.education.lu/zls/ORTHOGRAFIE

[2] https://infolux.uni.lu/schnessen/

[3] https://www.letzread.com

erature, transcripts of conversations, and media texts including articles from news outlets like RTL.lu, radio100,7, eldoradio, and social media). All texts are annotated and orthographically normalised. This corpus is owned by the University of Luxembourg and is for internal use only. Many lexical Luxembourgish-specific resources, including corpora, dictionaries, material for phonetics, applications, etc. are available at Infolux,[4] which is the research portal about Luxembourgish developed and maintained by the Institute of the Luxembourgish Language and Literature at the University of Luxembourg. Another important resource is the Luxembourgish Online Dictionary (LOD),[5] managed by the ZLS, a multilingual dictionary with 30,000 entries, in which Luxembourgish words are translated into German, French, English and Portuguese and illustrated by examples.

Among the recent research projects related to language technology (LT) including Luxembourgish are ENRICH4ALL, STRIPS, and Lingscape. ENRICH4ALL (E-goverNment [RI] CHatbot for ALL)[6] is a CEF-funded project (06/21-05/23) coordinated by the Luxembourg Institute of Science and Technology, and its objective is to have a multilingual chatbot through integrating the CEF AT core service platform eTranslation to existing AI-based chatbot technology. The chatbot service will be deployed in public administration in Luxembourg, Denmark, and Romania. STRIPS[7] was a three-year project (02/18-01/21), funded by the University of Luxembourg, that aimed to develop a semantic search toolbox for the retrieval of similar patterns in documents written in Luxembourgish. Lingscape[8] is a mobile application researching linguistic landscapes all over the world by collecting photos of signs and lettering on an interactive map.

3 Recommendations and Next Steps

Digitalisation plays a big role in the government of Luxembourg, the Ministry for Digitalisation was created on 11th December 2018. Luxembourg's national AI Vision initiative underlines the country's unique ability to become a living lab of real-world AI applications.

Mainly because of the lack of underlying data resources, there are gaps in many aspects of Luxembourgish Language Technology. What is currently missing are available bilingual corpora, e. g., Luxembourgish – English, German, French. The availability of such data sets would facilitate the development of many LT applications, such as named entity recognition, machine translation, virtual agents, recommender systems, etc. All of these applications are mainly statistically-based, so typically require a large amount of manually annotated training data. Regarding language mod-

[4] https://infolux.uni.lu

[5] https://www.lod.lu

[6] https://www.enrich4all.eu

[7] https://acc.uni.lu/Research/strips/

[8] https://lingscape.uni.lu

els which can be used for natural language understanding and generation, the multilingual BERT covers many languages, including Luxembourgish; however, a BERT model trained specifically on large Luxembourgish data would yield better results. Another important aspect is that written Luxembourgish is not well standardised; while both German and French are intensively taught in schools, Luxembourgish, although the first language of around 60% of the population, forms part of the school curriculum only rudimentarily. This has an impact on the correctness of Luxembourgish in the development of LT applications. It is noteworthy that Luxembourgish has become more important in secondary schools with changes incorporated for the academic school years 2021-2022 and 2022-2023.

Luxembourg needs united forces for efficient collaboration. Since most people are multilingual, various stakeholders do not see the need to invest either time or budget in creating or sharing Luxembourgish resources. The EU, the government, research institutions, and language service providers have to work together to achieve the desired results. Important action points to improve the Luxembourgish LT landscape are: 1. reaching a status that Luxembourgish can be used in many administrative procedures; 2. raising awareness among various stakeholders in public and private sectors about the impact of Luxembourgish data; 3. advancing the standardisation, use and study of Luxembourgish, and 4. dedicated and collaborative national and EU funding programmes for both basic and applied research on Luxembourgish.

References

Anastasiou, Dimitra (2022). *Deliverable D1.24 Report on the Luxembourgish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-luxembourgish.pdf.

Gilles, Peter (2019). "39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache". In: *Deutsch*. De Gruyter Mouton, pp. 1039–1060.

Gilles, Peter and Jürgen Trouvain (2013). "Luxembourgish". In: *Journal of the International Phonetic Association* 43.1, pp. 67–74. DOI: 10.1017/S0025100312000278.

Sirajzade, Joshgun and Christoph Schommer (2019). "The LuNa Open Toolbox for the Luxembourgish Language". In: *Advances in Data Mining, Applications and Theoretical Aspects, Poster Proceedings 2019*. Ibai publishing.

Chapter 27
Language Report Maltese

Michael Rosner and Claudia Borg

Abstract This chapter is a highly abbreviated version of an update (Rosner and C. Borg 2022) to the META-NET White Paper on Maltese (Rosner and Joachimsen 2012). Like its predecessor, the update forms part of a series for all European Languages. Section 1 provides a brief description of the language, its national status, its general typology as a language, and its current usage in the digital sphere. Section 2 gives an overview of technologies and resources that are currently available. Finally, Section 3 frames the main shortcomings of Maltese language technology in terms of fragmentation, and offers some recommendations on how that might be reduced.

1 The Maltese Language

Maltese (il-Malti) is an official EU language and the national language of the Maltese archipelago. 97% of the Maltese population (ca. 400,000 people) consider it their mother tongue. It is also spoken by communities in Australia, Canada, the USA and the UK. Maltese is derived from late medieval Sicilian Arabic with Romance superstrata, and is often referred to as a mixed language due to the large number of loan words from Italian, English and French. It shares characteristics with other Semitic languages, making use of root-and-template morphology whereby various forms of the same lexeme are formed by interdigitating vowels between a fixed sequence of root consonants. The main distinguishing characteristics of Maltese are free word order, mixed morphology, aspect-based temporal system, and lack of a morphological infinitive. Unlike other Semitic languages, the Maltese alphabet is based on the Latin one with the addition of some letters with diacritic marks and digraphs (ċ, ġħ, ż, ġ, ħ). It contains 24 consonants and 6 vowels. According to Fabri (2011), the writing systems used for Maltese were somewhat ad hoc before 1920, but a degree of consistency among writers and in publications became a reality in the 1950s.

Within the digital sphere, there have always been several Maltese language newspapers. The broadcast media (radio and TV) are almost exclusively in Maltese. Since

Michael Rosner · Claudia Borg
University of Malta, Malta, mike.rosner@um.edu.mt, claudia.borg@um.edu.mt

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_27

the previous report, there has been a general decline in hard-copy newspaper readership, as all the media are now available online and the majority of readers prefer the online version. Various online-only news websites have appeared, one of which (Newsbook) operates bilingually. The full Maltese character set is now universally used. Social media are extremely popular (97% of the population according to a 2021 survey). Facebook remains the most accessed, but there is a trend of increased usage of Instagram and YouTube. Unlike other EU countries, Twitter usage in Malta is remarkably low. The Maltese Wikipedia currently ranks at 204/325 (for comparison, English, Portuguese, Irish, Icelandic, Romansch rank at 1, 18, 93, 95, and 213, respectively). It contains nearly 4 million words distributed over 4,400 content pages (cf. 6.5 million for English). This compares to about 3,000 pages in 2011; there are ca. 19,000 registered users with only about 40 active users (making changes every 30 days or less. YouTube gives rise to localised content in many other countries but the local website still operates predominantly in English. In general, there tends to be a gap between social media content creators and non-creators. However, a renowned online page which has successfully bucked this trend is Kelma Kelma which started in 2013 as a Facebook page and gathers many interesting original contributions by locals about the Maltese language. The top-level country domain for Malta, .mt, is administered by the Malta Internet Foundation, has currently ca. 17,000 domain names and subdomains, more than three times the figure in 2010.

2 Technologies and Resources for Maltese

Rosner and Joachimsen (2012) describe the main enablers and contributions to Maltese Language Technology up to ca. 2011. 2012 marked the public release of the MSE speech synthesiser (M. Borg et al. 2014), whilst Gatt and colleagues began revamping the University's MLRS resource server (Rosner 2008; Gatt and Čéplö 2013) to include semi-automated data-collection, a tagger, Korpus Malti v3.0 (2016), containing ca. 250 million annotated tokens, pattern-based search facilities, CLEM, a 1 million token Corpus of Learner English in Malta, Ġabra, an Open Lexicon for Maltese, and a Dictionary of Maltese Sign Language.

Most available corpora are monolingual written text. A few are spoken, and fewer still are multimodal such as MAMCO (Paggio et al. 2018). Many monolingual corpora form part of unannotated *multilingual* collections. Others are by-products of projects and annotated for MWE identification (PARSEME) or POS Tagging (MLRS), anonymisation (MAPA), morphological analysis (UniMorph), NER (WikiAnn) etc. Bilingual/multilingual resources include the Laws of Malta, the Government Gazette, and the Acquis Communautaire.

Regarding tools and services, besides low-level text preprocessing for tokenisation, sentence and paragraph splitting and POS-tagging, the Ġabra dictionary has evolved into the online Dizzjunarju tal-Malti app. Machine translation for Maltese has improved not only through the availability of free tools like Google Translate, but also as a result of DGT's eTranslation platform whose increased takeup by pub-

lic administration officials followed a series of workshops organised through ELRC. Much recent effort has been focused on dependency parsing and ASR. There is now a 2000-sentence Universal Dependency Treebank for Maltese which has supported experiments (Zammit et al. 2019) aimed at delivering a prototype dependency parser in 2022. Similarly, for speech technology, the locally funded MASRI project has delivered a fully annotated speech corpus (Hernandez Mena et al. 2020). Most resources mentioned above are freely available through MLRS and also EU platforms.

Currently, the main drivers for the evolution of future Maltese LT are targeted national initiatives, against a mixed background of projects at EU level. At the national level, the National AI Strategy (2019) focuses on the creation of an AI ecosystem infrastructure including tools to enable Maltese Language AI solutions, with funds earmarked for Maltese LT resources. The Malta Digital Innovation Authority (MDIA) is committed to supporting Maltese LT tools which will focus on morphological analysis, dependency parsing, named entity recognition and POS tagging. In 2019, the Government also committed funds to the development of a spell checker. However, there is no information with respect to the progress of this important initiative. Meanwhile at the EU level Maltese participation in a wide range of projects, actions and initiatives including ELE, ELG, ELRC, DARIAH, LCT, LT-Bridge, MAPA, Nexus Linguarum, and NLTP, has ensured a level of Maltese presence on the European scene and also produced some specialised resources and tools.

3 Recommendations and Next Steps

Maltese LT is indeed alive, but manifests an important weakness: it is highly fragmented, in different ways: 1. between national efforts (small-scale, Maltese-focused) and international ones (large-scale, language-independent); 2. across resources/tools which are not necessarily compatible with each other; and 3. between users and developers of LTs (reduces the perceived relevance of the technologies developed). To address these requires further investigation of techniques like transfer learning, as seen, for example, in the MAPA project where general language models were successfully used for Maltese NER. Issue 2. can be reduced by insisting that such resources inhabit a framework which includes the necessary protocols to ensure interoperability, as seen in European infrastructures like ELG and NLTP, funded under CEF, aiming to build a National Language Platform for Maltese integrating eTranslation services developed by the European Parliament with fine-tuned local translation memories, and providing a central point for collecting different LT services together. 3. is in part the result of insufficient involvement of the IT industry in LT. Despite the latter being a major component of the local economy, the number of technical LT providers is very low. LT has a crucial role to play as a natural bridge linking IT, AI, communication and multilinguality. More needs to be done to support that role by encouraging participation in ELG by local IT players, among others. In 2016, the IT subcommittee of the Council for the Maltese Language had recognised the need for the long-term curation of resources, recommending the creation of a central

repository, and efforts to involve more stakeholders concerning the availability and importance of resources. Some progress towards the realisation of these recommendations has been made but the effort needs a substantial and sustained coordinated investment across the different sectors involved.

References

Borg, Mark, Keith Bugeja, Colin Vella, Gordon Mangion, and Carmel Gafa (2014). "Preparation of a Free-Running Text Corpus for Maltese Concatenative Speech Synthesis". In: *Perspectives on Maltese Linguistics, Studia Typologica 14*. Ed. by Albert Borg, Sandro Caruana, and Alexandra Vella, pp. 297–318.

Fabri, Ray (2011). "Maltese". In: *The Languages of the 25. Revue belge de Philologie et d'Histoire: RBPH*. Ed. by Christian Delcourt and Piet van Sterkenburg. Amsterdam, Philadelphia: John Benjamins, pp. 17–28.

Gatt, Albert and Slavomír Čéplö (2013). "Digital corpora and other electronic resources for Maltese". In: *Proceedings of Corpus Linguistics*. Ed. by Andrew Hardie and Robbie Love. University of Lancaster, UCREL.

Hernandez Mena, Carlos Daniel, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani (2020). "MASRI-HEADSET: A Maltese Corpus for Speech Recognition". In: *Proceedings of LREC 2020*. Marseille, France: ELRA, pp. 6381–6388.

Paggio, Patrizia, Luke Galea, and Alexandra Vella (2018). *Prosodic and gestural marking of complement fronting in Maltese*. DOI: 10.5281/zenodo.1181805.

Rosner, Mike (2008). "Electronic Language Resources for Maltese". In: *Proceedings of Bremen Workshop on Maltese Linguistics*. Springer.

Rosner, Mike and Claudia Borg (2022). *Deliverable D1.25 Report on the Maltese Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-maltese.pdf.

Rosner, Mike and Jan Joachimsen (2012). *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/maltese.

Zammit, Andrei, Slavomír Čéplö, Lonneke van der Plas, and Claudia Borg (2019). *A Dependency Parser for Maltese: Comparing the impact of transfer learning from Romance and Semitic Languages*.

Chapter 28
Language Report Norwegian

Kristine Eide, Andre Kåsen, and Ingerid Løyning Dale

Abstract The use of Language Technology (LT) has greatly increased in Norway in recent years, as have the linguistic resources needed to make them work. In the past 10 years, Norwegian has adopted new or improved versions of machine translation, speech technology, chatbots and digital assistants, and machine learning has improved. Nevertheless, LT for both written standards of the Norwegian language – the majority Bokmål and minority Nynorsk – is nowhere near the same level as that of major European languages such as English, German, French and Spanish.

1 The Norwegian Language

Norwegian is a North Germanic, verb second, SVO language, spoken by about five million people in Norway, with some additional speakers in the Norwegian diaspora in the US and South America. Norway is a highly digitalised society.

There is great dialectal variation in Norway, and dialects have much higher prestige than in the other Scandinavian countries. Unlike other official European languages, there is no official standard for spoken Norwegian. People tend to speak their own dialect, and expect to be understood. This dialectal variation as well as the pitch accent found in most dialects present the biggest challenges for Norwegian speech technology. While there is no official standard for the spoken language, there are two official written Norwegian languages, Bokmål and Nynorsk. The minority language, Nynorsk, has about 500,000 speakers. All public bodies at state level must be able to correspond with citizens in both written standards, and even though the linguistic differences between Bokmål and Nynorsk are rather small, most types of language technology, such as machine translation, chatbots, spellcheckers, speech-to-text and text-to-speech, need separate tools for each language. Both standards reflect dialectal variation and allow for large formal morphological as well as orthographic variation.

Kristine Eide
The Language Council of Norway, Norway, kristine.eide@sprakradet.no

Andre Kåsen · Ingerid Løyning Dale
The National Library of Norway, Norway, andre.kasen@nb.no, ingerid.dale@nb.no

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_28

With this variation, in combination with highly productive compounding, one single word can have a relatively high number of different spellings, which is a challenge for language technology (Smedt et al. 2012a,b).

2 Technologies and Resources for Norwegian

The overall accessibility of Language Resources (LRs) for Bokmål is fairly good (Eide et al. 2022). Size and contemporaneity are in place for unstructured and semi-structured data. With good linguistic insight, one can build several specialised applications and services from openly available resources. In contrast, most types of LRs and LTs are either scarce or lacking for Nynorsk, although both speech and text data have been added to the largest, open repository for language data (Språkbanken) in recent years. Domain-specific data is severely limited for both Bokmål and Nynorsk. This is also true for the spoken language with all its dialectal variation.

Awareness of the differences between Nynorsk and Bokmål is low outside Norway's borders. Norwegian can often be found in large, multilingual LR collections, and is available as a language choice also on large online platforms. However, both nationally and internationally developed tools and services cater first and foremost to the Bokmål written standard, or the Eastern Norwegian spoken dialect.

Speech technology development is challenged by the dialectal variation, in addition to the two orthographic standards that often allow for spelling variations. There are pronunciation lexicons which cover Bokmål and Nynorsk orthographic forms, and dialectal variation in pronunciation transcriptions is under development for both. Some speech corpora with dialectal variation and a mix of read and spontaneous speech exist, some have transcriptions in both standards. These corpora have proven useful in improving speech recognition scores, but they are either not large enough, or somewhat lacking in domain, style, societal or situational variation to train a robust general purpose speech recognition system. Until recently, speech processing tools have been almost non-existent for Nynorsk. Those that are deemed usable, for either of the written standards, are in general proprietary and not freely available.

The largest text corpus is the Norwegian Colossal Corpus (NCC), which comprises a majority of all Norwegian published works (digitised using OCR), in addition to several other corpora, including Wikipedia, legislation, newspapers, books, web content, etc. The more recently published texts are still copyright-restricted, which limits the availability of the full corpus. The NCC has texts in both written languages, but the Nynorsk proportion is significantly smaller (5-10%). To remedy the scarcity of Nynorsk text data, the Language Bank at the National Library harvests available legal documents from municipalities where Nynorsk is the main language.

There are three large language models (NorELMo, NorBERT, and Notram) for Norwegian, which have been trained on (parts of) the NCC. These models can be fine-tuned with annotated corpora to develop task-specific tools. The language models' embeddings are significantly less robust for Nynorsk than for Bokmål, again due to the disproportionate distribution of the languages in the training data.

Norway does not have access to the same amount of parallel data from the European institutions as the EU Member States. Even so, the ELRC initiative, in which Norway participates, has contributed to a growing awareness of the reusability of translations. Public administrations have contributed significant collections of Bokmål-English parallel data. Valuable translation memories for developing MT systems from English to Bokmål have also come out of EU-funded research projects, e. g., PRINCIPLE. There are very few translation memories between Nynorsk and English, but it is possible to use Bokmål as a pivot language when developing MT technology for English-Nynorsk. The most prominent Nynorsk-Bokmål corpus is the manually corrected output of the Nynorsk press agency Nynorsk Pressekontor's Apertium-based pipeline. Due to the similarities between Nynorsk and Bokmål, MT between the two written standards yields fairly good results.

The most important lexical resource for Norwegian is Norsk ordbank (the Norwegian Word Bank), a lexical database for Norwegian Bokmål and Nynorsk reflecting the official standard orthography as defined in the Norwegian dictionaries Bokmålsordboka and Nynorskordboka. Both are freely available for download and use in LT. While some domain-specific termbases exist for Bokmål, very few terms appear in their Nynorsk parallel, for instance in the national terminology portal Termportalen.

While there is no research programme in Norway aimed specifically at LT, several projects are in the process of filling some of the identified gaps in Norwegian LT and LRs. All major universities in Norway conduct research on LT and/or AI. Among the running projects, NorwAI aims at developing LTs for Scandinavian languages, including conversational search in natural language. SCRIBE seeks to develop an advanced speech-to-text transcription system for spontaneous speech. SANT (Sentiment Analysis for Norwegian Text) is to create open LRs for sentiment analysis for Norwegian. The public broadcasting corporation NRK and two private media groups contribute to the project. The Målfrid project collects all available digital texts from the public sector in Norway. An effort like this will ensure the availability of unstructured text data of a more recent date. CLEANUP aims to develop tools and techniques to automatically anonymise unstructured text data from an array of domains. The project Universal Natural Language Understanding builds upon the UD standard for syntactic treebanks. The goal of the project is to convert the syntactic representation to machine-readable semantic representation.

3 Recommendations and Next Steps

Even though the increase in data availability from 2018 to 2021 has been substantial, awareness of what language data is, what it can be used for and how it should be shared, needs to be raised in all sectors. Due to the lack of Nynorsk data and modern LTs' preference for big data, it must be a priority for decision makers to strengthen LT for the lesser-used language to avoid weakening its equal status. Public sectors must take on their new responsibility as required in the new language act and ensure parallel versions of Bokmål and Nynorsk LT in public procurement.

While there are certain synergies when developing parallel LT for both languages, there is also a need for parallel development of basic resources. The creation of missing tools and LRs must continue. There is a need for more text data for Nynorsk, more domain-specific data, and lexical/terminological resources, in particular for Nynorsk, as well as speech data that cover dialects and Nynorsk in addition to tools for semantic parsing. As for the quality of Norwegian LT, no overreaching assessment has been made of the improvement we assume has taken place. In particular, downstream (user-driven) quality assessment of Norwegian Nynorsk and Bokmål LT is needed, to compare their quality, as well as dialect understanding.

Political action is necessary to open up international platforms to include the possibility of introducing LTs for smaller languages such as Norwegian Nynorsk, even when the large platforms themselves do not offer LT for these smaller languages.

There must be sufficient funding for research and development for Bokmål and Nynorsk LT, and the extra cost of developing parallel versions of Bokmål and Nynorsk LT should be considered when funding future research programmes. A dedicated programme for LT should be considered. Participation in international research projects and programmes that focus on LT, should be encouraged.

References

Eide, Kristine, Andre Kåsen, and Ingerid Løyning Dale (2022). *Deliverable D1.26 Report on the Norwegian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-norwegian.pdf.

Smedt, Koenraad De, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard (2012a). *Norsk i den digitale tidsalderen (bokmålsversjon) – The Norwegian Language in the Digital Age (Bokmål Version)*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/norwegian-bokmaal.

Smedt, Koenraad De, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard (2012b). *Norsk i den digitale tidsalderen (nynorskversjon) – The Norwegian Language in the Digital Age (Nynorsk Version)*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/norwegian-nynorsk.

Chapter 29
Language Report Polish

Maciej Ogrodniczuk, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski

Abstract The quality of language technology (LT) for Polish has greatly improved recently, influenced by three independent trends. The first one is Poland-specific and concerns the increase in national funding of both scientific and R&D projects, resulting in the construction of The National Corpus of Polish and the development of the CLARIN-PL and DARIAH-PL infrastructures. Two other trends are global: the development of language resources (LRs) and tools by private companies and of course, the deep learning revolution which has led to enormous improvements in the state-of-the-art in all fields of language processing.

1 The Polish Language

Polish is a Slavic language of the Lechitic group, written in Latin script. It is the most spoken West Slavic language in the world. It is the official language of the Republic of Poland and since 2004, the sixth largest official language of the European Union. It is spoken by 10% of EU citizens: about 40 million native speakers and 10 million second language speakers worldwide. In Poland, it is the common spoken and written language and the native language of the vast majority of the population.

Polish exhibits some specific characteristics (Pisarek 2007), which contribute to the richness of the language but present a challenge for computational processing. Word order is relatively free, which is used mostly to stress the importance of information rather than simply following grammatical rules.

Maciej Ogrodniczuk
Inst. of Comp. Science, Polish Academy of Sciences, Poland, maciej.ogrodniczuk@ipipan.waw.pl

Piotr Pęzik
University of Łódź, Poland, piotr.pezik@uni.lodz.pl

Marek Łaziński
University of Warsaw, Poland, m.lazinski@uw.edu.pl

Marcin Miłkowski
Inst. of Philosophy and Sociology, Polish Academy of Sciences, Poland, mmilkows@ifispan.edu.pl

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_29

Polish is relatively morphologically rich, which means that for roughly 180,000 base forms of words, almost 4 million inflected word forms exist. The inflection paradigms are complex, and even their exact number is a matter of dispute, as single exceptions might even be thought to create a new paradigm. Even native speakers have problems with properly inflecting many words, and most speakers of Polish as a second language never completely master the complexities of the inflectional system. Polish syntax is similar to its neighbouring Slavic languages with a tendency to analyse constructions seen in gender marking, forms of address and the use of infinitive and impersonal constructions.

Polish is currently highly influenced by English, one of the biggest sources of neologisms and calques, in particular in science and technology. The number of words loaned from English into Polish is, however, much lower than in Dutch or German because of the problem with inflecting some words as well as differences in pronunciation systems. Other recent changes are the appearance of more direct forms of address and simplification of the traditional inflection patterns.

2 Technologies and Resources for Polish

The level of technology support for Polish is similar to that of many other official EU languages, with several available resources[1] and basic text processing tools obtaining satisfactory accuracy scores.[2] The current landscape of Polish language processing has been shaped by the following developments (see Ogrodniczuk et al. 2022; Miłkowski 2012): 1. The construction of the National Corpus of Polish[3] (NKJP; Przepiórkowski et al. 2012), a reference corpus containing over 1.5 billion words sampled from diverse sources such as classical literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived online texts, balanced with respect to gender, age and regional distribution of samples. The availability of the corpus, and particularly its manually annotated 1-million word sub-corpus, available under a CC-BY-licence, has boosted both research in the humanities as well as the development of many NLP tools. Since the completion of the NKJP in 2011, other reference corpora have been used to represent recent developments in Polish. The most significant examples are the MoncoPL monitoring corpus (Pęzik 2020) and the Corpus of the 2010s.[4] 2. The development of the CLARIN-PL[5] and DARIAH-PL[6] infrastructures, led to the development of many resources and tools such as Słowosieć, the Polish WordNet[7] (Dziob et al. 2019), Ko-

[1] http://clip.ipipan.waw.pl/LRT

[2] http://clip.ipipan.waw.pl/benchmarks

[3] http://nkjp.pl

[4] http://korpus-dekady.ipipan.waw.pl

[5] https://clarin-pl.eu, http://clarin.biz

[6] https://dariah.pl, https://lab.dariah.pl

[7] http://plwordnet.pwr.wroc.pl/wordnet/

rpusomat, a corpus creation tool[8] (Kieraś and Kobyliński 2021), COMBO, a neural tagger, lemmatiser and dependency parser[9] (Klimaszewski and Wróblewska 2021), or SpokesPL, a search engine for Polish conversational data.[10] 3. External funding in the form of grants, both European (Horizon 2020, Connecting Europe Facility) or national, distributed by the National Science Centre and National Centre for Research and Development, have allowed many research institutions and companies to increase the budgets of research projects by an order of magnitude, and thus react to commercial demands for speech recognition or dialogue systems. As a result, their NLP products are characterised by state-of-the art performance. 4. The PolEval evaluation campaign for NLP tools for Polish[11] started in 2017 as a practical exercise intended to advance the state-of-the-art with a series of tasks in which submitted tools compete against one another. This contest has brought the NLP community together and resulted in the development, enhancement and public release of reference datasets for tasks such as sentiment analysis, speech recognition and machine translation. 5. The latest Transformer models (HerBERT[12] and plT5[13]) trained by researchers from the company Allegro and the Institute of Computer Science of the Polish Academy of Sciences, based on several large corpora of Polish, including NKJP. Making these models freely available for the community has facilitated enormous progress. 6. Increased accessibility of multimodal spoken corpora and speech databases such as a large annotated corpus of phone-based customer support dialogues,[14] which boosts the development of goal-oriented chatbots and helps Polish ASR engines to be on par with solutions by global service providers. Nonetheless, many complex and labour-intensive resources such as audio-video corpora and corpora with discourse structure and semantic annotations are practically unavailable.

3 Recommendations and Next Steps

The national Polish AI strategy (Council of Ministers 2020) mentions the development of LT as a short-term goal, supported by national grants for projects related to Polish language processing based on world-leading algorithms. Notably, the document mentions the importance of language data: the need for the elimination of legal barriers to the exploration of language text corpora under copyright protection and awarding projects that make architecture, trained models and training data sets available for common use. This assumption is in line with findings from the Polish NLP community as well as international trends. What needs to be added to this plan is ac-

[8] https://korpusomat.pl

[9] https://github.com/360er0/COMBO

[10] http://spokes.clarin-pl.eu

[11] http://poleval.pl

[12] https://huggingface.co/allegro/herbert-large-cased

[13] https://huggingface.co/allegro/plt5-large

[14] http://pelcra.pl/new/diabiz

cess to common (national or European) computing power to boost the development and optimization of standard language models and secure stable funding for crucial LRs such as the National Corpus of Polish or the Great Dictionary of Polish.

However, there is also a new dimension of this plan, created by the Russian invasion of Ukraine. With 3 million Ukrainian refugees in Poland in 2022, bilingual public administration has become an important new role for the Polish LT community, and is boosting the development of bilingual Polish-Ukrainian resources and tools. On the European level, this new situation calls for the embracing of Ukrainian as one of the languages officially supported by the EU.

References

Council of Ministers (2020). *Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020 – The Policy for the development of AI in Poland from 2020*. https://www.gov.pl/web/ai/polityka-dla-rozwoju-sztucznej-inteligencji-w-polsce-od-roku-2020.

Dziob, Agnieszka, Maciej Piasecki, and Ewa Rudnicka (2019). "plWordNet 4.1 – a Linguistically Motivated, Corpus-based Bilingual Resource". In: *Proceedings of the 10th Global Wordnet Conference*. Global Wordnet Association, pp. 353–362.

Kieraś, Witold and Łukasz Kobyliński (2021). "Korpusomat – stan obecny i przyszłość projektu". In: *Język Polski* CI.2, pp. 49–58. DOI: 10.31286/JP.101.2.4.

Klimaszewski, Mateusz and Alina Wróblewska (2021). "COMBO: State-of-the-Art Morphosyntactic Analysis". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 50–62.

Miłkowski, Marcin (2012). *Język polski w erze cyfrowej – The Polish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/polish.

Ogrodniczuk, Maciej, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski (2022). *Deliverable D1.27 Report on the Polish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-polish.pdf.

Pęzik, Piotr (2020). "Budowa i zastosowania korpusu monitorującego MoncoPL". In: *Forum Lingwistyczne* 7, pp. 133–150. DOI: 10.31261/FL.2020.07.11.

Pisarek, Walery (2007). *The Polish Language*. Warsaw: The Council for the Polish Language.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, eds. (2012). *Narodowy Korpus Języka Polskiego*. Warsaw: PWN. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

Chapter 30
Language Report Portuguese

António Branco, Sara Grilo, and João Silva

Abstract This chapter provides an analysis of the level of technological preparation of the Portuguese language for the digital age, as well as the actions necessary for the consolidation of Portuguese as a language of international communication with global projection.

1 The Portuguese Language

Portuguese is the fifth most spoken language in the world, with around 280 million speakers (Instituto Camões 2021), of which 250 million are native speakers, spread over four continents: Africa, America, Asia and Europe. It is the official language of Angola, Brazil, Cape Verde, East Timor, Guinea-Bissau, Macau, Mozambique, Portugal, S. Tome and Principe, and Equatorial Guinea. All variants of Portuguese across the different continents are mutually understandable. Portuguese is an official language of the European Union, the Mercosul and the African Union. With the advancement of the alphabetisation in the African countries and in East Timor, Portuguese is confirming its growth potential in terms of the number of speakers. This chapter is partly based on Branco et al. (2022) and Branco et al. (2012).

Portuguese has a strong presence in social networks. For instance, a study of 100 million tweets reveals that Portuguese is the sixth most spoken language on Twitter, after English, Japanese, Spanish, Korean and Arabic.[1]

Portuguese is a Romance language, with most of its lexicon being derived from Latin. To a speaker not knowing Portuguese, the European variant of this language may often sound like a sequence of consonants. This is due to the fact that, differently from the other Romance languages, the Portuguese unstressed vowels are often weakened or even not pronounced. This vowel weakening is a late change in

António Branco · Sara Grilo · João Silva
University of Lisbon, Portugal, antonio.branco@di.fc.ul.pt, sara.grilo@di.fc.ul.pt, joao.silva@di.fc.ul.pt

[1] https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_30

European Portuguese and it did not affect the variety spoken in Brazil, which in this respect is closer to the Portuguese which was spoken some centuries ago.

The basic word order in Portuguese is subject-verb-object (SVO) (*ele leu o livro* / he read the book). Portuguese is a null subject language, where the subject of the sentence may not be realised by a phonetically overt expression (_ *li o livro* / [I] read the book). The inflection paradigm in Portuguese is very rich, especially in verbs. A verb with a regular inflection paradigm will have different markers for aspect, tense, mood, person, number or polarity, culminating in more than 160 different inflected verb forms, encompassing both simple and complex ones.

The advent of the digital age is a major challenge for the Portuguese language and its speakers. The scientific study and technological development of the Portuguese language, making it fit for the digital age, is thus an endeavour of utmost importance in order to ensure that its speakers can participate in the information society.

2 Technologies and Resources for Portuguese

The activity in Language Technology (LT) for the Portuguese language can be traced back to projects, programmes and initiatives carried out in the last decades.

One of the first important programs in this area was EUROTRA, an ambitious Machine Translation project established and funded by the European Commission from the late 1970s until 1994. The participation of Portugal in this project since 1986 was undertaken by ILTEC, specifically created for this purpose and involving mostly researchers from the Universities of Lisbon and Porto.

Another key European project was LE-PAROLE, developed in the late 1990s, with the participation of CLUL and INESC-ID. Its main achievement was the building of corpora and lexicons according to integrated models of composition and materials description. Part of this corpus was enriched and enlarged in the national project TagShare, conducted at the University of Lisbon, in the Department of Informatics (NLX Group) and in the Center of Linguistics (CLUL), in 2005. This project enabled the development of a set of language resources and software tools to support the computational processing of Portuguese. The result was a 1 million word corpus linguistically annotated and fully verified by experts, the CINTIL corpus, and a whole range of processing tools for tokenisation, morphosyntactic category (POS) tagging, inflection analysis, lemmatisation, multiword lexeme recognition, named entity recognition, etc., in the LX-* collection.

On the basis of these tools and resources, top-quality, manually verified treebanks, with syntactic and semantic grammatical analysis, and the companion computational grammar and parsers, have been also developed for the CINTIL-* and LX-* collections, in the national project SemanticShare at the Department of Informatics (NLX Group) of the University of Lisbon. The Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico), released in 2000, in turn, is a corpus of about 180 million words from excerpts of a Portuguese daily newspaper.

In the field of speech processing, it is worth noting the TECNOVOZ project, which started in 2006. This project was directed by INESC-ID and one of its major goals was to foster technology transfer to the business sector, having as partners companies like the public television RTP.

On the industry side, an important contribution to fostering an LT industry in Portugal was the establishment of the international Microsoft Language Development Center, near Lisbon, which lasted from 2005 to 2015. More recently, the two US-based startups DefinedCrowd and Unbabel have a significant presence in Portugal.

In Brazil, relevant efforts in LT for Portuguese have also been undertaken. To mention just a few illustrative examples, in the early 1990s, under the DIRECT project, the Bank of Portuguese was created at the Pontifical Catholic University of São Paulo. Since its inception, the Bank of Portuguese has been a source of data for corpus-based studies for several projects.

Also worth mentioning is the Summ-it corpus, built to support the study of summarisation along with the phenomena of anaphoric and rhetorical relations in Portuguese. This resource was developed under the PLN-BR project, by the Núcleo Interinstitucional de Lingüística Computacional (NILC), driven by the University of São Paulo and gathering researchers from seven other Brazilian institutions.

On par with these programmes and projects both in Brazil and in Portugal, it is worth underlining PROPOR as the key focal initiative of the research community working on Portuguese. PROPOR is the major international scientific conference devoted to the computational processing of Portuguese. The location of this biennial conference has been alternating between the two countries since 1993.

A landmark for the language technology for Portuguese landscape is the white paper *The Portuguese Language in the Digital Age* (Branco et al. 2012), produced in the scope of the European META-NET initiative.

As an outcome of the European CEF project ELRI, the Repository for Translation Resources (eTradução)[2] is available which has been maintained since 2019 by AMA, the government agency for the digital transformation of the Portuguese public administration. Several of its data sets are also distributed through ELRC-SHARE.

The major AI initiative specifically addressing the field of LT is the implementation (2017-2021) and operation (from 2021 onwards) of the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language.[3]

3 Recommendations and Next Steps

The development of technologies for Portuguese has progressed over the past decade. However, given that progress in LT has accelerated, the level of competitive technological preparation of Portuguese for the digital age has not changed significantly over this period when taking the best prepared language, English, as a reference.

[2] https://etraducao.gov.pt/pt-pt/

[3] https://portulanclarin.net

Some progress has been made in the area of text analytics and machine translation, thanks to further data collection and corpus creation through a number of initiatives funded by EU projects and national entities. Fundamental building blocks such as syntactic analysis tools have progressed significantly, but the underlying datasets still need to be enlarged to build more robust, reliable and application-ready systems.

There are still a large number of fundamental tools and datasets not yet available for Portuguese. While steps have been made towards speech corpus development, there is still no state-of-the-art automatic speech recognition system available for Portuguese as open-source software.

From a natural language understanding perspective, there is a lack of semantic-based datasets and tools. Critically, there is a severe lack of freely available large language models, also known as foundation models, based on deep language learning with artificial neural network technology. Such language models to support deep neural processing, including the development of large multimodal language models involving Portuguese, are thus very much needed, especially those openly available to be used in research and in innovation

The above considerations on the availability of data and tools for Portuguese clearly indicate the urgent need to direct substantially more funding and efforts to the preparation of Portuguese for the digital age. The scientific study and technological development of the Portuguese language is a crucial endeavour for its promotion, in order to ensure that its speakers can participate in the information society.

References

Branco, António, Sara Grilo, and João Silva (2022). *Deliverable D1.28 Report on the Portuguese Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-portuguese.pdf.

Branco, António, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, Vera Lúcia Strube de Lima, and Fernanda Bacelar (2012). *A língua portuguesa na era digital – The Portuguese Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/portuguese.

Instituto Camões (2021). *Português no Mundo*. https://pt.institutocamoes-praga.cz/centro-de-lingua-portuguesa-instituto-camoes/portugues-no-mundo/.

Chapter 31
Language Report Romanian

Vasile Păiş and Dan Tufiş

Abstract Since the previous META-NET report, there have been significant improvements (e. g., creation of a large Romanian national corpus, steady progress in written language technologies, LT, construction of a national LT portal for the Romanian language etc.), but things are far from what they should be. Support for LT and AI through national programmes is still modest, although there are signs of a more active involvement of policy makers in the strategic planning and funding programmes in this domain. Continued research is required to produce large language models, able to capture the characteristics of the Romanian language. Large language resources need to be created so that AI systems are able to learn from them.

1 The Romanian Language

The Romanian language which is an official language of the EU is also the official language of Romania. It is spoken by 19.4 million people in Romania and by about 3.5 million people in Moldova, where it is unofficially known as a Moldavian language. Speakers of Romanian in other European countries (Albania, Bulgaria, Croatia, Greece, Hungary, North Macedonia, Serbia, Ukraine and others) and communities of immigrants in Australia, Canada, Israel, Latin America, Turkey, USA and Asian countries total around 4 million Romanian native speakers.[1]

Romanian is an official language in the Autonomous Province of Vojvodina in Serbia. It is one of the languages spoken in the autonomous Mount Athos in Greece and a recognised minority language in Ukraine (Trandabăț et al. 2012). Romanian has four dialects: Daco–Romanian, Aromanian (about 500,000 speakers in Albania, Bulgaria, Greece and North Macedonia), Istro–Romanian (15,000 speakers in two small areas in the Istrian Peninsula, Croatia) and Megleno–Romanian (about 5,000 speakers in Greece and North Macedonia).

Vasile Păiş · Dan Tufiş
Romanian Academy, Romania, vasile@racai.ro, tufis@racai.ro

[1] https://en.wikipedia.org/wiki/Romanian_diaspora

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_31

The Romanian alphabet is based on the Latin script with five additional letters using diacritics (Ă, Â, Î, Ș, Ț and ă, â, î, ș, ț). Many digital texts are written without diacritics. The quotation marks use double low (left) and right marks („ and ”, respectively). However, especially in digital texts, the ASCII quotation mark character may be encountered. Dialogues are introduced using quotation dashes (–). The Oxford comma, used in certain English language documents, is considered incorrect in the Romanian language. In titles, only the first letter of the first word is capitalised, with the rest of the title making use of regular sentence capitalisation. Names of months and days, as well as adjectives derived from proper names are not capitalised, e. g., februarie (February), vineri (Friday), italian (Italian).

2 Technologies and Resources for Romanian

The availability of language-specific data has a direct impact on the quality of language-specific or cross-language tools. The availability of large pre-trained multilingual models that include representations for Romanian language, such as XLM-RoBERTa or mBERT, somewhat alleviates the problem of constructing compute-intensive contextual word representations. Nevertheless, monolingual representations such as RoBERT (Masala et al. 2020), DistilRoBERT (Avram et al. 2022), and ALR-BERT, led to increased performance of monolingual tools (Tufiș 2022). Static representations, such as CoRoLa-based word embeddings (Păiș and Tufiș 2018), are still used due to their lower compute requirements (Păiș and Tufiș 2022).

Word representations form only the basis of advanced language tools. In addition to language models, task-specific corpora are required to train and evaluate the tools. The vast majority of Romanian resources are multilingual, with some being bilingual, and only a few monolingual corpora exist. Compared to English, the available Romanian corpora represent around 10%. Available speech corpora with Romanian audio represent 5% of available English resources and about 50% when compared to neighbouring EU countries.

In spite of the reduced number of available language resources, applications for different NLP tasks exist for Romanian. These include lemmatisation, part-of-speech tagging, dependency parsing, named entity recognition, syllabification, speech recognition, text-to-speech, machine translation, punctuation restoration, terminology annotation, and text classification. The number of identified tools represents only 15% of the tools available for English.

Even if, in general, all LT fields are covered, certain fields are less developed or considered for the Romanian language by researchers and developers: language generation, dialogue management, multimodal corpus building, and social media aspects (including micro-blogging, social networks, and meme interpretation). Speech processing is much less mature than LT for written text, both in terms of corpora and instruments. Even though there has been much work on processing general Romanian language, more focus is needed for creating domain-specific resources and tools (especially for the biomedical, legal, economy and social media domains).

The Representative Corpus of Contemporary Romanian Language (CoRoLa)[2] (Tufiş et al. 2019) was created by the Romanian Academy as the largest IPR-cleared reference corpus of written and spoken Romanian. Texts cover four domains (arts and culture, science, society, nature), reflecting six styles (imaginative, journalistic, scientific, legal, administrative, memoirs) and different document types.

One of the largest Romanian speech corpora is RSC (Georgescu et al. 2020), containing 100 hours of audio files. The multilingual speech corpus VoxPopuli contains 83 hours of Romanian language speech. The speech component of the CoRoLa corpus (comprised of multiple smaller corpora together with additional audio files specifically obtained for inclusion in CoRoLa) totals 103 hours aligned with the text.

A number of Romanian LTs, covering different fields of research, are available within the RELATE[3] (Păiş et al. 2020) portal. The platform covers results derived from more than six national and international research projects.

3 Recommendations and Next Steps

Task-specific Romanian corpora (including multi-modal) are needed to enable new and complex language processing operations. In turn, these must lead to the development of new tools, finally working towards digital language equality. This requires dedicated long-term support at the national, regional and European levels. Furthermore, AI research should follow a human-centered approach. Biased or potentially harmful data in resources should be detected and addressed. This, together with following lawful and ethical principles, as well as robust implementations, should enable building Trustworthy AI (TAI)[4] applications for the Romanian language.

AI is an area of strategic importance and a key driver of economic development, providing solutions to many societal challenges. In this context, many EU countries prepared national plans for AI (e. g., the *Spanish National AI Strategy*[5] or the *French AI for Humanity*[6]). In Romania, however, there is currently no such national plan for AI. A strategy for AI[7] has been proposed recently within the RePatriot[8] project, but it was not adopted at national level. Furthermore, the strategy is not very concrete, it centres mostly on which Romanian sectors would benefit most from AI, and which steps are important for the process of developing Romanian AI initiatives, but it does not include any plans about how to accomplish these actions.

---

[2] https://corola.racai.ro

[3] https://relate.racai.ro

[4] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[5] https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/021220-ENIA.pdf

[6] https://www.aiforhumanity.fr/en/

[7] https://www.slideshare.net/MonicaIon1/strategy-romania-in-the-era-of-artificial-intelligence-rblrepatriot

[8] https://repatriot.ro

References

Avram, Andrei-Marius, Darius Catrina, Dumitru-Clementin Cercel, Mihai Dascalu, Traian Rebedea, Vasile Pais, and Dan Tufis (2022). "Distilling the Knowledge of Romanian BERTs Using Multiple Teachers". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 374–384. https://aclanthology.org/2022.lrec-1.39.

Georgescu, Alexandru-Lucian, Horia Cucu, Andi Buzo, and Corneliu Burileanu (2020). "RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 6606–6612. https://aclanthology.org/2020.lrec-1.814.

Masala, Mihai, Stefan Ruseti, and Mihai Dascalu (2020). "RoBERT – A Romanian BERT Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: International Committee on Computational Linguistics, pp. 6626–6637. DOI: 10.18653/v1/2020.coling-main.581. https://aclanthology.org/2020.coling-main.581.

Păiș, Vasile, Radu Ion, and Dan Tufiș (2020). "A Processing Platform Relating Data and Tools for Romanian Language". In: *Proceedings of the 1st International Workshop on Language Technology Platforms*. Marseille: European Language Resources Association, pp. 81–88. https://aclanthology.org/2020.iwltp-1.13.

Păiș, Vasile and Dan Tufiș (2018). "Computing distributed representations of words using the CoRoLa corpus". In: *Proceedings of the Romanian Academy Series A* 19.2, pp. 185–191.

Păiș, Vasile and Dan Tufiș (2022). *Deliverable D1.29 Report on the Romanian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-romanian.pdf.

Trandabăț, Diana, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiș (2012). *Limba română în era digitală – The Romanian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/romanian.

Tufiș, Dan (2022). "Romanian Language Technology – a view from an academic perspective". In: *International Journal of Computers Communications & Control* 17.1. DOI: 10.15837/ijccc.2022.1.4641.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei (2019). "Little Strokes Fell Great Oaks. Creating CoRoLa, The Reference Corpus of Contemporary Romanian". In: *Revue roumaine de linguistique* 64.3, pp. 227–240.

Chapter 32
Language Report Serbian

Cvetana Krstev and Ranka Stanković

Abstract Standard Serbian is the national language of Serbs and the official language in the Republic of Serbia. Although statistics show that the population of Serbia is well equipped to use IT, and although some important language resources and tools have been developed for Serbian, the language still lags significantly behind most European languages in terms of Language Technology (LT). This shows that a stable, dedicated and long-term investment in the development of LT for Serbian through national and international scientific and development projects is needed.

1 The Serbian Language

Standard Serbian is the national language of Serbs and the official language in the Republic of Serbia. Formed on the basis of Ekavian and Ijekavian Neo-Štokavian South Slavic Dialects, its form was determined by the reformer of the written language of Serbs Vuk Karadžić, who also reformed both the Cyrillic alphabet and orthography. In the 20th century, in the federal state of Yugoslavia, this language was officially encompassed by Serbo-Croatian, a name that implied a linguistic unity with Croats (and later with other nations whose languages were based on Neo-Štokavian dialects). In the 1990s, in Serbia the name Serbo-Croatian was replaced by the name Serbian. The Constitution of the Republic of Serbia from 2006 stipulates: "The Serbian language and the Cyrillic alphabet shall be in official use in the Republic of Serbia". However, the Latin alphabet is also in widespread use.

According to the 2011 census data published by the Statistical Office of the Republic of Serbia, the population of Serbia is 7,186,862, and Serbian is the mother tongue of 88.1% of the population. To this number, one should add the ethnic Serb population in other parts of former Yugoslavia (a number not easy to determine). The Serbian diaspora lives primarily in a number of countries of Central and Western Europe, in the US, Canada and Australia, and their knowledge of Serbian is mainly determined by the generation of immigrants they belong to.

Cvetana Krstev · Ranka Stanković
University of Belgrade, Serbia, cvetana@matf.bg.ac.rs, ranka@rgf.bg.ac.rs

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_32

The Statistical Office also collects data about the use of ICT in Serbia each year (Kovačević and Rajčević 2021). According to their data for 2021, published on 22 October 2021, the percentage of citizens between 16 and 74 years of age that used a computer regularly was 74.8%, while the internet was used regularly by 81.2% of citizens. Additionally, 76.7% of households possessed a computer in 2021, while 81.5% of all households had an internet connection. The internet was used for private purposes mostly for communicating with others, reading online news and magazines, and using social media. As for e-government, this study showed that 40% of internet users used online services instead of personally visiting public institutions and administrative bodies.

2 Technologies and Resources for Serbian

The variety of corpora as well as their availability has improved significantly in the last 10 years (Vitas et al. 2012; Krstev and Stanković 2022). Two corpora of contemporary Serbian are available online. The first, published in 2013 (SrpKor2013), contains more than 120 million words, while the second, published in 2021 (SrpKor2021), contains more than 600 million words. Both are annotated with part-of-speech and lemmas and contain a variety of text types, with literary text being particularly well represented. Along with the general corpus SrpKor2021, several large collections of domain texts were prepared that can be used within the same platform. Additionally, many text collections exist that contain data obtained from various news portals or by web crawling, some of which are represented as raw text, others are annotated with POS and lemmas, while a few are fully morphologically and/or NE-annotated. Some collections were prepared for a special purpose, such as sentiment analysis, text similarity and text paraphrasing analysis.

There are several bilingual, sentence-aligned corpora that include Serbian as one of the languages, with the other being English, French or German; texts are from various domains, including a large portion of literary texts. The digital library Bibliša supports online search of these corpora. Besides, there are numerous multilingual collections that include Serbian, with the majority of them being comparable.

By far the most comprehensive of the many lexical resources for Serbian is Serbian Morphological Dictionaries (SrpMD, Krstev 2008), covering both simple and multi-word units, general lexica, proper names, and domain-specific lexica. It covers morphological descriptions and, to a certain extent, semantics, usage, pronunciation, etymology, domains, derivational relations, etc. and it is being permanently updated. These dictionaries are open for search through the platform Leximirka at the site of JeRTeh,[1] while its largest part with full morphological description and restricted additional information is made public. There are also several monolingual and bilingual inflectional lexicons based on MULTEXT-East.

[1] The Association of Language resources and tools, http://jerteh.rs

Significant results have been achieved in the development of terminology resources for Serbian including simple- and multi-word terms from a wide range of domains. Part of these resources are bilingual (Sebian/English) or multilingual, and some of them can be searched on the platform Termi at JeRTeh. Several special purpose mono-, bi- and multilingual lexical resources have been built that include Serbian, primarily for sentiment analysis and hate-speech detection.

The Serbian WordNet, aligned to the Princeton WordNet 3.0, and SentiWordNet, is still underdeveloped. Formal domain ontologies for Serbian are rare.

There are a few language models and grammars for Serbian including Dict2Vec, an embedding model adapted for Serbian using the Serbo-Croatian Wikipedia and Wiktionary synonym pairs, and BERTić, a Transformer model pre-trained on eight billion tokens of crawled text from the Croatian, Bosnian, Serbian and Montenegrin domains. As for MT, we can only mention rudimentary attempts done in the scope of scientific research and products created by big technology enterprises.

Several taggers and/or lemmatisers for Serbian have been developed based on TreeTagger, spaCy, NLTK and others. Many of them are part of NLP suites that cover various tasks. Numerous local grammars (e. g., for compound verb forms, nominal phrases etc.) have been developed for Serbian texts using the Unitex/Gramlab corpus processing suite and SrpMD. Parsing of Serbian is possible online using Universal Dependency and CLASSLA piplines. The first NER system was the rule- and lexicon-based system SrpNER that tags fine-grained entities. It was used to produce training data for NER systems developed using various ML methods and tools. A web service was developed for the morphological and semantic query expansion that was incorporated into several online applications, such as the Bibliša digital library.

A substantial breakthrough in the area of speech processing was made by the AlphaNum company, a spin-off of the University of Novi Sad. They offer a large variety of commercial products and services: speech technologies, voice assistants, products for the disabled, etc.

The document *Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025* was adopted by the Government in 2019. As a result, the Institute for Artificial Intelligence was founded, with NLP as one of its research areas. However, there is still no LT-related funding in Serbia.

The strongest NLP/LT group consists of researchers from the University of Belgrade and JeRTeh. They started to work more than 40 years ago under the guidance of Prof. Duško Vitas, and they have produced by far the most resources and tools. The strongest group for speech technologies comes from the University of Novi Sad. In recent years, new NLP/LT research groups affiliated with different universities and research centres have emerged. Outside academia there are few LT providers.

3 Recommendations and Next Steps

According to recent statistical data provided by official authorities, Serbian citizens are equipped to live in the digital world and are ready to use LT. However, this

overview of LT for Serbian shows that some resources for Serbian are rich and diverse, while some types of resources are still rare, and some practically do not exist. This analysis of the availability of resources, tools and services shows that Serbian is only weakly or fragmentarily supported. It also reveals that although languages close to Serbian (geographically, historically and by the number of speakers) such as Bulgarian, Slovene and Croatian lag behind English, they have better LT support than Serbian. The policies taken in these countries to promote and support LT can serve as a guideline on how to improve LT for Serbian.

Despite the valuable achievements documented here, Serbian is still a disadvantaged language, with the risk that in a few years Serbian speakers will not benefit from the AI/LT revolution. To prevent this from happening, there is a need for more dedicated LT funding, on both the national and international level. This is especially important bearing in mind that in the past, as well as today, researchers working on NLP/LT for Serbian are mostly affiliated with state universities, which require stable and adequate levels of funding. At the international level, Serbian and other weakly supported languages would benefit from more knowledge transfer projects that would not merely aim at mirroring existing solutions for English, but rather support the production of adequate resources and tools for endangered languages.

References

Kovačević Miladin, Vladimir Šutić and Uroš Rajčević (2021). *Употреба информационо-комуникационих технологија у Републици Србији, 2021. (The use of ICT in the Republic of Serbia in 2021)*. https://publikacije.stat.gov.rs/G2021/Pdf/G202116016.pdf.

Krstev, Cvetana (2008). *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Belgrade: University of Belgrade, Faculty of Philology.

Krstev, Cvetana and Ranka Stanković (2022). *Deliverable D1.35 Report on the Serbian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-serbian.pdf.

Vitas, Duško, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević (2012). *Српски језик у дигиталном добу – The Serbian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/serbian.

Chapter 33
Language Report Slovak

Radovan Garabík

Abstract For Slovak, all the fundamental NLP building blocks for basic applications exist, but they are often of lesser quality and lower accuracy than those of other languages. The availability of free and open tools and data is rather low, with most of the resources proprietary. Compared to neighbouring languages of similar levels of NLP development (Czech, Polish, Hungarian), Slovak is positioned toward the lower end of this group. Slovak language support by "big players" in the LT industry is comparable to other European languages with similar size; speech recognition and synthesis work acceptably while machine translation between Slovak and English is almost good enough to be used by professionals as a source for post-editing. Spell checkers, LT-assisted mobile phone input, OCR and lemmatised fulltext search are taken for granted, although their quality is significantly lacking compared to bigger European languages.

1 The Slovak Language

Slovak is the official language in the Slovak Republic. Since May 2004 it has also been one of the administrative languages of the European Union. According to the 2021 census data, out of 5.4 million inhabitants of Slovakia, 4.7 million people have Slovak as their mother tongue.[1] Other estimates (perhaps overly optimistic) claim that Slovak is spoken by more than one million emigrants in the United States, about 300,000 people in the Czech Republic, and smaller groups in Hungary, Romania, Serbia, Croatia, Bulgaria, Poland and other countries. A fact which is not well known is that there exists another written variant of (Eastern) Slovak, using Cyrillic script. This variant is used around Ruski Krstur in Serbia by a few thousand speakers, but thanks to historical religious circumstances it is generally considered a dialect of the Rusyn language, not Slovak. As such, it is almost completely ignored in all aspects concerning Slovak linguistics.

Radovan Garabík
Slovak Academy of Sciences, Slovakia, radovan.garabik@kassiopeia.juls.savba.sk

[1] Corrected for the inhabitants with an unidentified mother tongue.

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_33

As a typical Slavic language Slovak is moderately inflected with a complex morphology and relatively flexible word order. It has three or four[2] genders, two grammatical numbers, three tenses and prominent aspectual pairs. It belongs (together with Polish, Czech, Lower and Upper Sorbian) to the West branch of Slavic languages. In the 16th to 18th centuries, Czech was used as the cultural language in Slovakia, together with several types of cultural Slovak, and the modern standard of the language dates to the second half of the 19th century.

Slovak is generally considered to be mutually intelligible with Czech, with some caveats regarding different inflection of pronouns, some lexical and terminological differences and differences in verb conjugations. Czech enjoys a unique sociolinguistic status in Slovakia; the population is widely exposed to the Czech language in media (TV, movies, internet, and literature). As a result, Czech is widely understood in Slovakia above the level of natural mutual intelligibility. Note that the opposite – exposure of Czech Republic inhabitants to the Slovak language – is only marginal. Despite this, the visible influence of Czech on Slovak is limited to some lexical items and syntactical constructions, often regarded as "incorrect".

The language is written using the Latin alphabet with additional diacritical marks, marking palatalisation of consonants, postalveolars, and phonemic length of vowels and consonants. The Slovak alphabet has the distinction of having the greatest number of characters (43, or 46 including digraphs) among European languages.

On the web, Slovak is a sharply localised language, closely interwoven with the .sk top-level domain (TLD). Distribution (as of 2021) of the most frequent top-level domains of web pages in the Slovak language from the Araneum Slovacum VI Maximum Beta web corpus (Benko 2014) shows that 76.6% of documents in Slovak are from the .sk TLD; 8.8% from the .com TLD, 3.8% from .cz, 2.9% from .eu, 2.0% from .net and the rest from other, less frequent domains.

2 Technologies and Resources for Slovak

Slovak language NLP and LT[3] lag behind that of neighbouring languages of similar status (i. e., Czech, Polish and Hungarian). Predominantly developed in academic environments (Šimková et al. 2012), Slovak language technologies used to be mostly limited to lemmatisation and morphosyntactic analysis, with some limited industry interest in other tools (e. g., NER). The situation has somewhat changed in recent years, with industry more interested in deep learning models. Nevertheless, the availability of huge language corpora and lexical resources available for Slovak is comparable to similar languages (Aldabe et al. 2022).

The main institution tasked with compiling and curating big, representative corpora is the Slovak National Corpus (SNK)[4] department of the Ľ. Štúr Institute of

[2] Masculine is sometimes analysed as two genders; masculine animate and masculine inanimate.

[3] See, for example, https://github.com/essential-data/nlp-sk-interesting-links

[4] https://korpus.sk

Linguistics, Slovak Academy of Sciences. SNK was also active in developing basic digital language resources of the contemporary language, but also parallel corpora, spoken, dialect and historical corpora and lexicographical databases (Garabík 2010) and in digitalisation of linguistic research in Slovakia.

Corpora compiled at SNK have formed an indispensable part of linguistic research in Slovakia for a number of years, together with the ARANEA family of huge web corpora for more than 20 languages (Benko 2014).[5] Currently, the main Slovak language corpus, prim-10.0, contains about 1.7 billion words.[6] The web corpus Araneum Slovacum VI Beta contains about 4.4 billion words. In NLP and LT industry, companies usually use in-house collected web corpora.

Official Slovak translations of various EU texts (such as Acquis communautaire, EU parliament proceedings, Official Journal of the EU etc.) make up the bulk of available, unrestricted by copyright, parallel corpora suitable for MT-related tasks.

All building blocks of basic NLP processing for Slovak are covered: lemmatisation (since Slovak is a moderately inflected language, lemmatisation is often indispensable for any subsequent language processing), and morphological analysis, including POS tagging and syntactic parsing. Spell checkers, LT-assisted mobile phone input, OCR, and lemmatised fulltext search are hidden parts of the technological background that is already taken for granted, although their quality and accuracy are lacking compared to bigger European languages. In recent years, deep learning language models appeared on the Slovak NLP scene, often adopted from comparable work for other languages (Pikuliak et al. 2021).

Recently, chatbots have noticeably penetrated many areas of human-computer interaction, as the first line of contact in customer support, and although primarily used in English-speaking countries, they are now used in other countries as well, including Slovakia, where chatbots (in written communication mostly) are used by many companies. However, since poorer accuracy of Slovak analysis leads to mixed results and the chatbots are deployed at least partly for public relations reasons, quite often these are just menu-driven FAQs (or an expert system in disguise) camouflaged by an animated head or similar graphical element, without deeper NLP processing.

3 Recommendations and Next Steps

In Slovakia, academic research and industry dealing with NLP and LT function rather separately. The academic sphere often reacts rather slowly to real demands, and instead often explores tasks with little immediate business application; the industry is mostly interested in specific tools and generally does not do NLP-related research, although there are a few companies which are active in applied NLP research.

Since many resources are not reusable due to copyright issues, clarification (i. e., opening) of the licensing of many existing datasets would be helpful for further NLP

[5] http://aranea.juls.savba.sk/aranea_about/

[6] https://korpus.sk/prim-10-0/

development. Many resources remain at the "proof of concept" stage and dedicated effort is needed to bring them up to proper levels of usability. This is also connected with the issue of sustainability of existing resources, many of which were developed as a result of specific research grants, and once the financing stopped, the resources were basically abandoned and no new development is taking place.

The Action Plan for the digital transformation of Slovakia for 2019-2022 (AP 2019) describes a centralised coordinated approach and cooperation between academic and commercial sectors in NLP. It is written only in general terms, without specific steps to be taken; the lack of computational linguists in Slovakia is not addressed (e. g., by promoting university education). The change of government after parliamentary elections in February 2020 and the COVID-19 pandemic have led to the NLP section of the Action Plan not having been acted upon at all.

References

Aldabe, Itziar, Georg Rehm, German Rigau, and Andy Way (2022). *Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (second revision)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-strategic-documents-v3.pdf.

AP (2019). *Action plan for the digital transformation of Slovakia for 2019 – 2022*. https://www.mirri.gov.sk/wp-content/uploads/2019/10/AP-DT-English-Version-FINAL.pdf.

Benko, Vladimír (2014). "Aranea: Yet another family of (comparable) web corpora". In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 247–256.

Garabík, Radovan (2010). "Slovak National Corpus tools and resources". In: *Proceedings of the 5th Workshop on Intelligent and Knowledge oriented Technologies*. Institute of Informatics, Slovak Academy of Sciences, pp. 2–7.

Pikuliak, Matúš, Marián Šimko, and Mária Bieliková (2021). "Cross-lingual learning for text processing: A survey". In: *Expert Systems with Applications* 165, p. 113765. DOI: 10.1016/j.eswa.2020.113765.

Šimková, Mária, Radovan Garabík, Katarína Gajdošová, Michal Laclavík, Slavomír Ondrejovič, Jozef Juhár, Ján Genči, Karol Furdík, Helena Ivoríková, and Jozef Ivanecký (2012). *Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/slovak.

Chapter 34
Language Report Slovenian

Simon Krek

Abstract Around 2.5 million people around the world speak or understand Slovene, with the vast majority of them living in the Republic of Slovenia where it is the official language. The constitution grants the right to use their mother tongue to Italian and Hungarian minorities in certain municipalities. In terms of Language Technology, the Slovene CLARIN.SI consortium plays the key role in the community; all major Slovene institutions involved in the development of LT resources, tools and services are members of the consortium. In contrast, the number of private companies in Slovenia specialising in LT for Slovene remains low, and most of the LT products come either from the (Slovene) academic sphere via national or EU funding, or from the big international IT companies that cover a large number of languages.

1 The Slovenian Language

Slovene is a member of the South Slavic language family and is spoken mainly in Slovenia and the neighbouring areas in Italy, Austria, Hungary and Croatia. In the national census of 2002, the last one that recorded the number of native speakers of different languages, 87.8% of the population – of a total of just under 2 million at the time – declared Slovene to be their mother tongue, with another 3.3% claiming that they use Slovene as the language of their everyday communication at home, which amounts to 91.1% of the population using Slovene as their first language. This number puts Slovenia in the group of EU states with the most homogeneous linguistic situation. Among other linguistic groups, native speakers of languages of the former Yugoslavia were the largest in 2002, with 3.3% of them using a combination of Slovene and their mother tongue for everyday communication, and another 1% using only their mother tongue: Bosnian, Croatian, Serbian or Montenegrin. Other smaller communities included speakers of Albanian, Macedonian and Romani.

Slovene is the official language in the Republic of Slovenia. The constitution grants the right to use their mother tongue to the two minorities declaring that "in

Simon Krek
Jožef Stefan Institute, Slovenia, simon.krek@ijs.si

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_34

those municipalities where Italian or Hungarian national communities reside," Italian or Hungarian are also official languages. In 2002, it was recorded that Hungarian is the mother tongue of 0.4% of the population, and Italian of 0.2%.

According to legislation in Slovenia, all education and teaching provided as part of the current state curriculum, from preschool through to university level, must be in Slovene. In preschool, primary and secondary education, Italian is used in the schools of the Italian minority community, while Hungarian and Slovene are used in bilingual schools where the Hungarian minority is found. Special arrangements exist for children whose mother tongue is not Slovene, for the education of Roma children, children of foreign citizens and children of people without citizenship.

2 Technologies and Resources for Slovenian

A useful place to discover Slovene corpora are the CLARIN.SI NoSketch Engine[1] and KonText[2] concordancers.[3] At the time of writing, there are 76 corpora of varying sizes containing Slovene data in the repository, and 59 corpora in the concordancers. Most of them are available for download under open licences. The more important families of corpora cover general written standard language (Gigafida), Slovene Web and social media (slWaC, Janes), academic discourse (KAS), parliamentary transcriptions (siParl, ParlaMint), Slovene Wikipedia (CLASSLAWiki-sl), historical texts (IMP), literature (MAKS, ELTeC-slv), specialised domains (KoRP, DSI, Konji, etc.), and school essays (Šolar, SBSJ). There are also various manually annotated training and evaluation corpora available (ssj500k, etc.).

The GOS (GOvorjena Slovenščina, Spoken Slovene) family of corpora contains transcriptions of spoken Slovene. The original GOS includes about 120 hours of transcripts from various situations: radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations, etc.

In terms of parallel data, Slovene has benefited from its status as one of the official EU languages since 2004 and is included in the standard multilingual parallel data sets produced either by EU institutions (JRC-Acquis, DGT-Acquis, DCEP, DGT-TM, EAC-TM, ECDC-TM, JRC-Names) or by EU-funded or other projects (INTERA, WIT3, ParaCrawl, CommonCrawl, OpenSubtitles etc.), which are available either from OPUS or from repositories such as ELG. Two TM corpora produced by the Secretariat-General of the Slovene government were made available in the context of the ELRC project and are uploaded in the ELRC-SHARE repository.

There are 82 lexical/conceptual resources with Slovene data in the CLARIN.SI repository available under open access licences. Those that deserve special mention due to their size or importance are: Sloleks – morphological lexicon contain-

[1] https://clarin.si/noske/

[2] https://clarin.si/kontext/corpora/corplist

[3] https://clarin.si/info/about/

ing around 100,000 most frequent Slovene lemmas, their inflected or derivative word forms (2.7M) and the corresponding grammatical description; sloWNet is the Slovene WordNet developed in the expand approach: it contains the complete Princeton WordNet 3.0 and over 70,000 Slovene literals; Dictionary of the Slovenian Normative Guide is a normative orthographic dictionary of Slovene standard language. It contains 140,266 lemmas and sublemmas in 92,617 entries; Thesaurus of Modern Slovene is an automatically created thesaurus from Slovene data available in a comprehensive English–Slovene dictionary, a monolingual dictionary, and a corpus. It contains 105,473 entries and 368,117 synonym pairs.

In terms of language models, the most recent one is the Slovene RoBERTa model. The corpora used for training the model contain 3.47 billion tokens in total. The subword vocabulary contains 32,000 tokens.[4] Multilingual models are also available, e. g., a trilingual BERT model, trained on Croatian, Slovene, and English data.[5]

The standard and most accurate text processing tool for Slovene is the CLASSLA fork of the Stanza pipeline.[6] It supports processing of both standard and non-standard Slovene at the level of tokenisation and sentence segmentation, part-of-speech tagging, lemmatisation, dependency parsing and named entity recognition.

There are some Slovene LT companies that develop speech-to-text and text-to-speech tools.[7] Slovene is also available in speech technology services offered by large enterprises such as Microsoft and Google, as well as by other companies specialising in speech technology.[8] These solutions have also found their way into some specialised devices covering many languages.[9] At the University of Ljubljana, a system has been developed for automatically translating lectures from Slovene to other languages in real time, in the context of the Online Notes project.[10]

Machine translation services for Slovene are available through more or less the same stakeholders: some Slovene LT companies,[11] the large enterprises such as Microsoft and Google, and some other international companies specialising in machine translation technology or general translation services.[12] As an official EU language, Slovene is included in the eTranslation service offered by the European Commission.

The biggest investment in LT for Slovene is the Development of Slovene in Digital Environment project financed by the Slovene Ministry of Culture between 2020-2023.[13] The project will significantly upgrade existing LT resources, tools and services, or produce many of those that do not exist yet. The results of the project are

[4] http://hdl.handle.net/11356/1397

[5] http://hdl.handle.net/11356/1330

[6] https://github.com/clarinsi/classla, https://pypi.org/project/classla/

[7] Amebis, Alpineon: eBralec, https://ebralec.si; Vitasis: Truebar, https://vitasis.si

[8] NEWTON Technologies, https://www.newtontech.net; Sonix: https://sonix.ai

[9] Pocketalk: https://europe.pocketalk.com/languages-countries/

[10] https://www.cjvt.si/en/infrastructure-support/tolmac/

[11] Vitasis: Truebar, https://vitasis.si; Aikwit, https://aikwit.com; Taia, https://taia.io

[12] DeepL Translate, https://www.deepl.com; Pangeanic, https://pangeanic.com/languages/slovenian-translation-services/, etc.

[13] Razvoj slovenščine v digitalnem okolju (RSDO): https://www.slovenscina.eu

expected to be published on the CLARIN.SI and GitHub repositories in November 2022 and February 2023.

3 Recommendations and Next Steps

In general, one can conclude that 1. the support for Slovene is comparable with other languages with a similar status (Krek 2022, 2012), 2. there is a general awareness in governmental bodies that LT for Slovene should be supported in the future, 3. the LT community is growing, also through new educational initiatives such as the MA study of Digital Linguistics (Faculty of Arts, University of Ljubljana), and 4. there is infrastructural support, mainly through the CLARIN.SI infrastructure at the Jožef Stefan Institute, which also covers all other stakeholders through the CLARIN.SI consortium. However, more efforts are needed in the future to bring the existing support closer to those available for other (official EU) languages.

References

Krek, Simon (2012). *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/slovene.

Krek, Simon (2022). *Deliverable D1.31 Report on the Slovenian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-slovenian.pdf.

Chapter 35
Language Report Spanish

Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca Calvo, Mar Rodríguez, and Marta Villegas

Abstract Spanish, one of the most spoken languages in the world, is not threatened by globalisation in the way other languages are and is well-supported by big technological companies, albeit still a long way from English. The number of available language resources (text, and to a lesser extent speech) in Spanish is quite large, but there is still a lack of high-quality, well-curated, annotated resources, available under open-access conditions. Initiatives at the national level, such as the Plan de Impulso de las Tecnologías del Lenguaje, have already started to address this gap.

1 The Spanish Language

The Spanish language, also known as Castilian, is the most spoken Romance language and the fourth most spoken language in the world. Spanish is the official language of Spain, where it originated as an evolution of Vulgar Latin, but most Spanish speakers are in the Americas. It is spoken natively by about 473 million people across 21 countries, where it shares territory with a multitude of languages. Spanish is the third most used language on the internet[1] and this use is steadily growing due to the progressive incorporation of Latin American users. Its growth potential is still very high due to the limited access still seen in some Spanish-speaking countries (the average internet penetration in the Americas is only 67% vs. 92.6% in Spain). Currently, Spanish ranks second on the most popular social networks (Facebook, Instagram, Twitter) and streaming platforms (Netflix, Youtube). Youtube, in particular, has now become one of the main dissemination channels for popular culture in Spanish. It has made consumers of audiovisual products in Spanish much less confined to their geographical area of reference, favouring an unprecedented transfer of linguistic phenomena between the different varieties of Spanish. In contrast, the Spanish Wikipedia ranks only ninth in the number of articles, behind not only some

Maite Melero · Pablo Peñarrubia · David Cabestany · Blanca Calvo · Mar Rodríguez · Marta Villegas
Barcelona Supercomputing Center, Spain, maite.melero@bsc.es, pablo.penarrubia@bsc.es, david.cabestany@bsc.es, blanca.calvo@bsc.es, mar.rodriguez@bsc.es, marta.villegas@bsc.es

[1] https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2021.pdf

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_35

big languages like German and French, but also much smaller ones like Swedish and Dutch. With regard to AI applications that use Spanish, most of the solutions offered by big companies (Google, Amazon, Facebook, Apple, Microsoft) have a Spanish version. Some of them even offer support to dialectal varieties, like Mexican Spanish or peninsular Spanish. Most of these products offer less functionality than their English counterparts, and the quality is lower but keeps improving with each release.

2 Technologies and Resources for Spanish

The Spanish language extends over a very large geographical area and, consequently, many research centres across this area are devoting efforts to developing resources and tools for Spanish, although Spain still leads these efforts. As a global language with hundreds of millions of speakers, the number of unannotated resources (text, and to a lesser extent speech) in Spanish is quite large. However, although good progress has happened since the last survey (Melero et al. 2012), there is still a lack of high-quality, well-curated, annotated and open-access resources.

There are over 20 textual corpora exceeding 100 million words in Spanish, with half of them reaching a billion words, such as the Now Corpus,[2] or the BNE Corpus (Melero et al. 2022). Most of these are automatically cleaned and tagged web corpora, but some come from well-edited sources such as newspapers, scientific journals, collections of published books, or Wikipedia. In some cases, they can be queried but not downloaded, like Codicach[3] or CORPES.[4] Additionally, it should be noted that only half of the Spanish corpora contain linguistic annotations. The most common annotations are morpho-syntactic tags, like part-of-speech and lemma. The number of corpora in Spanish for the different domains varies greatly. Thus, while a sizeable amount of corpora on legal and administrative language can be found, other domains are under-represented. Spanish also appears in many multilingual corpora, together with European languages or with the three other major languages in Spain (Catalan, Basque, Galician). In contrast, there is a lack of parallel corpora with other minority languages of Spain, such as Asturian, Aragonese, Mirandese and Romani, and very few with indigenous languages of the Americas, such as Nahuatl, Guarani, Quechua or Aymara. There is also a lack of bilingual corpora with languages of migrants. As for Spanish Sign Language (LSE), it is estimated that there are more than 100,000 signers of LSE, 20–30% of whom use it as their second language. At least three LSE corpora as well as lexicons and learning resources have been documented.

In the last couple of years, several large language models (LLMs) have been trained for Spanish. RoBERTa-bne and BETO are the most popular BERT-based ones; GPT2-2-bne is the only generative LLM to date.[5] Even though applications

[2] https://www.corpusdelespanol.org/now/

[3] http://sadowsky.cl/codicach.html

[4] https://www.rae.es/banco-de-datos/corpes-xxi

[5] https://github.com/PlanTL-GOB-ES/lm-spanish

based on LLMs tend to be trained end-to-end, limiting the relevance of typical NLP low-level tasks, such as word tokenisation, segmentation, part-of-speech tagging, parsing, etc., those tasks remain important components of many applications. There are a number of toolkits and packages that gather and maintain these tools, like Freeling, SpaCy, UDPipe, LIMA and Connexor, all including Spanish. There are also numerous tools for common end-user tasks in Spanish, such as spellcheckers, grammar-checkers, style-checkers, etc. which can be integrated into most content management systems. Other tools deal with stylometry, plagiarism, information extraction, sentiment analysis, automatic transcription, etc. Spanish is also well served by popular machine translation platforms, such as Google Translate, DeepL or Bing. In addition, Apertium[6] has built downloadable translation models to translate from Spanish into other languages of Spain (Catalan, Basque, Galician), and eTranslation,[7] the EC service provided to public administrations and SMEs, offers neural-based translation between all official European languages, including Spanish. Speech recognition and synthesis are behind some of the most iconic AI applications, such as virtual assistants and dialogue agents. There are close to a hundred speech technology tools documented for Spanish, including text-to-speech (TTS), automatic speech recognition (ASR), and speaker recognition (SR).

Public research centres and universities play an important role in developing language technologies for Spanish. They are responsible for creating and distributing many of the tools and resources mentioned above. In Spain, the Plan de Impulso de las Tecnologías del Lenguaje[8] plays a central role in promoting the development of language resources for Spanish, but also for the other official languages of Spain. The Plan is supported by the Secretary of State for Digitalisation and Artificial Intelligence, and through its collaboration with the Text Mining Unit in the Barcelona Supercomputing Center, it has produced several relevant assets in the biomedical text mining domain, machine translation, and LLMs.[9] Another project, Spanish Language and Artificial Intelligence (LEIA),[10] is also currently underway between the Real Academia Española de la Lengua, the institution entrusted with the stability of the Spanish language, and the big enterprises (Microsoft, Amazon, Google, Twitter, Facebook) with the objective of ensuring high quality coverage of the Spanish language by their AI products. Aside from the big companies in the technology industry, there are many SMEs developing solutions in Spanish. The top services offered include customised chatbots, machine translation systems, speech technologies, spellcheckers and specialised tools for linguistic information extraction and management. Finally, mention should be made of the Spanish Society for Natural Language Processing (SEPLN),[11] a non-profit organisation supported by research groups and the NLP industry, created back in 1983 to promote teaching, research and development

[6] https://www.apertium.org

[7] https://ec.europa.eu/digital-building-blocks/wikis/display/CEFDIGITAL/eTranslation

[8] https://plantl.mineco.gob.es

[9] https://github.com/PlanTL-GOB-ES/lm-spanish

[10] https://www.rae.es/noticia/que-es-leia

[11] http://www.sepln.org/en/sepln

of Spanish NLP, and to organise an annual conference, regularly attended by a number of research groups and companies working in the field.

3 Recommendations and Next Steps

Despite its privileged position as a global language, more effort needs to be devoted for Spanish to realise its full technological potential. Spanish is included in many multilingual projects and is well-supported by large industrial corporations and projects, although the gap in the number and quality of resources and tools compared to English is still quite large. There are many resources documented for Spanish, but there is still a lack of high-quality, well-curated, annotated and open-access resources. Moreover, much more should be done to identify untapped data silos in the public administration, both textual and speech, and facilitate its exploitation, following the European directives on the reuse of public sector information. National initiatives such as the Plan de Impulso de las Tecnologías del Lenguaje need a more sustained effort, capable of 1. filling the gaps in the available resources, 2. ensuring well-regulated access to language data, 3. increasing the innovation capacity of Spanish public services through Language Technologies, 4. promoting research in Spanish NLP and translation technologies and, finally 5. helping bring research solutions to the market, and to the public.

References

Melero, Maite, Toni Badia, and Asunción Moreno (2012). *La lengua española en la era digital – The Spanish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/spanish.

Melero, Maite, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas (2022). *Deliverable D1.32 Report on the Spanish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-spanish.pdf.

Chapter 36
Language Report Swedish

Lars Borin, Rickard Domeij, Jens Edlund, and Markus Forsberg

Abstract Swedish speech and language technology (LT) research goes back over 70 years. This has paid off: there is a national research infrastructure, as well as significant research projects, and Swedish is well-endowed with language resources (LRs) and tools. However, there are gaps that need to be filled, especially high-quality gold-standard LRs required by the most recent deep-learning methods. In the future, we would like to see closer collaborations and communication between the "traditional" LT research community and the burgeoning AI field, the establishment of dedicated academic LT training programmes, and national funding for LT research.

1 The Swedish Language

Swedish is the main language of Sweden and also a constitutional official language of Finland. There are about 10 million native speakers of Swedish, the vast majority of which are Swedish citizens (Parkvall 2019), and an estimated additional 3 million second-language speakers. Swedish is spoken at all levels of government and education in Sweden and to some extent in Finland. Its vitality is strengthened by its closeness to the languages spoken in Norway and Denmark: speakers of Swedish, Norwegian and Danish are able to communicate with relative ease (Haugen and Borin 2018). These languages have around 20 million native speakers in total.

Swedish is written using a modified Latin script with a 29-letter alphabet (the 26-letter Latin alphabet is extended with the vowel characters å, ä, ö). The writing system is in the mid-range of orthographic transparency. It is a relatively normal Germanic (and European) language. Its most "exotic" aspects are found in the domain of phonology, such as: a phonemic pitch accent system; an unusually large

Lars Borin · Markus Forsberg
University of Gothenburg, Sweden, lars.borin@svenska.gu.se, markus.forsberg@svenska.gu.se

Rickard Domeij
Institute of Languages and Folklore, Sweden, rickard.domeij@isof.se

Jens Edlund
KTH Royal Institute of Technology, Sweden, edlund@speech.kth.se

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_36

vowel system, including front rounded vowels (where the high vowels display a notable two degrees of rounding); and rather liberal phonotactics with CCC onsets and CCCC codas. Structurally, Swedish generally follows the patterns typical of Germanic languages, including V2 word order, rich nominal compounding (orthographically written without spaces), and a propensity for forming lexicalised particle (or phrasal) verbs, which appear in speech and text as discontinuous multiword expressions. Among more unusual traits we find a third-person reflexive possessive (i. e., a special possessive form used only if the possessor is co-referential with the subject), and Swedish stands out in relation to its Germanic relatives through the recent introduction (and wide adoption) of a consciously coined gender-neutral third-person singular personal pronoun (*hen*, 'he/she').

Approx. 95% of the Swedish population use the internet at least once a week. In 2020, 86% of Swedish households were connected to 100 Mb or faster fibre optic and 90% of the population used a smartphone. Over the last five years, the .se country top-level domain together with the popular .nu domain have had around 2 million registered domain names. Swedish web pages are overwhelmingly produced in Swedish, often with a parallel English version. The majority of mainstream software such as operating systems, word processors, etc., are localised to Swedish.

2 Technologies and Resources for Swedish

There is a wealth of monolingual text corpora with automatic linguistic annotations available for Swedish, comprising billions of tokens in a variety of genres and text types (Borin et al. 2022, 2012). In contrast, there is a notable lack of gold-standard text corpora, in particular corpora that reflect the present-day language and text genres. Notably, there is currently an ongoing national collaboration with the aim of creating a Swedish natural language understanding benchmark like the English (Super)GLUE,[1] called SuperLim.[2]

There are few publicly available collections of transcribed speech, and there is also a distinct lack of publicly available large multimodal corpora specifically designed or curated for speech technology (ST) and/or LT purposes. Hence, a number of initiatives aim to record and make available speech corpora. Notably, an ASR corpus is being created with 100 speakers recorded in a studio setting, as well as recordings for a male and a female TTS voice, and the Finnish Language Bank is recording Finnish Swedish voices donated by the public. Furthermore, the lack of freely available recordings of real-world speech is an inhibiting factor for ST development beyond relatively simple and controlled applications and domains. While the availability of unannotated audio and video recordings on the internet is greater than ever before, the legality and circumstances under which the use of such data is permissible are unfortunately especially unclear when speech is involved.

[1] https://gluebenchmark.com, https://super.gluebenchmark.com

[2] https://spraakbanken.gu.se/en/resources/superlim

The Sign Language Research Unit at Stockholm University provides access to a Swedish Sign Language (SSL) corpus, with close to 200k annotated tokens.[3]

LR compilation and LT for written Swedish started in the 1960s largely motivated by lexicographic considerations. For this reason, Swedish is well-equipped with high-quality lexical and conceptual resources.[4] A notable lacuna in this context is a Swedish wordnet, which is still pending.

For text processing, grammar-based LT has now largely yielded ground to deep neural machine learning approaches. Drawing on its vast text holdings, the National Library of Sweden has taken a leading role in training large neural language models (LLMs) for Swedish.[5] For Swedish speech processing, several acoustic models for Kaldi and wav2vec are available. Notable Swedish tools for speech include Wavesurfer[6] and the Snack Sound Toolkit.[7]

There is academic research as well as commercial initiatives on several LT component technologies for Swedish, such as tools for text and speech processing, machine translation, computer-aided translation, spoken dialogue systems, language generation and text summarisation, while information retrieval and information extraction for Swedish are primarily being developed by commercial companies, e. g., as parts of proprietary business intelligence and intranet search applications. Notable is the work at Stockholm University on developing LT tools for (transcribed) SSL.

There is no dedicated national LT research funding programme, but several projects have recently been funded. The Wallenberg AI, Autonomous Systems and Software Program supports projects that benefit LT, such as the building of Swedish LLMs and improved ST algorithms. Outside academia there is great interest in LT and language-centric AI from commercial enterprises and public agencies; Sweden has a modest but thriving spectrum of companies offering various LT and AI solutions. Within academia, the research infrastructure Nationella språkbanken[8] 'the Swedish Language Bank' – funded jointly by the Swedish Research Council and ten universities and cultural heritage institutions – collects, develops, manages and distributes LTs and LRs for research, notably including resources and tools for historical stages of Swedish, where we do not expect commercial initiatives to materialise. Nationella språkbanken also coordinates the Swedish membership in CLARIN ERIC.

3 Recommendations and Next Steps

For most of its long history, Swedish academic LT has been pursued by a well-balanced and mutually complementary mix of researchers from computer science

[3] https://www.ling.su.se/teckensprakskorpus, http://sts-korpus.su.se

[4] https://spraakbanken.gu.se/en/research/themes/swedish-framenet-plus-plus

[5] https://huggingface.co/KBLab

[6] https://sourceforge.net/projects/wavesurfer/

[7] https://en.wikipedia.org/wiki/Snack_Sound_Toolkit

[8] https://www.sprakbanken.se

and linguistics (engineering and phonetics in the case of ST). However, recent years have seen a clear shift towards LT researcher teams having a strong or pure computer science background, with an accompanying lack of awareness of many important linguistic aspects of LT research problems.

The Swedish academic LT expertise represents seventy years of accumulated knowledge and experience, which should not be allowed to go to waste. In the short term, the best way of ensuring this is to focus on further LR development for Swedish. Well-designed gold-standard corpora for fine-tuning LLMs and evaluating LT systems require exactly this kind of expertise for their construction, not least in order to avoid pitfalls such as models making undesirable biased predictions that risk perpetuating gender roles or leading to unfair treatment of minority groups. In the medium term, we should aspire to understand current LLMs – which typically come across as black boxes – in order to be able to exploit already existing linguistic knowledge (e. g., information about words collected in a lexical or conceptual resource) when training LLMs, which potentially will reduce their training data requirements, thus putting state-of-the-art LT tools in reach of lower-resourced languages.

This calls for the establishment of closer collaborations and communication between the "traditional" LT research community and the new AI field, e. g., through dedicated LT training opportunities and earmarked funding for LT research.

References

Borin, Lars, Martha D. Brandt, Jens Edlund, Jonas Lindh, and Mikael Parkvall (2012). *Svenska språket i den digitala tidsåldern – The Swedish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/swedish.

Borin, Lars, Rickard Domeij, Jens Edlund, and Markus Forsberg (2022). *Deliverable D1.33 Report on the Swedish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166 ELE. https://european-language-equality.eu/reports/language-report-swedish.pdf.

Haugen, Einar and Lars Borin (2018). "Danish, Norwegian and Swedish". In: *The World's Major Languages*. Ed. by Bernard Comrie. 3rd ed. London: Routledge, pp. 127–150.

Parkvall, Mikael (2019). *Den nya mångfalden*. Stockholm: Makadam.

Chapter 37
Language Report Welsh

Delyth Prys and Gareth Watkins

Abstract In this chapter, based on Prys et al. (2022), an update to the META-NET White Paper (Evas 2014), we present Language Technology (LT) for the Welsh language, providing an overview of the status of Welsh in Wales and a summary of the Welsh writing system and typology. We describe key tools and our recommendations for Welsh LT and associated resource development.

1 The Welsh Language

Welsh is mainly spoken in Wales, together with a small population in Argentina. A minoritised language (Prys 2006), Welsh is considered "vulnerable" (Moseley 2010). Welsh has official status in Wales (National Assembly for Wales 2011). The 2011 census reported that there were 562,000 Welsh speakers in Wales (19% of the population). The Welsh Government aim to almost double that figure by 2050 and recognise that technology is key to this ambition (Welsh Gov. 2017).

The Welsh alphabet contains 29 letters, including eight digraphs (e. g., ch) and the letter j borrowed from English to represent the borrowed /dʒ/ consonant phoneme. V, x and z are not used in Welsh, but are included with the alphabet for computer use as they often appear in named entities such as foreign placenames. Welsh belongs to the insular Celtic branch of Indo-European languages. It is verb initial, following a VSO order. It has consonant mutations at the beginning of words. Accented characters are common over vowels. Welsh has a continuum of other registers, with colloquial or informal registers differing markedly from the standard written form. It has many local dialects, with the main difference between those of north and south Wales. Welsh has two methods of verb formation, utilising concise forms or periphrastic forms, using auxiliary verbs. Guidelines to the latest version of the modern Welsh orthography, first standardised in 1928, were published in 1987 (Prys 2006). In 2021 a new Welsh Orthography Panel was established by the Welsh Government, which aims to resolve minor inconsistencies in the orthography.

Delyth Prys · Gareth Watkins
Bangor University, United Kingdom, d.prys@bangor.ac.uk, g.watkins@bangor.ac.uk

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_37

2 Technologies and Resources for Welsh

According to Cunliffe et al. (2021), "on the Digital Language Vitality Scale […], Welsh is 'Developing', arguably tending towards 'Vital' in some aspects". 90% of the 2019/2020 National Survey for Wales' respondents used the internet (Welsh Gov. 2021). However, English is the dominant online language among Welsh speakers (Welsh Gov. 2015). A lack of language tools for Welsh and inequality or lack of equivalence to English language provision exacerbates the problem.

The major paper dictionaries have been digitised and made available online, and ongoing lexical work now occurs natively in a digital environment. In contrast to traditional descriptive dictionaries, terminology work in Welsh is concept based, held in databases, and published in many formats. These resources have been re-used in lexicons for various purposes, including spelling and grammar checkers.

Monolingual, bilingual and multilingual text corpora, as well as speech corpora, mainly in the standard or neutral language register, have been curated. The Language Technologies Unit at Bangor University holds the largest collection of corpora, at over 700 million tokens, including the Cysill Ar-lein Monitor Corpus (Prys et al. 2016). The CorCenCC (Knight et al. 2020) corpus is the largest annotated, balanced general corpus to date, with 11 million tokens. Crowdsourcing has been successfully used to gather large speech corpora of recorded prompts, currently using Mozilla Common Voice. Recordings of voice talents, collected specifically for building synthetic voices, have been released under the CC0 licence. Intellectual Property and licensing issues are of utmost concern when assessing the suitability of these corpora for use and reuse and can hamper their open distribution.

In terms of speech technology, a Welsh personal assistant (Jones 2020) has been developed as has the first Welsh speech-to-text transciber. Synthetic voices have been created for Welsh using older diphone technology, with newer, more natural sounding unit selection voices becoming available under open licences. A voice banking initiative, Lleisiwr, a joint venture between Bangor University and NHS Wales, has been created for bilingual Welsh/English speakers about to lose their speech capabilities, and is one of the most innovative services established to date.

Acoustic and language models for Welsh are being developed. Some of these are part of multilingual sets, which are of variable quality compared to those developed specifically for Welsh. A Welsh part-of-speech tagging model has been developed for spaCy, unlocking the potential to perform many other NLP tasks on Welsh texts. Welsh has NLP tools for text analysis, anonymisation, and information extraction.

In terms of translation, a commercial Welsh–English translation system exists and MT for Welsh is offered by some major companies such as Google and Microsoft. Moses has been used to develop SMT for Welsh. Newer neural net engines are being used, and the first domain-specific MT engine for health launched. Welsh/English translation memories can be shared on the Open Translation Memories site, emulating the ELRI project. An overview of these LT tools and resources may be found on the Welsh National Language Technologies Portal (Prys and Jones 2018).

While the UK LT industry is mostly focused on the English language, Welsh language LT provision is mainly driven forward by the higher education sector. Wales

has vibrant creative technology, media and translation sectors which make use of the government-funded open source LT created by universities. The main hub for LT research in Wales is Bangor University, notably its Language Technologies Unit. Relevant research is also undertaken at the universities of Cardiff, Swansea and South Wales. Efforts have also been made to improve teaching digital technologies in schools and universities. The current Welsh Government's Welsh language strategy states that "We must ensure that high-quality Welsh language technology becomes available […] to support education, workplaces and social use of Welsh" (Welsh Gov. 2017). This was further elaborated in the Government's Welsh Language Technology Action Plan (Welsh Gov. 2018). After years of small-scale and fragmented initiatives, the publication of this plan provides a coherent, planned way forward for the development of Welsh LT resources, tools and services.

3 Recommendations and Next Steps

There has been much progress in Welsh LT in recent years, but further work needs to be done if the Welsh language is to thrive in the digital world. While FAQ generation is used for the Welsh language, the development of more sophisticated chatbot systems would further benefit Welsh speakers. There is no published research on Welsh language knowledge graphs, nor what they have to offer to Welsh. Limited research has been conducted on Welsh language sentiment analysis. A key new area for development is bilingual models to aid minoritised languages where users constantly have to switch between their own language and the majority language or code-switch within the minoritised language. Promising work has been done for Welsh in developing a bilingual model for text-to-speech. Similar work for speech recognition is underway, where pre-trained multilingual acoustic models can provide useful crosslingual speech representations that can be fine-tuned for effective bilingual Welsh and English speech recognition. There are many other bilingual situations where a similar approach could be explored.

In order to fill these gaps Welsh needs to be able to join in large-scale multinational and multilingual research and development programmes of the type previously reserved for official EU languages. Also, in common with other minoritised languages, Welsh needs a space within the European community where special attention can be paid to up-resourcing these languages and up-skilling their communities. Minoritised European languages often also belong to the economic periphery in Europe, and using LT for economic regeneration in those areas would have a positive effect on their economic, social and linguistic well-being.

It is often more attractive to court new and exciting project ideas. Funding opportunities are often prejudiced in favour of such ventures, but attention also needs to be paid to maintaining, improving, consolidating and further developing existing tools and resources. At the same time minoritised languages need to take full advantage of any emerging innovations, playing their full part in the LT developments for Europe.

References

Cunliffe, Daniel, Andreas Vlachidis, Daniel Williams, and Douglas Tudhope (2021). "Natural language processing for under-resourced languages: Developing a Welsh natural language toolkit". In: *Computer Speech & Language* 72.

Evas, Jeremy (2014). *Y Gymraeg yn yr Oes Ddigidol – The Welsh Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/welsh.

Jones, Dewi Bryn (2020). "Macsen: A Voice Assistant for Speakers of a Lesser Resourced Language". In: *Proceedings of the 1st SLTU-CCURL workshop*. Marseille, France: European Language Resources Association (ELRA).

Knight, Dawn, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, and Enlli Môn Thomas (2020). *The National Corpus of Contemporary Welsh: Project Report; Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect*. https://corcencc.org/wp-content/uploads/2020/06/CorCenCC-report_2020_en.pdf.

Moseley, Christopher (2010). *Atlas of the World's Languages in Danger*. Paris: UNESCO.

National Assembly for Wales (2011). *Welsh Language (Wales) Measure 2011*. https://www.legislation.gov.uk/mwa/2011/1/contents/enacted.

Prys, Delyth (2006). "Setting the Standards: Ten Years of Welsh Terminology Work". In: *Terminology, Computing and Translation*. Ed. by Pius ten Hacken. Tübingen: Narr.

Prys, Delyth and Dewi Bryn Jones (2018). "National Language Technologies Portals for LRLs: A Case Study". In: *Human Language Technology. Challenges for Computer Science and Linguistics*. Cham: Springer.

Prys, Delyth, Gruffudd Prys, and Dewi Bryn Jones (2016). "Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled from an On-line Spelling and Grammar Checker". In: *Proceedings of LREC 2016*. Portorož, Slovenia: European Language Resources Association (ELRA).

Prys, Delyth, Gareth Watkins, and Stefano Ghazzali (2022). *Deliverable D1.34 Report on the Welsh Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-welsh.pdf.

Welsh Gov. (2015). *Welsh language use in Wales, 2013–15*. https://www.gov.wales/sites/default/files/statistics-and-research/2018-12/160301-welsh-language-use-in-wales-2013-15-en.pdf.

Welsh Gov. (2017). *Cymraeg 2050: A million Welsh speakers*. https://www.gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf.

Welsh Gov. (2018). *Welsh language technology action plan*. https://www.gov.wales/sites/default/files/publications/2018-12/welsh-language-technology-and-digital-media-action-plan.pdf.

Welsh Gov. (2021). *Internet skills and online public sector services (National Survey for Wales): April 2019 to March 2020*. https://www.gov.wales/internet-skills-and-online-public-sector-services-national-survey-wales-april-2019-march-2020-html.

Part II
European Language Equality: The Future Situation in 2030 and beyond

Chapter 38
Consulting the Community: How to Reach Digital Language Equality in Europe by 2030?

Jan Hajič, Maria Giagkou, Stelios Piperidis, Georg Rehm, and Natalia Resende

Abstract This chapter describes the community consultation process carried out in the European Language Equality (ELE) project concerning the future situation in 2030. Due to its central status for the future-looking activities within the project, this chapter introduces the second part of the present book. We gathered, analysed and structured the views, visions, demands, needs and gaps of European Language Technology (LT) developers, both industry and academia, and European LT users and consumers. Additionally, based on these collected findings and other evidence, we attempted to derive a thorough description of the steps to take to reach Digital Language Equality (DLE) in Europe by the year 2030 and, moreover, what the field of LT will look like in Europe in about ten years from now.[1]

1 Introduction

The goal of WP2, "European Language Equality – The Future Situation in 2030" of the European Language Equality (ELE) project was the collection of a vast amount of input for the Strategic Research, Innovation and Implementation Agenda (SRIA) and Roadmap and the production of several reports by a broad and diverse spectrum of stakeholders – from research through industry to users – about their views, visions, demands, needs and gaps related to LT, language-centric AI and DLE, while at the same time anticipating the expected developments over the next ten years. The activities in the project put a special focus upon ways and means of achieving DLE by 2030

Jan Hajič
Charles University, Czech Republic, hajic@ufal.mff.cuni.cz

Maria Giagkou · Stelios Piperidis
R. C. "Athena", Greece, mgiagkou@athenarc.gr, spip@athenarc.gr

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

Natalia Resende
Dublin City University, ADAPT Centre, Ireland, natalia.resende@adaptcentre.ie

[1] This chapter is an abridged version of Hajič et al. (2021).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_38

through the development, implementation and use of LT, in order to make Europeans of all regions and origins truly equal when accessing and interacting with education, business, governments and public services in their own language. A large part of the information eventually integrated into the SRIA was collected through carefully designed surveys distributed to researchers, developers, innovators and users and their communities as well as through reports produced by a number of ELE consortium partners. This chapter describes the overall methodology of the community consultation approach applied in the project and the various reports produced. The collected findings, presented in the subsequent chapters, have been used as input for the development of the SRIA (see especially Chapter 45 and the other chapters of Part II of the present book).

Section 2 provides a description of the overall methodology. The following two sections specify how the consortium conducted consultations with the European LT developers (Section 3), European LT users (Section 4) and European citizens (Section 5). Section 6 describes the preparation process of the four technology deep dives (included in this book in Chapters 40 to 43). Section 7 explains the instruments used for the collection of additional input and feedback. Section 8 concludes the chapter.

2 Methodology

Our primary objective in the ELE project was the preparation of the *Strategic Research, Innovation and Implementation Agenda and Roadmap for achieving full DLE in Europe by 2030* (see Chapter 45). Since the overarching goal of achieving DLE involved a large number of stakeholders, the process of preparing, discussing and finalising the different parts of the strategic agenda and roadmap was carried out by all 52 partners of the consortium and the wider European LT community, which we involved via the consortium partners' networks and connections.

The project made use of the support of the consolidated European LT research and industry community – brought together through previous projects such as META-NET and CRACKER – and produced a convincing, sustainable and evidence-based agenda and roadmap. Only with the input and feedback from experts working in different areas of our core field of Computational Linguistics and LT and also on the borders to other fields such as, among others, Cognitive Science, AI, Machine Learning, Data Science and Knowledge Technologies, could the agenda and roadmap be prepared in a way that was goal-oriented, all-encompassing, realistic, supported and overall meaningful. Only with the inclusion of representatives from various different companies active in the field did the involvement of industry make sense in the grander scheme of things, especially regarding the inclusion of their needs and goals. The same holds for the non-industrial, but important stakeholders as users and consumers, in areas such as Digital Humanities/Social Science and Humanities (DH/SSH) research, policymaking, normative language policy (including minority ones), education and others.

At the most abstract level, our main approach was twofold: we distinguished between input for the agenda and roadmap generated within the consortium, and input generated by organisations not participating as partners in the ELE project (through surveys, interviews, external consultation meetings, etc.). When putting the consortium together, we opted for a large number of partners that cover many relevant areas that needed to be taken into account for the development of the strategic agenda. The consortium-internal and consortium-external stakeholders' input and feedback was systematically collected, structured and included in the agenda and roadmap development process, resulting in an all-encompassing, coherent and convincing strategic roadmap with agreed-upon research questions and research goals, realistic timing, and a meaningful plan.

To come up with suggestions and recommendations on how to achieve full DLE in Europe by 2030, we distinguished between two main stakeholder groups: 1. LT developers (industry and academia) and 2. LT users and consumers. Both groups were represented in ELE with several networks, initiatives and associations that produced one report each, together with their respective constituencies, highlighting their own individual views, needs, wishes, demands and contributions towards DLE. The industry partners of the ELE consortium generated, in various tandem groups, four technology deep dives to provide, similarly, the views, needs, wishes, demands and contributions of the European LT industry, structured into 1. Machine Translation (see Chapter 40), 2. Speech (see Chapter 41), 3. Text Analytics (see Chapter 42) and 4. Data and Knowledge (see Chapter 43). We also carried out additional surveys and consultation meetings as well as interviews with stakeholders who were not represented in the consortium.

The methodology applied was based on a number of stakeholder-specific surveys (inspired by Rehm and Hegele 2018) as well as collaborative document preparation that also involved technology forecasting. Both approaches were complemented with the collection of additional input and feedback through various online channels (see Figure 3 in Chapter 1 on page 7). As Table 1 illustrates, the two main targeted stakeholder groups differ in one substantial way: while the group of commercial or academic LT developers was, in a certain way, *closed* and well represented through relevant organisations, networks and initiatives in the ELE consortium, the group of LT users is an *open* set of stakeholders that was only partially represented through relevant organisations, networks and initiatives in the consortium. Both stakeholder groups were addressed with targeted and stakeholder-specific surveys that were distributed to the relevant stakeholders through the responsible ELE partners. In addition, we communicated with additional stakeholders, primarily through interviews.

3 The Perspective of European Language Technology Developers

One mission-critical aspect when it came to consulting the community was the collection of views, demands, needs, ideas and visions with regard to the wider topic of DLE from the community of European LT developers and also the highly diverse

Stakeholder Group
Task 2.1 The perspective of European LT developers (industry and research)
European LT developers (industry and academia): *Closed set* that is well represented through relevant organisations, networks and initiatives in the ELE consortium ⇒ Instruments: Surveys, interviews ⇒ Approach further detailed in Section 3 of this chapter ⇒ Results reported in Thönnissen (2022), Eskevich and Jong (2022), Rufener and Wacker (2022), Hajič et al. (2022), Hegele et al. (2022)
Task 2.2 The perspective of European LT users and consumers
All potential European LT users: *Open set* that is only partially represented through relevant organisations, networks and initiatives in the ELE consortium ⇒ Instruments: Surveys, interviews ⇒ Approach further detailed in Section 4 of this chapter ⇒ Results reported in Gísladóttir (2022), Kirchmeier (2022), Hicks (2022), Blake (2022), Hrasnica (2022), Heuschkel (2022)
Task 2.3 Science – Technology – Society: Language Technology in 2030
Prominent companies of the European LT developer landscape, all represented in the ELE consortium: *Closed set* ⇒ Instrument: Collaboratively created technology deep dives ⇒ Approach further detailed in Section 6 of this chapter ⇒ Results reported in deliverables Bērziņš et al. (2022), Backfried et al. (2022), Gomez-Perez et al. (2022), Kaltenböck et al. (2022)

Table 1 Stakeholder groups and instruments relevant for the three tasks in WP2

group of European LT users. This section describes the process for engaging with LT developers (supply side) while Section 4 describes how we collaborated with LT users (demand side) with regard to their visions for 2030; as such, these sections are follow-ups that cover the forward-looking projections of the same stakeholder groups whose views as of 2022 are presented in Chapter 4 (Section 3, p. 84 ff.).

We analysed the views of European LT developers and providers, i. e., representatives both from industry and academia to investigate their ideas, demands, visions and predictions with regard to DLE going towards 2030. We explored the factors that drive their development plans and investments (e. g., market demand, number of speakers, available funds etc.) and the perceived obstacles that should be overcome to achieve DLE. The main instrument for collecting the LT developers' views was a set of surveys, which were distributed through the established research and industry networks of the ELE consortium to their members. In addition, the survey was forwarded to other pan-European initiatives, thus covering the widest possible range from generic AI to media- and language-related infrastructures. The data collection

activity was supplemented by focused meetings and interviews with targeted informants which were selected based on either the quality of their input to the survey or their prominence in and impact on the European LT landscape. The collected feedback of the European LT developers was augmented with additional input produced by the networks, analysed and consolidated in five reports (see Table 1).

3.1 Stakeholders

The European LT developers are a diverse group of stakeholders, comprising *academic and industrial researchers* in the field of LT/NLP. In addition to conducting research, the members of this group also develop pre-commercial prototypes, algorithms, applications and systems. They can also be *innovators and entrepreneurs* who productise and commercialise LTs to address, among others, the needs for digital content analysis and generation as well as for pertinent content transformation and dissemination. An initial grouping is, thus, *LT research (academia)* and *LT industry* (also see Chapter 4).

Europe has a long-standing tradition in LT with over 800 centres (Rehm et al. 2023a, 2020) performing excellent, highly visible and internationally recognised research on almost all European and also many non-European languages. The European LT industry has been estimated to comprise 473 LT vendors in the EU26 plus Iceland and Norway in 2017 (Vasiljevs et al. 2019). The ELG catalogue comprises more than 800 commercial entities, also including integrators and a certain number of user companies (Rehm et al. 2021, 2023a). While LT is at the intersection of Linguistics and Computational Linguistics, Computer Science and Artificial Intelligence, we also take relevant neighbouring fields into account, especially Digital Humanities/Social Science and Humanities (DH/SSH).

With the aim of informing the ELE SRIA with the opinions, views and demands of the widest possible group of these stakeholders, we mobilised existing European networks, associations, initiatives and projects. Some of the well-established and long-standing pan-European LT networks were represented in the ELE consortium (Table 2). The ELE partners that represented these initiatives contributed their views to the project and also facilitated access to and elicitation of the views of their constituency and members with regard to how DLE can be achieved by 2030. They coordinated the distribution of a questionnaire to their members, conducted interviews and focused consultation meetings, where needed and appropriate (see Section 3.2 and Table 1).

While these stakeholders already represented a significant part of the European LT community, we engaged additional initiatives in the consultation process (see Hajič et al. 2021, for further details).

Initiative	Description	Stakeholder Group
META-NET	The META-NET Network of Excellence consists of 60 research centres in 34 European countries. It develops the technical foundations of a multilingual, inclusive and innovative European society, supporting all European languages.	European LT community (especially research)
ELG	The European Language Grid (ELG) project developed a cloud platform and marketplace for the whole European LT community. The shared platform includes language resources, datasets and services to benefit European society and industry. It addresses the fragmentation of the European LT landscape.	European LT community
LT-Innovate	LT-Innovate is the European LT industry association with more than 200 members. It supports its members by promoting the industry as a whole in the most promising target markets.	European LT industry
CLARIN	The European Research Infrastructure for Language Resources and Technology consists of more than 20 national consortia, which themselves consist of multiple partners. CLARIN makes language resources available to researchers and students from all disciplines, especially in the humanities and social sciences, through single sign-on access.	European DH, NLP, SSH community
CLAIRE	The Confederation of Laboratories for AI Research in Europe has 394 members in 36 countries. CLAIRE seeks to strengthen European excellence in AI research and innovation across all of AI, for all of Europe, with a human-centred focus. It is now supported by nine EU Member State governments.	European AI community

Table 2 LT developer communities represented in the ELE consortium who shared their views in dedicated reports

3.2 Instruments

To collect and analyse the LT developers' views, demands, visions and predictions, we adopted an inclusive and participatory approach, through which every voice was enabled to find its way into the SRIA. We reached out to as many representatives of the LT community as possible and elicited their educated views in a structured, yet flexible, way. Two main instruments were used to collect the views of the European LT developers: surveys (Section 3.2.1) as well as interviews and focused consultation meetings (Section 3.2.2).

3.2.1 Survey

The LT developer survey attempted to elicit views in a structured way that lent itself to the efficient analysis, consolidation and integration of the feedback in the respective project reports, which, in turn, were fed into the SRIA (Chapter 45). Driven by the envisaged topics that the final SRIA intended to cover, the survey encompassed closed and open-ended questions to inquire about the LT developers' future predictions and visions. The overall structure of this online survey is described in Chapter 4

(Section 3, p. 84 ff.), and the forward-looking questions, in particular, were gathered in a specific part, as follows:

- **Predictions and visions for the future**: This part of the stakeholders survey was forward-looking and investigated ideas, predictions and wishes of the LT community about how the LT field as a whole will be able to equally support all European languages by 2030, i. e.,
  - policies or instruments that could contribute to speeding up the effective deployment of LT in Europe equally for all languages;
  - prediction of future opportunities for LT in basic and applied research (scientific vision) and in innovation and industry;
  - expectations with regard to the challenges a large-scale, long-term ELE programme can address by 2030.

3.2.2 Interviews and focused consultation meetings

To supplement the survey responses and to collect more detailed feedback, where appropriate, we conducted interviews and consultation meetings with targeted informants who were selected based on either the quality of their input to the survey or their prominence in and impact on the European LT landscape. Operationally, the selection of stakeholders to be interviewed was based on the following criteria.

1. The respondent had partially filled in the survey and some essential input was missing in order to have a more complete understanding of his/her views; or
2. No member of a network or association (see Section 3.1) had filled in the survey.

In the first case, we asked for a short and focused meeting with the respondent to elicit the missing information. In the second case, when a network or association that was considered a stakeholder for ELE was not represented, we identified key persons and conducted an interview. The key details of the respondents are described in Chapter 4 (Section 3, p. 84 ff.), while the results and findings of the survey and consultations with LT developers concerning the future situation in 2030 are discussed in Chapter 39 (Section 2, p. 246 ff.); their views have been taken on board in the ELE SRIA (Chapter 45).

4 The Perspective of European Language Technology Users

This section describes our approach to gathering the voices of the highly heterogeneous and diverse group of European LT users and consumers as the final "beneficiaries" of LT with regard to the necessary and desired developments supporting DLE for all European languages by 2030. This activity required engagement with individuals, representative public bodies and government units, organisations and

businesses, including SMEs as well as larger companies, that use LT. We also explored the factors that can promote language equality in the users' and consumers' view, especially with regard to encouraging the uptake of missing or poor LTs that can solve real communication problems for the members of all European language communities. Special attention was paid to the speakers of lesser-served languages, particularly those that face digital extinction or neglect, eliciting from the LT users of such language communities indications of necessary or desirable developments that are expected to put their own languages on an equal footing with the dominant ones by 2030. A complementary focus considered the perceived obstacles that hinder full DLE, so that effective remedial action can be promptly taken. We followed the same approach as for the supply side (Section 3), i. e., based on surveys and structured templates several reports have been produced by the ELE consortium members who represented relevant stakeholder groups.

4.1 Stakeholders

LT users and consumers comprise a broad group of stakeholders from a wide variety of domains and sectors. We reached out to representatives from public administration (public bodies and government units), organisations and businesses, including SMEs as well as larger companies, that currently use and benefit from LT, as well as individuals. Six stakeholders are represented in the ELE consortium with a special focus on speakers of lesser-served languages, particularly those that face digital extinction or neglect (see Table 3).

In addition to the reports produced by these six core representative bodies and ELE partners, other relevant external stakeholders were consulted as well. The inclusion of additional groups ensured the widest possible coverage and promoted our inclusive approach to build a comprehensive, accurate and all-encompassing SRIA and roadmap towards achieving full DLE in Europe by 2030 (presented in Chapter 45).

4.2 Instruments

In a similar way as described in Section 3.2 for the stakeholder class of LT developers, surveys and focused consultation meetings were used to collect and analyse the perspective of European LT users, i. e., their views, ideas, demands, future visions and predictions with regard to DLE. Our goal was to consult with as many representatives of this stakeholder class as possible to collect their opinions in a structured, yet unconstrained, way.

Initiative	Description	Stakeholder Group
ECSPM	The European Civil Society Platform for Multilingualism is an alliance for the languages spoken in Europe (national/official, minority, regional and autochthonous, as well as the languages of immigrant communities). It includes networks of more than 200 European associations, societies and organisations that view multilingualism as an asset for European economic, social and cultural development, as well as a facilitator for intellectual and personal growth. It is a fervent voice of Europe's civil society promoting languages, language policies and research on multilingualism.	European Platform for Multilingualism
EFNIL	The European Federation of National Institutions of Language is a pan-European organisation that was founded in 2003. EFNIL has 41 members from 27 countries and provides a forum for these institutions to exchange information about their work and to gather and publish information about language use and language policy within the EU.	European National Languages
ELEN	The European Language Equality Network is an international NGO for the protection and promotion of European lesser-used languages gathering 166 member organisations representing 46 languages in 23 European states. Founded in 2012, it represents the voice of grassroots European RML civil society.	European Regional, Minority and Endangered Languages
LIBER	The Association of European Research Libraries is Europe's principle association of research libraries, consisting of nearly 450 national, university and other libraries from more than 40 countries. LIBER helps European research libraries to ensure the preservation of European cultural heritage, to improve access to collections, and to provide more efficient information services. Enabling Open Science is a major priority, as is promoting innovative scholarly communication, fostering digital skills and services, and engaging with world-class e-infrastructures.	European Research Libraries
NEM	New European Media is the leading European Network for Media and Creative Industries with the mission to foster the impact of interactive technologies on the future of new media through interaction between media, content, creative industries, social media, broadcasting and telecom sectors as well as consumer electronics, represented by more than 1,000 members. The application of the newest technologies in respect to equal access to media for all is one of its higher priorities.	European New Media Community
Wikipedia	Wikimedia Deutschland is an independent, charitable membership-based non-profit organisation that serves as the German chapter of the global Wikimedia movement. With more than 140 employees it is the oldest and largest of about 40 independent chapters.	European Free Knowledge Community

Table 3 LT users and consumers represented in the ELE consortium who shared their views in dedicated reports

4.2.1 Survey

Similarly to the survey for LT developers, feedback from the LT users and consumers was collected in a structured way that lends itself to the efficient analysis, consolidation and integration of the feedback into the ELE SRIA (see Table 1). Driven by the envisaged topics that the final strategic agenda would cover, this survey encompassed closed and open-ended questions to understand the LT users' and consumers' future predictions and visions with regard to DLE in Europe. The survey had four parts and encompassed 63 questions in total. Some of the questions depended on previous answers. As a result, a respondent was presented with 30 (minimum) to 63 (maximum) questions, including the "if other" questions. If presented the maximum set of questions, 46 questions were mandatory, and 33 of them were closed (single or multiple choice). In particular, beyond the preliminary sections covering demographic information and the language(s) for which the respondents used LRTs, the last part of the questionnaire is of interest here, as it focused on the forward-looking opinions of the LT users going towards full DLE in Europe by 2030:

- **Predictions and visions for the future**: This part of the online survey for LT users investigated ideas, predictions and wishes about how DLE can be achieved in Europe by 2030.
  - policies or instruments that could contribute to speeding up the effective deployment of LTs in Europe equally for all languages;
  - expectations with regard to the challenges that a large-scale long-term ELE programme can address by 2030.

The survey was circulated through the networks and associations described in Section 4.1 (also see Table 3) and through additional channels (see Section 7). It was set up as an online form for easy distribution as well as analysis of responses.

4.2.2 Interviews and focused consultation meetings

To complement the survey responses of the six LT user and consumer stakeholder groups represented in the ELE consortium, we conducted consultation meetings with targeted informants. The approach was similar to the one described in Section 3.2.2.

5 The Perspective of Europe's Citizens

In addition to the consultation with the more focused stakeholder groups (Sections 3 and 4), a large-scale, online and multilingual survey targeting Europe's citizens was carried out with the aim of taking into account their opinions, individual needs, wishes and general demands as well as to make sure that their voices play a decisive role in the pursuit of full DLE. This consultation with a larger and more diverse cohort of LT consumers allowed us to obtain an accurate picture of the current scenario

in terms of LT support across European languages and have a more representative basis for a technological and scientific forecasting on how LTs can be deployed and applied in Europe by 2030 to the benefit of all European citizens.

Different survey platforms were tested to choose the most suitable one for our needs. After setting up the survey in the platform of our choice, it was disseminated in 28 European countries and in 38 European languages from January 2022 to 01 May 2022. The survey included a total of 11 questions, four of which were single-choice questions, six were multiple-choice and one open-ended question which allowed respondents to include any comments or feedback they had. These 11 questions could be answered in about five minutes via computers or mobile devices. More details concerning the translation of the online survey into several languages and its careful and well-balanced distribution are given in Chapter 4 (Section 3, p. 84 ff.).

After a few initial survey items that aimed at understanding the level of familiarity of respondents with terms from the field of LTs, the respondents' profiles and language backgrounds were checked through a multiple-choice question that asked them to select the terms (e. g., "Information Retrieval", "Natural Language Processing", "Natural Language Understanding") that they were familiar with or could immediately recognise. The questions of particular interest here were the final two about the future of LTs in Europe, which "requested respondents to indicate the tools they would like to use in the future if not currently available in their languages and also to rate the top three advantages of improving LTs for all languages".

6 Predicting Language Technology in 2030: Technology Deep Dives

The ELE project also attempted to assess and predict, in a dedicated forward-looking task, what the field of LT will look like in 2030. To this end, we collected, analysed and consolidated the views of European LT industrial and academic stakeholders on anticipated future technological progress, innovations and impact on society in the coming decade, with a special emphasis on technologies, resources, approaches, coverage and performance needed to achieve DLE by 2030.

The task was set up to seek agreement among these stakeholders in terms of pinpointing novel or significantly extended or adapted technologies that would ultimately enable or contribute to DLE, and consequently help bring about true digital equality in European society. To achieve these goals, such new technologies would have to take into account the state-of-the-art in various LT and AI areas, including the reasons why current technologies do not perform equally well for all languages (e. g., due to lack of data, poor-quality data, language properties, knowledge collectively and indirectly acquired for only some languages in the past, etc.) as well as the reasons for biased results in some areas. Focusing on possible methods, technologies and processes for bringing all European languages on par both technologically and in consumer applications, there was a unifying theme, namely, to discover and explore ways to convert the unique challenges of a diverse European multilingual

society into opportunities and technologies, processes and services superior to those developed in the context of largely homogeneous linguistic societies.

We also took a fresh look at deployment, i. e., how LTs would be made available to the different stakeholders and end-users, from machines to household appliances to mobile devices and perhaps even "invisible" devices. To achieve these goals, structured document templates and also surveys oriented towards technological development and technology forecasting along the aforementioned lines (see Sections 3 and 4, respectively) were used by both industrial and academic stakeholders, and then assembled into four project reports, reflecting the major technology areas (Machine Translation, Speech Technologies, Text Analytics, Data and Knowledge).

Four ELE partners were selected to lead the development of these technology deep dives, which are presented in abridged form in Chapter 40 (p. 263 ff.) on Machine Translation, Chapter 41 (p. 289 ff.) on Speech Technologies, Chapter 42 (p. 313 ff.) on Text Analysis, and Chapter 43 (p. 337 ff.) on Data, Knowledge and Language Resources. They collaborated closely with other ELE partners who also work in the respective fields. The four authoring teams made use of existing scientific publications, reports and foresight studies as well as science and technology predictions. In this way, the respective groups of experts developed a consolidated opinion with regard to the direction in which the relevant field is moving or should be moving, what the current gaps and roadblocks as well as the industry's needs from research are, and what they can contribute to DLE.[2]

7 Collecting Additional Input and Feedback

Complementing the instruments described above, we set up additional ways of collecting input for the emerging SRIA. We wanted to enable all stakeholders to communicate with ELE easily so that their opinions and ideas could be integrated into our recommendations. Over several months throughout 2022, the emerging ELE results were disseminated through various channels (e. g., website, publications, presentations, social media, etc.), and we solicited input by actively asking stakeholders for feedback, or by actively listening, especially on social media, to identify additional opinions regarding our topic (see Rehm et al. 2023b).

7.1 Conferences and Workshops

ELE results were presented and discussed at many different conferences and workshops. One example was the presentation of the pre-final ELE recommendations at META-FORUM 2022 in June 2022, which resulted in a valuable discussion with

[2] This approach was inspired by the methodology followed in META-NET, in which "vision groups" worked on similar documents ("vision papers", see, for example, Rehm and Uszkoreit 2013).

the audience in terms of, among others, additional aspects to take into account.[3] The final recommendations were presented at the STOA workshop "Towards full digital language equality in a multilingual European Union" held at the European Parliament in November 2022.[4]

7.2 Project Website

An interactive contact form was implemented on the ELE website through which interested stakeholders could – and still can – communicate with the ELE team.[5] We also distributed all reports through the website to enable others to provide feedback.[6]

7.3 Social Media

Social media activities in ELE concentrated on LinkedIn and Twitter. We used LinkedIn[7] to address professional stakeholders including LT developers and users. In contrast, while Twitter[8] was primarily used for reaching out to European citizens, it was also used by many stakeholders for professional communication purposes. The social media activities of the ELE project were planned and executed in close collaboration with the ELG project. To be able to disseminate news about both activities through these joint channels, we subsumed the two initiatives under the title "European Language Technology" (ELT, for more details see Rehm et al. 2023b), and a biweekly newsletter with updates and highlights from the ELE SRIA was circulated to a large and diverse audience of around 4000 recipients, also inviting input and feedback.[9]

8 Summary and Conclusions

This chapter describes the consultation process carried out under the umbrella of WP2, "European Language Equality – The Future Situation in 2030", in the ELE project. It is meant to be a brief summary that illustrates the guidelines as well as instructions specified with regard to the implementation of our internal processes

[3] https://www.european-language-grid.eu/events/meta-forum-2022

[4] https://www.europarl.europa.eu/stoa/en/events/details/towards-full-digital-language-equality-i/20220711WKS04301

[5] https://european-language-equality.eu/contact/

[6] https://european-language-equality.eu/deliverables/

[7] https://www.linkedin.com/company/european-language-technology/

[8] https://twitter.com/EuroLangTech

[9] https://www.european-language-technology.eu

and instruments applied by all actively involved partners, especially with regard to reaching out to and gathering feedback and input from European LT developers and European LT users and consumers, but also with regard to technology forecasting through the four technological deep dives. These activities had an important role within the ELE project: they defined all aspects of the future situation with regard to DLE by 2030. Due to this important, mission-critical role in the project, all involved stakeholders were made aware of the different aspects and dimensions the project needed to provide input for when it came to assembling the final recommendations for the SRIA.

The main findings of the consultation process briefly summarised in the present chapter are presented in the subsequent chapters. Chapter 39 presents the results of the different surveys. Abridged versions of the four technology deep dives are presented in Chapters 40 (Machine Translation), 41 (Speech Translation), 42 (Text Analytics) and 43 (Data and Knowledge Technologies). Finally, a compact but comprehensive summary of the ELE SRIA and Roadmap is presented in Chapter 45.

References

Backfried, Gerhard, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz (2022). *Deliverable D2.14 Technology Deep Dive – Speech Technologies*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/speech-deep-dive.pdf.

Bērziņš, Aivars, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, and Andy Way (2022). *Deliverable D2.13 Technology Deep Dive – Machine Translation*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/MT-deep-dive.pdf.

Blake, Oliver (2022). *Deliverable D2.10 Report from LIBER*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LIBER.pdf.

Eskevich, Maria and Franciska de Jong (2022). *Deliverable D2.3 Report from CLARIN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLARIN.pdf.

Gísladóttir, Guðrún (2022). *Deliverable D2.7 Report from ECSPM*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ECSPM.pdf.

Gomez-Perez, Jose Manuel, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn (2022). *Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/text-analytics-deep-dive.pdf.

Hajič, Jan, Maria Giagkou, Stelios Piperidis, Georg Rehm, and Natalia Resende (2021). *Deliverable D2.1 Specification of the consultation process*. European Language Equality (ELE); EU project

no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-process.pdf.

Hajič, Jan, Tea Vojtěchová, and Maria Giagkou (2022). *Deliverable D2.5 Report from META-NET*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-META-NET.pdf.

Hegele, Stefanie, Katrin Marheinecke, and Georg Rehm (2022). *Deliverable D2.6 Report from ELG*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELG.pdf.

Heuschkel, Maria (2022). *Deliverable D2.12 Report from Wikipedia*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-Wikipedia.pdf.

Hicks, Davyth (2022). *Deliverable D2.9 Report from ELEN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELEN.pdf.

Hrasnica, Halid (2022). *Deliverable D2.11 Report from NEM*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-NEM.pdf.

Kaltenböck, Martin, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg (2022). *Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs*. European Language Equality (ELE); EU project no. LC-01641480 – 1010-18166. https://european-language-equality.eu/reports/data-knowledge-deep-dive.pdf.

Kirchmeier, Sabine (2022). *Deliverable D2.8 Report from EFNIL*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-EFNIL.pdf.

Rehm, Georg and Stefanie Hegele (2018). "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs". In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA, pp. 3282–3289. https://aclanthology.org/L18-1519.pdf.

Rehm, Georg, Katrin Marheinecke, Rémi Calizzano, and Penny Labropoulou (2023a). "Language Technology Companies, Research Organisations and Projects". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 171–185.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Katrin Marheinecke, and Jens-Peter Kückens (2023b). "European Language Technology Landscape: Communication and Collaborations". In: *European Language Grid: A Lan-*

guage Technology Platform for Multilingual Europe. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 189–204.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg etc.: Springer. http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

Rufener, Andrew and Philippe Wacker (2022). *Deliverable D2.4 Report from LT-innovate*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LTInnovate.pdf.

Thönnissen, Marlies (2022). *Deliverable D2.2 Report from CLAIRE*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLAIRE.pdf.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI 10.2759/142151. A study prepared for the European Commission, DG Communications Networks, Content & Technology by Crosslang, Tilde, ELDA, IDC.

Chapter 39
Results of the Forward-looking Community-wide Consultation

Emma Daly, Jane Dunne, Federico Gaspari, Teresa Lynn, Natalia Resende, Andy Way, Maria Giagkou, Stelios Piperidis, Tereza Vojtěchová, Jan Hajič, Annika Grützner-Zahn, Stefanie Hegele, Katrin Marheinecke, and Georg Rehm

Abstract Within the ELE project three complementary online surveys were designed and implemented to consult the Language Technology (LT) community with regard to the current state of play and the future situation in about 2030 in terms of Digital Language Equality (DLE). While Chapters 4 and 38 provide a general overview of the community consultation methodology and the results with regard to the current situation as of 2022, this chapter summarises the results concerning the future situation in 2030. All of these results have been taken into account for the specification of the project's Strategic Research, Innovation and Implementation Agenda (SRIA) and Roadmap for Achieving Full DLE in Europe by 2030.[1]

1 Introduction

Within ELE three complementary online surveys were designed and implemented in order to consult the Language Technology (LT) community with regard to the current state of play and the future situation in about 2030 in terms of Digital Language Equality (DLE). While Chapter 38 provides a general overview of the community consultation process and methodology and Chapter 4 in Part I gives a brief account

Emma Daly · Jane Dunne · Federico Gaspari · Teresa Lynn · Natalia Resende · Andy Way
Dublin City University, ADAPT Centre, Ireland, emma.daly@adaptcentre.ie, jane.dunne@adaptcentre.ie, federico.gaspari@adaptcentre.ie, teresa.lynn@adaptcentre.ie, natalia.resende@adaptcentre.ie, andy.way@adaptcentre.ie

Maria Giagkou · Stelios Piperidis
R. C. "Athena", Greece, mgiagkou@athenarc.gr, spip@athenarc.gr

Tereza Vojtěchová · Jan Hajič
Charles University, Czech Republic, vojtechova@ufal.mff.cuni.cz, hajic@ufal.mff.cuni.cz

Annika Grützner-Zahn · Stefanie Hegele · Katrin Marheinecke · Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, annika.gruetzner-zahn@dfki.de, stefanie.hegele@dfki.de, katrin.marheinecke@dfki.de, georg.rehm@dfki.de

[1] This chapter summarises results reported in Way et al. (2022a) and Way et al. (2022b).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_39

of the results with regard to the current situation in 2022/2023, the present chapter summarises our results concerning the future situation. All of these results have been taken into account for the specification of the project's strategic recommendations (see Chapter 45).

Section 2 summarises the future-looking results with regard to the stakeholder group of European LT developers, introduced in Chapters 4 and 38, whereas Section 3 reports the findings with regard to the stakeholder group of European LT users and consumers. Section 4 describes the findings of the survey in which we reached out to Europe's citizens to gauge their expectations and desires in terms of DLE by 2030 (see Chapter 4, Section 3, p. 84 ff., and Chapter 38, Section 3, p. 231 ff.). Section 5 concludes the chapter.

2 The Perspective of European Language Technology Developers

The survey targeting LT developers and researchers generated a large number of responses between June and October 2021, representing more than 200 different organisations and more than 30 countries. The survey investigated topics like language coverage and evaluation of the current situation but also predictions and visions for the future. Detailed breakdowns of the results can be found in various ELE project reports (Thönnissen 2022; Eskevich and Jong 2022; Rufener and Wacker 2022; Hajič et al. 2022; Hegele et al. 2022). In addition to the survey, expert interviews with selected representatives from initiatives such as, among others, ELG and META-NET were conducted. The interviewees shared details on their work and related challenges, elaborating on how to do justice to all European languages, ways to position European LT on a global level and the key challenges towards establishing a long-term European LT programme.

2.1 Respondents' Profiles

One major goal of this survey was to bring the European LT community together and to reach a wide and demographically distributed audience. In total, the LT developers survey was filled in by 321 different respondents who represent 223 different organisations: 73% of the organisations were research or academic institutions (63% universities, 10% research centres) and 22% were companies (17% SMEs, 5% large enterprises). In 5% of responses the type "other" was indicated, i. e., freelancer, private practitioner, government agency, not-for-profit organisation, etc.

The headquarters of these organisations are located in 32 different countries, covering all EU member states and other European countries, such as the UK, Switzerland, Serbia, etc., but also other global regions, e. g., Brazil, the US and Israel. Most responses were contributed from Spain, Germany, Greece, the Czech Republic, and the Netherlands. The respondents cover a wide spectrum of the targeted groups

of stakeholders, as apparent from the range of networks, associations and relevant projects ongoing at the time the survey was circulated. The most established research networks in LT/AI, i. e., META-NET, CLARIN and CLAIRE are well represented in the survey responses with about 40 to 90 respondents each. ELG, ELE's sister project, is represented with more than 50 participants. Other related projects and networks focusing on LT or on neighbouring fields, such as AI4EU, ELISE, ELEXIS, and Nexus Linguarum are represented with around 10 to 25 survey respondents each (Table 1). Additional networks, associations and projects indicated by the respondents include ELRC, ELRA, ACL, EAMT, DARIAH and others.

Initiative	Responses	Interviews
CLAIRE	37	3
CLARIN	90	4
ELG	54	20
LT-Innovate	18	29
META-NET	61	5
AI4EU	16	–
BDVA	12	–
DIH4AI	1	–
ELEXIS	19	–
ELISE	4	–
HumanE AI	11	–
Nexus Linguarum	25	–
TAILOR	9	–
Other	31	–
None of the above	115	–

Table 1 LT developers survey – survey responses and interviews collected through the participating initiatives

The respondents were mainly active in the following areas: 1. Basic natural language processing services (POS tagging, parsing, named entity recognition etc.), 2. Text analytics and mining, information extraction, text classification, and 3. Language resources (LRs): data production, data aggregation (Figure 1).

The technologies, products or services offered by the respondents' organisations are used in various domains, a finding that demonstrates the applicability of LT in practically all economic sectors. The top three domains indicated by the respondents were 1. Information and communication technologies (ICTs), 2. Digital humanities (DH), arts, culture and other services and 3. Education.

2.2 Language Coverage

The respondents listed a wide range of languages they actively include in their research and development work and for which they offer services, software, resources, models etc. All official EU languages are covered as well as other state official, re-

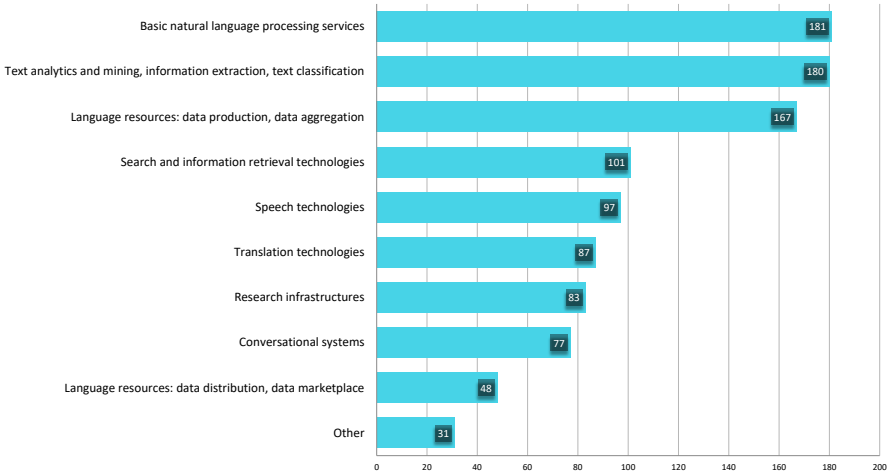

Fig. 1 LT areas in which the respondents conduct research or develop tools and services

gional or co-official European languages (see Figure 6 in Chapter 4, p. 86). The five most frequently mentioned languages are English, German, Spanish, French and Italian. A total of 80 respondents indicated "other" languages they support in their products or research, languages spoken in the Middle East and Asia with Arabic, Chinese, Japanese, Russian and Turkish being the five most frequently mentioned ones. Sign languages were also mentioned.

To get an idea about the focus of future work, the respondents were asked about the languages their organisation does not yet support, but plans to support in the next three years. Apart from some of the big languages, the respondents' future plans additionally include some regional and minority languages (RMLs), such as Basque, Catalan, Breton, Mirandese, Romani or Aromanian. Sign languages were mentioned five times, and it is worth noting the presence of regional and dialectal varieties in the respondents' future plans, e. g., Pontic Greek or Spanish varieties.

When considering the top three drivers for the decision to support additional languages (Table 2), the most frequently selected factor is research interest (212 mentions), followed by the availability of LRs (144) and market interest or demand (138). As expected, the prioritisation of these factors is different when the type of organisation the respondent represents is taken into account. For industry (including large enterprises and SMEs) market interest or demand by users or consumers play a pivotal role, while the availability of LRs follows at a distance. For research organisations and SMEs, more than big organisations, funding and investment opportunities are also to be considered. In terms of "other" reasons, these were often specified with an appeal for equality and the need for preserving all languages in the digital age, as for instance in the following answers: "Need for equality", "Ensure language rights in the digital economy, services, applications", "Supporting under-represented language communities to work towards the knowledge equity goals".

Drivers	Research organisation	Industry	Other	Total
Research or scientific interest	196	12	4	212
Availability of language resources	108	29	7	144
Market interest or demand	65	66	7	138
Available funding or investment	107	18	3	128
Availability of human experts	60	12	3	75
Availability of technologies or tools	44	18	5	67
Other	69	14	4	87

Table 2 LT developers survey – the top drivers for the decision to support additional languages

2.3 Predictions for the Future

We were also interested in the respondents' views on the measures and instruments that are deemed effective as well as the key challenges that a future large-scale ELE programme should address. The participants had the option to rate a number of policies and instruments as either very effective, effective, slightly effective or not effective at all. In addition, respondents were given the opportunity to elaborate on other policies or instruments, which they consider effective in speeding up the development and deployment of LT in Europe equally for all languages. The responses were provided as free text.

A critical aspect of the respondents' visions for DLE, as brought up in multiple answers, is the availability of resources. By 2030 all European languages should have developed the critical mass of resources needed for developing LTs. These include not only raw data, but also large multilingual language models. The issue of data availability is often mentioned in relation to the legal framework for sharing them. Large amounts of data for all languages are expected not only to be available by 2030, but also available for free or at a reasonable cost for research and commercial purposes. Standardised training and evaluation data for all languages are deemed critical. In parallel, according to the survey respondents, LT developers will be working towards automated procedures for the construction, annotation and curation of language data, as well as to address the issue of data bias. Such achievements, combined with continuous work on improving transfer learning methods, are expected to contribute to a situation in which all languages, including small, minority and regional ones, enjoy technology support and a level of presence and use in the digital sphere that will ensure their preservation and prosperity.

A shared scientific goal of the LT community is the achievement of *Deep Natural Language Understanding by 2030*, brought up in numerous responses with various phrasings: "hybrid intelligence", "cognitive AI", "symbolic AI", etc. Nonetheless, all these mentions converge on the description of a future status of LTs where the leap from superficial language *processing* to language *understanding* has been achieved and seamless human-like interaction, viable discourse interpretation and ubiquitous natural language interfaces are a reality for all Europeans in their own language.

With respect to measures and instruments that can be employed to help achieve these goals and realise the visions, the respondents evaluated the effectiveness of a set of proposed measures. A long-term programme of ten or more years can potentially lead to groundbreaking research and subsequently to the desired leap from simple language processing to deep language understanding according to almost all respondents (average score 4.2 on a five-point Likert scale with 5: very effective and 1: not effective at all). Continuous investment in existing research infrastructures (RIs) that support LT was considered equally effective (average score 4.2). Among others, access to data and tools via distributed RIs is argued to allow for optimising both the storage space and processing power, as well as to compare the LTs in terms of their computational footprint.

At the technological level, investing in the development of new scientific methodologies for the transfer or adaptation of resources or technologies to other domains and languages is considered an effective measure to boost the digital readiness of less supported languages (average score 4.0). Given the importance of a strong foundation in basic research, it does not come as a surprise that a large majority of over 86% of respondents welcome an increase in the availability of qualified LT personnel and incentives for talent retention. This also included reinforcing training and education initiatives, including undergraduate and Master's programmes.

A number of elaborate answers focused on funding instruments as leverage to help Europe achieve global excellence and leadership in LT. Funding and investments should concentrate not only on the applied (computational) aspects of LT but also on basic research in linguistics and computational linguistics. Support of LR creation and sharing is an issue in many responses. With respect to the beneficiaries of funding, a number of respondents and interviewees expressed the opinion that incentives should be provided to language communities that strive to preserve their cultural and linguistic identities, especially with regard to enhancing a language's presence on the internet. Businesses and industry-research collaborations are noted as an additional target group.

In this context, some respondents perceive the role of national centres of excellence in LT as critically important. Such centers could collect and boost the voices of local players at a national level and increase industry visibility nationally and at the European level. Apart from designing the national research agendas in LT, they should be responsible for the collection, curation, sharing and standardisation of language data, and for following and implementing the European Data Strategy.

Regulatory aspects pertinent to the LT field, in the form of regulations, recommendations or guidelines, have additionally been highlighted. These include, e. g., the adoption of the FAIR principles in Europe, a revised legislative framework for facilitating the use of language data and the application of data mining techniques for both research and commercial purposes, guidelines for procurement beneficiaries and for public bodies to release their funded or public data, recommendations for big technology companies to open up their platforms for the lesser spoken languages and for the public and private sectors equally to provide multilingual websites. It could be also beneficial to impose content accessibility regulations, e. g., for multimedia subtitling, readability, dubbing, etc.

The role of the research community is often criticised for its bias towards publications on a small number of the world's languages. Raising awareness of equality issues in international LT fora and incentivising Open Access journals and conferences dedicated to less supported languages are among the suggested measures.

Awareness raising of the importance of LT for digital interactions and the role of training young LT professionals is mentioned in numerous responses. Finally, the social dimensions of DLE have been emphasised by respondents who argued that linguistic and social diversity go hand in hand: the more diverse our society is, the more there is an actual need for multilingual resources and technologies. Thus, large-scale policies against racism and discrimination are considered essential. In parallel, engaging minoritised language communities and supporting community building is argued to benefit the LT field, as it will increase demand for and the impact of LT.

European LT should foster and support multilingualism while strictly adhering to European values such as privacy by design, transferability, fairness, diversity and openness, transparency and accountability, public wealth, individual rights and collective purposes. Europe's strengths lie in catering for multilingual solutions covering all the European languages and serving all citizens of Europe. By supporting its linguistic diversity, Europe can achieve digital self-determination and sovereignty.

3 The Perspective of European Language Technology Users

For LT users, a similar survey was set up (see Chapter 4, Section 3, p. 84 ff., and Chapter 38, Section 4, p. 235 ff.) and generated almost 250 responses. Similarly to the LT developers survey, numerous additional interviews were conducted for more in-depth insights.

The survey brought together diverse groups of stakeholders including representatives of communities of LT users, academic and commercial stakeholders, language professionals (e. g., translators, lecturers and professors in the fields of linguistics and computational linguistics) and stakeholders from different economic sectors (e. g., banking, health, public administration, language services). The survey was disseminated mainly via email by the relevant ELE partners, namely, ELEN, LIBER, ECSPM, NEM, EFNIL and Wikipedia as well as through social networks. Table 3 shows the breakdown of responses collected through the survey.

3.1 Respondents' Profiles

Responses came from a diverse range of sectors and professional activities; most of the respondents work in the education and research sector with 130 responses (53%) out of 246, that is, most respondents were researchers, university professors, assistant professors, lecturers or held other academic positions. The survey was also filled out by representatives of non-governmental organisations (NGOs), large en-

Initiative	Responses	Interviews
ECSPM	10	2
EFNIL	28	6
ELEN	7	19
LIBER	29	3
NEM	29	6
Wikipedia	22	3
Other (e. g., social media)	121	–
Total	246	39

Table 3 LT users survey – survey responses and interviews collected through the participating initiatives

terprises, SMEs, government departments and independent contractors and consultants in diverse economic sectors. The 15 (6%) respondents who selected the option "other" represented non-governmental bodies, non-profit organisations, public sector organisations, social organisations and independent government departments (see Figure 2).

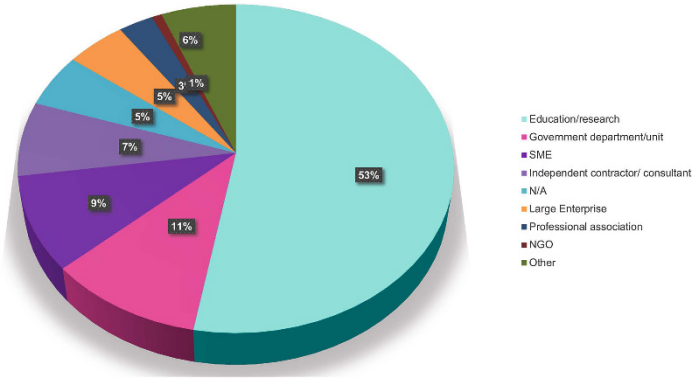

Fig. 2 LT users survey – types of sectors and professional activities

Contributions to the survey came from all over Europe and, due to social media sharing, some responses were provided by people based outside European countries such as the US, the Democratic Republic of Congo and the Russian Federation. In Europe, the most represented countries were Croatia (33 responses), Spain (23 responses), the UK (23 responses), Ireland (17 responses), Germany (16 responses) and France (14 responses).

3.2 Language Coverage

A total of 74% of the respondents indicated that they work with English, which is the dominant language followed by a well-balanced group of languages composed of German (31%), French (31%) and Spanish (30%). At the other end of the spectrum, many other European languages (e. g., Welsh, Catalan, Basque, Luxembourgish, Galician) are under-represented as few respondents (between one and three) indicated they work with them. Respondents who selected "other", mentioned that they work with Basque, Catalan, Macedonian, Luxembourgish, Moldovan, Welsh and Galician. Among the non-European languages respondents mentioned Japanese, Chinese (or Mandarin) and Russian. Figure 3 shows the breakdown of European languages the respondents work with in absolute numbers.

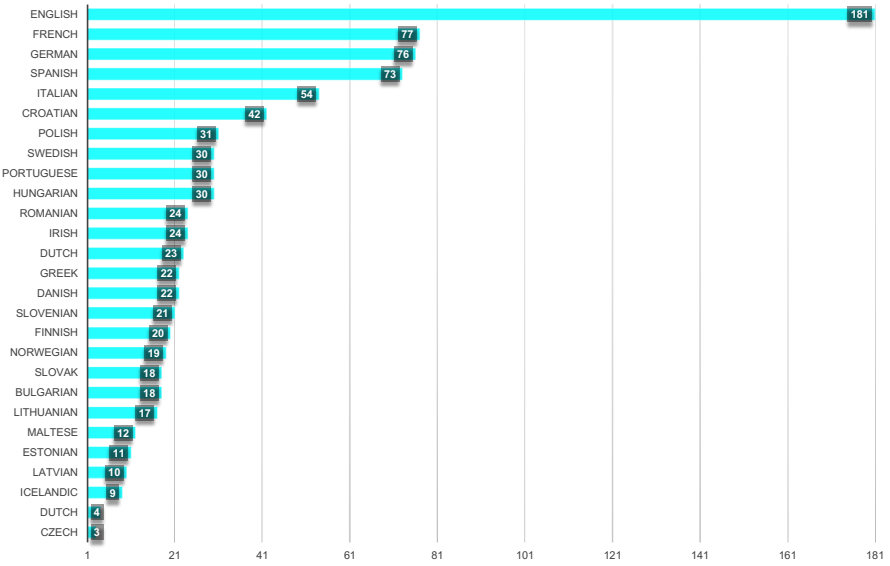

Fig. 3 LT users survey – European languages respondents work with (based on a set of 246 responses)

In relation to the languages respondents intend to include in their workflow, 50 respondents (20%) indicated that they plan to include English, German, Spanish and French. The survey shows, again, the English predominance over all languages followed by German, Spanish and French. Other official EU languages were mentioned by only a few respondents (between two and three respondents only) such as Italian, Portuguese and Greek as well as some minority, regional, and lesser-used languages such as Breton, Catalan, Faroese but only by one respondent each. These findings suggest a worrying scenario, where, in a multilingual and multicultural Europe, most minority, regional, lesser-used languages are disregarded either for not being commercially interesting or simply for lack of institutional investment.

3.3 Predictions for the Future

With regard to their predictions for the future, the range of opinions was very broad. In general, most respondents (68%) are confident that in the next ten years, there will be higher-quality tools for all European languages including minority, regional, and lesser-used languages and that there will also be a wider range of tools for all European languages (83%). However, fewer respondents (46%) believe that LTs will help to prevent linguistic loss, although 65% think that LTs can help to prevent RMLs from disappearing. Most respondents (64%) also agree that LTs can increase individuals' exposure to these languages and 60% believe that LTs can increase engagement with social, leisure and work activities in their own languages. Among other benefits mentioned in the open questions, respondents think that LTs can improve medical interactions between patients and clinicians and improve medical documentation. One respondent highlighted that LTs can help with the preservation of cultural heritage and improve its visibility. Another respondent pointed out that LTs can improve online and print publishing in minority, regional, and lesser-used languages, including academic publications and works of fiction.

The survey also looked into the respondents' ideas for the future of LT. They had the chance to indicate applications that could potentially use LT they want to see that are not currently available for the languages they work with. There were several interesting responses. In general, we can see respondents wish for higher-quality tools for certain languages such as "better parsing of Danish than currently available" or the availability of tools that do not yet exist for some languages but exist for others such as "speech recognition for Welsh", "speech recognition for Catalan", "free spell check for Irish", "more reliable speech recognition, information extraction, summarisation, semantic parsing and semantic search for Greek", "a good Georgian-English Translator" and "better MT for Croatian". Other respondents indicated that they would like to see some of the existing tools and technologies available in more languages, for instance, "Text-To-Speech for low resource languages" or "more accurate speech2text, decent text summarization, GPT2 for Finnish".

Some ideas for new (currently non-existent) LTs were also provided. For instance, "case-sensitive tools or the creation of a tool that might provide more context, or warn the user if the same word means something completely different depending on the context. A tool that would be sensitive to connotative meanings" or "tools for collecting lexical data and speed up the process of dictionary building".

We can conclude that the most important finding of this survey is the respondents' concern regarding the differences in technological support between European languages, specifically the poor technological support of minority, regional and lesser-used languages. The differences in support are mainly reflected in differences in the quality and performance of tools between the languages as well as in the availability of tools for a small group of low-resource languages, while these same tools do not exist for many other European languages. In order to achieve full DLE as a crucial step to maintain linguistic diversity, the survey shows the necessity for action and an implementation agenda with the objective of fostering and supporting a multilingual and linguistically inclusive Europe that brings solutions to all European citizens.

4 The Perspective of Europe's Citizens as Consumers of LTs

The ELE project has made an effort to ensure that all voices were heard and taken into account in the preparation of the SRIA. With the support of social media campaigns and an agency specialising in survey dissemination, we were able to reach thousands of EU citizens to hear their thoughts on how well they feel their languages are digitally supported. The European Citizen survey included a total of 11 questions, six multiple-choice questions, four single-choice questions and one open-ended question which allowed respondents to include any comments or feedback they had. The survey was designed to take less than five minutes to fill in (see Chapter 4, Section 3, p. 84 ff., and Chapter 38, Section 4, p. 235 ff.). It was translated into 35 languages. To ensure the reliability of the survey data captured, a number of data cleaning steps were taken to remove responses that were deemed noisy or at risk of skewing the survey results. We analysed a total number of 20,586 valid responses, the largest public survey ever conducted to date among European citizens concerning LRTs.

4.1 Respondents' Profiles

We collected (anonymous) demographic information from respondents with the objective to ensure our sample was representative enough of the population for generalisation purposes. We asked respondents to state their level of education, age group and country of residence. We collected responses from 28 countries, and Figure 4 shows the breakdown of contributions per country.

The demographic of the respondents is as follows: 27% of the respondents were between 25-34 years old. A total of 23% accounted for both the 18-24 and 35-44 age brackets. The rest of the respondents were 45+ years old, 1% of the respondents preferred not to say. In terms of education, 35% of the respondents had reached high school level, 23% held a Bachelor's Degree, 17% held a Master's Degree, with the rest reporting vocational training (11%), only some high school completion (7%) and holding a PhD (5%), 2% declined to say.

4.2 Language Coverage

We asked respondents to select the languages they use both socially and professionally. Overall, results show that many respondents use their native language in addition to English even if they are not based in English-speaking countries. Therefore, we once again see a dominance of English over all other languages. Following English, German and French also appear as languages frequently used in non-German or non-French speaking countries. Figure 5 illustrates the comparison of the most represented languages in the survey.

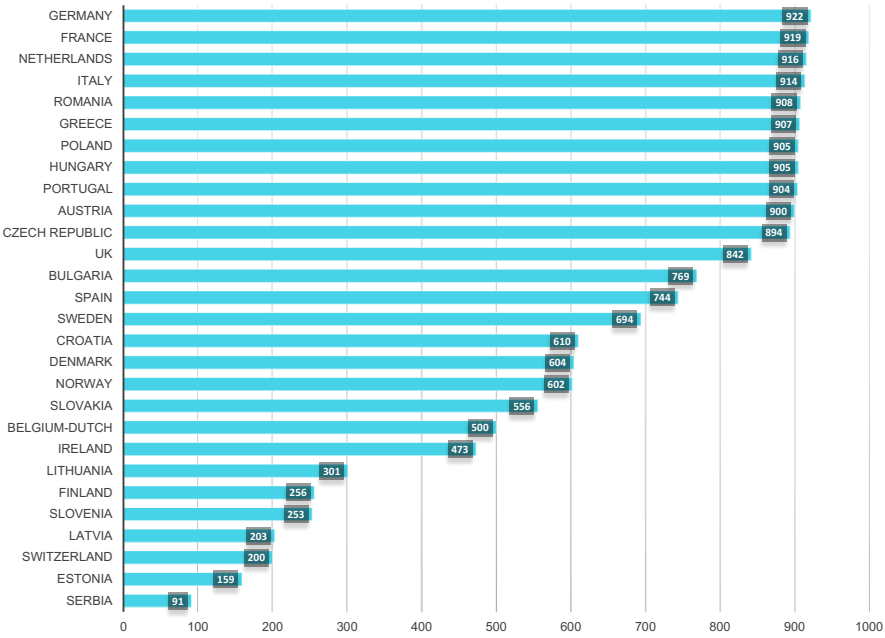

Fig. 4 European citizens survey – number of responses collected

4.3 Predictions for the Future

The following discussion concentrates on the forward-looking questions of the EU citizens survey and the responses concerning anticipated or hoped for future developments with regard to the development and consolidation of LTs for Europe's languages. In one question we asked the respondents "What would be the top 3 advantages of improving apps and tools for all languages? Please select the three most important advantages in your opinion." The purpose of this question was to assess respondents' views on the benefits of LTs. Notably, as seen from Figure 6, LTs are regarded as key to enhancing multilingual societies from a linguistic diversity perspective. Of seemingly less importance to the average citizen is the economic advantage that arises from LT support.

With regard to the question "What holds you back from using some of these apps or tools in your languages?", based on the answers received, it is reasonable to assume that if the reported barriers that are currently holding users back from using apps or tools in their languages were removed, and tools more adequately supported, then there would be more uptake in the number of people using language tools in their own preferred language (see Figure 7). It was somewhat surprising that the top response was "I don't need to use any apps or tools for this language", which might suggest that the poor support for some languages may condition users into believing that technologies do not apply to some chronically underserved languages. This

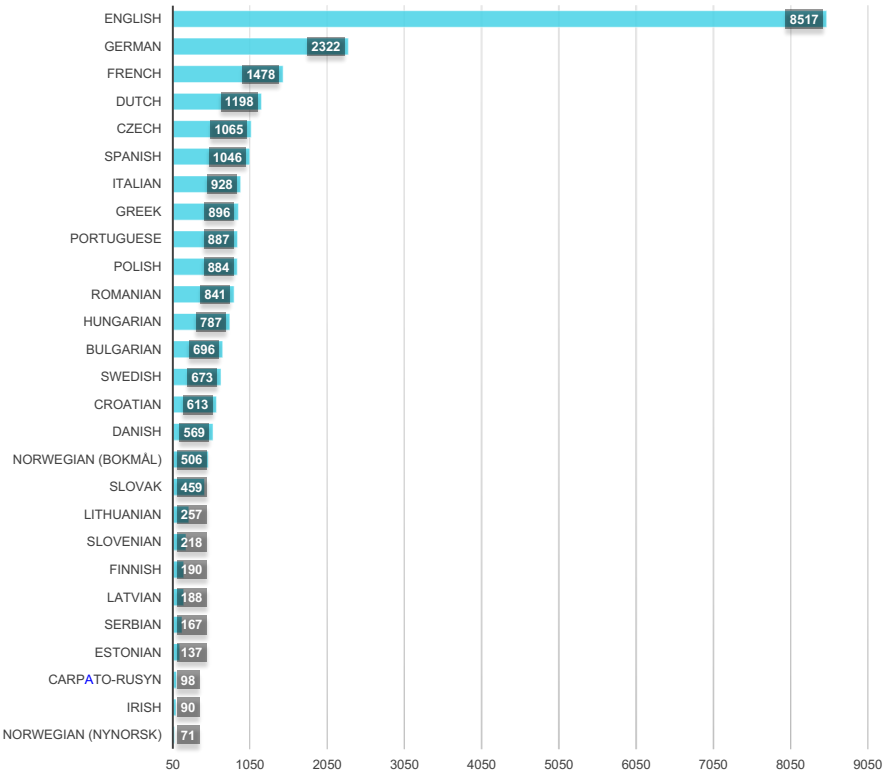

Fig. 5 European citizens survey – most represented languages

may apply in particular to users who also speak a dominant language that is well supported by tools and apps, in addition to one that is scarcely supported.

In other words, these responses suggest that there is a real risk that some users have become so accustomed to using apps in or for better supported languages that they no longer see the need for similar apps to be developed and made available in or for their own language; at the same time, this disappointing perception may stabilise a situation where users default to using apps and tools in an additional language that is better supported, also due to their overall superior quality. Another popular response was “Issues with the quality of the available apps or tools”, indicating that people will not use an app or tool if they perceive its quality to be insufficient or inadequate. This suggests that once the quality of the tools is improved to a sufficient standard, more people would be inclined to use the app or tool in their language in the future.

Concerning the query “Please select the tools that you currently do not use but would like to use in the future.”, one tool that people are calling for in particular among those to be made available for their languages is automatic subtitling (Figure 8). Having this available for more languages would improve communication

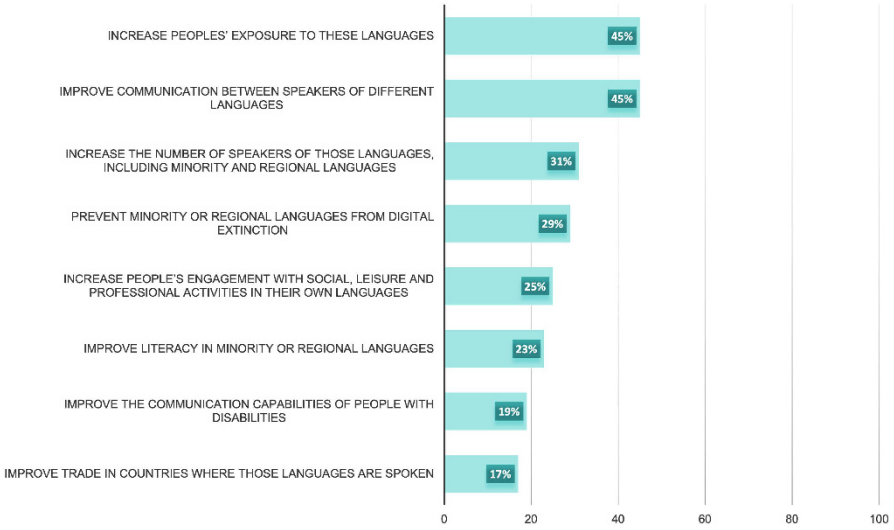

Fig. 6 Responses to the question "What would be the top 3 advantages of improving apps and tools for all languages?" in the EU citizen survey

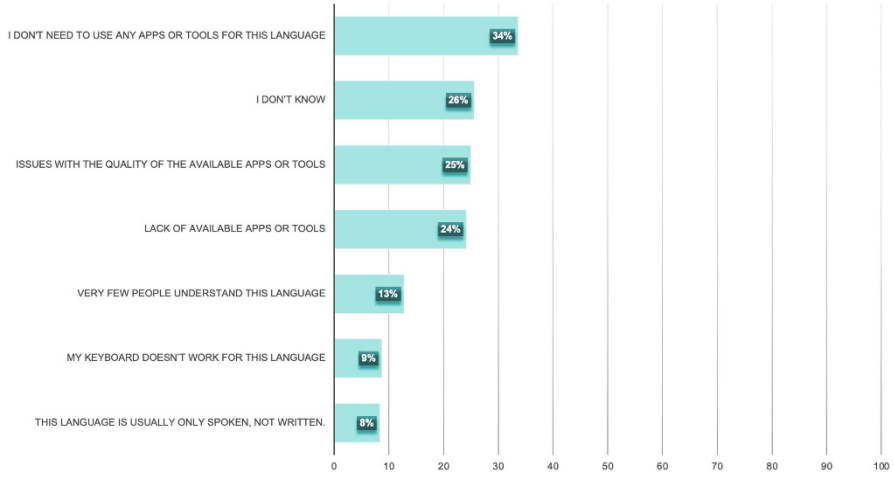

Fig. 7 Responses to the question "What holds you back from using some of these apps or tools in your languages?" in the EU citizen survey

and accessibility of multimedia content for an ever-increasing range of European citizens (e. g., disabled people, elderly users, etc.). Relevant examples include automatic subtitles being made available to those who are hearing-impaired, so they can watch videos and read subtitles in their own language. Translation apps are also in very high demand, which is not particularly surprising. However, even for those language-pairs that are serviced by MT, we need to be vigilant as many of the freely

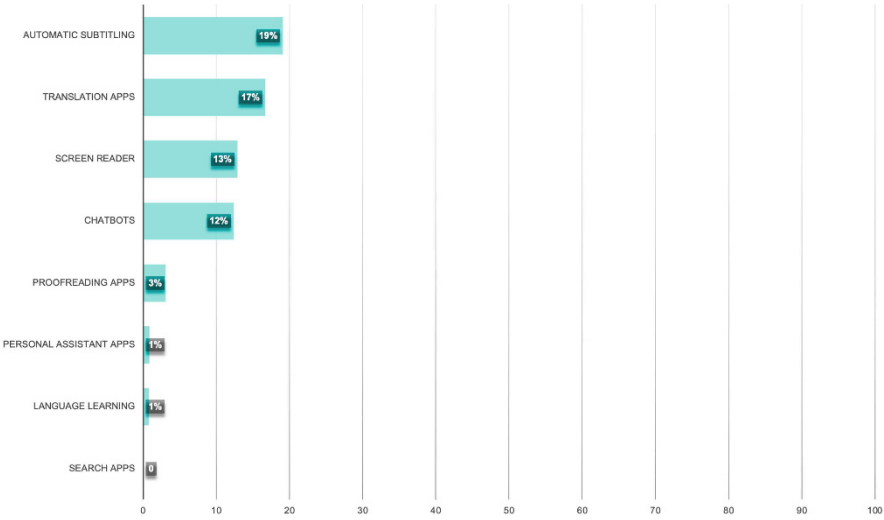

Fig. 8 Responses to the question "Please select the tools that you currently do not use but would like to use in the future." in the EU citizen survey

available translation tools are not owned or resourced by EU companies. Screen readers are another tool that is quite popular, with obvious relevance to visually impaired people. If screen readers were available in more languages, accessibility would be substantially increased for several language communities across Europe.

Finally, in the analysis of the responses to the survey, a number of interesting comments made by ordinary EU citizens were found in the section that elicited more general reactions at the end of the questionnaire. In particular, the very last question of the survey asked the participants to enter any comments they had about the survey or LTs in general. Here follows a selection of the most insightful comments that we feel encapsulate some of the most relevant opinions on the matter.

- "No language is inferior to others. All languages are worthy of survival as long as there is at least one person who speaks that language."
- "Usually I google things in English because more information is available in English."
- "It is extremely important to have more language technology tools for the national minority languages in Sweden. It is a rights issue to access everything from speech synthesis, machine translation, language apps, proofing programs, etc. At the moment, there are no opportunities for this for Roma, Meänkieli and to some extent for Sami and Yiddish."
- "It would be great to have a little more guidance on what ordinary people (without great technological resources such as universities and companies) can do to 'feed' or develop those technological resources for our minority languages."

These comments clearly indicate that some European citizens are eager to have more LT tools and apps made available to them in their language in the future, as

this is related to the role that individual speakers and their communities can play going forward in the digital age in the interest of equality. At the moment many people seem to be resorting to using search apps and personal assistants particularly in English or other well-resourced languages, as they are currently unavailable in their own language or are not perceived to perform equally well. This suggests that if required LTs were developed and made available as tools or apps, people would use them in their own language rather than English; at the very least they would have a choice, depending on the type of tasks that they need to perform in different circumstances (e. g., for professional purposes as opposed to personal or social reasons, with colleagues, within the family or with circles of friends and acquaintances, etc.).

The survey also revealed that some European citizens want to see technology for their languages improved and maintained, and some are willing to get involved themselves, as shown by the comment asking what the ordinary citizen can do to help the development of these much-needed technologies. Overall, citizens are concerned about the technological status of their language, and are willing to help to ensure that their language is technologically well supported in the future for the digital age, especially if otherwise there is a threat of extinction. We were particularly pleased at respondents' willingness to take ownership of these issues, and act not only as users of tools but also as developers. We take this as a strong endorsement of the ELE project, and further evidence of the need for the ELE programme to be fully funded throughout Europe to ensure DLE for all Europeans, as reflected in the ELE SRIA.

5 Summary and Conclusions

The surveys and expert interviews discussed here targeted LT developers, users and the EU citizens. We investigated language coverage and encouraged participants to share their predictions and visions for the future of LTs in Europe with respect to achieving full DLE. The results show that there is still a huge gap between the LT support for English and all other European languages, with dramatic differences in several cases. Even though there is an increased interest in bridging this gap and in expanding technological support to more languages, limited funding, demand and obstacles with regard to available resources make it a challenging endeavour. While basic research is still urgently needed, the last decade has seen progress on a larger scale than could have been imagined ten years ago. Many experts highlight European excellence, also on a global level and consider leadership in LT and language-centric AI to be possible if the necessary conditions are created by political decision-makers.

The LT developers survey addressed the European LT community, reaching a wide and demographically distributed audience. It was answered by 321 respondents who represent 223 organisations in 32 countries. The respondents were recruited by the research networks, i. e., META-NET, CLARIN and CLAIRE, projects like ELG and other related initiatives focusing on LT or neighbouring fields, such as ELISE, ELEXIS, and Nexus Linguarum. Additional networks, associations and projects represented by the respondents include ELRC, ELRA, ACL, EAMT, DARIAH and oth-

ers. The areas in which the respondents are active covered the full range of LT. The languages they focus on have a skewed distribution that reflects current imbalances in the field in Europe as well as elsewhere, with English first by a large margin, followed by the big official EU languages. The two main concerns expressed were the insufficient support for basic research in NLP and LT and the fierce competition of non-EU companies with the market disruption they cause. The survey answers to the open-ended questions and views of the interviewed experts brought a host of opinions and suggestions in several important directions, in particular: the higher and even elementary education area, research funding, legal and regulatory obstacles, biases and privacy issues of various types, commercialisation difficulties and ways of supporting such efforts, the need to coordinate efforts between national centres of excellence vs. pan-European ones, etc.

The LT users and consumers survey brought together academic and commercial stakeholders, language professionals and stakeholders from different sectors. It was disseminated by the relevant ELE partners, i. e., ELEN, LIBER, ECSPM, NEM, EFNIL and Wikipedia who promoted the survey targeting representatives of organisations and communities of users and consumers. Based on the results, it can be concluded that the most important finding is the respondents' concern regarding the differences in technological support between Europe's languages, specifically the poor technological support of minority, regional and lesser-used languages. The differences in support are mainly reflected in differences in the quality and performance of tools between the languages as well as in the availability of tools for a small group of languages, while these same tools do not exist for many other European languages. To achieve full DLE as a step to maintain and promote linguistic diversity, the survey shows the necessity for action and calls for an implementation agenda with the objective of fostering and supporting a multilingual and linguistically inclusive Europe that brings solutions to all European citizens that are relevant in the digital age.

An additional survey was carried out targeting EU citizens with the aim of taking into account their opinions, individual needs, wishes, general demands and, importantly, to make sure that their voices play a decisive role in the pursuit of full DLE supported by LT. The survey was disseminated in 28 countries with the help of a service provider. Additional dissemination was carried out with the help of ELE partners who promoted the survey on social media, within their networks and through the ELE project website. While structured very differently than the stakeholder group surveys, there are several similarities not only in terms of the scope of the analysis, but also of the key results that were obtained: languages other than English are poorly supported (with only a few exceptions) – something evident even from the distribution of languages that the respondents considered in their responses. These answers show that raising awareness for the LT potential in Europe on a political and institutional level is more important now than ever before. The European LT community is in a position where change is needed in order to compete with innovative systems and tools built elsewhere. On a political level, this involves more commitment from the European institutions as well as those of the Member States.

References

Eskevich, Maria and Franciska de Jong (2022). *Deliverable D2.3 Report from CLARIN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLARIN.pdf.

Hajič, Jan, Tea Vojtěchová, and Maria Giagkou (2022). *Deliverable D2.5 Report from META-NET*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-META-NET.pdf.

Hegele, Stefanie, Katrin Marheinecke, and Georg Rehm (2022). *Deliverable D2.6 Report from ELG*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELG.pdf.

Rufener, Andrew and Philippe Wacker (2022). *Deliverable D2.4 Report from LT-innovate*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LTInnovate.pdf.

Thönnissen, Marlies (2022). *Deliverable D2.2 Report from CLAIRE*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLAIRE.pdf.

Way, Andy, Georg Rehm, Jane Dunne, Maria Giagkou, Jose Manuel Gomez-Perez, Jan Hajič, Stefanie Hegele, Martin Kaltenböck, Teresa Lynn, Katrin Marheinecke, Natalia Resende, Inguna Skadina, Marcin Skowron, Tereza Vojtěchová, and Annika Grützner-Zahn (2022a). *Deliverable D2.18 Report on the state of Language Technology in 2030*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-in-2030.pdf.

Way, Andy, Georg Rehm, Jane Dunne, Jan Hajič, Teresa Lynn, Maria Giagkou, Natalia Resende, Tereza Vojtěchová, Stelios Piperidis, Andrejs Vasiljevs, Aivars Berzins, Gerhard Backfried, Marcin Skowron, Jose Manuel Gomez-Perez, Andres Garcia-Silva, Martin Kaltenböck, and Artem Revenko (2022b). *Deliverable D2.17 Report on all external consultations and surveys*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/external-consultations.pdf.

Chapter 40
Deep Dive Machine Translation

Inguna Skadiņa, Andrejs Vasiljevs, Mārcis Pinnis, Aivars Bērziņš, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García-Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, and Andy Way

Abstract Machine Translation (MT) is one of the oldest language technologies having been researched for more than 70 years. However, it is only during the last decade that it has been widely accepted by the general public, to the point where in many cases it has become an indispensable tool for the global community, supporting communication between nations and lowering language barriers. Still, there remain major gaps in the technology that need addressing before it can be successfully applied in under-resourced settings, can understand context and use world knowledge. This chapter provides an overview of the current state-of-the-art in the field of MT, offers technical and scientific forecasting for 2030, and provides recommendations for the advancement of MT as a critical technology if the goal of digital language equality in Europe is to be achieved.[1]

Inguna Skadiņa · Andrejs Vasiljevs · Aivars Bērziņš · Mārcis Pinnis
Tilde, Latvia, inguna.skadina@tilde.com, andrejs.vasiljevs@tilde.com, aivars.berzins@tilde.com, marcis.pinnis@tilde.com

Nora Aranberri · Iakes Goenaga
University of the Basque Country, Spain, nora.aranberri@ehu.eus, iakes.goenaga@ehu.eus

Joachim Van den Bogaert
CrossLang, Belgium, joachim.van.den.bogaert@crosslang.com

Sally O'Connor · Riccardo Superbo
KantanMT, Ireland, sallyoc@kantanai.io, riccardos@kantanai.io

Mercedes García-Martínez · Manuel Herranz
PANGEANIC, Spain, m.garcia@pangeanic.com, m.herranz@pangeanic.com

Jan Hajič · Martin Popel · Rudolf Rosa
Charles University, Czech Republic, hajic@ufal.mff.cuni.cz, popel@ufal.mff.cuni.cz, rosa@ufal.mff.cuni.cz

Christian Lieske
SAP SE, Germany, christian.lieske@sap.com

Maja Popović · Sheila Castilho · Federico Gaspari · Andy Way
Dublin City University, ADAPT Centre, Ireland, maja.popovic@adaptcentre.ie, sheila.castilho@adaptcentre.ie, federico.gaspari@adaptcentre.ie, andy.way@adaptcentre.ie

[1] This chapter is an abridged version of Bērziņš et al. (2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_40

1 Introduction

Machine translation (MT) was one of the first application areas of natural language processing (NLP). Starting from the first attempts to apply dictionary-based approaches right up to modern neural network-based systems, MT has aimed to provide automatic translation from one natural language into another.

Today, MT has become an important asset for multilingual Europe, allowing citizens, governments and businesses to communicate in their native languages, breaking down language barriers and supporting the implementation of the European digital single market. For example, the eTranslation automated translation tool,[2] developed by the European Commission, and its various adoptions (e. g., EU Council Presidency Translator, Pinnis et al. 2021)[3] provide reasonably good MT service in 24 EU official languages for governments, the public sector and SMEs.[4] However, MT support and the quality of its output still differ from language to language, and from domain to domain. In particular, MT quality drops significantly when translation concerns less-resourced languages, speech or terminology-rich domains with limited available data.

1.1 Scope of this Deep Dive

In 2012, the META-NET White Paper series (Rehm and Uszkoreit 2012) presented a thorough analysis of Language Technology (LT) support for 31 European languages. According to this study, for MT *good support* only applied to English and *moderate support* to only two widely spoken languages (French and Spanish), leaving the remaining 28 European languages in clusters of *fragmented* or *weak or no support*.

This chapter focuses on the MT landscape a decade after the publication of the META-NET White Papers. We analyse progress in MT, identify the main gaps and outline visions, the breakthroughs needed and development goals towards Digital Language Equality (DLE) and Deep Natural Language Understanding (NLU) by 2030. We look at the current services and technologies offered by MT providers in the European market. The dominance of global companies in the free online translation market and the risks for Europeans caused by this dependence are among the key topics discussed in this chapter, especially to identify solutions going forward.

The main gaps are identified for four dimensions of MT: data, technology, approaches and legislation. We focus not only on data availability and usability and the need for less-resourced technologies, but also discuss limitations related to multimodal MT. While MT technologies today are available for most European languages, many of these languages are less attractive from a business point of view, and con-

[2] https://webgate.ec.europa.eu/etranslation/public/welcome.html

[3] https://www.eu2020.de/eu2020-en/presidency/uebersetzungstool/2361002

[4] As of February 2022, eTranslation was used by 108 projects – 87 projects reusing eTranslation and 21 projects committed to analysing or reusing eTranslation.

sequently they are not so well equipped with MT tools. Throughout the chapter, language coverage is addressed as a key dimension for DLE. We also discuss legal and ethical aspects related to the development, production and use of MT systems and services. We analyse IPR and GDPR restrictions and the 'fair use' principle from the developer's perspective, and privacy and security issues from the user's perspective. Finally, all these aspects are taken into consideration from the perspective of their impact on society, with a focus on Europe. The chapter provides a series of recommendations on how to address the current limitations of MT technologies and how to contribute to DLE as a crucial goal for Europe and its citizens.

1.2 Main Components

While different MT types (e. g., rule-based, example-based, statistical, hierarchical) have been investigated, in this subsection we will focus only on the recent development of Neural MT (NMT), based on an overview by Popel (2018). We present the main MT components of the general NMT architecture and the currently most popular example: Transformer (Vaswani et al. 2017). There are many other components related to MT, which are not described here, e. g., automatic speech recognition[5] and speech synthesis, which are needed in the speech-to-speech translation pipeline; cross-lingual information retrieval; multilingual summarisation; integration into production systems and multilingual websites using suitable metadata formats.[6]

In NMT, each input sentence is first tokenised into a sequence of tokens. The most popular approach today is to split words into subword units (subwords, which need not be actual words of the language or even morphemes). For example, the German word *Forschungsinstituten* ('research institutes') may be encoded with three subwords: *Forsch* + *ungsinstitu* + *ten_*. There are several algorithms for training subword models (e. g., Sennrich et al. 2016b). NMT based on subwords shows better results than early approaches based on words and recent approaches based on characters (Libovický et al. 2022). Each token is represented as a real-value vector, called (subword/word) embedding. Most NMT systems initialise embeddings randomly and train them jointly with the whole translation, but pre-trained (contextual) embeddings may be used as well, especially in low-resource settings.

NMT systems are based on an encoder-decoder architecture. The encoder maps the input sequence to a vector of hidden states (sometimes called continuous representation or sentence embedding). The decoder maps the hidden states into the output sequence (of target-language tokens). Each hidden state usually corresponds to one position (token) in the input sequence, so in general, the vector of hidden states has a variable length. Early NMT systems (Sutskever et al. 2014) used only the last hidden vector as an input for the decoder. Thus, the training was forced to encode all the information about the input sentence into a fixed-length vector. Bahdanau et al.

[5] See, for example, the reports of the ELITR project at https://elitr.eu.

[6] https://www.w3.org/TR/mlw-metadata-us-impl

(2015) introduced an encoder-decoder attention mechanism, where the decoder has access to all of the encoder's hidden states. This way, when generating each output token, the decoder can *attend* to different parts of the input sentence. The encoder-decoder attention mechanism circumvents the fixed-length sentence-representation restriction and improves translation quality, especially on longer sentences.

The process of translating sentences (at test time) with a trained NMT model is usually called inference. Most NMT systems use auto-regressive inference. This means that the output sentence is generated token by token and after each token is generated, its embedding is used as input for generating the next token. Decoding finishes once the decoder generates a special end-of-sentence token.

The advantage of NMT systems is that all their components can be trained in an end-to-end fashion unlike earlier data-driven approaches, where most components had to be trained separately. NMT is usually trained using backpropagation optimising the cross-entropy loss of the last decoder's softmax layer, which predicts output token probabilities; there are also NMT systems optimising sentence-level metrics (e. g., BLEU, Papineni et al. 2002, or simulated human feedback) with reinforcement learning techniques (e. g., Nguyen et al. 2017). NMT usually uses teacher-forcing: when generating the next word during training, it uses the previous word from the reference translation as the input instead of using the previously predicted word.

The Transformer architecture follows the general encoder-decoder architecture, but unlike earlier recurrent-networks it uses self-attention and feed-forward layers in both the encoder and decoder. This allows training and partially also the decoding process to be sped up thanks to better use of parallelisation.

Self-attention is based on a compatibility function which assigns a weight to each pair of tokens, more precisely, to their vector representation on each layer. Transformer uses multi-head self-attention, so multiple versions (heads) of the self-attention function are trained for each layer. Figure 1 shows an example of visualisation for different heads.

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

Deep learning techniques have given a major boost to the area. The application of neural networks to MT has opened the path to developing a universal engine whose ultimate goal is a single model to translate between any arbitrary language pair. The effects of different advanced approaches for multilingual MT models have been investigated by Yang et al. (2021), for example. They first explore how to leverage the large-scale language models created from the publicly available DeltaLM-Large multilingual pre-trained encoder-decoder model (Ma et al. 2021) to initialise the model. For efficient training, they apply progressive learning (e. g., Zhang et al. 2020) to create a deep model from a shallow one. Additionally, they implement multiple rounds of back-translation (e. g., Dou et al. 2020) for data augmentation purposes. While the

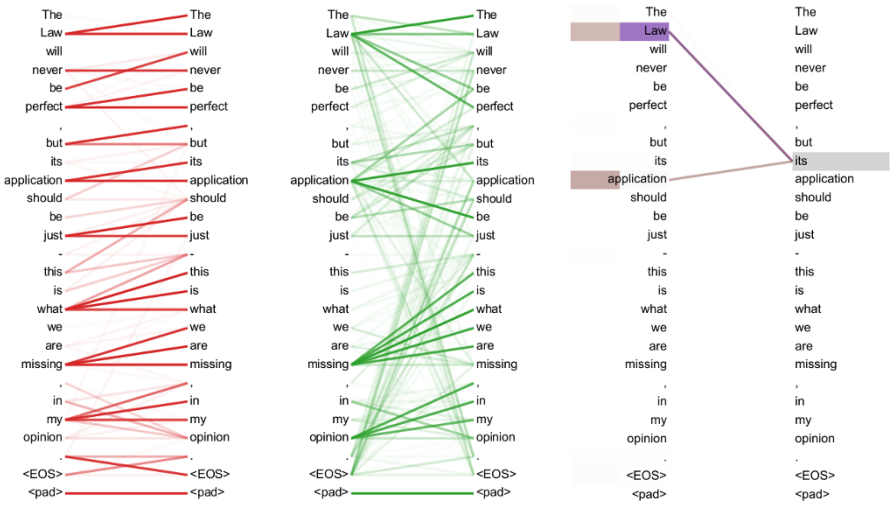

Fig. 1 Visualisation of self-attention in a Transformer model trained on English → German translation (adapted from Vaswani et al. 2017). Each head is visualised in a different colour and edge weight is indicated by thickness. Each of the figures shows another attention head in encoder layer 5 (out of 6). The words in the left column in each of the three visualisations represent vectors corresponding to these words on the input to the fifth layer of the encoder. The right-most figure shows two attention heads, but focusing only on the word '*its*' and illustrating coreference resolution.

results are very promising, they reflect a worrying trend: when English is involved in the translation process either as a source or target language, the BLEU scores are rather high. However, the results worsen considerably when translation in language pairs without English is considered.

If we turn to the goal of achieving language equality, one of the most interesting approaches is unsupervised MT (e. g., Artetxe et al. 2018) where no bilingual parallel data is needed to train a fully working system. In recent years, this approach has slowly been catching up with the translation quality obtained by supervised systems. For instance, Han et al. (2021) build a state-of-the-art unsupervised NMT system derived from a generative pre-trained language model. Their method is a concatenation of three steps: few-shot amplification, distillation and back-translation (Sennrich et al. 2016a). They first use the zero-shot translation ability of a large pre-trained language model (GPT-3) to generate translations for a small set of unlabeled sentences. In the next step they amplify these zero-shot translations by using them as few-shot demonstrations for sampling a larger synthetic data set, which is then distilled into a smaller model via fine-tuning to obtain a new state-of-the-art in unsupervised translation on the WMT14 English-French benchmark. While still restricted to a well-resourced language pair, learning outcomes are promising for lower-resource pairs.

Within the industrial context, a look at providers' solutions gives a clear overview of the strengths of each company, as well as the issues that remain relevant regarding the successful implementation of the technology. A key aspect that most companies emphasise is the capacity for domain adaptation. This allows for engines that learn

from domain-specific texts, avoiding the noise that expressions from other fields might introduce in the learning process (e. g., Pangeanic, RWS, Tilde, Welocalize). Further customisation is also highly valued, most frequently by refining their own generic or domain-specific engine with a customer's own data (e. g., Across, Language I/O, Tilde). Alternatively, do-it-yourself MT opportunities are provided where customers build their own system from scratch using just their own data.

The text type involved is also distinctive across companies, with some pushing for real-time adaptive MT for email and chat (e. g., Language I/O), while others emphasise multimodality. When a level of accuracy and/or cultural adaptation is required, MT is coupled with post-editing, which is implemented with functionalities directed at professional translators or crowd-sourcing platforms (e. g., Lengoo, Unbabel).

Apart from the quality of the technology itself, seamless integration within existing localisation workflows is paramount for its successful adoption, as well as scalability (e. g., KantanMT, Lilt, Tilde), open-source technology (e. g., Pangeanic, Apertium) and speech MT (e. g., Papercup, Tilde). Additionally, privacy and security are of huge interest as texts often include sensitive product or customer information. The lack of understanding of how MT works and the unclear legal rights, obligations and consequences of misuse cause clients to seek secure solutions (e. g., Across, Language Weaver, Pangeanic, Tilde).

There are numerous European companies providing MT tools and services, each with their own strengths and limitations. However, it is tech giants such as Amazon, Facebook, Google, Microsoft who set the standards and best practices for LT development and provision. Most such companies are headquartered outside Europe and so have business and societal objectives that do not always align with European needs and goals. The dominance of those global companies exposes Europe's lack of market power which results in increasing market disparities.

The absence of a clear roadmap and support for LT at the European level results in a disjointed European market with disparate support for the language communities of Europe. Such a roadmap is crucially important now that MT is playing a key role in communication activities across the globe. As a result, the demand for translated content has reached an all-time high, but seems set to rise for the foreseeable future.

Nowadays, there are countless online MT sites for general use that offer access to MT either from companies that make the systems freely available with some usage restrictions (Amazon, Google, Microsoft, DeepL, and Tilde among others) or from public bodies that facilitate their custom-based MT capabilities (the European Commission and the Basque and Latvian governments, among others, Skadins et al. 2020). People use these tools to translate a very diverse range of texts. While access is fast and straightforward, they do present privacy risks and cultural bias. To this day, the legal boundaries of text ownership and use are not fully regulated across Europe. Also, the array of languages available is increasing, but it is the major languages that benefit from the advances first and foremost, with small and minority languages often suffering from uneven and generally low quality.

MT has been available to the video game localisation industry for years without much success given the need for highly creative and culturally adapted options, often with constraints dictated, for example, by available on-screen space. For current

online collaborative games, in-game dialogue has become critical, as has the need for instant translation between multiple languages. This has motivated some game developers to explore the potential of MT in their localisation processes.

Medical translation is highly sensitive and requires the utmost precision. Given the serious consequences of mistranslations, MT has been largely absent from this area. However, it is time to push for MT accuracy and consistency, and accept nothing short of high-quality translation (Haddow et al. 2021). MT could prove of great assistance not only for written text but also in doctor-patient communication. While medical interpreters remain the go-to specialists, often their services are not available. To facilitate this type of communication, systems that can specifically tackle the local languages and those of the immigrants are essential. There are now a number of success stories that demonstrate the utility of MT in this field. For example, in 2020 SDL made their MT system available to all engaged in COVID-19 medical research;[7] NAVER LABS Europe released an MT model for COVID-19 research;[8] and, to make emergency and crisis-related content available in as many languages as possible, Translators without Borders and several academic and industry partners prepared COVID-19 materials for training MT models for nearly 90 languages.[9]

Public Administration – Making legal and administrative documents available in at least the official languages of Europe is an obligation of national governments. Given the intricacies of the texts, MT is not yet central in the translation process. However, several initiatives such as ELRC[10], ELRI[11] and ELG[12] (Rehm et al. 2023) have curated and shared LRs that can improve MT services. Along the same lines, the availability of high-quality NMT at different levels of public bodies, Member States and public administrations has been put forward as a key priority for the European Commission, particularly for under-resourced EU languages (see, e. g., the projects NTEU and iADAATPA, Bié et al. 2020; Castilho et al. 2019). An excellent example of the use of MT by EU Council Presidency staff members and public administration translators is demonstrated by the eight EU Council presidencies that used the EU Council Presidency Translator (Metuzale et al. 2020). The challenge is the provision of this type of service not only for the 24 official languages, but for all languages in Europe, promoting citizen equality and European cohesion, which are key to a stable and unified view in the region.

To increase customers' understanding of a product and to build trust, global content on an eCommerce website should be translated into the target customer's language. eCommerce companies require a mix of technical, highly accurate yet informal, creative, and culturally aware translations. While that can be challenging for MT, there are many companies (e. g., Lionbridge, Protranslating, Simultrans, Smartling) that can help online business owners to make their content multilin-

[7] https://www.biospace.com/article/releases/sdl-offers-machine-translation-free-of-charge-to-health-science-professionals-

[8] https://europe.naverlabs.com/blog/a-machine-translation-model-for-covid-19-research

[9] https://tico-19.github.io

[10] https://www.lr-coordination.eu

[11] http://www.elri-project.eu

[12] https://www.european-language-grid.eu

gual, with multiple plugins compatible with common Content Management Systems (CMS) and eCommerce solutions in the market (WordPress, Drupal, Joomla, Magento and WooCommerce).

This short review shows that the current shortcomings of MT technology and areas where effort should concentrate revolve around aspects that help increase trust through increased accuracy, as well as through high cultural adaptation and creativity. It is high time MT quality and suitability are accounted for not only by means of usage-agnostic metrics, but also by customer experience measurements. It is clear that a scenario where all citizens feel equal, with the same quality of language access to resources, services and commerce, will considerably boost European cohesion.

2.2 Main Gaps

Data Availability and Data Quality – As stated in the EU Charter and the Treaty on the EU, all 24 official EU languages are granted equal status. However, the META-NET White Paper Series found that 21 of the 30 European languages investigated were at risk of digital extinction. In addition to the official languages, there are over 60 regional and minority languages, as well as migrant languages and sign languages, spoken by 40 to 50 million people. The negative consequences of this lack of resources are twofold: 1. Europeans are not receiving the digital resources they are entitled to; and 2. there is a lack of language data to train MT engines to mitigate this problem. The Open Data Directive (2019/1024/EU) does not recognise language data as a high-value data category. This means that it may not be clear what language data exists for at-risk languages, or how data can be used for MT/LT development. Moreover, availability does not guarantee usability. To be considered usable, language data must meet certain criteria. For instance, to train high-performance NMT systems, bilingual data needs to be clean and correctly aligned.

Domain-specific Data – NMT systems benefit from exposure to a wide variety of data, including style and content variety. Likewise, while domain specificity is important to tune an engine towards a particular field or subfield, expanding the domain coverage usually brings benefits to the training of an NMT system. This means that domain availability is almost as relevant as language availability. While categories such as legal, financial, and technical are usually well covered in terms of availability and suitability for a number of languages and language pairs, more specific or uncommon domains may not have comparable amounts of training data available. Moreover, there is generally a disparity between publicly available and proprietary bilingual corpora. As a result, there is a gap in the availability of domain-specific language data both in official and minority languages, which could lead to the centralisation of some specialised fields over others, excluding speakers of less supported languages in the long term.

The Compute Divide – With the paradigm shift to NMT, MT has become increasingly computationally intensive. Access to hardware, experts, and involvement in research has also shifted in such a way that elite universities and larger enterprises

have an advantage due to their relative ease of access to compute power. According to the ELE analysis on strategic documents and projects (see Chapter 44, p. 361 ff.), there is a lack of necessary resources (experts, High Performance Computing, capabilities, etc.) in Europe compared to large US and Chinese IT corporations that lead the development of new LT systems. Furthermore, there is an uneven distribution of resources, including scientists, experts, computing facilities, and companies, across countries, regions and languages in Europe (cf. Rehm et al. 2023).

Multimodal MT – MT is commonly thought of as translating text to text, but multimodal MT is also possible, although it is still in its early stages. Fields in which further technological innovation would increase potential use-cases for MT include image recognition, speech synthesis and automatic speech recognition. Image-to-text translation makes use of Optical Character Recognition (OCR) to isolate text in images. This technology is quite effective, and nowadays smartphone and tablet users can generally avail of image translation services free of charge. However, OCR software is not as widespread as standard text-to-text translation. Multiple factors affect OCR accuracy, including coloured or decorative backgrounds, blurred texts, non-Latin alphabets, larger or smaller letters, look-alike characters, and handwritten text, all or any of which may result in nonsensical translations. Combining OCR with text prediction may improve the accuracy of this technology. Audiovisual media is playing an increasingly central role in our lives thanks to AI-powered virtual assistants and online streaming services. For this reason, the ever-growing demand for translation of audiovisual content has sparked interest in the development of MT-centric text-to-speech and speech-to-text applications. Moreover, the need for accessible content in the form of subtitles and audio descriptions for those who are visually impaired, deaf, or hard of hearing has the potential to drive innovation in MT. The Strategic Research Agenda developed by New European Media[13] provides a number of recommendations related to MT, including 1. streamlining the circulation of audiovisual (or video) programs through MT, while humans focus on the quality of work, for example; 2. encouraging synergies and convergence between subtitling and the development of multilingualism or the integration of foreign migrants, for example; 3. developing AI tools for automatic translation from speech to subtitles, and text to/from sign language; and 4. developing AI tools for robust automatic translation of subtitles. Training high-performance MT systems to translate subtitles is particularly challenging. Rigid copyright laws in Europe forbid the use of translations of copyrighted movies and audiovisual material, despite the fact that this may constitute fair use. Compared to technical language, subtitles are often more creative and idiomatic in nature, increasing the difficulty of translation and the need for high volumes of good-quality training data.

Different Types of End Users – The language industry is often faced with pressure to provide discounts when using MT under the premise that MT boosts productivity, allowing linguists to post-edit more words per hour than if they were to translate from scratch. While the advent of MT has allowed translators and linguists to spend less time on repetitive content, productivity gains still depend on several other fac-

[13] https://nem-initiative.org

tors, including the quality of the MT output and the complexity of the content or domain. The pricing pressure often arises from a lack of consideration of these extra factors which make post-editing a more complex task than it initially appears. Providing industry with the resources to better communicate these factors could be a step towards relieving pricing pressure. Furthermore, LT has changed the role of the translator.[14] There tends to be a generational divide in attitudes towards the adoption of MT in translation workflows among linguists, with some older linguists fearing that MT threatens their job security. Younger linguists tend to have more positive dispositions due to proper training in such technologies being included in their higher education courses. However, linguists play an important role in the assessment and continuous improvement of MT engines, because there is no universal way to automatically evaluate MT quality. Therefore, while the role of traditional translators might have changed, demand for linguists has remained high alongside the developments of MT. At the other end of the spectrum, the hype about the advancements of AI and MT might convince people with low levels of expertise into thinking that MT is infallible (for clear demonstrations that the 'human parity' claims were less than watertight, see Läubli et al. 2018; Toral et al. 2018). The wide availability of MT applications coupled with the sometimes deceptive fluency of NMT output may lead users to avail of MT uncritically, without always understanding its pitfalls. Another step in this direction includes educational publications, which address the technical foundations of machine learning as used in MT as well as the ethical, societal, and professional implications of its use (Kenny 2022).

Automated Evaluation of MT – Automated metrics are a cost-effective way of assessing the quality of MT output. Research in the field focuses heavily on developing metrics that are able to show higher and higher correlations with human judgement. As a result, different metrics are presented at conferences around the world every year. Despite (or as a result of) their abundance, there is still a lack of agreement among the MT community on a single metric which can be used universally to assess the quality of MT engines prior to deployment. Adopting a single metric as a standard would possibly allow for a widespread benchmarking of MT across Europe.

Bilingual Evaluation Understudy (BLEU, Papineni et al. 2002), for example, has enjoyed perhaps the broadest use in the MT industry, despite its known shortcomings with regards to neural MT. Many other metrics have been developed since BLEU, and while they all have their pros and cons, the widespread use of BLEU has proven that metrics can serve a purpose without being scientifically infallible.

Licensing – Translation memory and terminology data is often licensed for non-commercial use only. When commercial licences do exist, their prices are often prohibitively high. This acts as a major barrier to SMEs developing MT applications, especially when there is a limited amount of data available.

Copyright – Copyright laws pose a further barrier in Europe. While copyright law is subject to fair-use exceptions in countries such as the US, European law is far less flexible, and severely restricts the use of parts of copyright works for purposes such as data mining. If lawmakers could agree that using aligned translations of

[14] We use the word *linguist* to refer to language professionals who translate, post-edit, and evaluate LT among other tasks

copyrighted data constitutes fair use, as far as it in no way impairs the value of the materials and does not curtail the profits reasonably expected by the owner, LT stakeholders could avail of this high-quality language data for the immediate benefit of European language communities.

Legislative and Adoption Gaps – Despite the widespread celebration of multilingualism in the EU, there is no common policy addressing language barriers as of yet. We now provide a few examples of scenarios where multilingualism acts as a barrier to people in times of crisis. It is fair to say that current legislation does not account for these scenarios, resulting in critical gaps in services for communities in the EU. Adopting MT in these areas could mitigate the difficulty sometimes caused by language barriers, strengthening the position of multilingualism as a facet of European identity. 1. the COVID-19 pandemic has shown the need for rapid dissemination of information and guidelines in times of crisis. To give one example, in Ireland, the provision of multilingual information was seen to be slow, and reactive, with even the provision of information in Irish and Irish Sign Language being slow in the early stages. The first recommendation made (O'Brien et al. 2021) is for state departments to implement a coordinated approach to the provision of translated content in crises; 2. the requirement for all translations of personal documents to be stamped by a sworn translator can increase the stress on civilians, adding costs and waiting times. The repetitive nature of documents like these as well as their standardised terminology are particularly well-suited to MT; 3. just as the Audiovisual Media Services Directive boosted demand for text-to-speech and speech-to-text technologies, there could be an increase in the demand for MT if policies necessitating the translation of certain audiovisual material into all 24 official languages were introduced. While EU law requires that the product descriptions of goods sold within the EU be translated into the Member State's official language, as of yet there are no such regulations regarding product descriptions for cross-border eCommerce; 4. there is a gap in publicly available MT services which cater specifically to the needs of people in Europe. Users can globally avail of free-of-charge MT services but the multinationals who provide the services could withdraw or start charging for them at any time. Moreover, they do not cater specifically to the needs of European citizens.

Training NMT engines is resource intensive and has a heavy carbon footprint. One area where the law is perhaps too relaxed is in relation to carbon emissions in the field of AI research and development. Researchers have warned of the marginal performance gains associated with expensive compute time and non-trivial carbon emissions. Strubell et al. (2019) recommend that time spent retraining should be reported for NLP learning models and that researchers should prioritise developing efficient models and hardware. The EU has the opportunity to be a pioneer in training and developing green LT by following and enforcing these recommendations.

3 The Future of the Area

In this last section, we will examine the contribution of MT to DLE (Section 3.1), briefly sketch the main breakthroughs needed (Section 3.2), discuss our main technology development goals and visions (Section 3.3) and describe the next steps towards Deep NLU (Section 3.4).

3.1 Contribution to Digital Language Equality

Nowadays, due to globalisation, MT is essential for the development of society. People can access MT allowing for the democratisation of information in many languages. MT directly impacts the economy and cultural exchange between countries. In various scenarios, human translators cannot meet the huge demand for translations in a short time and at low cost. In such cases, MT is much faster and may require less effort to post-edit than translating from scratch.

Massive amounts of parallel data are required to build solid MT systems. Parallel data creation is costly in terms of time and resources. We contend that work done for or by public administrations might offer a solution in this regard. The NEC TM project,[15] for example, calculated in its market study that European public administrations spend about 300 million Euros p. a. in translation contracts with language vendors. This parallel data is mostly not requested back by institutions, many of which operate in low-resource languages, but it should be made publicly available. Data availability directly affects the availability and quality of MT, as well as the contribution it can make to DLE and the wider society. These data pipelines can improve local (national) technology, raise awareness of the fact that citizens are also data producers, and improve and increase the availability and quality of MT. For example, in the case of Catalan, having co-official status (in three Spanish regions) kickstarted a series of administrative decisions that facilitated the creation of more and more parallel data, which has been utilised by local MT companies. Societies that care about data sovereignty and establish language data policies can facilitate the growth of LT companies, which in turn can positively impact those societies.

Uses of MT are very varied, from customer reviews on travel sites to legal document translation for public administrations. None of those uses and the business intelligence that can be derived from them can happen without translation. MT not only works for equality on dispute resolution or as a source of information for insights at scale irrespective of the source, but also enables businesses to build on those services, impacting the society they belong to. We cannot separate the use and availability of the technology from its societal impact.

The ubiquity of MT services is an indisputable fact of current European digital societies. It is now embedded in many services as a real-time high-quality commodity. The ELE consortium has identified several day-to-day uses which illustrate how

[15] https://www.nec-tm.eu

MT is used in very different spheres, including: 1. civil servants verify the national legislation of other EU Member States by machine-translating it; 2. citizens communicate via MT when visiting other countries; 3. the general public use MT to understand social media conversations; 4. students machine-translate research papers; 5. eCommerce websites offer products online to consumers in multiple languages; and 6. public administrations translate documentation for information exchange.

All these use-cases generate massive amounts of online data, that is not reused by EU businesses and research groups. Worse still, it can happen that it is generated for the benefit of the (non-European) free online tools providers to make their technology more accurate. Access to massive amounts of data that is freely available and provided by general users has scaled a lot of MT research, whilst it has provided little in terms of open-source, generally available resources.

Whilst the majority of the talent in NLP and AI has been European, large-scale developments are foreign to the EU or the result of private sponsorship. Heavy investment in MT research at universities over the years has created the know-how and technical knowledge which has only rarely been exploited commercially (e. g., KantanMT, Iconic). The question for Europeans remains on the privacy of the data used and how this data is transmitted. The MT landscape is dominated by large non-European players and technology companies. DeepL is the only significant EU-based provider, being sponsored by a German initiative born as a result of parallel text data collection over many years (Linguee). Most European MT companies remain fairly small and have much less impact (visibility) on society beyond professional-level usage. The EU's own service (eTranslation) is available for free to public administrations and it also opened its services to SMEs in 2021.

A good example of increasing concerns comes from Switzerland, where DeepL and Google Translate were recently banned at Swiss Post as external tools amid concerns of privacy and data exploitation (access was later reopened, though). Swiss Post declared that its staff should only use its own MT technology, so no private data or data belonging to the organisation would be sent to third parties.[16] GDPR has the potential to change things as privacy concerns become relevant to institutions and enterprises, with EU projects such as MAPA[17] providing accurate, open-source anonymisation for public administrations. It remains to be seen how this potential is exploited so that MT and general NLP solutions permeate and help create a more data-based Europe, based on intelligent solutions with the citizen at its core.

3.2 Breakthroughs Needed

According to a competitiveness analysis ordered by the European Commission, the position of the European MT market, as compared to that of North America and Asia, is excellent for research and innovation, while it lags behind in terms of in-

[16] https://slator.com/swiss-post-bans-deepl-backs-down-after-staff-uproar/

[17] https://mapa-project.eu

vestment, infrastructure and industry implementation (Vasiljevs et al. 2019). At the same time, the study highlights that the market is fragmented, which causes serious issues for the level of intensity at which LT research can be conducted. While in North America and Asia resources can be allocated to only a limited number of languages, in Europe, resources must be distributed across a multitude of official and unofficial EU languages. As a result, the scale at which European research can be conducted is limited. Considering the massive infrastructure that is required to train very large state-of-the-art MT/LT systems, Europe starts with a systemic handicap. Looking forward to 2030, we expect the movement towards more efficient and real-time translation to continue. Europe's strong foundation in research and innovation can compensate for the disadvantage European organisations have with respect to infrastructure, provided that a concerted effort is undertaken in researching the development of new hardware platforms and AI training paradigms.

For Europe, a breakthrough in these fields is needed to remain on par with the rest of the world. Breakthroughs in the development of hardware platforms and training paradigms are also warranted by several EU policies. Through the European Green Deal[18] and the Horizon Europe Work Programme (European Commission 2021), the European Commission has committed to making "Europe the world's first climate-neutral continent by 2050", i. e., the economy must be transformed with the aim of climate neutrality. More efficient AI infrastructure can help in reducing the amounts of energy that are required for data storage and algorithm training. If we want MT to become ubiquitous, especially in embedded devices, the hardware on which it runs must be scaled down and the models that run on it must be adapted accordingly. Such adaptation must occur with a minimal loss of quality, while increasing translation speed and reducing power consumption. To achieve this, a breakthrough in MT hardware and software codesign is required; both need to be developed in cooperation to ensure that the capabilities of the hardware are aligned with the needs of MT training and inference.

An equally fundamental breakthrough is needed in the understanding of how our current algorithms work. Many NLP systems today are based on large pre-trained language models which have demonstrated outstanding results on different tasks. However, a boost in performance comes with a cost in efficiency and interpretability, which "is a major concern in modern Artificial Intelligence and NLP research, as black-box models undermine users' trust in new technologies" (Fomicheva et al. 2021). The EU Coordinated Plan on Artificial Intelligence (ECPAI, European Commission 2018) recognises this problem and advocates the need for trustworthy AI, mainly from the perspective of the end-user, but interpretability and explainability of AI models are also of great importance for the scientific community. If researchers wish to improve their algorithms, they must gain a deeper understanding of what causes models to behave the way they do, in order to prevent models from performing poorly or from acting in a gender- or culturally-biased manner.

The ECPAI correctly states that "[f]urther developments in AI require a well-functioning data ecosystem built on trust, data availability and infrastructure", but

[18] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640

it underestimates the effect that one of its cornerstones has had on data collection in the field. According to the plan, "[GDPR] is the anchor of trust in the single market for data. It has established a new global standard with a strong focus on the rights of individuals, reflecting European values, and is an important element of ensuring trust in AI. […] The Commission would like to encourage the European Data Protection Board to develop guidelines on the issue of the processing of personal data in the context of research. This will facilitate the development of large cross-country research datasets that can be used for AI." (European Commission 2018).

Unfortunately, GDPR has had an adverse effect on a large part of the European LT industry. Stakeholders in data management, publication and collection have come to *incorrectly* assume that all data is personal by default, as an overly cautious measure to comply with GDPR. This is especially true for human language data, since it has no fixed schema indicating when personal details may occur. As a result, expensive legal counsel and tools for anonymisation are applied in situations where they could be avoided or are not necessary at all. In addition, non-European AI companies have been able to operate without GDPR restrictions, which has given them a considerable competitive advantage over EU companies.

Although the ECPAI has foreseen a framework for the free flow of non-personal data in the European Union (European Union 2018b), including the creation of common European data spaces in a number of areas, and a proposal for a directive on the reuse of public sector information (European Union 2018a), the process of obtaining linguistic data that has been created using public funding is currently far too cumbersome and pull-oriented. The data resulting from public procurement procedures has a tendency to remain locked up in privately-owned data silos, while the research community and LT industry must go to great lengths to identify and reconstruct the public part of this data using NLP tools (see, for example, Koehn 2005). A crucial breakthrough could be achieved if existing policy frameworks were adapted to make it mandatory for Member States to make all data in natural language-related workflows publicly available. It is the LT industry's mission to reconstruct human thought processes in an automated way. Human operations on linguistic data such as translation, revision and correction of translations, summarisation, etc. can provide the necessary data points to train AI algorithms to achieve this mission. A policy-inspired push model would be greatly beneficial for the development of all related research domains. As a first step, public service administrators should be made aware of the value of their human workflows. As a second step, the IP resulting from public service workflows should be publicly released by default. Finally, workflow data should be made discoverable in a publication/subscription manner, so it can be easily picked up by interested parties.

Although MT has taken a big leap forward with the advent of neural systems, some types of translation remain very difficult. If we want MT to become pervasive for problematic text types (spreadsheets with tabular data, metadata fields, etc.), the problem of context modelling needs to be addressed. For textual translation, incorporating ontological information may help. Continued development on multilingual lexical resources will be required for this. For multimodal settings, extra-lingual context must be incorporated to improve results. Context modelling is not only required

to deal with short sentences or phrases, but also to obtain more cohesive translation across larger volumes of text. NMT systems have improved over SMT, but have not yet succeeded in efficiently incorporating basic grammatical relations between sentences and paragraphs. Since the majority of human language is produced outside of written texts, extra-lingual cues are often required to decode a message adequately and to translate it correctly. To enable better modelling of multimodal environments, we not only need research into how modalities can enrich one another, but also in how training and test sets can be constructed to achieve better modelling.

In terms of the development of data, two important breakthroughs which must be achieved are 1. the creation of new data sets, and reiteration over existing data sets; and 2. policy support for public data reuse. Ideally, new data annotation efforts should build upon existing work. For example, for document-level NMT this can be done with limited effort, as demonstrated in the WMT19 campaign (Barrault et al. 2019). For video and audio content, it will most definitely require more work, but with existing NLP technology it is not unthinkable that EU Parliament sessions could be semi-automatically linked with related video and audio content to create an annotated corpus that can be used for both building new NMT systems and analysing the contribution of multimodal features towards translation quality.

There are various other fields and areas in which further breakthroughs are needed, some of which are novel methods for document-level MT (with a focus on coherent translations of whole texts and documents), the integration of visual and audio features into MT approaches and engines as well as improved explainability (see Bērziņš et al. 2022). Another field is quantum computing, where more research is needed on how MT, and NLP in general, can be reframed as a quantum computing problem. Current work is still laying the foundation for future developments, because the hardware needed is not available yet. But it is important to note that the first theoretical steps towards reformulating MT and NLP as quantum computing problems have already been made.

3.3 Technology Visions and Development Goals

The strategy of building huge MT models by collecting all available data coming from many different domains (and also languages in current multilingual systems) should be complemented by developing smaller models, too. These small(er) models should be trained using the largest possible set of available information, helping under-resourced languages and domains by appealing to knowledge from higher-resourced ones. One of the current problems is that if this results in a single huge model, most practitioners cannot run the model owing to hardware constraints, so smaller models adapted to particular language pairs and domains need to be made available. This would have several benefits: such models would be easy to integrate and use on any device, provide high-quality translations for all domains and languages, and also be greener by requiring fewer computational resources.

The future publicly available MT systems should be less dependent on large companies, especially those which are not European. The risk is that what is freely available now could (easily) be taken away if those companies – none of them MT companies per se, note – find a way to increase revenue in other directions, so that they deprecate their MT offerings, as has happened with other services provided by these large corporations.

Another challenge of the current systems is represented by various biases in the models, such as gender, racial and ethnic bias (Vanmassenhove et al. 2019). Such biases replicate regrettable patterns of socio-economic domination that are conveyed through language, since these biases are present in the training data and are then amplified by models which tend to choose more frequent patterns and discard rare ones. In the future, ethical and fair MT should not further propagate notions of inequality, but rather foster an inclusive society based on acceptance and respect.

More and more NMT systems are being developed which go beyond the single sentence level (e. g., Lopes et al. 2020), using a variety of different approaches: taking into account source- or target-language context, or both. Another interesting avenue being pursued is that different context spans have been investigated, ranging from a single preceding sentence to the entire 'document'. While this might be straightforward for news articles and user reviews, the situation is different for literary texts or movie subtitles, to name but two. Future systems should be able to identify which sentences benefit from the availability of context, and then find that context. This task is far from trivial because relevant information can be found in different places, sometimes even beyond the given text, such as the topic of the text, the gender of the writer/speaker, or even general world knowledge.

Such external information can go beyond text data and include images, videos, tables, etc. by developing multimodal MT systems (Yao and Wan 2020). Such systems currently include image information to help in the translation of image captions. Future systems should combine sources of information which go beyond this, so that an image of a product can help disambiguate words in the description or review of the said product, for example. Multimodal models should also include sign language translation, which currently relies mainly on computer vision methods. Sign language MT should use models based on both images and natural language.

Training data, crucial to building models, should receive more attention. Currently, the majority of MT systems are trained on large amounts of data covering only a small amount of languages, language pairs and domains. While progress in MT is mainly measured under high-resource conditions, the majority of domains and languages, including many of those spoken in Europe, are under- or low-resourced. Future systems should be able to cover all European languages as well as language pairs (not always including English or some other higher-resourced language), and be trained on many different domains and genres. For this to work for all – and not only for big companies and leading research teams – the availability and quality of training data should be increased. Attention should also be given to languages where there is no written tradition, in which case spoken-language data needs to be sourced.

While techniques such as multilingual models, unsupervised MT, synthetic data, and transfer learning are all helping, if there is not enough good-quality data for

a language (pair), then such methods will not reach the goal of high-quality MT, in which case novel methods and research breakthrough will be needed in this direction.

The test sets used for assessing MT systems should receive more attention, too. Currently, a large number of research publications use news articles coming from shared tasks. Researchers test their systems on these texts and report improved automatic scores. However, some of the human translations in these test sets used as references for automatic scores are of poor quality (Toral et al. 2018). The shared task organisers cannot be blamed for this situation, as they do the best that they can with the limited budgets that they have. Still, these human translations should be thoroughly examined in order to discard the inappropriate ones and keep only the good ones for long-term testing. Note that in light of the comparison between MT outputs and human translations carried out in recent years where claims of "human parity" have been investigated, the quality of human translations used in MT evaluation has to be high (Toral et al. 2018; Läubli et al. 2018).

In addition, other test sets coming from different genres and domains need to be more widely used. A vast amount of systems are currently tested only on a limited set of domains, news being the predominant one, while many genres and domains are as yet hardly covered by current research, such as user-generated content (which itself is not a homogeneous genre), despite having great potential for future growth. In the long run, we strongly contend that MT systems should be tested on a large number of different domains and genres, and for an ever-increasing range of languages in order to help facilitate DLE. In this regard, the rise of NMT and its increasing quality have led to more and more challenge test sets (or test suites). These specified test sets enable better understanding of certain (linguistic) aspects which cannot be properly assessed in standard 'natural' test sets. The development and creation of such test sets necessitate a large amount of human expertise, time and effort. In the future, they should be easy and fast to create for any language pair.

As for the evaluation process itself, automatic metrics remain invaluable tools for the rapid development and comparison of MT systems. They have been developed and improved constantly, with more and more metrics coming onstream each year. However, a number of challenges remain. Perhaps the most significant is that the community still relies to a large extent on BLEU, despite there being a large body of research pointing out its drawbacks. Future systems should be evaluated by new metrics which represent better approximations of human judgments and also ideally abandon the dependence on human reference translations, which is a serious limitation. Recently, more and more metrics based on neural networks and/or word representations have emerged which show better correlation with human judgment and do not require reference translations. However, these metrics have another limitation: they require labelled training data which as we have pointed out are available only for a limited number of language pairs and domains. Future automatic metrics should be equally valid without such constraints. In addition, all future automatic metrics should be able to evaluate MT output taking the context into account in order to be more reliable (Läubli et al. 2018; Castilho 2021).

Manual evaluation of translation quality, despite its disadvantages (time- and resource-intensive, as well as being subjective), remains the gold standard, both for

evaluating MT systems and for developing suitable automatic metrics. That being said, the design of experiments and the standard method of reporting the results is far from perfect. Different papers use the same quality criterion name with different definitions, or the same definition with different names. Furthermore, many papers do not use any particular criterion, asking the evaluators only to assess "how good" the output is. We assert that any idea of a single standard general unspecified notion of quality should be abandoned, and factors like the context in which MT is to be used together with appropriate quality aspects should be considered, as pointed out by Way (2013) and Mason (2019). These aspects might include adequacy/accuracy, readability/comprehension, appropriate register, correct terminology, or adequately fulfilling a particular task. Consequently, metrics should be created with such criteria designed in from the outset, and not only to provide a general unspecified score which is meaningless to most people.

Furthermore, recent research has found that readers tend to fully trust fluent translations as well as comprehensible translations even if they contain severe adequacy errors which change the actual content and deliver completely different information (Popović 2020; Martindale et al. 2021). Therefore, future automatic metrics should provide confidence indicators for translations in order to inform users about the level of trust they should have in the MT output they are reading.

Allowing users to interact naturally with machines via speech has the potential to greatly transform, enhance and empower work, leisure and social experiences. The increasing quality of MT and the expanding preference (especially among younger users) for voice-based interaction with devices points to more and more applications for speech-to-text and speech-to-speech translation. This means, of course, not only that spoken language input should become more and more a topic of close attention, but also that more data of exactly the right type needs to be available. By 2030, it is likely that the Automatic Speech Recognition-MT-Speech Synthesis pipeline will have been replaced by more direct approaches which model spoken language translation as an end-to-end process (Gangi et al. 2019), but clearly more work needs to be done in this regard.

Sign language translation should be widely available for many domains to break down language barriers for deaf and hearing-impaired users so that they can access information like the rest of society. For this to be done properly, sign language translation needs to include language features in addition to image features. In addition, it should not only be translated from/into text but also from/into speech.

It is more and more the case that MT is being used for expanding other NLP tasks (e. g., text classification, topic modelling, sentiment analysis) to multiple languages. Usually, full translation is carried out and then the labels for the original source language together with the translations are used for training classifiers in the new target language. However, for such tasks, where the translated text is not used directly, quality criteria might be rather different, and full translation might not be necessary. Extracting different representations from various layers could be even better suited for certain tasks, so this option should be made easily available in future MT systems.

3.4 Towards Deep Natural Language Understanding

Applying a purpose- and communication-oriented view on MT allows us to discuss the extent to which MT needs (deep) NLU, since it helps to put the prevailing MT-related metrics – not related to purpose and communication aspects – in perspective. Accordingly, claims related to MT reaching parity with human translations are misleading since the metrics to measure this via reference translation data are too limited to address whether the intended communication has fulfilled its purpose when this is related to reader impression and style.

With a view on communication success, it becomes obvious that MT – core technology, evaluation methodologies, metrics and data for training and evaluation – needs NLP that goes beyond traditional capabilities such as detection of terms, keywords, labels, entities, relations, and sentiments. These capabilities – often referred to as 'deep' NLU – will be aware of context and able to consider annotations/metadata. Context and annotation awareness will allow MT to generate texts that are faithful to the intended communication (input view), take translation purpose/specifications/requirements into account (sender view), and show consideration of the reader/listener (output/consumer view).

Only MT with deep NLU will, for example, be able to efficiently support a human-to-human or human-to-machine conversation that exhibits qualities like being contextualised, adaptive, personalised, and knowledge-rich. The following ingredients currently seem to emerge as important elements for next-generation MT (based on Deep NLU): 1. existing standards related to annotations; 2. the FAIR data principles as backbones of investment protection, and 'responsible MT'; 3. experts like translators, domain specialists, modellers, data scientists for curation; 4. more open, standardised, flexible and robust technologies for all dimensions of data management; and 5. large, multilingual translation models that are safe to use and can easily be adapted for resource-sparse computing environments, to specific tasks and domains, and for low-resource languages.

4 Summary and Conclusions

Nowadays MT is widely used by the general public, public sector and government agencies, SMEs, LSPs and many other industries. This will continue to grow, covering new application areas to support Europe's digital single market as well as DLE. Looking forward to 2030, we expect the movement towards deep NLU to enable efficient, real-time translation to support human-to-human or human-to-machine communication.

Despite the widespread celebration of multilingualism in the EU, there is no common policy addressing language barriers. So far, the absence of a clear roadmap and support for LT at European level has led to an incohesive, fragmented European market with disparate language support for the language communities of Europe. We hope that the ELE SRIA (Chapter 45) will have positive effects in this regard.

There is also a gap in publicly available MT services which cater specifically to the needs of people in Europe. Users around the world avail of free-of-charge MT services provided by global companies. The risk is that what is freely available now could (easily) be taken away if those companies find a way to increase revenue in other directions. The future publicly available MT systems should not depend on non-European multinationals.

With the help of neural networks, MT has recently improved significantly in its quality, consistency and productivity. However, in many cases the focus of new technologies is still on well-resourced languages, limiting diversity and reinforcing existing disparities. Furthermore, explainable and interpretable machine learning is attracting more and more attention, and a fundamental breakthrough is needed in the understanding of how current MT algorithms work.

The increasing quality of MT and the expanding preference for voice-based interaction points to applications for speech-to-speech translation and multimodal MT in order to break the language barrier for human communication.

Publicly available multilingual data should include a greater diversity of domains and languages, so that building high-quality MT systems becomes an option for all. Collection of usable language data is particularly important. If lawmakers could agree that using aligned translations of copyrighted data constitutes fair use, LT stakeholders could immediately avail of this high-quality language data. There is also a disparity between publicly available and proprietary bilingual data. A crucial breakthrough could be achieved if policy frameworks make it mandatory for Member States to make all data in natural language-related workflows publicly available.

Increased attention should be paid to the human judgments used for tailoring the automatic metrics, as well as to manual evaluation in general. There is also a lack of necessary resources (experts, HPC capabilities, etc.) compared to large US and Chinese IT corporations. There is also an uneven distribution of resources across countries, regions and languages.

Finally, the hardware on which MT runs must be scaled down. By ensuring that the capabilities of the hardware are aligned with the needs of MT training and inference models, smaller models would be easy to integrate and use on any device and also be greener by requiring fewer resources. The EU has the opportunity to be a pioneer in green LT by developing efficient models and hardware.

At the level of policies/instruments, much more synchronisation of activities between national and international bodies is necessary. A desirable approach for the efficient and homogeneous implementation of policies towards DLE would be more equal support for all EU languages, including equal involvement of national research communities.

References

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018). "Unsupervised Statistical Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, Brussels, Belgium, October 31 - November 4, 2018. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Association for Computational Linguistics, pp. 3632–3642. https://doi.org/10.18653/v1/d18-1399.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1409.0473.

Barrault, Loïc, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri (2019). "Findings of the 2019 Conference on Machine Translation (WMT19)". In: *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*. Ed. by Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Marco Turchi, and Karin Verspoor. Association for Computational Linguistics, pp. 1–61. https://doi.org/10.18653/v1/w19-5301.

Bērziņš, Aivars, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, and Andy Way (2022). *Deliverable D2.13 Technology Deep Dive – Machine Translation*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/MT-deep-dive.pdf.

Bié, Laurent, Aleix Cerdà-i-Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis (2020). "Neural Translation for the European Union (NTEU) Project". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 477–478. https://aclanthology.org/2020.eamt-1.60.

Castilho, Sheila (2021). "Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Online: Association for Computational Linguistics, pp. 34–45. https://www.aclweb.org/anthology/2021.humeval-1.4.

Castilho, Sheila, Natália Resende, Federico Gaspari, Andy Way, Tony O'Dowd, Marek Mazur, Manuel Herranz, Alex Helle, Gema Ramírez-Sánchez, Víctor Sánchez-Cartagena, Mārcis Pinnis, and Valters Šics (2019). "Large-scale Machine Translation Evaluation of the iADAATPA Project". In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, pp. 179–185.

Dou, Zi-Yi, Antonios Anastasopoulos, and Graham Neubig (2020). "Dynamic Data Selection and Weighting for Iterative Back-Translation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, pp. 5894–5904. https://doi.org/10.18653/v1/2020.emnlp-main.475.

European Commission (2018). *Coordinated Plan on Artificial Intelligence*. COM(2018) 795 final. https://digital-strategy.ec.europa.eu/en/policies/plan-ai.

European Commission (2021). *Horizon Europe Work Programme 2021-2022*. European Commission Decision C(2021)4200 of 15 June 2021. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents.

European Union (2018a). *Proposal for a Directive of the European Parliament and of the Council on the re-use of public sector information (recast), COM(2018) 234 final*.

European Union (2018b). *Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union*.

Fomicheva, Marina, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao (2021). "The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results". In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 165–178.

Gangi, Mattia Antonino Di, Matteo Negri, Roldano Cattoni, Roberto Dessì, and Marco Turchi (2019). "Enhancing Transformer for End-to-end Speech-to-Text Translation". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*. Ed. by Mikel L. Forcada, Andy Way, Barry Haddow, and Rico Sennrich. European Association for Machine Translation, pp. 21–31. https://aclanthology.org/W19-6603/.

Haddow, Barry, Alexandra Birch, and Kenneth Heafield (2021). "Machine Translation in Healthcare". In: *The Routledge Handbook of Translation and Health*. Routledge, pp. 108–129.

Han, Jesse Michael, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever (2021). "Unsupervised Neural Machine Translation with Generative Language Models Only". In: *CoRR* abs/2110.05448. https://arxiv.org/abs/2110.05448.

Kenny, Dorothy, ed. (2022). *MultiTraiNMT: Machine Translation for Multilingual Citizens*. In preparation. Berlin: Language Science Press.

Koehn, Philipp (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pp. 79–86. https://aclanthology.org/2005.mtsummit-papers.11.

Läubli, Samuel, Rico Sennrich, and Martin Volk (2018). "Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation". In: *Proceedings of EMNLP*. Brussels, Belgium, pp. 4791–4796.

Libovický, Jindřich, Helmut Schmid, and Alexander Fraser (2022). "Why don't people use character-level machine translation?" In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 2470–2485.

Lopes, António V., M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins (2020). "Document-level Neural MT: A Systematic Comparison". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*. Ed. by Mikel L. Forcada, André Martins, Helena Moniz, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof Arenas, Mary Nurminen, Lena Marg, Sara Fumega, Bruno Martins, Fernando Batista, Luísa Coheur, Carla Parra Escartín, and Isabel Trancoso. European Association for Machine Translation, pp. 225–234. https://aclanthology.org/2020.eamt-1.24/.

Ma, Shuming, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei (2021). "DeltaLM: Encoder-Decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders". In: *CoRR* abs/2106.13736. https://arxiv.org/abs/2106.13736.

Martindale, Marianna, Kevin Duh, and Marine Carpuat (2021). "Machine Translation Believability". In: *Proceedings of the First Workshop on Bridging Human – Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, pp. 88–95. https://aclanthology.org/2021.hcinlp-1.14.

Mason, Sarah Bawa (2019). "Joss Moorkens, Sheila Castilho, Federico Gaspari, Stephen Doherty (eds): Translation quality assessment: from principles to practice – Machine Translation: Technologies and Applications, Volume 1, Springer International Publishing, Heidelberg & Berlin, xii + 287 pp, ISBN 978-3-319-91240-0 (hardcover), 978-3-030-08206-2 (paperback), 978-3-319-91241-7 (eBook)". In: *Machine Translation* 33.3, pp. 269–277. https://doi.org/10.1007/s10590-019-09241-w.

Metuzale, Kristine, Alexandra Soska, and Marcis Pinnis (2020). "A Tale of Eight Countries or the EU Council Presidency Translator in Retrospect". In: *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, AMTA 2020 - Volume 2: User Papers, Virtual, October, 2020*. Ed. by Janice Campbell, Dmitriy Genzel, Ben Huyck, and Patricia

O'Neill-Brown. Association for Machine Translation in the Americas, pp. 525–546. https://aclanthology.org/2020.amta-user.25/.

Nguyen, Khanh, Hal Daumé III, and Jordan Boyd-Graber (2017). "Reinforcement Learning for Bandit Neural Machine Translation with Simulated Human Feedback". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1464–1474. https://www.aclweb.org/anthology/D17-1153.

O'Brien, Sharon, Patrick Cadwell, and Alicia Zajdel (2021). *Communicating COVID-19: Translation and Trust in Ireland's Response to the Pandemic*. Tech. rep. School of Applied Language and Intercultural Studies, Dublin City University. https://www.dcu.ie/sites/default/files/inline-files/covid_report_compressed.pdf.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, pp. 311–318. https://aclanthology.org/P02-1040/.

Pinnis, Mārcis, Stephan Busemann, Arturs Vasilevskis, and Josef van Genabith (2021). "The German EU Council Presidency Translator". In: *KI – Künstliche Intelligenz*.

Popel, Martin (2018). "Machine Translation Using Syntactic Analysis". PhD thesis. Praha, Czechia: MFF UK.

Popović, Maja (2020). "Relations between comprehensibility and adequacy errors in machine translation output". In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 256–264.

Rehm, Georg, Katrin Marheinecke, Rémi Calizzano, and Penny Labropoulou (2023). "Language Technology Companies, Research Organisations and Projects". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 171–185.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016a). "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. https://doi.org/10.18653/v1/p16-1009.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016b). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. https://doi.org/10.18653/v1/p16-1162.

Skadins, Raivis, Marcis Pinnis, Arturs Vasilevskis, Andrejs Vasiljevs, Valters Sics, Roberts Rozis, and Andis Lagzdins (2020). "Language Technology Platform for Public Administration". In: *Human Language Technologies – The Baltic Perspective*. Ed. by Utka Andrius, Vaicenoniene Jurgita, Kovalevskaite Jolantai, and Kalinauskaite Danguole. Vol. 328. FAIA. IOS Press, pp. 182–190.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 3645–3650. https://doi.org/10.18653/v1/p19-1355.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*, pp. 3104–3112.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way (2018). "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". In: *Proceedings of WMT*. Brussels, Belgium, pp. 113–123.

Vanmassenhove, Eva, Dimitar Sht. Shterionov, and Andy Way (2019). "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation". In: *Proceedings of Machine Trans-*

lation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019. Ed. by Mikel L. Forcada, Andy Way, Barry Haddow, and Rico Sennrich. European Association for Machine Translation, pp. 222–232. https://aclanthology.org/W19-6622/.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem*. DOI: 10.2759/142151. https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

Way, Andy (2013). "Traditional and Emerging Use-Cases for Machine Translation". In: *Proceedings of Translating and the Computer*. Vol. 35. London.

Yang, Jian, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei (2021). "Multilingual Machine Translation Systems from Microsoft for WMT21 Shared Task". In: *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*. Ed. by Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Tom Kocmi, André Martins, Makoto Morishita, and Christof Monz. Association for Computational Linguistics, pp. 446–455. https://aclanthology.org/2021.wmt-1.54.

Yao, Shaowei and Xiaojun Wan (2020). "Multimodal Transformer for Multimodal Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4346–4350. https://aclanthology.org/2020.acl-main.400.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich (2020). "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault. Association for Computational Linguistics, pp. 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148.

Chapter 41
Deep Dive Speech Technology

Marcin Skowron, Gerhard Backfried, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernáez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz

Abstract This chapter provides an in-depth account of current research activities and applications in the field of Speech Technology (ST). It discusses technical, scientific, commercial and societal aspects in various ST sub-fields and relates ST to the wider areas of Natural Language Processing and Artificial Intelligence. Furthermore, it outlines breakthroughs needed, main technology visions and provides an outlook towards 2030 as well as a broad view of how ST may fit into and contribute to a wider vision of Deep Natural Language Understanding and Digital Language Equality in Europe. The chapter integrates the views of several companies and institutions involved in research and commercial application of ST.[1]

Marcin Skowron · Gerhard Backfried
HENSOLDT Analytics GmbH, Austria, marcin.skowron@hensoldt.net, gerhard.backfried@hensoldt.net

Marek Kováč · Johan Rohdin · Petr Schwarz
Phonexia, Czech Republic, kovac@phonexia.com, rohdin@phonexia.com, schwarz@phonexia.com

Eva Navas · Inma Hernáez · Jon Sanchez · Ibon Saratxaga
University of the Basque Country, Spain, eva.navas@ehu.eus, inma.hernaez@ehu.eus, jon.sanchez@ehu.eus, ibon.saratxaga@ehu.eus

Aivars Bērziņš
Tilde, Latvia, aivars.berzins@tilde.com

Joachim Van den Bogaert
CROSSLANG, Belgium, joachim.van.den.bogaert@crosslang.com

Franciska de Jong
CLARIN ERIC, The Netherlands, franciska@clarin.eu

Andrea DeMarco · Michael Rosner
University of Malta, Malta, andrea.demarco@um.edu.mt, mike.rosner@um.edu.mt

Peter Polák
Charles University, Czech Republic, polak@ufal.mff.cuni.cz

[1] This chapter is an abridged version of Backfried et al. (2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_41

1 Introduction

Speech – as the most natural manner for humans to interact with computers – has always attracted enormous interest. Speech Technology (ST) has been a focus of research and commercial activities over the past decades. From humble beginnings in the 1950s, they have come a long way to current state-of-the-art approaches.

Stimulated by a shift towards statistical methods, the 1980s witnessed an era of Hidden-Markov-Models (HMM), Gaussian-Mixture-Models (GMM) and word-based *n*-gram models combined into speech recognition engines employing ever more refined data structures and search algorithms (Jelinek 1998). The availability of data to train these systems was limited to only a few languages, often driven by security and commercial interest. Even then, work on neural networks (NN) was already being carried out and viewed by many as the most promising approach. However, it was not until later (2000s) that the availability of training data paired with advances in algorithms and computing power finally began to unleash the full potential of NN-based ST. Especially over the past couple of decades, ST has evolved dramatically and become omnipresent in many areas of human-machine interaction. Embedded into the wider fields of Artificial Intelligence (AI) and Natural Language Processing (NLP), the expansion and scope of ST and its applications have accelerated further and gained considerable momentum. Recently, these trends were complemented by a paradigm shift related to the rise of language models (Bommasani et al. 2021), such as BERT (Devlin et al. 2019) or GPT-3 (Brown et al. 2020): models trained on a broad scale, adaptable via fine-tuning and able to perform very well on a wide range of tasks. Substantial advances in algorithms and high-performance hardware have led to massively increased adoption and further technological improvements. With speech and natural language forming fundamental pillars of human communication, ST may now even be perceived as "speech-centric AI".

With the emergence of intelligent assistants, ST has become ubiquitous, yet many ST systems can only cope with restricted domains and can be used only with the most widely spoken languages. For languages with a low number of speakers, ST systems are still all but absent or severely limited in their scope. Recent advances in Machine Learning (ML) and ST have begun to enable the creation of models also for such less well-resourced languages. However, these approaches are generally more complex, expensive and less suitable for wide adoption. While recently presented results indicate that novel approaches could indeed be applied to address some of the challenges related to low-resourced languages, the scope of their application and inherent limitations are still the subject of ongoing research (Lai et al. 2021).

STs have been investigated and researched in their own right. However, their full potential often only becomes evident when combined with further technologies forming intelligent systems capable of complex interaction, encompassing a diverse set of contexts and spanning multiple modalities. To the casual user, individual components then become blurred and almost invisible with one overall application acting as the partner within an activity which may otherwise be carried out together with a fellow human being. In this setting, the aggregation of technologies goes beyond narrow and highly specialised systems towards combined and complex systems, pro-

viding a notion of a more general and broader kind of intelligence. Speech and language, as the most natural vehicles for humans to communicate with machines, thus become the gatekeepers to and core of a broader kind of AI.

1.1 Scope of this Deep Dive

The scope of this deep dive encompasses a wide range of STs including language identification, speaker recognition, automatic speech recognition, technologies addressing paralinguistic phenomena as well as text-to-speech. It gathers and synthesises the perspectives of European research and industry stakeholders on the current state of affairs, identifies several main gaps affecting the field, outlines a number of breakthroughs required and presents the technological vision and development goals for the next years. In line with the other deep dives in this book, we adopt a multidimensional approach where both market/commercial as well as research perspectives are considered and concentrate on the following aspects: technologies, models, data, applications and the impact of ST on society. The tendency for the combination of technologies into more powerful systems, encompassing several individual technologies and models, has become apparent and is reflected throughout this chapter.

1.2 Main Components

STs encompass technologies on the recognition as well as production side of speech. They comprise a wide spectrum of sub-fields such as automatic speech recognition (ASR), the identification of language or dialects, speaker recognition/identification (SR/SID), the detection of age and gender, emotions, paralinguistic traits and the production of synthesised speech (often called text-to-speech).

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

Traditional ASR systems consist of components for audio pre-processing, an acoustic model, a pronunciation model as well as a language model defined over units of a lexicon. Within a search algorithm, these elements are combined to produce the most likely transcript given the input audio. In this scheme, models generally are of a generative nature and optimised individually. Since the early 2000s, these components are being replaced with deep neural networks (DNNs). This change was made possible by advances in algorithms and models as well as the massive increase in available training data and computing power (GPUs). As a result, word error rates

(WERs) could be reduced considerably in many domains and languages. However, the performance of ASR systems still varies dramatically depending on the domain and language, with low-resource languages still exhibiting WERs resembling those of English many years ago.

For applications in practice ("ASR in the wild"), hybrid systems combining elements such as HMMs and DNNs still dominate the state of play. As such, they can still be regarded as state-of-the-art outside of research labs. Toolkits like Kaldi provide a sound basis for the development of systems for research as well as commercial environments. Novel approaches in the area of self-supervised learning, e. g., Wav2Vec 2.0 by Facebook (Baevski et al. 2020), focus on leveraging vast amounts of unlabelled data. Latent representations of audio are produced representing speech sounds similar to (sub-)phonemes which are then fed into a Transformer network. This approach has been shown to outperform other typical paths of semi-supervised methods, while also being conceptually simpler to implement and execute. The possibility to employ smaller amounts of labelled data as well as being able to train multilingual models provide strong arguments for such approaches.

Typically, ASR outputs unstructured and normalised text without punctuation marks. This is not problematic in use-cases where the user input is short and concise, e. g., when asking a question to a virtual assistant. However, when generating transcripts for longer speech, it is crucial to restore punctuation to improve readability and provide structure to the transcript. Moreover, punctuation is relevant for further downstream tasks such as named-entity recognition (NER), part-of-speech (POS) tagging and machine translation (MT). Recognition errors introduced by ASR may lead to cascaded errors in these tasks, e. g., for MT (Ruiz et al. 2019).

State-of-the-art SR systems use neural networks to extract a representation (embedding) for the speaker in an utterance. The input to the network typically consists of features extracted from frames of 20-30ms, although there are also ongoing efforts to take the raw waveform as an input. Embeddings are then compared in order to decide whether they are from the same person or not. Typical NN architectures for embedding extraction are TDNN, ResNet, or LSTM. The standard choice of backend is a generative model: Probabilistic Linear Discriminant Analysis (PLDA). Recently, using cosine similarity plus an affine transform has proven to yield competitive performance. An advantage of generative backends is that scoring with different numbers of enrolment utterances becomes trivial. In addition to variations of the embedding extractor architecture, many recent research efforts have focused on the training objective. If the task at hand is verification, the most intuitive manner would be to train the extractor for this task. However, in practice, it often works better to train the extractor for classification. That is, for a training utterance the network should classify who among the speakers in the training set speaks in the utterance.

State-of-the-art language identification (LID) systems are based on DNNs ingesting sequences of frame-level features as input, processing them and applying a pooling mechanism to obtain an utterance level representation which is eventually classified. During training, this whole chain is performed in an end-to-end (E2E) fashion. In testing, either the trained DNN is used directly for classification or the utterance

level representations can be extracted and used in a simple backend for classification, e. g., a Gaussian linear classifier.

In the field of Speech Emotion Recognition (SER), a wide range of methods have been used to extract emotions from signals. Similar to other ST domains, Deep Learning is rapidly becoming the method of choice and several E2E models have been proposed (Tang et al. 2018). Unlike ASR, these have not yet become part of our everyday lives. To achieve this goal, SER systems require more accurately labelled data to improve training accuracy, more powerful hardware to speed up processing, and more powerful algorithms to improve recognition rates. In addition, further insights from fields such as psychology or neurology may be required. Detecting the cognitive states and reactions of a user is a step towards designing proactive systems capable of adapting to the user's needs, preferences and abilities. As in other related ST-fields, the detection of personality traits, mood disorders, signs of depression and other medical conditions has found its application in recent years. Techniques based on automatic processing of the voice signal have been used for language and cognitive assessments. These approaches provide the means for quantifying signal properties relevant for the detection of specific pathologies. Due to the development of automatic methods facilitating the evolving control of a wide population suffering from Alzheimer's disease, a number of industry applications aimed at the detection of neurodegenerative disorders have been introduced.

Neural networks have greatly impacted the speech synthesis field by improving the quality and naturalness of synthetic voices compared to traditional systems and by enabling training in an E2E fashion. While traditional multi-stage pipelines are complex and require extensive domain expertise, E2E systems reduce the complexity by extracting the audio directly from the input text without requiring separate models. E2E text-to-speech (TTS) systems have shown excellent results in terms of audio quality and naturalness. However, they usually suffer from low training efficiency, requiring large sets for training. Full E2E architectures have been proposed, e. g., FastSpeech 2 (Ren et al. 2021). These systems produce spectrograms from text by applying an encoder-decoder architecture that produces a latent representation of the input text (or phonetic transcription) which is subsequently transformed into spectrograms. These systems provide outstanding results in terms of the quality and naturalness of the generated voices but require large amounts of high-quality recordings to be trained properly. Efforts are being made to deploy these systems for low-resource languages by improving data efficiency, applying transfer learning or training multilingual models. Other areas of intense research activity are style transfer, controllable and expressive voice generation, new efficient neural vocoders and speaker adaptation with a reduced amount of data. Regarding expressive speech synthesis, Global Style Tokens (Wang et al. 2018) represent one of the most common approaches. It consists of a reference encoder, encoding the speech Mel-spectrogram, and a style token layer, learning different prosodic aspects in a set of trainable embeddings. The reference embedding is compared with each style token with the help of a sequence-to-sequence multi-head attention module, forming a weighted sum of the style tokens called "style embedding". This style embedding is then concatenated to the text encoder output, thus conditioning the Mel-spectrogram

synthesis on both text and encoded prosody of the speech. Other popular methods include Flowtron (Valle et al. 2021), Mellotron (Valle et al. 2020), and Ctrl-P. Developing high-quality synthetic voices with DNN-based techniques requires large amounts of high-quality recordings from a single speaker. This requirement is often difficult to fulfil, especially for minority languages and dialectal speech. The generation of new synthetic voices is also hindered by this extensive data requirement. Efforts are being made to share data among languages and speakers in order to train the common aspects more robustly. Multi-speaker and multi-language modelling is a common strategy in DNN-based TTS synthesis to achieve improved voice quality with a reduced amount of data from a single speaker. However, the quality of these voices is not yet comparable to those obtained with large databases.

2.2 Main Gaps

While ST has found its way into a series of application fields, various important issues have not been addressed thoroughly and remain active areas of research. In the following, we review the main gaps and present them in the context of global and regional business activities, requirements related to the availability of qualified personnel, privacy and trust concerns, as well as technical and end-user perspectives.

Effects of scale – A trend towards increasingly complex E2E systems can be observed in all areas of ST. Due to the extreme demand on resources, e. g., data, compute, energy, or infrastructure, the construction of such models is limited to a handful of actors. The activities to make pre-trained language models available for transfer learning and fine-tuning and to allow others to also participate in major advances are certainly beneficial. However, the extent of this transfer and level of control in the hands of a few institutions poses a risk to other actors, to the market and potentially even to innovation in the sector as a whole. Compared to the US and China, European players are at a stark disadvantage concerning resources, i. e., data, technology and funding. Academic institutions risk lagging behind industrial research due to a lack of resources and may have to rely on national initiatives to keep up.

Trained personnel and expertise – A further gap, concerning all areas of speech processing, can be identified in the scarcity of trained personnel and expertise as well as the risk of losing emerging talent to innovative power-players outside of Europe (with possibilities and employment conditions which generally cannot be matched by European players). Even in light of the democratisation of technology and auto-ML, allowing a much broader audience to create models and deploy these for use, respective educational programmes in speech (and NLP/LT) technologies form the foundation for future European success in these areas and may hinder it if not appropriately established and strengthened.

Privacy and trust – Data leaks and scandals in recent years have spurred the interest of individuals as well as of policy-makers. Concerns have arisen regarding trust, privacy, intrusion, eavesdropping, or the hidden collection and use of data. These

concerns have been recognised by many actors but are only addressed to a very limited extent, as they often counteract commercial interests.

Technical perspectives – The focus of many ST fields on rather constrained conditions has left gaps in more diverse settings such as: processing of distant speech; noisy environments; accented speech, non-native speech, dialectal speech, code-switching, spontaneous, unplanned speech, emotional speech and connected aspects concerning sentiments expressed; the integration of ST into collaborative environments, multiple, simultaneous speakers engaged in vivid discussions; as well as the integration of paralinguistic aspects and technologies.

Group settings, multiple-user scenarios – While most research focuses on a single user's interactions, STs embodied in virtual assistants are becoming increasingly popular in social spaces. This highlights a gap in our understanding of the opportunities and constraints unique to multiple user scenarios. These include detecting whether users are addressing the system or other participants, speaker diarisation, aspects of social dynamics, and finding interaction barriers. Due to these factors, the usefulness of voice interfaces in group settings is still restricted.

Interdisciplinary research work (Digital Humanities and Social Sciences and the Humanities, SSH) – While the connection to the field of digital humanities and computational social sciences is not firmly established yet, it could be beneficial to set up collaborative links with a range of disciplines and domains working with spoken data. In particular, the insights and requirements stemming from the needs for transcription workflows and audio mining tools of communities producing and (re)using oral history data and interview recordings may help identify gaps in language resources for model training and domain adaptation (Draxler et al. 2020). It could be beneficial to identify imbalances in language-specific support for the recognition, annotation and retrieval of the types of structured conversational speech that are used in interview settings in SSH and beyond (Pessanha and Salah 2022).

Challenges related to an increased modelling power – The increase in modelling power and performance achieved over the last years also comes with some drawbacks and challenges. These include a need for even more data, respectively a lack of interest and work on the creation of new paradigms using less data. Current approaches include shallow and deep fusion, but the question of how to optimally combine language models (LMs) and DNN structures has still not been addressed comprehensively. Models requiring the complete input sequence for processing do not match well with requirements to perform causal processing. Several attempts to enable causal processing are being explored, among them the use of neural transducers running processing at regular intervals. The extent of context may also incur additional processing costs which need to be balanced and mitigated.

Models: interoperability and transparency – Models are not transparent and thus hard to interpret. This is partly due to the fact that previously individual components have been combined into single models. The complex process of hyper-parameter tuning is often too resource-intensive and thus has not been addressed in many instances. Elements of input/output like byte-pair-encodings (BPE) have been suggested but these contradict the idea of genuine E2E processing. Integration of several components into one model prompts the question of whether further downstream

technologies will also become part of such integrated models. The combination in turn raises questions about the interpretability and transparency of such systems.

Explainability and transparency for critical methods and technologies – While in the last decade, ST research has achieved improvements in terms of performance, progress in terms of understanding of the architectures used and of the nature of the data and task has been limited. This is partly due to the fact that the NNs used in modern systems are harder to understand than the generative models of previous generation systems. It is also due to a lack of interest from the industry and funding agencies to support this type of research. Students are also generally inclined to work on topics that mainly aim at improving performance since this increases their chances of obtaining a well-paid job in the industry after graduation.

End-users' perspective – STs have made a leap in becoming adopted in many settings for commercially attractive languages. Especially the proliferation of intelligent Voice Assistants (VAs) has made speech a common mode of interaction. However, several issues limiting the further adoption and widespread use of ST remain: these include problems in accurately recognising accented speech, a lack of trust in VAs to execute more complex or sensitive tasks, and concerns related to privacy and data collection. This issue is further exacerbated by the fact that systems often operate in the cloud rather than on-premise. Many VAs may already be utilised in languages other than English, but coverage and supported functionality vary greatly. The gaps in support create barriers for users whose primary language is not fully catered for, or supported only to a limited extent, forcing them to communicate in a non-native language or risk being excluded from using the ever more popular systems and services. This way, non-native users are pushed to develop different strategies and modes of interaction, including a reduced level of language production and more frequent use of visual feedback.

Data: availability, diversity – The main challenge related to data concerns its availability, i. e., adequate datasets for low-resource languages of an appropriate amount and quality. Various efforts aim to mitigate this fact by focusing on transfer learning and fine-tuning of models. However, whereas this approach is certainly beneficial, it generally does not yield models of equal performance as for languages equipped with large amounts of training data. The lack of data for low-resource languages effectively excludes certain approaches from being applied.

Data: diversity of voices – Some public databases available to train DNN-based TTS systems are only useful for building monolingual neutral voices for a number of major languages. The availability of open data free of restrictions such as copyright and limitations due to GDPR regulations in the remaining major languages and all minority languages would allow the development of TTS systems for these languages too. Databases with more expressive and spontaneous recordings are needed to build TTS systems suitable for more emotion-demanding applications like audiobook reading, movie dubbing and HCI. The vast majority of datasets correspond to adult voices and there is a lack of data to generate child and elderly voices. As the voice is an important component of our identity, more diverse datasets are needed to generate personalised voices that can suit any user.

Accuracy: reaching usable thresholds for applications – The single most frequently mentioned hindering factor for the broad adoption of ST is one that has been mentioned for the past 40 years, namely accuracy. The perceived accuracy and its exact meaning have changed dramatically: from individual words being misrecognised to intentions that are not correctly interpreted in complex situations. For example, WER as an evaluation measure has had its merits in measuring progress in ASR (and still does). However, more comprehensive approaches to measuring the impact of ASR performance on downstream tasks and actual deployments may require novel measures. WER alone clearly does not provide the full picture when it comes to the perceived performance and usability of complete systems comprising several kinds of STs and LTs. Regarding TTS, accuracy translates to a lack of naturalness and robustness of the synthesised speech. Different approaches have been taken, some of them focused on designing robust attention mechanisms, others including alignment information at the input, or substituting the attention mechanism with networks that can predict the estimated duration of the input phonemes. However, the problem has not been solved completely yet and keeps hindering the practical application of TTS systems in many instances. For SR, technologies have already reached acceptable performance for many applications. However, this does not mean that there is no need or opportunity for further research. All applications of SR would benefit from better core performance and increased robustness to different acoustic conditions and other variables occurring in real-world speech data.

Dialectal speech and multilingual training – Most ST systems process speech only in the main variety of languages. To date, little attention has been devoted to dialectal speech. Certain STs can be used in languages different from the one(s) they were originally designed for. However, the performance of such systems typically deteriorates. Some progress has been made to make systems more language-independent (e. g., multilingual training, adversarial adaptation), but there is still ample room for improvement. The effectiveness of such approaches for languages that differ substantially from those used in training has not been investigated thoroughly and warrants further work.

3 The Future of the Area

3.1 Contribution to Digital Language Equality

Purely technological systems alone do not exist – they are always embedded in a social context and should thus always be viewed as socio-technical systems. The applications of ST have diverse and multifaceted impacts on several key aspects for societies. Technologies reaching performance levels resembling those of humans may in many aspects lead to a humanisation of technology, ascribing human attributes to system behaviour. Patterns of human-to-human (H2H) interaction may be applied to human-to-machine (H2M) interaction leading to heightened expectations and potentially to subsequent disillusion.

Digital language inequality – The unbalanced availability and quality of ST resources strongly impact the performance of systems for different groups of languages. For languages supported to a lesser extent, performance and accuracy are typically significantly lower compared to resource-rich languages. In extreme cases, selected functionalities or support for such languages may not be available at all. In addition, language varieties, dialects or accents may not be supported or only supported on very limited levels. STs are thus not accessible nor available to everyone on an equal level. The lack of commercial interest in the long tail of "small languages" translates to a significantly slower pace of ST improvements and commercial adoption for the latter group. For native speakers of these languages, these imbalances lead to wider usage of the better-supported major languages, such as English. Motivating speakers to use these major languages more frequently creates a new set of challenges related to handling accented and non-native speech. Compared to the level of service and the support provided for native speakers, this results in lower performance, weakened experience and reduced usability, rendering ST less useful or even useless in the extreme case.

Energy consumption and sustainability – The growing energy consumption required for the ever-expanding amount of data being processed and the tendency towards continuously more complex ST models have become evident since the race for the largest models has been going on. Due to the extreme demand on resources, the generic construction of complex AI, NLP and ST systems is typically limited to a few actors. Surging interest in sustainability may cause actors to reconsider the massive increase in energy consumption that currently often accompanies progress in ST. An opportunity (and marketing advantage) may arise from directing efforts towards the creation of high-performance/low energy-consumption ST, exploring the capacities of E2E or novel direct speech-to-speech systems to lower the energy consumption by avoiding a separate, cascading training of sub-systems.

Labour market – A further economic aspect concerns the impact of ST on automation and as a consequence on the job market as a whole. As technologies such as chatbots are being adopted in pursuit of efficiency, they also perform an increasing number of tasks previously reserved for humans. ST and AI thus blur the boundary between humans and technology leading to shifts in jobs and even entire industries. Clearly, a message of cooperation and support rather than of rivalry and replacement needs to be communicated and acted upon.

Politics and democracy – It has been pointed out that language strongly influences the manner in which we think and argue about political issues. Language causes mental frames to be activated and form our portfolio of ideas. Politicians and influencers have long discovered these mechanisms and are applying them actively to push their respective agendas. Having this central and immediate effect on cognitive mechanisms, linguistic plurality also forms the basis of cognitive plurality and as such plays a fundamental role in securing diverse and democratic values. Limitation to a few individual languages – such as may happen due to limited digital support for certain languages – impoverishes and reduces this variety, the flexibility and spectrum for expression of thoughts and (political) ideas.

Biases and ethical issues – Several ST systems have been shown to be less accurate for female speakers than for males. This is not because women are underrepresented in the training data but more likely due to the properties of female and male voices. Various ethnic groups may be underrepresented in datasets and consequently, performance becomes less accurate. It should also be noted here that being in a group for which a system performs worse can be either an advantage or a disadvantage depending on the application and the type of error the system tends to commit more often (false positives or false negatives). Another ethical concern pertaining to ST is due to possible privacy breaches through mass surveillance. TTS systems have reached a quality level and degree of similarity with the voice of humans that could be used to generate deep-fake voices or voices of deceased persons. Despite this scope for misuse, most of the possible applications of high-quality voices are positive, and people with speech disorders, visual impairment and other disabilities could greatly benefit from them. However, deep-fakes could also be employed for illegal activities such as committing fraud or discrediting people. New regulations and the development of ad hoc legislation are critical to mitigating this pernicious effect. Tools able to detect speech deep-fakes need to be produced, and anti-spoofing techniques that discriminate synthesised from natural speech must be developed in close collaboration with teams working in ST.

Users with special needs – While ASR systems achieve great accuracy on standard speech, they perform poorly on disordered speech and other atypical speech patterns. Personalisation of ASR models, a commonly applied solution to this problem, is usually performed on servers posing problems related to data privacy and data transfer. While on-device personalisation of ASR has recently shown promising results in a home automation domain for users with disordered speech (Tomanek et al. 2021), more research is required to increase performance for these groups of users and provide support for open conversations. TTS is considered an assistive technology and as such, it may contribute to the integration of individuals with visual impairments or learning disabilities. By developing robust TTS systems, these people could enjoy the same advantages as any person without a disability. It also facilitates equal access to education and supports foreigners who may struggle with the language. ST can contribute to the integration of immigrants by making it easier to learn local languages and can help people with literacy issues and pre-literate children to access content presented in written form. ST may also prove helpful in times of aging populations with degrading eyesight. Integrated into virtual assistants, STs are able to provide support to elderly people, assisting them with reminders of appointments and medication needs, providing access to online information and improving both their ability to live by themselves and strengthen their autonomy. Another particular benefit of TTS relates to orally impaired people. Voice is an essential component of our identity that we usually take for granted. However, losing it can affect how others perceive us and our own sense of who we are. TTS technology is able to provide a voice for those who have lost their own via personalisation suiting the characteristics desired by each user.

Privacy and trust – As technologies are entering the homes and offices of users on a broad scale, an enhanced level of attention to privacy concerns, ethics and policy

is essential. Policymakers, policy watchdogs, the media and consumers alike need to assume the role of gatekeepers. Trust is viewed as the main currency and key to the adoption and acceptance of technologies. Scandals and opaque behaviour on the part of ST providers may have detrimental effects. Whenever ST is linked to a person's identity and used for access control or authorisation, the issue of trust becomes especially important. For example, STs are used to authorise access to resources such as a bank account or building. In surveillance applications, it is used for detecting and identifying criminals. In forensics, SR is used for comparing a voice recording from a crime scene with the voice of a suspect or a victim. For voice assistants, SR can be essential to make sure that certain requests are fulfilled only if made by the owner of the respective device or commodity. All of the above applications rely on high-performance and trusted ST, and can benefit tremendously in commercial terms if applied within these contexts. Many applications of ST store audio in the cloud. It is essential to secure guarantees regarding how data is used or will be used in the future by cloud service providers (the risk of leaking always remains). In the long run, the question will be whether any possible breaches, leaks or scandals involving ST will erode trust to a level that users will no longer volunteer to provide their data. Of course, the distrust will be weighed against the commodity of using certain devices and platforms whose terms of use may simply require the user to do so. Opting out may not always be a realistic option.

Unlawful surveillance – A further area of concern is the extent of unlawful surveillance by governments, state agencies or corporations, infringing citizens' rights, liberties, adversely affecting public discourse, democratic values and influencing the political powers (Stahl 2016). The concerns comprise privacy invasion, accountability of intelligence and security services, and the (non-)conformity of mass surveillance activities with fundamental rights (Garrido 2021). Their effects on the social fabric of nations can only be considered and analysed jointly with the rapidly extending technological capacities and the pervasiveness of devices able to capture, process and transmit relevant data. Regardless of the form of government, the growing extent of mass surveillance and especially its unlawful application may lead to the erosion of public trust in governments and state agencies (Westerlund et al. 2021).

3.2 Breakthroughs Needed

In the context of Digital Language Equality (DLE), the main challenges are linked to the inferior support and resources available for less common languages, and a need for improving the performance and capabilities of ST for these languages. The proliferation of ST, including areas with a high potential impact on individuals and large groups of users, also has to be considered in a wider context of policies governing ST and relevant fields and calls for major breakthroughs in terms of explainability for the critical methods and technologies. Policies and governance concerning the use of ST and data – in particular personal data – need to be kept up to date and on par with rapidly developing technologies and applications. In order to democratise

STs and to strengthen their position within LT and AI, the base of users should be widened. An increase in educational programmes, including in general AI, ML, NLP, and inter-disciplinary projects, is necessary for the continuous training of experts in these fields able to draw upon expertise in voice technologies but at the same time also in domain-specific fields, thus forming the links between them.

Training paradigms – For approaches requiring large amounts of annotated data, strategies and frameworks for joint (potentially distributed) data collection, improved annotation, and joint provision are needed. This not only concerns the collection but equally the storage and provision of such resources. A lack of commercial interest needs to be alleviated by public efforts to jump-start and boost efforts in low-resource languages to limit the threat of digital language extinction. From the perspective of data augmentation, the generation and use of synthetic data may provide a complementary alley in the creation or extension of datasets. Efficient use of transfer learning and fine-tuning, as well as work on algorithms and methodologies that use less data or provide more robust models with lower amounts of data, present promising alternatives to relieve the lack-of-data challenge. For specific fields of ST, improved use of unlabelled data in an unsupervised or semi-supervised manner (pre-training, self-supervised training) provides further possibilities (Lai et al. 2021). For several technologies, making better use of the hierarchical structure and relatedness of languages may be beneficial. Methods like one-shot learning or few-shot learning likewise provide promising approaches.

Access to and discoverability of training data – The need for large amounts of data severely limits the possibilities for small companies and niche players to compete and be able to develop their own solutions. A plethora of licensing agreements pose further obstacles to access datasets and resources. Simplification and harmonisation of these mechanisms would be highly beneficial. In the larger context of open data sharing and bringing digital technology to businesses, citizens and public administrations these issues connect with the EU's Digital Europe Programme.

Support for low-resourced languages – To provide first-rate ST in any language, additional high-quality datasets are essential. Creating a wide set may not be feasible in general, but could be achieved at least for several major European languages. New techniques for transfer learning and model adaptation from systems trained for resource-rich languages to systems able to function in languages with more reduced quantities of available data should enable the development of cutting-edge ST systems also for these languages. New architectures allowing the combination of resources from several languages in such a way that their commonalities are learned in a more robust way (by cross-lingual knowledge-sharing) and methods for the creation of multilingual or language-agnostic models which can be applied to a number of different languages are of utmost importance.

Confluence and context information integration – A tendency towards confluence – the combination of technologies and inclusion of a larger context – can be observed and also be assumed to play a more pronounced role in the future. The increased presence of conversational interfaces, a proliferation of chatbots combining ASR, NLP and TTS with an ever-increasing presence of AI in general, has modified not only the technical and commercial landscape but also the expectations of users, which have

been accelerated by increased time spent in home-office setups and virtual meetings. More powerful tools and greater capabilities also prompt the integration of upstream technologies such as summarisation or sentiment analysis with voice technologies. Speech synthesis is bound to become as emotional and persuasive as the human voice itself. Automatic translation may be used to bridge language barriers. Technologies will need to be integrated in a manner allowing for feedback loops and adaptation seamlessly. Models need to be dynamic and methods allowing for dynamic adaptation – learning and unlearning certain features – will need to be developed to account for flexible and continuously changing conditions. Areas of linguistics such as pragmatics or paralinguistics will need to be considered and integrated to a much higher extent to allow for more natural and human-like interaction. Adding emotions and affections into the recipes for HCI, recognising intent and taking into account a broad variety of contexts holds the potential to turn these interactions into truly human-like experiences. The components related to emotional understanding and empathy are especially relevant for systems functioning in social domains, such as healthcare, education, and customer service.

Explainability, transparency and privacy concerns – Trust in STs and in the use of data obtained by interacting with them may become a decisive factor in the adoption of technologies and success of individual market players. An increased interest in the transparency of data use and system functionality can be observed across the board in many areas of ML and AI. A fundamental question to be answered by providers will be where processing is performed and to what extent and purpose data is used to modify models. One end of the spectrum of processing is large, anonymous data-centres spread around the globe, the other is formed by strictly local processing on personal devices. On-premise solutions provided by companies or institutions form an intermediate setting. In all of these setups, the balance between capabilities and the requirements to achieve these capabilities will need to be determined and balanced against ethical concerns and personal and privacy-preserving arguments. The extent and amount of end-user control will be a crucial factor. Approaches like privacy-by-design accompanied by high ethical and legal standards may be determining factors in enabling trust, fostering adoption and leading to economic success.

Performance, robustness and evaluation paradigms – Driven by various national and international evaluations, standard performance measures have been defined on standard test sets. Current measures like the standard WER only take certain performance aspects into account and may need to be reconsidered, extended or complemented. Robustness and generalisability of ST components and models as well as standard evaluation sets for multiple languages and evaluation sets allowing the parallel evaluation of several technologies (all on the same dataset) should be devised. The topics of ageing and recency of data for evaluation sets need to be taken into consideration. In general, evaluation (as well as training) datasets should be viewed more as work in progress than static artefacts. Extension to further languages and language varieties, dialects and speaking conditions likewise should receive further attention to ensuring broad availability and adoption. Another needed innovation is a method for objectively measuring TTS results; such systems are currently assessed by means of subjective evaluations which are time-consuming and laborious.

Outreach – communities, non-experts – Recent years have witnessed an increase in interest in the democratisation of AI. The widespread application of ML and the well-known fact that experts in ML and AI have become scarce resources has led to the desire to empower a wider set of individuals to participate in the creation and use of these technologies. Toolkits and *do-it-yourself modelling* form part of the trend to democratise voice technologies. Approaches like Auto-ML aim to provide access to ML also for non-experts and align with strategies to allow a wider audience to participate in the process. As LTs are aggregated and applied to more complex settings, inter-disciplinary research and activities (for instance) from fields in the social sciences are becoming more relevant and synergies become apparent. Programmes and funding schemes to actively engage these communities and foster inter-disciplinary research would further boost developments.

Alignments with EU policies and policy breakthroughs needed – Copyright legislation is more restrictive in Europe than in other economic regions and countries, e. g., utilising closed captions from TV broadcasts or subtitles from a copyrighted film to train and evaluate ST models could enable access to high-quality language data if lawmakers could agree that training of models on copyrighted data constitutes fair use, as long as it does not diminish the value of the assets or reduce the profits reasonably expected by the owner. The pace of ST development in Europe could be further increased by introducing changes that enable the re-use of existing data, while at the same time ensuring that the value of the copyright owners is not impaired. GDPR introduced a new global standard that places an emphasis on individual rights and reflects European values, and as such contributes to building trust in AI. GDPR has had a *negative* impact on the majority of Europe's LT business and research activities (Smal et al. 2020). Furthermore, non-European AI firms have been able to operate free of GDPR constraints since then, giving them an economic advantage. One of the required breakthroughs relates thus to ensure that while individual rights are protected, the extent of these – in particular, in practical settings and day-to-day operations – does not go beyond the intended scope. Automatic, efficient and free anonymisation tools are required for all European languages.

3.3 Technology Visions and Development Goals

ST: the interface of the future – In many settings, voice provides the most natural way to interact with devices and appliances. The coming years will witness an increased advance in voice technologies to the point that interacting with automated systems will be virtually indistinguishable from communication with human beings in many cases. Interfaces predominately relying on typing, clicking and swiping will gradually transform into multimodal, or fully virtual interfaces including voice, shifting the task of adaptation from human users to computer systems. Compared to the other modalities currently dominating the HCI landscape, communication will encompass richer kinds of (linguistic and paralinguistic) information, including gender, age, emotional or cognitive state, health conditions or speaker-specific traits allow-

ing for more sophisticated and accurate speaker identification, modelling, adaptation and personalisation. These factors and their integration into HCI – as beneficial and powerful as they may be – also give rise to privacy and ethical concerns. They prompt questions of control, user understanding and intent when it comes to sharing information and the extent to which different kinds of information are transmitted and used in the future. Ensuing risks and the potential impact need to be carefully met and balanced with measures to increase security and trust through technical means as well as policy and legislative measures. Striking this balance will affect the adoption of a wide range of devices and services: from VAs in homes and phones, navigation and control systems in cars to cooperative office and work environments and systems supporting a wide range of business and leisure activities.

User and application contexts – A trend towards the integration of richer context is to be expected, regardless of the sub-field of voice processing. This concerns individual technologies and their combination. For TTS, to have a truly interactive experience when dealing with our devices, the integration of context will play a major role. To give just one example, the correct way to pronounce a message should be inferred from the context or the previous steps of a dialogue. Technologies will need to be sensitive to the user's character, state, mood and needs and adapt themselves accordingly. Potentially, they will also need to take into account other participants' states in case of group activities such as business meetings. Topics of pragmatics will be reflected by all technologies. Rather than individual communication turns, complete conversations with history and context will be the norm.

Addressing existing technological gaps – Continued efforts towards better understanding and modelling human speech perception might result in sophisticated ASR addressing several of the limitations and gaps identified in current approaches. Improved handling of audio conditions currently perceived as difficult (e. g., multiple simultaneous speakers in noisy environments speaking spontaneously and highly emotionally in a mix of languages) will be possible thanks to such advances. A wider deployment and further popularisation of ST will require solutions that offer high robustness, low latency, efficient customisation and the ability to provide possible equal support for a diverse set of speakers.

ST integration – An intimate relation of ASR, SID and TTS with downstream Natural Language Understanding (NLU) technologies is needed to allow the correct interpretation of the input. A combination of technologies to interact in multimodal ways (including visuals) and the efficient combination of inter-linked models will be able to guarantee the best experience possible. The successful combination will result in an enhanced easiness and naturalness of use, hiding individual components and allowing systems to be perceived as assistants using natural language much in the way that human assistants would.

Multimodal models – Recently introduced NN architectures support encoding and decoding schemes of various modalities, e. g., Perceiver IO (Jaegle et al. 2021). Despite being task-agnostic, the model provides competitive results on modalities such as language, vision, multimodal data, and point clouds. In the near future, this type of architecture is expected to be used in a range of applications where multimodal content needs to be jointly analysed. Furthermore, a future line of work that can eas-

ily be envisaged is the training of a single, shared NN encoder on several modalities at the same time, and only using modality-specific pre- and post-processors.

Development pace – The pace of development in voice-based technologies is driven by general advances in ML and associated hardware as well as domain-specific advances in speech perception and production. The former can be expected to accelerate even more due to general interest in ML and AI from a wide portfolio of domains. Advances in transfer learning, reinforcement learning, fine-tuning, the use of pre-trained models and components as well as the arrival of platforms such as Hugging Face have created additional momentum. The extension of GPU capabilities can likewise be expected to continue at a fast pace.

Training and evaluation – Further improvements introduced in the process of creation and distribution of ever-growing, ever more coherent and diverse datasets can be expected. These will include large, multilingual, multi-domain and multimodal datasets, which will become de facto standard sets for training and evaluation. We will witness an increase in labelling efficiency, a wider adaptation of continuous learning, self-adaptation and self-modification paradigms. While datasets will continue to grow, the quality and amount of data of high- versus low-resourced languages are unlikely to converge in the short term. The development of more complex and multifaceted datasets calls for more comprehensive evaluation and quality criteria: a shift that would change the focus from an individual technology to an end-user assessment of an experience while conducting a specific task in a non-laboratory environment and within a specific operational and personalised contexts.

Infrastructure, hardware – Extrapolating from the current trends a further rapid increase in the capacities of ST-related hardware and infrastructure can be foreseen (faster communication networks, higher bandwidths). Further popularisation of ST solutions in the context of the Internet of Things (IoT), and a new set of voice-enabled devices will be available to users at work, leisure and commercial settings. These developments create additional challenges related to load and scalability of the underlying infrastructure, hardware and networks. Moving computation to edge devices will also continue to be a trend in the near future.

Privacy, accountability and regulations – The future development of ST and the wider LT field will be strongly influenced by the regulations governing the collection, storage, transmission, and use of personal data. In the context of European AI companies and research institutes, the pace of development appears to be particularly influenced by current regulation schemes. Lawmakers' decisions will thus have to consider the wide and profound impact of their regulations: on the protection of citizens' personal data and privacy on the one hand, and on the wider field of AI technologies and the comparative advantages and disadvantages vis-à-vis other geopolitical regions on the other. Extrapolating from current regulations concerning user privacy, and differences in data collection and use, it seems probable that the divide between the EU and non-EU countries will continue to grow. It is unlikely that a consensus or standardisation between competing regions will be found. With the growing presence of ST and AI in general, increased concerns about hidden flaws, shortcomings and baked-in biases of such systems are gaining momentum. Whereas citizens and academia may work towards enhancing transparency and mechanisms

that may be able to avoid certain phenomena, the industry may work towards obfuscation and hindrance of these mechanisms. A sequence of scandals and growing interest in issues of ethics and privacy have led to an increased awareness in society of this issue. Trust in technology is a key ingredient for the adoption of technologies by a large portion of the population. Transparency in how privacy is integrated into technologies is a crucial ingredient to earning trust. Privacy-by-design beyond mere statements may become a decisive factor for technology uptake and market success.

Disclosure of the use of AI/ST – Due to the ever more human-like nature of ST, the use of AI technologies should be disclosed at the earliest stage possible for all transactions and applications. Making users aware of what they interact with can be regarded as a fundamental step in the creation of more transparency. This will not prevent humans from attributing personhood to machines or hinder human-like communication, but present an ethical and transparent frame around such settings.

Audits of algorithms and models – Auditors will have to be independent for this to make sense and not open the door to even more secretive and evasive behaviour by companies. Federal agencies or boards may be required to preside over such activities. Standard test sets and tests may have to be created and applied.

Impact assessments of the introduction of such technologies – The concept of measuring impact and potential harm is firmly established in fields such as the environment. Similarly, algorithmic impact assessments need to cover a broad range of factors, with ST and NLP focusing on language- and language use-related aspects.

Public repositories of incidents where AI/NLP caused harm – Public repositories and ways to report problematic uses of AI would allow the identification of repeat offenders and act in case of recurring problems. Furthermore, making such cases known publicly may serve as an incentive to correct or prevent them.

Effects on society, workplace – The discussion about which jobs or areas within domains are likely candidates to be replaced by AI carries over to the domain of speech processing – as well as to NLP in general – as they form a core element of AI. Issues concerning automation and job replacement and the ensuing policy-making and social ramifications thus also directly concern ST and their perception.

Pervasiveness – A further spread and ubiquitous presence of voice-based technologies, and wider deployment of ST across a multitude of services and devices due to a reduction in size and integration into wearable and virtual environments can be expected. This may also concern further persons being in the vicinity of such deployments who may be involved indirectly by someone else's use of ST.

Future applications – ST in combination with other NLP and AI technologies will pave the way for intelligent applications with human-like capabilities and the potential for disruptive innovation in various sectors. Intelligent assistants and chatbots currently provide the leading paths towards general and broad adoption. Future applications will be expected to understand a user's intents over sequences of interactions, completely eliminating perceived boundaries between individual technologies. STs are already being used by multiple industries to increase self-service functionalities, reduce average handling time, increase availability and reduce employee costs.

Personalised Voices – Voices for TTS will be generated for any language and be fully customisable. In the same way as we can now personalise avatars in video

games, we will be able to set every aspect of the synthetic voice to suit the characteristics we prefer for each situation. Moreover, TTS technology will extend, and speech will be generated not only from text but also from other input information that could be more convenient for some users who do not have easy access to text or for some situations (e. g., requiring privacy). Multi-modal systems will allow the generation of speech from lip-reading, articulatory data acquired by diverse technologies such as electromyography, permanent magnet articulography and other silent speech interfaces, and even cerebral activity with brain-computer interfaces.

Ambient intelligence – Viewing ST as a means for intelligent interaction, integrating nuanced and fine-grained context and input from multiple modalities can be expected to lead to more human-like systems where the perception of individual components will blur into an overall experience for end-users. Such combinations may be a step towards a broader kind of AI as opposed to the narrow, highly-specialised versions in use today.

3.4 Towards Deep Natural Language Understanding

In many instances, the most natural manner for humans to interact with machines is through voice, for issuing commands or queries as well as generating responses and statements. Certain types of scenarios (e. g., limiting the interaction to small, handheld devices) may call for voice-only interaction, whereas others (e. g., allowing for feedback via large screens, augmented- or virtual-reality environments) may favour multimedia settings, permitting the flow of information across different modalities in parallel. Other scenarios may ask for communication completely without the use of audio, in particular when considering special needs and inclusive communication.

STs play a role in the ingestion of information, by acting as a kind of sensor conveying linguistic as well as paralinguistic inputs and converting them into structured information. Equally, their use concerns the output of information in auditive form (speech, but also non-speech, e. g., confirmations) to communicate with human users. Both directions of the flow of information apply to HCI as well as H2H interaction in the case of groups of human users interacting with each other or with computers, e. g., during meetings with intelligent assistants for transcription, translation and summarisation. STs thus form an intermediate interface layer between humans and machines. Inbound (auditive) information is captured and enriched by ST before being passed on to downstream NLU processing. Outbound information is enriched, transformed and eventually realised as audio based on content, structure and metadata provided by semantic components. The semantics and interpretation of utterances as well as the generation of appropriate responses based on a logical representation and state of a conversation fully reside within the scope and components of NLU and technologies such as dialogue managers (to carry out conversations) or knowledge graphs (networks for semantic representations). As such, STs provide essential contributions to the functioning of NLU in the input and output directions but they do not perform any semantic processing (understanding) themselves.

Visual cues such as gestures or manual articulation (sign language) may replace the audio-element of ST when operating in noisy environments or involving hearing-impaired or deaf people. Visual processing technologies assume the roles of ST in these cases. The combination of modalities is also possible and may be appropriate or imperative depending on the actual context, such as working environments requiring a hands-free operation. The contribution of ST towards achieving deep NLU may thus lie in the improvement and extension of the individual technologies (both from accuracy as well as a language- and domain-coverage perspective), their integration into E2E systems allowing for joint operation and optimisation, including different kinds of knowledge sources and their flexible and dynamic configuration depending on the state and context of an application or user. Approaches including the combination of several modalities for input and output may likewise prove beneficial in the context of achieving deep NLU. In many cases, the real power of NLU will become clear when it is part of a complex system functioning as a human-like counterpart in communication: exhibiting context, history and elements of general intelligence. However, it may also come about that NLU is overshadowed by the cognitive downstream processing and eventually perceived as a mere commodity. The element of admiration and awe on the part of the user will then concern the complete system performance, with NLU itself disappearing in importance as a small part of a much larger and more complex intelligent system.

4 Summary and Conclusions

The substantial advances made in the field of STs over the past decades hold the potential for disruptive innovation in many areas and application domains. Combined with the progress of related fields, they provide the basis for the broad adoption of speech and voice as the primary modality for interacting with computer systems as part of larger and more complex systems modelling human-like communication and interaction. This chapter outlined several research fields and business domains that provide promising areas for the use of ST and their inclusion into larger solutions yielding more natural means of communication. Several issues and challenges have been identified which need to be resolved to make this promise materialise. Below we summarise the key elements identified and provide recommendations for possible future actions. All these strands of progress can aid in supporting the overarching goal of achieving DLE in Europe by providing services made possible by these technologies to larger multilingual audiences at similar levels of scope and performance.

Training data is still a key factor as long as supervised paradigms prevail. Accessibility is often limited, or even locked, with individual actors amassing massive amounts of data, effectively creating monopolies for certain markets. Licences and regulation as well as interoperability and compatibility of data resources and providers remain obstacles that need to be overcome. Methods not relying on vast amounts of data are an active area of research.

Even though the range of languages supported by ST has increased dramatically over the past decades, English still holds a privileged position. The creation of resources for further languages and dialects (some may only be spoken) is ongoing; the investigation of phenomena that are only present in other language families is also an active area of research. The creation of multilingual or language-agnostic models provides further avenues for improvement.

A trend of integrated E2E models into one combined overall model can be observed. Training takes place in a single framework rather than individually, capitalising on joint factors. Considerable progress in performance has been made through this approach which can be expected to continue. The integration of semantic components such as NLU or knowledge graphs into these frameworks may provide additional elements required for intelligent interaction.

In current applications, different components operate in an independent and isolated manner. The dynamic inclusion and integration of context would allow STs to operate on a significantly higher level of accuracy, eliminating errors and narrowing down alternatives. Various ways for the fusion of information have been investigated but have not effectively come to fruition. Parallel systems for multiparty conversations and multimodal approaches may provide ways forward.

STs primarily address the voice modality for interacting with computers. Combining STs with multimodal inputs and outputs may provide a basis for next-generation HCI. The inclusion of gestures, facial expression, emotions or haptics, and the generation of multimodal outputs reflecting these elements may result in a richer and more natural user experience and lead to wider adoption and acceptance of ST.

Although established measures allow quantification of progress in ST, they may only tell part of the story when it comes to real-world applications and the combination with downstream processing. In many fields of ST, performance has reached (near-)human levels under controlled conditions with progress being significant in theory but often only marginal when translated into reality. A shift towards increasing robustness and generality of results may prove beneficial at this stage.

Recent progress and an abundance of ST in chatbots may evoke expectations of ST being a mere commodity and raise unrealistic expectations on the part of users. STs perform considerably worse when applied to conditions unlike those for which they were originally created. Accordingly, adaptation and customisation to special domains provide opportunities for specialists. Expectation management and open communication about the possibilities but also limitations from the ST community may help set expectations to realistic and practical levels.

The interest and concern about fairness and biases of models and ethical issues relating to their use have been receiving increased attention. Methods for detecting biases and de-biasing need to be improved and are expected to become a more active area of development. Furthermore, access to ST for people with disabilities and impairments needs to be extended. Triggered by an increased interest in the fairness of AI systems (e. g., assessments of job applications, prison-parole, credits), applications continue to be subjected to scrutiny. Users demand explanations on the capabilities and functioning of ST. Results are questioned with some application areas demanding audits of models and algorithms. Technical issues need to be

addressed and accompanied on the policy-making and legislative levels. Standardisation of evaluations and publication of results may function as motivating factors for providers to address these issues more thoroughly.

With the current and near-future state of ST, many businesses, political parties and ideological movements may develop conversational agents as a ubiquitous representatives to convey their agenda and sway public opinion to get support for their cause. Situations where the agents' identity is known or hidden should be clearly distinguished. Cases where a company or party is represented by a single conversational agent, or by hundreds or even thousands to create a representation of mass support, should be marked. Scandals, data leaks and an increase in cyber-crime have brought issues of security and privacy to the fore. Devices are ever more pervasive, taking ST into people's offices and homes. IoT and wearables further accelerate this trend. Users are becoming increasingly wary of the risks and undesired effects related to the introduction of ST. Clandestine ways of data collection and eavesdropping infringing privacy are rightly exposed and castigated by the media. Actors risk suffering dire consequences if they do not respond and put corrective measures into place. The balance between convenience and privacy will remain a fluid one to be negotiated repeatedly and on multiple levels.

The legislation governing the acquisition, storage, transmission, and use of personal data has a significant impact on the future of ST and the wider LT area. Extrapolating from current trends, the gap between the regulations used in different regions will continue to widen. As AI technologies play a critical role in creating competitive advantages across a wide range of human activities, it is unlikely that competing countries and regions will be able to reach a broad, far-reaching agreement, resulting in one standardised set of regulations. Lawmakers' decisions will thus have to consider a wide and profound impact of their regulations, on the protection of citizens' personal data and privacy on the one hand, and on the pace of development in the broader field of AI technologies on the other: research, development and application and the comparative advantages and disadvantages vis-à-vis other regions and global centres of AI technology development.

As technologies need to be accepted by society in order to be adopted, advancements as described in this chapter are not exclusively technical ones, but need to be accompanied by progress from the humanities. Multi-disciplinary approaches, as demonstrated by the rise of the digital humanities, may prove advantageous also in these scenarios. As systems become natural companions, the fields of psychology, neuroscience and philosophy bring new aspects and visions to the agenda and inspire novel approaches. Fear and anxieties generated by overly aggressive marketing, science-fiction and disinformation need to be met with prudent transparency, adequate management of expectations and accompanying policy measures. An inclusive approach akin to making ST (and AI) visible, transparent and understandable to a larger public – a kind of AI-literacy in the sense of media-literacy – may be a strong supporting topic for all the above-mentioned domains. People have always tended to humanise machines. Powerful systems formed by the combination and integration of technologies and components described above may effectively be attributed human-like qualities and personhood by their users. Ethical aspects of such interaction must

be addressed in parallel with technological progress. Transparency (e. g., chatbots introducing themselves as machines) and openness are among the key factors to be considered when leaving users a freedom of choice rather than imposing technology on them. This certainly reaches far beyond ST but rather concerns AI in general.

References

Backfried, Gerhard, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz (2022). *Deliverable D2.14 Technology Deep Dive – Speech Technologies*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/speech-deep-dive.pdf.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *NIPS'20: Proc. of the 34th Int. Conf. on Neural Information Processing Systems*, pp. 12449–12460.

Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258 [cs.LG]. https://arxiv.org/abs/2108.07258.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Draxler, Christoph, Henk van den Heuvel, Arjan van Hessen, Silvia Calamai, and Louise Corti (2020). "A CLARIN Transcription Portal for Interview Data". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3353–3359. %7Bhttps://aclanthology.org/2020.lrec-1.411%7D.

Garrido, Miguelángel Verde (2021). "Why a Militantly Democratic Lack of Trust in State Surveillance can Enable Better and More Democratic Security". In: *Trust and Transparency in an Age of Surveillance*. Routledge, pp. 221–240.

Jaegle, Andrew, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira (2021). "Perceiver io: A General Architecture for Structured Inputs & Outputs". In: *arXiv preprint arXiv:2107.14795*.

Jelinek, Frederick (1998). *Statistical Methods for Speech Recognition*. Cambridge: MIT Press.

Lai, Cheng-I Jeff, Yang Zhang, Alexander H Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and Jim Glass (2021). "PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition". In: *Advances in Neural Information Processing Systems* 34, pp. 21256–21272.

Pessanha, Francisca and Almila Akdag Salah (2022). "A Computational Look at Oral History Archives". In: *Journal on Computing and Cultural Heritage* 15.1.

Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu (2021). "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Ruiz, Nicholas, Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico (2019). "Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors". In: *CoRR* abs/1904.10997. arXiv: 1904.10997. http://arxiv.org/abs/1904.10997.

Smal, Lilli, Andrea Lösch, Josef van Genabith, Maria Giagkou, Thierry Declerck, and Stephan Busemann (2020). "Language Data Sharing in European Public Services – Overcoming Obstacles and Creating Sustainable Data Sharing Infrastructures". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association, pp. 3443–3448. https://aclanthology.org/2020.lrec-1.422/.

Stahl, Titus (2016). "Indiscriminate Mass Surveillance and the Public Sphere". In: *Ethics and Information Technology* 18.1, pp. 33–39.

Tang, Dengke, Junlin Zeng, and Ming Li (2018). "An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals". In: *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*. ISCA, pp. 162–166. https://doi.org/10.21437/Interspeech.2018-2581.

Tomanek, Katrin, Françoise Beaufays, Julie Cattiau, Angad Chandorkar, and Khe Chai Sim (2021). "On-Device Personalization of Automatic Speech Recognition Models for Disordered Speech". In: *arXiv preprint arXiv:2106.10259*.

Valle, Rafael, Jason Li, Ryan Prenger, and Bryan Catanzaro (2020). "Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens". In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, pp. 6189–6193.

Valle, Rafael, Kevin J. Shih, Ryan Prenger, and Bryan Catanzaro (2021). "Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Wang, Yuxuan, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous (2018). "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research, pp. 5167–5176. http://proceedings.mlr.press/v80/wang18h.html.

Westerlund, Mika, Diane A Isabelle, and Seppo Leminen (2021). "The Acceptance of Digital Surveillance in an Age of Big Data". In: *Technology Innovation Management Review* 11.3.

Chapter 42
Deep Dive Text Analytics and Natural Language Understanding

Jose Manuel Gómez-Pérez, Andrés García-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiljevs, and Teresa Lynn

Abstract In this chapter, we present a comprehensive overview of text analytics and Natural Language Understanding (NLU) from the perspective of digital language equality (DLE) in Europe. We focus on the research that is currently being undertaken in foundational methods and techniques related to these technologies as well as on the gaps that need to be addressed in order to offer improved text analytics and NLU support across languages. Our analysis includes eight recommendations that address central topics for text analytics and NLU, e. g., the role of language equality for social good, the balance between commercial interests and equal opportunities for society, and incentives to language equality, as well as key technologies like language models and the availability of cross-lingual, cross-modal, and cross-sector datasets and benchmarks.[1]

1 Introduction

Text analytics tools have been in the market for a long time and have proven useful for extracting meaningful information and insights from documents, web pages and social media feeds, among other text sources. Text analysis processes are designed to gain knowledge and support strategic decision-making that leverages the informa-

Jose Manuel Gómez-Pérez · Andrés García-Silva · Cristian Berrio
Expert.AI, Spain, jmgomez@expert.ai, agarcia@expert.ai, cberrio@expert.ai

German Rigau · Aitor Soroa
University of the Basque Country, Spain, german.rigau@ehu.eus, a.soroa@ehu.eus

Christian Lieske · Johannes Hoffart · Felix Sasaki · Daniel Dahlmeier
SAP SE, Germany, christian.lieske@sap.com, johannes.hoffart@sap.com, felix.sasaki@sap.com, daniel.dahlmeier@sap.com

Inguna Skadiņa · Aivars Bērziņš · Andrejs Vasiljevs
Tilde, Latvia, inguna.skadina@tilde.com, aivars.berzins@tilde.com, andrejs.vasiljevs@tilde.com

Teresa Lynn
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie

[1] This chapter is an abridged version of Gomez-Perez et al. (2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_42

tion contained in the text. Typically, such a process starts by extracting relevant data from text that is later used in analytics engines to derive additional insights. Nowadays text analysts have a wide range of accurate features available to help recognise and explore patterns when interacting with large document collections.

Text analysis is an interdisciplinary enterprise involving computer science techniques from machine learning, information retrieval, and particularly natural language processing (NLP). NLP is concerned with the interactions between computers and human (natural) languages, and, in particular, with programming computers to fruitfully process large natural language corpora. Challenges in NLP frequently involve natural language understanding (NLU), natural language generation, connecting language and machine perception, dialogue systems, and their combination.

Recent breakthroughs in deep learning have resulted in impressive progress in NLP. Neural language models like BERT and GPT-3 are able to infer linguistic knowledge from large collections of text that can then be transferred to deal effectively with NLP tasks without requiring too much additional effort. Neural language models have had a positive impact on key tasks of text analytics and NLU, such as syntactic and semantic analysis, entity recognition and relation extraction, text classification, sentiment analysis, machine reading comprehension, text generation, conversational AI, summarisation, and translation, among others.

The success of machine and deep learning has caused a noticeable shift from knowledge-based and human-engineered methods to data-driven architectures in text processing. The text analytics industry has embraced this technology and hybrid tools are emerging nowadays, combining or replacing robust rule-based systems that used to be the norm in the market with machine learning methods. Nevertheless, despite all the hype about data-driven approaches to text processing and particularly Transformer-based language models like BERT (Devlin et al. 2019), which might lead non-experts to think that everything is already solved in text analysis and NLU, many gaps still need to be addressed to make state-of-the-art language technologies (LTs) fully operational and benefit all European languages. Especially relevant is the fact that data-driven approaches require very large amounts of data for training.

Language models have lessened the requirement of labelled data to address downstream tasks, but the need for such data has not disappeared. Beyond general purpose datasets, labelled data is scarce, labour-intensive and thus expensive to produce. Access to labelled data is one of the major hurdles in leveraging data-driven approaches in business applications, and is especially problematic for under-resourced languages for which such data does not exist in sufficient quantities, and there is little interest from technology providers to produce it. Moreover, neural language models work as black boxes that are hard to interpret. This lack of transparency makes it difficult to build trust between human users and system decisions. Lack of explanatory capability is a major obstacle to bringing such technology in domains where regulation demands systems which can justify every decision they make. Furthermore, language models pose ethical challenges including gender and racial biases that are learned from biases present in the data the models are trained on, thus perpetuating social stereotypes.

While the progress made in the last years is undeniably impressive, we are still far from having perfect text analytics and NLU tools that provide appropriate coverage for all European languages, particularly for minority and regional languages. Thus, one of the main goals of this chapter is to outline how the European text analytics industry and research community can address the shortcomings by building on the strengths of current text analytics and NLU tools. We call for human-centric text analysis where people's knowledge, emotions and needs are put at the centre of the design and learning process of the next generation of tools. Other topics in the research agenda are hybrid approaches combining existing rule-based and data-driven systems, multilingualism in text analytics, multimodal analysis of information, and a new generation of benchmarks.

1.1 Scope of this Deep Dive

To better understand how text analytics and NLU technologies are currently being made available to end users, stakeholders and society, we adopt a multidimensional approach where both a market and research perspective are considered, as well as the key domains and applications related to text analytics and NLU. We look at the current service and tool offerings of the main text analytics and NLU providers in the European market. This analysis also includes recent findings in related research areas, such as NLP/NLU, machine learning, and information retrieval, where language understanding tasks that not long ago were the subject of study in research laboratories are now part of the text analytics market. This is as a result of recent breakthroughs in deep learning, structured knowledge graphs and their applications.

Conventional text analytics services available in the market include syntactic analysis, extractive summarisation, key phrase extraction, entity detection and linking, relation extraction, sentiment analysis, extraction of personal identifiable information, language detection, text classification, categorisation, and topic modelling, to name but a few. Also, conversational AI services and tools, including chatbots and virtual agents, are frequently offered under the umbrella of text analytics. More recent additions to the text analytics catalogue are machine reading comprehension services based on tasks such as extractive question answering, which are usually marketed as part of both virtual agents and intelligent search engines to provide exact answers to user questions.

In addition to general-purpose text analytics, we also consider specific domains where such technologies are particularly important. For example, there is a significant number of specific text analytics tools focused on health, including functionalities such as extraction of medical entities, clinical attributes, and relations, as well as entity linking against medical vocabularies. Other use-cases for text analytics tools include customer and employee experience, brand management, recruiting, or contract analysis. An exhaustive account of each sector and use-case, and their relevance for text analytics, is out of scope of this chapter.

Text analytics tools and services are available for widely spoken languages or otherwise strategic languages where the market is big enough for companies to make a profit. Unfortunately, other languages may be less attractive from a business point of view and consequently they are not equally covered by the current text analytics tools. This chapter addresses language coverage as another key dimension for the analysis of text analytics and NLU tools when considering DLE.

We include recent research breakthroughs associated with the text analytics services mentioned above. Many applications of text analytics can be effectively solved using classical machine learning algorithms, like support vector machines, logistic regression or conditional random fields, as well as rule-based systems, especially when there is little or no training data available. However, more sophisticated approaches are needed as we transition towards scenarios involving a deeper understanding of text in order to solve increasingly complex tasks like abstractive summarisation, reading comprehension, recognising textual entailment, or stance detection. Therefore, this chapter puts a special emphasis on deep learning architectures, like Transformer language models, and their extensions.

Of particular interest for language equality are different means to deal with data scarcity for low-resource languages. Self-supervised, weakly supervised, semi-supervised, or distantly supervised algorithms reduce the overall dependence on labeled data, but even with such approaches, there is a need for both sufficient labeled data to evaluate system performance and typically much larger collections of unlabeled data to support data-hungry machine learning techniques. Also in this direction, we include a discussion on hybrid approaches where knowledge graphs and deep learning are used jointly in an effort to produce more robust, generalisable, and explainable tools. Another important area of research that we cover deals with leveraging other modalities of information in addition to text.

All such aspects are considered from the perspective of their combined impact on society. We provide recommendations to address the current limitations of text analytics and NLU technologies in the interest of promoting DLE in Europe.

1.2 Main Components

The goal of text analytics is to discover novel and interesting information in documents and text collections that can be, among others, useful for further analysis or strategic decision-making. Text analytics tools support a wide range of functionalities to process, leverage and curate texts. Most of these functionalities can be broadly categorised into syntactic analysis, information extraction (e. g., key phrases, entities, relations, and personal identifiable information), text classification, sentiment and emotion analysis, and conversational AI functionalities. Recently, question answering, a functionality that requires machine-reading comprehension, has made the transition from research labs to production systems.

The challenges involved in NLP and NLU have different levels of complexity, and as a result, the solution to each of the many challenges is at a different level

of progress. For example, natural language generation is one such challenge, where recent advances like GPT-3 are heralded as a key enabler for a new generation of language applications.[2] Therefore, in addition to functionalities that are already available in the market, there are others which the research community is currently working on. Some advanced functionalities involve reasoning, such as *multi-hop question answering* where systems need to gather information from various parts of the text to answer a question, and *textual entailment*, where the goal is to determine whether a hypothesis is true, false, or undetermined given a premise. Moreover, with the advent of *generative models* like GPT-3, new opportunities have arisen to address hard problems involving text generation, e. g., *abstractive text summarisation*, where the system generates a summary of a text rather than extracting relevant excerpts, or *data to text generation*, where the goal is to generate text descriptions from data contained in tables or JSON documents.

Recently, commercial text analytics providers have started supporting the customisation of functionalities, e. g., users can define classes, entity and relation types, or sentiment scores. This is possible thanks to supervised machine learning making use of user-generated examples. The user only provides examples while the text analytics tool handles all the complexity of the machine learning process. Thus, end users do not need a background in ML to customise their own services. However, some basic knowledge is required to understand how the trained models are evaluated and how to generate a balanced set of examples. The most common customisable text analytics services are classification and entity extraction, but providers typically offer support for sentiment analysis and relation extraction, too. To customise a text classifier users need to provide examples of text labeled with classes, for entity extraction the text is labeled with entity types, for relation extraction relations between entities are indicated, and for sentiment analysis documents are labeled with a sentiment score.

To study the language support of existing text analytics technologies and NLU tools, we look in two main directions: 1. the catalogue of services of global technology providers, which provides us with a notion of what is being currently made available and marketed to the public; and 2. European initiatives that offer repositories of language resources and tools (LRTs), like the European Language Grid (ELG, Rehm 2023). At the time of writing, the ELG catalogue holds more than 11,500 metadata records (Labropoulou et al. 2020), including both data and tools/services, covering almost all European languages.[3] The ELG platform was populated with more than 6,000 additional language resources identified by language informants in the ELE consortium and harvests major EU LRT repositories such as CLARIN[4] and ELRC-SHARE.[5] The observations and figures included in this chapter have been extracted from ELG, which aims at concentrating all available resources, tools and services and making them available in a single platform. Our goal with this chapter is not

[2] https://openai.com/blog/gpt-3-apps/

[3] https://www.european-language-grid.eu

[4] https://www.clarin.eu

[5] https://elrc-share.eu

to provide an exhaustive account, for which such figures could be complemented with additional information from other European infrastructures like the ones mentioned above, but rather to provide an up-to-date indication of the support that each European (and non-European) language enjoys.

For commercial text analytics services, we draw on reports from key players in market intelligence such as Gartner Magic Quadrant for Insight Engines[6] and the Forrester Wave: AI-Based Text Analytics Platforms 2020.[7] A mandatory requirement for providers to be included in this study is for service documentation be publicly available. We study services and languages supported by Azure Text Analytics, IBM Watson, Expert.ai and SAS Visual Text Analytics. In addition, we include other recognised providers, like Amazon Comprehend and Google Natural Language API. To simplify the analysis of the language support we use the following groups:

- A – Official EU Languages (24): Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish
- B – Other European languages; languages from EU candidate countries and Free Trade Partners (11): Albanian, Basque, Catalan, Galician, Icelandic, Norwegian, Scottish Gaelic, Welsh, Serbian, Turkish, Ukrainian
- C – Languages spoken by immigrants in Europe; languages of important trade and political partners (18): Afrikaans, Arabic, Berber, Cebuano, Chinese, Hebrew, Hindi/Urdu, Indonesian, Japanese, Korean, Kurdish, Latin, Malay, Pashto, Persian (Farsi), Russian, Tamil, Vietnamese

A summary of our findings follows. A small set of services including entity extraction, key phrase extraction, and syntactic analysis, offered by global text analytics providers, have a large coverage, above 80%, of EU official languages in category A. Nevertheless, the support of the languages in category A provided by the rest of the services is poorer, ranging from 20% to 45%. The situation of other European languages in category B is actually the worst: the language support of the functional services is scarce or non-existent. Languages in category C also have low coverage across all functional services. In contrast, custom entity extraction has almost perfect support of the languages across all categories. However, custom classification, custom sentiment analysis, and custom relation extraction have a language coverage similar to off-the-shelf text analytics services, covering less than half of the languages in categories A and C, and barely any language at all in category B.

According to the ELG catalogue, syntactic analysis services (language identification, tokenization, etc.) are available for nearly all languages in category A. However, the language support of such services drops to 63% of languages in category B, and 72% in category C. Named entity recognition has moderate support across all language categories reaching 66% for category A, 54% for category B and 61%

[6] https://www.gartner.com/en/documents/3999454

[7] https://www.forrester.com/report/The-Forrester-Wave-AIBased-Text-Analytics-Platforms-Document-Focused-Q2-2020/RES159887

for category C. From there, language support for text analytics services such as keyword extraction, sentiment analysis, summarisation, and entity linking is poor or non-existent in every language category.

Our analysis shows that official EU languages are covered by a subset of text analytics services including syntactic analysis, key phrase extraction, and entity extraction. However, only a small fraction of category A languages are supported by the remaining services. For other European languages in category B, global players offer scarce support or none at all, and for languages in category C support is also low. In ELG the picture changes a little for category B languages since the number of supported languages increases for some of the functional services. However, overall support of languages in categories B and C is still low, i. e., global players plan their offerings based on the volume of the potential market for each language.

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

LRTs have increased and improved since the end of the 1990s, a process further catalysed by the advent of deep learning and neural networks and lately with large pre-trained language models. Today, NLP practitioners find themselves in the midst of a paradigm shift. This revolution has brought noteworthy advances to the field. However, this transformative technology poses problems from a research advancement, environmental, and ethical perspective. Furthermore, it has also laid bare the acute digital inequality that exists between languages. Many sophisticated NLP systems are unintentionally exacerbating this imbalance due to their reliance on vast quantities of data derived mostly from English-language sources. Other languages lag far behind in terms of digital presence. Moreover, the striking asymmetry between official and non-official European languages with respect to available digital resources is worrisome.

Unfortunately, European DLE is failing to keep pace with these rapidly evolving changes. Neural language models and related techniques are key to NLP progress and so being able to build them for target languages with the same quality as English is key if language equality is to be achieved. Now is the moment to seek balance between European languages in the digital realm. There are ample reasons for optimism. Although there is more work that can and must be done, Europe's leading LRT repositories, platforms, libraries, models and benchmarks have begun to make inroads. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models and self-supervised systems opens up the way to leverage NLP for less developed languages.

We are moving from a methodology in which a pipeline of multiple modules was the typical way to implement NLP solutions, to architectures based on complex neural networks trained with vast amounts of data. This rapid progress in NLP has been possible because of different factors: 1. mature deep learning technology; 2. large

amounts of data (including multilingual text data); 3. increase in HPC (GPUs); 4. application of simple but effective self-learning and transfer learning approaches using Transformers. The NLP community is currently engaged in a paradigm shift with the production and exploitation of large, pre-trained Transformer-based language models (Han et al. 2021; Min et al. 2021).

2.2 Main Gaps

We focus on eight main areas related to text analytics and NLU that have an impact on digital language equality: data, legal aspects, limitations, benchmarking, conformance, and domain experts' tooling.

Data – The availability of suitable data for training and evaluating NLP tools is crucial. Unfortunately, current language data for text analytics suffers from several shortcomings. Labelling data can be a lengthy operation that requires skilled domain expertise, which is costly and hard to find. Data and language coverage is a concerning issue as the majority of datasets that are relevant to Europe are general-purpose datasets based on major languages such as English, German, Spanish and French. However, under the EU Digital Europe Programme, new common Data Spaces, including a Language Data Space, will be created. Quality is also important: reliable (misinformation-free), balanced (no bias) and clean content (non-toxic/hate-speech). Machine learning models are notoriously sensitive to bias and noise within datasets. Thus, there is a clear need for reliable bias and toxicity detection tools.

Legal aspects – Since text can often include personal data, data protection and privacy (DPP) policies can put limits on the type of data that can be made available for text analytics. GDPR, the EU's General Data Protection Regulation, while important for EU citizens' protection, significantly hampers language data sourcing and reuse for machine learning-based tools in Europe. The principles of DPP and legal provisions such as GDPR stipulate that data should only be used for a priori defined narrow purposes and that these purposes must be made transparent to the data subject upfront. This proves problematic when dealing with induced models or datasets from web sources that have been reused without website owners' or individuals' consent. European-based researchers and LT developers cannot, therefore, use, share, modify or build upon many of these datasets, which sets DPP-compliant players in this field at a competitive disadvantage.

NLU limitations – Most of today's text analytics solutions are language-specific. Challenges arise in many contexts (business, personal, governmental), where the multilingual requirements of customers and users from across Europe and around the globe need to be served. As we have seen, data availability is already a general problem, but when it comes to lesser-spoken languages with lower amounts of digital content, such scarcity is compounded. Similarly, key pieces of contextual information such as the author, intended audience, societal factors and the purpose of communication also need to be considered. As such, there is much scope for improving contextualised and personalised analytics. One growing area of research is

multimodal NLP, which aims to capture these contextual features to make better judgements or predictions. One priority for many businesses and organisations is to build trust and confidence in AI models. As a result, there has been a notable increase in attention given to the area of explainable AI. In cases where decisions are made based on AI model prediction, it is important that businesses can assess these models' level of accuracy, fairness and transparency. Finally, further exploration is required into extensibility methods to include domain-specific knowledge (e. g., when large corpora are not available), allowing LT providers to easily build custom extensions for machine learning-based systems.

Benchmarking – In language technology (and NLU in particular), a wide range of benchmarking frameworks exists depending on the task at hand. Evaluation metrics also vary depending on the task, ranging from reporting on precision, recall and F1 scores for classification tasks, to exact matching or, say, SacreBLEU[8] scores for dialogue systems. Current NLU benchmarks include widely adopted ones like GLUE and and SuperGLUE.[9] In terms of the nature of datasets used in benchmarking, realistic data is lacking. Therefore, the increasing trend for creating (often general purpose) synthetic data proves to be problematic. Some evaluation datasets are also often criticised in academic shared tasks, where they are sometimes referred to as 'toy' examples that are not applicable to real-world problems. There is a clear need for an increase in diversity, relevance and suitability of annotated test data.

Conformance – A dimension related to standards concerns conformance, namely "the fulfillment of specified requirements by a product, process, or service."[10] While such requirements are not so crucial for academic research, they are highly relevant to enterprise language technology development as they assure quality standards for consumers. Accordingly, requirement statements are needed for any text analytics artefact. For entity detection, this requirement statement could, for example, mention that a conformant application must be able to detect any of the entity types of the Common Locale Data Repository[11] in Spanish and Portuguese.[12] In particular, in the context of regulated industries, certification may need to be considered.

Domain experts tooling – Today, most work in LT based on ML requires expert level skills in tools related to data management, data science and NLP. This creates bottlenecks since it does not allow domain experts (e. g., experts in finance) to become actively involved without extensive tool training or understanding of the underlying technology. This setup causes overhead and delays since work between tool experts and domain experts needs to be coordinated. What is lacking as a way to address this is the availability of consumer-grade, highly usable, low code or no code tools for domain experts. Ideally, such tools should be developed in collaboration with usability specialists, to allow domain experts to play a more active role in the development of solutions for application scenarios they are familiar with.

[8] https://huggingface.co/metrics/sacrebleu

[9] https://gluebenchmark.com, https://super.gluebenchmark.com

[10] https://www.w3.org/TR/qaframe-spec/#specifying-conformance

[11] https://cldr.unicode.org/index/downloads

[12] https://www.w3.org/TR/its20/#conformance and http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#Conformance for sample conformance clauses.

3 The Future of the Area

3.1 Contribution to Digital Language Equality

Today, text analytics tools can help societies and individuals in various ways by supporting tasks that involve the discovery of information (facts, rules, relationships) in text. There are widely-used and indispensable applications available to businesses, consumers, citizens and governments that cover a wide range of usage scenarios, starting from recommendation and sentiment analysis tools to intelligent virtual assistants, business intelligence tools, predictive analytics, fraud management, risk management, and cybercrime prevention. Text analytics tools are also widely used in online and social media data analysis of use to both businesses and governments.

Currently, however, all of these advances and digital innovations are really only supporting major well-resourced languages (i. e., English, French, German, Spanish). Adapting these technologies to support other languages across Europe is not a trivial task of simply localising software or connecting existing technology to local databases or information sources. Languages differ significantly in many ways, not just in words but also inflectional nature (e. g., plural forms of nouns or tenses of verb), sentence structure (word order), idiomatic uses, semantic variability, and so on. To that end, applications need to be built upon systems that understand the underlying patterns in each language that requires support. As today's NLP techniques are increasingly data-driven, this means that sufficient amounts of data need to be made available in order to adapt technologies to these languages. However, even here, it may not be as simple as plugging in new datasets to existing technologies; due to the fact that languages and domains can differ so significantly, various types of parameter tuning, system adaptation or hybrid implementation may also be required to achieve robust and reliable technologies in new languages and scenarios.

Text analytics and NLU can play a major role in overcoming current language and technology barriers that prevent the flow and accessibility of information and knowledge across Europe. From an economic perspective, this language barrier has an impact on the Digital Single Market (European Parliament 2018). Europe's Single Market seeks to guarantee the free movement of goods, capital, services, and people. The role of technology in this is key as countries seek to ensure continued access to this single market, including product information, national and local policies, education information, trade information, financial information, and so on. Such information needs to be accessible to all EU citizens. Text analytics tools (together with machine translation solutions and other cross- and multi-lingual solutions) are key for accessing this information and knowledge across Europe.

The META-NET White Paper Series (Rehm and Uszkoreit 2012) reported on an analysis of LRTs available for EU languages. The results showed that with respect to text analytics, *good support* only applied to English, and *moderate support* to five widely spoken languages: Dutch, French, German, Italian and Spanish. This meant that the other 24 (out of 30) European languages in this study were clustered under *fragmented* as well as *weak or no support*. Today, all 24 official EU languages

benefit from basic tools: tokenizers, lemmatizers, morphological analysers, part-of-speech tagging tools, and syntactic parsers. While the quality, reliability or robustness of these tools vary across languages, their existence represents a step in the right direction. In contrast, more sophisticated tools and services (e. g., summarisation tools) are available only for a small number of languages.

Some of the main reasons that prevent sophisticated text analytics techniques from being available for many EU languages (Rehm et al. 2020) are lack of data and data sparsity (especially for morphologically rich languages) for training and testing text analytics technologies, and the complexity of technology adaptation in low-resource settings. For instance, in the case of dialogue systems and chatbots, analysis of available datasets for dialogue modelling clearly demonstrates a gap for less-resourced languages (Serban et al. 2018; Leonova 2020).

Gartner (2021) forecasts the worldwide AI software revenue to $62.5 billion in 2022, an increase of 21.3% from 2021. Intelligent, AI-based, virtual assistants are already in demand in the digital market and their use in the workplace is growing. Gartner (2020) predicts that by 2025, 50% of knowledge workers will use a virtual assistant on a daily basis, up from 2% in 2019. For the public sector and businesses, this provides an opportunity to use intelligent virtual assistant technology to take care of more repetitive and auxiliary business processes. Gartner (2019) predicts that decision support/augmentation will be the largest area of AI by 2030, accounting for 44% of business value, with agents representing 24%.

For countries with lesser-spoken languages, these predictions only hold if technology exists to support them, of course. If not, an economic divide will emerge, as countries with sufficient language technologies will gain (further) advantage.

3.2 Breakthroughs Needed

Various global enterprises from the US and Asia have started deploying large pre-trained neural language models in production. However, despite their impressive capabilities, large language models raise severe concerns. Currently, we have no clear understanding of how they work, when they fail, and which emergent properties they present. As argued by Bender et al. (2021), it is important to understand the limitations of language models, which they call “stochastic parrots”, and put their success in perspective. There are also worrying shortcomings in the text corpora used to train these Anglo-centric models, ranging from a lack of representation of low-resource languages, to harmful stereotypes, and to the inclusion of personal information. Moreover, these models are costly to train and develop, both financially and environmentally. This also means that only a limited number of organisations with abundant resources in terms of funding, computing capabilities, NLP experts and corpora can currently afford to develop them (Ahmed and Wahed 2020).

To tackle these questions, much more critical interdisciplinary collaboration and research are needed. In Europe there is a lack of necessary resources (experts, data, computing facilities, etc.) compared to large US and Chinese IT enterprises that lead

the development of these new systems. In particular, the computing divide between large firms and non-elite universities increases concerns around bias and fairness within this technology breakthrough, and presents an obstacle towards democratising NLP. In fact, in the EU there is an uneven distribution of resources (funding, open data, language resources, scientists, experts, computing facilities, IT companies, etc.) by country, region and language. We note with concern a tendency to focus on state-of-the-art results exclusively with the help of leaderboards, without encouraging a deeper understanding of the mechanisms by which they are achieved. We believe that such short-term goals can generate misleading conclusions and direct resources away from important efforts that facilitate long-term progress towards efficient, accurate, explainable, ethical and unbiased multilingual language understanding. Progress in these fields will help achieve DLE in Europe in all aspects of society, from government to businesses to the citizens themselves. Next, we focus on some of these key technical areas.

Recent work has shown that pre-trained language models can robustly perform NLP tasks in a few-shot or even in zero-shot fashion when given an adequate task description in its natural language prompt (Brown et al. 2020; Ding et al. 2022). *Prompting* is a technique that involves adding a piece of text (prompt) to the input examples to "encourage" a language model to bring to the surface the implicit knowledge the user is interested in, i. e., guiding the language model to perform the task at hand. Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al. 2021; Sanh et al. 2022; Tafjord and Clark 2021). The application of zero-shot to few-shot transfer learning with multilingual pre-trained language models, prompt learning, and self-supervised systems opens up opportunities for less developed languages in NLP.

Integrating common sense knowledge and reasoning in NLP systems has traditionally been seen as a nearly impossible goal. Now, research interest has sharply increased with the emergence of new benchmarks and language models (Mostafazadeh et al. 2016; Talmor et al. 2019; Sakaguchi et al. 2021; Ma et al. 2021; Lourie et al. 2021). This renewed interest in common sense is encouraged by both the great empirical strengths and limitations of large-scale pre-trained neural language models. This motivates new, relatively under-explored research avenues in common sense knowledge and reasoning. Combining large language models with symbolic approaches (knowledge bases, knowledge graphs), which are often used in large enterprises because they can be easily edited by human experts, is a non-trivial challenge. It is worth investigating ways to leverage structured and unstructured information sources and to enhance contextual representations with structured, human-curated knowledge (Peters et al. 2019; Colon-Hernandez et al. 2021; Lu et al. 2021). Despite perhaps overly optimistic claims of human parity in many tasks, *Natural Language Understanding is still an open research problem* far from being solved since all current approaches have *severe* limitations. Language is grounded in our physical world, as well as in our societal and cultural context. Knowledge about it is required to properly understand natural language (Bender and Koller 2020).

While NLP systems based on deep learning obtain remarkable results on many tasks, the output provided by NLP models, particularly those models that generate text, is still far from perfect. For example, the textual snippets generated by advanced language models such as GPT and successors are formed by syntactically correct sentences that seem to talk about a particular topic, however, there is often a lack of coherence among them and humans still need to monitor and adapt the output of such systems. There is a growing body of research of human-in-the-loop NLP frameworks, where model developers continuously integrate human feedback into the model deployment workflow. These feedback loops cultivate a human-AI partnership that enhances model accuracy and robustness and builds users' trust in NLP systems (Z. J. Wang et al. 2021). In the foreseeable future we expect more such interactions, as AI and NLP become embedded in everyday work processes.

While the NLP community is fully committed to the open-source culture, the aspect of reproducibility has been less of a concern, although the topic is becoming a central one in NLP. Nowadays the majority of scientific articles are accompanied by the source code and data required to reproduce the experiments. Leaderboards such as NLP-progress,[13] Allen Institute of AI leaderboard,[14] Papers with code,[15] or Kaggle[16] encourage participation and facilitate evaluation across many different tasks and datasets. As a result, the NLP community has considerably increased access to publicly available and easily accessible models and datasets. This culture focused towards sharing fosters opportunities for the community to inspect the work of others, iterate, advance upon, and broaden access to the technology, which will in turn strengthen the collective skill sets and knowledge. Open-source libraries such as Transformers[17] may open up these advances to a wider LT community. This library consists of carefully engineered state-of-the art Transformer architectures under a unified API and a curated collection of models (Wolf et al. 2020a). Following up on the success of the Hugging Face platform (Wolf et al. 2020b), the BigScience project took inspiration from scientific creation schemes such as CERN and the LHC, in which open scientific collaborations facilitate the creation of large-scale artefacts that are useful for the entire research community.[18]

3.3 Technology Visions and Development Goals

In this section, we provide an overview of the main technological visions for NLP and NLU, which will contribute to achieving DLE in Europe by 2030. We have identified developments for increasing the language support of such technologies, putting

[13] http://nlpprogress.com

[14] https://leaderboard.allenai.org

[15] https://paperswithcode.com/area/natural-language-processing

[16] https://www.kaggle.com/datasets?tags=13204-NLP

[17] https://huggingface.co

[18] https://bigscience.huggingface.co

users' needs at the centre of any breakthroughs involving language technologies, the integration with other modalities of information in addition to text, the hybridisation of symbolic AI and neural systems, and the need for a new benchmarking approach.

Language support beyond widely spoken languages, including minority and under-resourced languages, is still a pending issue in text analytics and NLU. The investment of LT providers in such languages is inhibited most probably due to a comparatively lower profitability in this space compared to mainstream languages, considering the number of potential users. Nevertheless, the current trend in LT relying on neural language models and research on unsupervised and zero-shot learning opens up new possibilities to increase the coverage of minority and under-resourced languages in the text analytics industry. Language models have shown promising results in zero-shot settings in a wide range of tasks (Radford et al. 2019; Brown et al. 2020; Gao et al. 2021). This is primarily due to the fact that language models learn to perform tasks from patterns occurring in text, eliminating or reducing to a great extent the need for additional labeled data which is a scarce resource for many languages.

Despite their dominance in current NLP pipelines, language models have mainly been addressed as a one-size-fits-all approach, offering almost no customisation beyond the data used to fine-tune (Devlin et al. 2019) or prompt (Brown et al. 2020) models for downstream tasks. Current research focused on unsupervised and zero-shot learning (Gao et al. 2021) delves into this issue since users have little to say in the learning process. Moreover, the data-driven approach and race for accuracy have yielded opaque tools that are hard to interpret, and biased tools that perpetuate social stereotypes related to gender, race and ethnicity in text collections. The lack of transparency makes it difficult to build trust between users and system predictions, having negative consequences for technology adoption. Biased tools have a direct impact on society, especially for marginalised populations (Sheng et al. 2021).

We advocate for a *next generation of language tools that care about end user needs and expectations*, making them part of the design and learning process. These tools will be human-aware, encompass human emotions, and be trustworthy, avoid bias, offer explanations, and respect user privacy. Moreover, human intelligence will be used together with machine learning techniques to produce better LRTs. Human feedback will be a guide in the learning process, informing the machine as to what users want or do not want. Reinforcement learning from human feedback is a promising research avenue (Stiennon et al. 2020; Li et al. 2016) to use human intelligence to improve NLP tools. Also, interactivity with domain experts and users (e. g., Shapira et al. 2021) is a key area for further advances beyond the usual supervised paradigm.

As practitioners come to realise the inevitable limitations of purely end-to-end deep learning approaches, which increase in the case of under-represented languages (both in terms of available language models and suitable training corpora), the *transition to hybrid approaches involving different ways of combining neural and symbolic approaches* becomes an alternative that appears more and more tangible. Therefore, it is important that we exhaustively discuss the components necessary to build such systems, how they need to interact, and how we should evaluate the resulting systems using appropriate benchmarks. The field of neurosymbolic approaches will be increasingly important in order to ensure the integration of existing knowledge bases

within our models, as already shown by approaches like KnowBert (Peters et al. 2019) and K-Adapter (R. Wang et al. 2021), not only to make NLU models aware of the entities contained in a knowledge base and the relations between them from a general point of view, as provided by resources like Wikipedia or Wikidata, but also when it comes to quickly incorporating existing resources from vertical domains and custom organisations into our models in a fast, scalable way. Some, e. g., Sheth et al. (2017) and Shoham (2015), argue that knowledge graphs can enhance both expressivity and reasoning power in machine learning architectures. Others (Gómez-Pérez et al. 2020) propose a working methodology[19] for solving NLP problems that naturally integrate symbolic approaches based on structured knowledge with neural approaches. These are the first practical steps in this direction. Many more are needed, particularly in a multilingual and language equality scenario.

Different modalities can be combined to provide complementary information that may be redundant but can help to convey information more effectively (Palanque and Paternò 2000). For example, multimodal analysis has allowed machines for the first time ever to pass a test from middle school science curricula involving questions where it was necessary for the model to understand both language and diagrams in order to answer such questions (Gomez-Perez and Ortega 2020). This convergence across modalities requires synergies from AI research fields that until now have been conducted individually such as NLP, automatic speech recognition and computer vision. Deep learning techniques will play an important role in multimodal analysis. Recently, Transformer architectures (Devlin et al. 2019), initially proposed for NLP, have been used for image processing (Dosovitskiy et al. 2021) and cross-modal information processing including images and text (Hu and Singh 2021). Other approaches based on contrastive language-image pre-training, like CLIP (Radford et al. 2021), emphasise the relevance of zero and few-shot scenarios. CLIP shows that scaling a simple pre-training task is sufficient to achieve competitive zero-shot performance on a great variety of image classification datasets by leveraging information from text. Unfortunately, such text is in English only, showing how language inequality also impacts language-vision tasks.

Benchmarking aligns research with development, engineering with marketing, and competitors across the industry in pursuit of a clear objective. However, for many NLU tasks evaluation is currently unreliable and biased, with plenty of systems scoring so highly on standard benchmarks that little room is left for researchers who develop better systems to demonstrate their improvements. The recent trend to abandon independent and identically distributed benchmarks in favour of adversarially constructed, out-of-distribution test sets ensures that current models will perform poorly, but ultimately only serves to obscure the abilities that we want our benchmarks to measure. Adversarial data collection, understood as the process in which a human workforce interacts with a model in real time, attempts to produce examples that elicit incorrect predictions, but does not meaningfully address the causes of model failures, as shown, for instance, by Kaushik et al. (2021) for question answering. Restoring a healthy evaluation ecosystem will require significant progress

[19] Methods, resources and technology on Hybrid NLP, https://github.com/expertailab/HybridNLP

in the design of benchmark datasets, the reliability with which they are annotated, their size, and ways in which they handle social bias. This is even more important when we expand our view to the multilingual landscape, such as the European multilingual reality. Furthermore, much more emphasis will need to be given to typical realistic settings (Church et al. 2021), in which large training data for the target task is not available, like few-shot and transfer learning. Moreover, while measuring performance on held-out data is a useful indicator, held-out datasets are often not comprehensive, and contain the same biases as the training data, as illustrated by Rajpurkar et al. (2018) inter alia. Recht et al. (2019) also showed that this can lead to overestimating real-world performance. Approaches like Ribeiro et al. (2020) advocate for a methodology that breaks down potential capability failures into specific behaviours, introducing different test types, such as prediction invariance in the presence of certain perturbations and performance on a set of sanity checks inspired in software engineering. Two requirements must be compulsory for such benchmarks: On the one hand, they will need to cover a representative sample of the key sectors in the European economy, including among others finance, health, tourism, manufacturing, and the corresponding added value chains. In contrast, such benchmarks need to be multilingual by design and cover each economic sector for each of the European languages, guaranteeing language equality regardless of the size of the market associated with each language.

3.4 Towards Deep Natural Language Understanding

Much has been said about the impact of intelligent systems on our lives. Today's large amounts of available data, produced at an increasing pace and in heterogeneous formats and modalities, have stimulated the development of means that extend human cognitive and decision-making capabilities, alleviating such burdens and assisting our drivers, doctors, teachers and scientists. In scientific disciplines like biomedical sciences, some like Kitano (2016) even propose a new grand challenge for this kind of systems: to develop an AI that can make major scientific discoveries that are eventually worthy of a Nobel Prize. This suggests the time is ripe for a shared partnership with machines, where humans can benefit from augmented reasoning and information management capabilities. Through such a partnership, we foresee a virtuous circle of data collection, active learning, and interactive feedback, which will result in adaptive, ever-learning systems.

We have already seen signs of such a partnership, e. g., in the application of generative models like GPT-3 to produce text given a prompt, with applications in different business sectors. Based on these developments, some suggest[20] that the future of AI lies in the development of systems that allow maintaining a conversation with a computer. This scenario should go beyond current and past chatbots, able to copy form without understanding meaning but nevertheless capable of creating a dialogue

[20] https://www.theverge.com/22734662/ai-language-artificial-intelligence-future-models-gpt-3-limitations-bias

with the user. However, this often seems to be missing from AI systems like facial recognition algorithms, which are imposed upon us, or self-driving cars, where the public becomes the test subjects in a potentially dangerous experiment. Language will require advances in knowledge representation, true understanding of meaning and pragmatics, and the ability of models to explain and interpret their predictions in ways that humans can understand and relate to.

The AI community and particularly the areas related to text understanding also need to address issues like fairness in ways that tangibly and directly benefit disadvantaged and misrepresented populations. We have spent large amounts of effort discussing fairness and transparency in our algorithms. At the algorithmic level, fairness has to do with the absence of bias in the models that for example in NLU are used to address tasks that may range from the evaluation of mortgage applications or insurance policies to medical examinations and career recommendations. If algorithms are biased, so are their predictions, in which case inequalities would be perpetuated as AI technologies are deployed more and more in society.

This is essential work. The lack of resources in a specific language to train an NLU model in that language can be seen as another source of discrimination. A very visual example in a related domain has to do with the use of a smartphone navigation app in a wheelchair, only to encounter a stairway along the route. Even the best navigation apps pose major challenges and risks if users cannot customise suggested routes in order to avoid insurmountable obstacles. Similarly, the lack of availability of service functionalities in all languages will have an unwanted effect in the respective populations. Accessibility, education, homelessness, human trafficking, misinformation, and health among others are all areas where AI and text understanding can have a really positive impact on people's quality of life. So far, we have only started to scratch the surface.

4 Summary and Conclusions

We finish this chapter with a list of recommendations and guidelines that address central topics for text analytics and NLU. Among others, we emphasise the role of language equality for social good, the balance between commercial interests and equal opportunities for society, and incentives to help bring about language equality. We also focus on key technologies like neural language models and the availability of multilingual, cross-sectorial datasets and benchmarks.

1. *Language equality in text analytics is a transformative and integrative force for social good* that can stimulate development in such important aspects for our societies as access to health, public administration services for everyone, better education and more business opportunities. These will contribute to more developed societies, which in turn will encourage progress and prosperity, creating new markets for text analytics and other areas related to AI and LT across Europe. However, this is not yet a common scenario for all European languages. The question we should ask ourselves is: what is the alternative? What will the

social cost be if the required policies do not effectively reach *all* European languages until 2030?

2. *The balance between legitimate commercial interests and equal access to opportunities is fragile* when it comes to DLE in text analytics. We have shown how global providers tend to concentrate their offerings and investment in more widespread languages, neglecting a long tail of languages with smaller populations. In contrast, European initiatives such as ELG (Rehm 2023) provide a more equitable coverage. Two reflections emerge. First, it is a European priority to ensure that all European languages are properly covered. Therefore, European companies and also European research organisations in the text analytics space should benefit from incentives that allow them to focus on such languages. Such incentives should naturally come from a thriving market demanding these services in Europe, but also in other forms, like – for companies – tax breaks associated to language services for less represented languages or – for research organisations – specific regional or national funding that can only be used for developing tools or resources for the national or regional language. Second, to create traction this effort should involve European technology providers but also consumers of such services at the different levels of the European public administration and large European companies.
3. *Possible incentives to language equality in text analytics and NLU are not just financial.* Acknowledging that we are working on a particular language conveys the opportunity to stress that research is language-specific. Conversely, neglecting to state that a particular piece of research worked on, say, English language data gives a false veneer of language independence (Bender 2011). Incentives need to be provided for Text Analysis research to cover *all* European languages.
4. *Neural language models are a cornerstone of most NLU and text analytics pipelines now, and this will continue in the next few years.* However, current methods to create such models are hardware-intensive, require vast amounts of text data, and the training comes at the cost of high energy consumption and a large carbon footprint. Because of this, most of the language models available nowadays (like BERT, RoBERTa, T5, GPT-3, etc.) have been trained on general-purpose documents collected from the internet and freely available resources, which hinders their application in vertical domains, requiring additional pre-training on relevant data that is not easy to find.
5. *Data is key.* Without sufficient amounts of good-quality data, language models and text analytics solutions based on ML approaches cannot be trained. However, suitable data and particularly multilingual text is hard to find and expensive to annotate in order to enable subsequent fine-tuning of pre-trained language models on tasks like classification, sentiment analysis, etc. While much progress has been made in creating large-scale labeled data sets for the major languages, it is not yet feasible, especially from a business-driven perspective, to do this for all European languages, let alone the literally thousands of languages spoken on the planet. As suggested in the previous item, there is little or no doubt that enough general-purpose data can be collected in the different European languages that will suffice to pre-train language models for each of our

languages following self-supervised approaches. The problem comes in satisfying the needs of domain- and task-specific data to adapt such models to solve real-life problems in each of the different business sectors and languages.

6. *Data tends to be locked in regulatory and corporate silos.* Research and solutions for LTs that address problems of business and social relevance is underdeveloped. A major reason is that enterprise data is not available to researchers in academia. As enterprise data is by nature confidential and companies need to respect data protection regulations, the barriers for making data available are high. The idea to create data spaces through which companies can make data available under certain terms still needs to crystallise into a dynamic ecosystem that can be compared to generally available text analytics and NLU datasets and models. To address this bottleneck, further collaboration is required between industry, academia and European institutions that facilitates the creation of multilingual text data spaces across the different strategic business sectors. This effort would benefit from an improved balance between European regulations like GDPR and the use of data for research purposes. Currently, companies abiding by GDPR face restrictions and demands that impose some burdens. To be competitive, European companies may need to use neural language models built by third parties in the US or China that are not subject to such regulations.
7. *Benchmarking is inadequate and needs to be fixed and updated.* For many NLU tasks evaluation is currently unreliable and biased, with plenty of systems scoring so highly on standard benchmarks that little room is left for better systems to demonstrate their improvements. The recent trend to abandon traditional, independent and identically distributed benchmarks in favour of adversarially-constructed, out-of-distribution test sets means that current models will perform poorly, and ultimately only obscures the abilities that we want our benchmarks to measure. Restoring a healthy evaluation ecosystem, particularly one involving a vision for DLE, will require significant progress in the design of benchmark datasets, the reliability with which they are annotated, their size, and the ways they handle social bias. However, if we want to make well-grounded progress it is crucial that improved benchmarking considers not only technical but also ethical and societal issues. Benchmark design needs to fit realistic data compositions, rather than synthetic ones within our comfort zone. Addressing such shortage of real-life benchmarks will require significant collaboration between European industry and academia.
8. *Text does not live in isolation. Information is cross-modal.* Text is rarely found in isolation in real-life. Addressing many of the market and societal challenges towards DLE will benefit from taking into account cross-modal scenarios to leverage additional sources of free supervision. Recent advances like OpenAI's CLIP and Meta's Data2Vec[21] seem promising. However, perhaps not surprisingly, all such models are currently available in English only.

[21] https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language

Finally, we would like to emphasise two points that are particularly critical to ensure DLE in Europe. First, *neural language models and related techniques are at the core* of sustaining progress in LT in modern NLP. Therefore, being able to build language models for target languages with the same quality as English is key for language equality. Second, *multilingual data is the key element* to train such models in the target languages. We should not assume that large amounts of publicly available corpora of good quality can be readily obtained for all European languages, but rather the contrary. The effort to ensure that all languages have large amounts of publicly available corpora of good quality, taking into account fairness issues, should be at the centre of any future efforts striving for DLE.

References

Ahmed, Nur and Muntasir Wahed (2020). "The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research". In: *CoRR* abs/2010.15581. https://arxiv.org/abs/2010.15581.

Bender, Emily M. (2011). "On Achieving and Evaluating Language-Independence in NLP". In: *Linguistic Issues in Language Technology* 6.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada, pp. 610–623.

Bender, Emily M. and Alexander Koller (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. https://aclanthology.org/2020.acl-main.463.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Church, Kenneth, Mark Liberman, and Valia Kordoni (2021). "Benchmarking: Past, Present and Future". In: *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Online: Association for Computational Linguistics, pp. 1–7. DOI: 10.18653/v1/2021.bppf-1.1.

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal (2021). "Combining Pre-Trained Language Models and Structured Knowledge". In: *arXiv preprint arXiv:2101.12294*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Ding, Ning, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun (2022). "OpenPrompt: An Open-source Framework for Prompt-learning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demon-*

strations. Dublin, Ireland: Association for Computational Linguistics, pp. 105–113. https://aclanthology.org/2022.acl-demo.10.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs.CV].

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Gao, Tianyu, Adam Fisch, and Danqi Chen (2021). "Making Pre-trained Language Models Better Few-shot Learners". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 3816–3830. https://aclanthology.org/2021.acl-long.295.

Gómez-Pérez, José Manuél, Ronald Denaux, and Andrés García-Silva (2020). *A Practical Guide to Hybrid Natural Language Processing - Combining Neural Models and Knowledge Graphs for NLP*. Springer. DOI: 10.1007/978-3-030-44830-1.

Gomez-Perez, Jose Manuel, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn (2022). *Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/text-analytics-deep-dive.pdf.

Gomez-Perez, Jose Manuel and Raúl Ortega (2020). "ISAAQ – Mastering Textbook Questions with Pre-trained Transformers and Bottom-Up and Top-Down Attention". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5469–5479. https://aclanthology.org/2020.emnlp-main.441.

Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu (2021). "Pre-Trained Models: Past, Present and Future". In: *AI Open* 2, pp. 225–250. https://www.sciencedirect.com/science/article/pii/S2666651021000231.

Hu, Ronghang and Amanpreet Singh (2021). "Transformer is all you need: Multimodal multitask learning with a unified transformer". In: *arXiv preprint arXiv:2102.10772* 2.

Kaushik, Divyansh, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih (2021). "On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6618–6633. https://aclanthology.org/2021.acl-long.517.

Kitano, Hiroaki (2016). "Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery". In: *AI Magazine* 37, pp. 39–49. DOI: 10.1609/aimag.v37i1.2642.

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva (2020). "Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3421–3430. https://www.aclweb.org/anthology/2020.lrec-1.420/.

Leonova, Viktorija (2020). "Review of Non-English Corpora Annotated for Emotion Classification in Text". In: *Databases and Information Systems – 14th International Baltic Conference, DB&IS 2020, Tallinn, Estonia, June 16-19, 2020, Proceedings*.

Li, Jiwei, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao (2016). "Deep Reinforcement Learning for Dialogue Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1192–1202. https://aclanthology.org/D16-1127.

Lourie, Nicholas, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2021). "UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.15, pp. 13480–13488.

Lu, Yinquan, Haonan Lu, Guirong Fu, and Qun Liu (2021). "KELM: Knowledge Enhanced Pre-Trained Language Representations with Message Passing on Hierarchical Relational Graphs". In: *arXiv preprint arXiv:2109.04223*.

Ma, Kaixin, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari (2021). "Knowledge-Driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering". In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, pp. 13507–13515. https://ojs.aaai.org/index.php/AAAI/article/view/17593.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth (2021). "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey". In: *arXiv preprint arXiv:2111.01243*.

Mostafazadeh, Nasrin, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen (2016). "A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 839–849. https://aclanthology.org/N16-1098.

Palanque, Philippe and Fabio Paternò, eds. (2000). *Interactive Systems: Design, Specification, and Verification, 7th International Workshop DSV-IS, Limerick, Ireland, June 5-6, 2000, Proceedings*. DOI: 10.1109/ICSE.2000.870518.

Peters, Matthew E., Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith (2019). "Knowledge Enhanced Contextual Word Representations". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 43–54. https://aclanthology.org/D19-1005.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. PMLR, pp. 8748–8763.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (2018). "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 784–789. https://aclanthology.org/P18-2124.

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2019). "Do ImageNet Classifiers Generalize to ImageNet?" In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach. https://proceedings.mlr.press/v97/recht19a/recht19a.pdf.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh (2020). "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 4902–4912. https://www.aclweb.org/anthology/2020.acl-main.442.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2021). "WinoGrande: An Adversarial Winograd Schema Challenge at Scale". In: *Communications of the ACM* 64.9, pp. 99–106. https://doi.org/10.1145/3474381.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Michael Mckenna, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *ICLR 2022 – Tenth International Conference on Learning Representations*. Online. https://hal.inria.fr/hal-03540072.

Serban, Iulian, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau (2018). "A Survey of Available Corpora for Building Data-Driven Dialogue Systems". In: https://arxiv.org/abs/1512.05742.

Shapira, Ori, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan (2021). "Extending Multi-Document Summarization Evaluation to the Interactive Setting". In: *Proceedings of the 2021 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online, pp. 657–677. DOI: 10.18653/v1/2021.naacl-main.54.

Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng (2021). "Societal Biases in Language Generation: Progress and Challenges". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 4275–4293. https://aclanthology.org/2021.acl-long.330.

Sheth, Amit, Sujan Perera, Sanjaya Wijeratne, and Krishnaprasad Thirunarayan (2017). "Knowledge Will Propel Machine Understanding of Content: Extrapolating from Current Examples". In: *Proceedings of the International Conference on Web Intelligence*. Leipzig, Germany: ACM, pp. 1–9. DOI: 10.1145/3106426.3109448.

Shoham, Yoav (2015). "Why Knowledge Representation Matters". In: *Communications of the ACM* 59.1, pp. 47–49. DOI: 10.1145/2803170.

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano (2020). "Learning to Summarize with Human Feedback". In: *Advances in Neural Information Processing Systems* 33, pp. 3008–3021.

Tafjord, Oyvind and Peter Clark (2021). "General-Purpose Question-Answering with Macaw". In: *ArXiv* abs/2109.02593.

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant (2019). "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158. https://aclanthology.org/N19-1421.

Wang, Ruize, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou (2021). "K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 1405–1418. DOI: 10.18653/v1/2021.findings-acl.121. https://aclanthology.org/2021.findings-acl.121.

Wang, Zijie J., Dongjin Choi, Shenyu Xu, and Diyi Yang (2021). "Putting Humans in the Natural Language Processing Loop: A Survey". In: *Proceedings of the First Workshop on Bridging Human – Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, pp. 47–52. https://aclanthology.org/2021.hcinlp-1.8.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2021). "Finetuned Language Models Are Zero-Shot Learners". In: *arXiv preprint arXiv:2109.01652*. arXiv: 2109.01652 [cs.CL]. https://arxiv.org/abs/2109.01652.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020a). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (2020b). "Transformers: State-of-the-art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6.

Chapter 43
Deep Dive Data and Knowledge

Martin Kaltenböck, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg

Abstract This deep dive on data, knowledge graphs (KGs) and language resources (LRs) is the final of the four technology deep dives, as data as well as related models are the basis for technologies and solutions in the area of Language Technology (LT) for European digital language equality (DLE). This chapter focuses on the data and

Martin Kaltenböck · Artem Revenko
Semantic Web Company, Austria,
martin.kaltenboeck@semantic-web.com, artem.revenko@semantic-web.com

Khalid Choukri · Victoria Arranz
Evaluations and Language Resources Distribution Agency, France,
choukri@elda.org, arranz@elda.org

Svetla Boytcheva
Ontotext, Bulgaria, svetla.boytcheva@ontotext.com

Christian Lieske
SAP SE, Germany, christian.lieske@sap.com

Teresa Lynn
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie

German Rigau · Aritz Farwell · Itziar Aldabe · Ainara Estarrona
University of the Basque Country, Spain, german.rigau@ehu.eus, aritz.farwell@ehu.eus,
itziar.aldabe@ehu.eus, ainara.estarrona@ehu.eus

Maria Heuschkel
Wikimedia Deutschland, Germany, maria.heuschkel@wikimedia.de

Gareth Jones
Bangor University, United Kingdom, g.jones@bangor.ac.uk

Katrin Marheinecke
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany,
katrin.marheinecke@dfki.de

Stelios Piperidis
R. C. “Athena”, Greece, spip@athenarc.gr

Vincent Vandeghinste
Dutch Language Institute, The Netherlands, vincent.vandeghinste@ivdnt.org

Claudia Borg
University of Malta, Malta, claudia.borg@um.edu.mt

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_43

LRs required to achieve full DLE in Europe by 2030. The main components identified – data, KGs, LRs – are explained, and used to analyse the state-of-the-art as well as identify gaps. All of these components need to be tackled in the future, for the widest range of languages possible, from official EU languages to dialects to non-EU languages used in Europe. For all these languages, efficient data collection and sustainable data provision to be facilitated with fair conditions and costs. Specific technologies, methodologies and tools have been identified to enable the implementation of the vision of DLE by 2030. In addition, data-related business models and data-governance models are discussed, as they are considered a prerequisite for a working data economy that stimulates a vibrant LT landscape that can bring about European DLE.[1]

1 Introduction

Digital language equality (DLE) as well as the European data economy rely on the availability, the interoperability and the form of (unstructured, semi-structured, structured) data as a basis for further innovation and improved technological development, especially for trustworthy AI "made in Europe" and powerful language technology (LT) that respects and reflects European values. Data spaces,[2] data sharing and exchange platforms[3] and marketplaces are enablers, key to unleashing the potential of such data. However, data sharing and interoperability are still in their infancy. The diffusion of platforms for data sharing and availability of interoperable datasets is one of the key success factors which may help to drive the European data economy and industrial transformation.

The European Digital Single Market strategy that was adopted on 6 May 2015[4] has been built on three pillars: access, environment, and economy & society. The latter aims at maximising the growth potential of the digital economy, inspired by the 2018 Commission Communication "Towards a common European data space",[5] which provides guidance on B2B data sharing, bringing together data as a key source of innovation and growth from different sectors, countries and disciplines, into a common data space. Overall, the EU has specified its ambition[6] to become the world's most secure and trustable data hub.

This chapter provides insights into: 1. the main components of this deep dive, 2. the current state-of-the-art, 3. the main gaps identified in the field, 4. its contri-

[1] This chapter is an abridged version of Kaltenböck et al. (2022).

[2] Next-generation data acquisition and processing platforms as exemplified, among others, by the BDVA reference model: https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf.

[3] Data sharing and exchange platforms, through which data is commercialised using open data, monetised data and trusted data sharing mechanisms.

[4] https://ec.europa.eu/commission/presscorner/detail/en/IP_15_4919

[5] https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-232-F1-EN-MAIN-PART-1.PDF

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

bution to DLE and the impact on society, 5. an analysis of the main breakthroughs needed in the area of data, language resources (LRs), and knowledge graphs (KGs), 6. the main technology visions and development goals identified to help achieve deep natural language understanding (NLU), all closed by 7. a summary and conclusions section.

1.1 Scope of this Deep Dive

This deep dive covers a relatively wide range of technologies in the area of LT, including machine translation (MT, Chapter 40), speech technologies (Chapter 41), text analytics and NLU (Chapter 42) as well as content management and knowledge management systems, text generation, and language learning systems, as data and LRs are the backbone for all these technologies as well as many more. In addition, the area of KGs plays an important role in this deep dive as KGs provide powerful mechanisms and principles to interlink and enrich data in a high-quality manner. KGs can build a powerful and relatively easy to maintain network of interlinked data – including and combining structured, semi-structured and unstructured data – that can be seen as a crucial element of the data infrastructure required to develop future LT solutions, which require not only a single underlying dataset but in addition a wide range of meaningful and contextualised data. Furthermore, the integrated data models inside of KGs (taxonomies, vocabularies and ontologies) allow the training of algorithms for LT solutions with higher precision requiring smaller amounts of training data.

The topic of metadata and data in this chapter is always related to LT, language understanding and DLE in Europe. Accordingly, metadata and data in this respect concern (mostly, but not exclusively) LRs, (annotated) corpora, translation memories, dictionaries and lexicographic resources, as well as other LRs and relevant data that is required for powerful multilingual LT. Such data and metadata constitute a strong enabler of AI and machine learning (ML), methodologies that have enabled innovative approaches and advances in the field of LT (Elliot et al. 2021).

In addition to these principal components, a number of related methodologies and tools are currently on the rise, and these form part of the technological vision for 2030 in this deep dive. The subject of data-related business models is tackled throughout the chapter, as functioning, sustainable data-related business models are a prerequisite for a thriving data economy and ecosystem that in turn stimulates and fosters those data-related components listed above, to enable a working LT landscape that can deliver European DLE.

1.2 Main Components

The main components of our analysis related to data, LRs, and KGs include: 1. availability of data and metadata, 2. accessibility of data, 3. quality of data, 4. data interoperability , 5. licensing and data-related regulations, 6. data and ethics, and 7. data literacy. At the same time, the following related concepts, methodologies and tools also need to be considered: 8. data infrastructures, data spaces and data markets; 9. data at scale; 10. KGs; 11. semantic AI (statistical and symbolic AI in combination); and 12. innovative data and metadata management tools.

These main components always include structured data, semi-structured data and/or unstructured data, which can apply to different modalities, e. g., written, spoken, signs, etc. In addition, as for other technology areas, the data for LT may be available as raw data and/or curated data, at varying levels of quality.

With the rise of AI, the importance of large language models (such as, e. g., BERT[7] or GPT-3[8]), and comprehensive and multilingual KGs – all based on a broad range of domains and/or languages – is continuously increasing. For all LRs and data types there is the requirement for domain-specificity, so that domain- and industry-specific applications can be developed where specialised language and terminology are realised, e. g., in industries such as health, pharmaceuticals or finance. Let us now examine each of these aspects in detail:

Availability of data and metadata – As data and metadata form the backbone of any LT, the availability of data and metadata is the overall basis to enable such technologies and services. Availability therefore impinges on data collections, data types available, and how to find and explore such data.

Accessibility of data – The accessibility of data is crucial, it is also reflected in the FAIR principles (Wilkinson et al. 2016), initially advocated for research data management and stewardship in order to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. Since 2007, accessibility has also been one of the initial eight key principles of open (government) data.[9]

Quality of data – When data is available and accessible, users often consider additional attributes and components, one being quality of data. As the value of data is based on its fit for certain use-cases and business cases, data quality is a crucial issue reflecting and impacting the respective data value. Dimensions to measure data quality often include – but are not limited to – completeness, validity, timeliness, consistency, and integrity (Sebastian-Coleman 2012). Reliability is also an important factor of data quality, although it is hard to measure. When all things are considered, the quality of an LT application is often based largely on the quality of the underlying data used to train the system.

Data interoperability – Data interoperability is defined as[10] "addresses[ing] the ability of systems and services that create, exchange and consume data to have clear,

[7] https://en.wikipedia.org/wiki/BERT_(language_model)

[8] https://en.wikipedia.org/wiki/GPT-3

[9] https://opengovdata.org

[10] https://datainteroperability.org

shared expectations for the contents, context and meaning of that data." Interoperability ensures the seamless interplay of different LT systems regarding both APIs and data exchange. Not unexpectedly, it is often connected with and facilitated by the specification and adoption of related standards in the field.

Licensing and data-related regulations – Relevant data often comes from different owners and publishers, such as companies, public administrations or citizens, with different licences. Accordingly, proper licence clearing is a crucial task for all data-related activities in LT. The licences on data that are usually specified by data owners/publishers need to be taken into account as an important component, as well as the applicable laws and regulations around data, such as those concerning data privacy, security, processing and protection of personal identifiable information (PII), as laid out, for instance, in the General Data Protection Regulation (GDPR). National and regional as well as international regulations and policies around data use and re-use should also be taken into account.

Data and ethics – The rise of AI and ML has led to an increase in both data collection and processing, so the issue of data and ethics has become more and more important. It is closely connected to data-related regulations. Language, by its very nature, can be ambiguous and the associated interpretations can easily represent and expose bias. Accordingly, ethics plays a crucial role regarding the use of data in LTs and impacts equality in general, including language equality.

Data literacy – Gartner Research[11] defines data literacy as "the ability to read, write and communicate data in context, including an understanding of data sources and constructs, analytical methods and techniques applied, and the ability to describe the use-case, application and resulting value."

Data infrastructures, data spaces, data markets – The ideas behind data spaces and data markets follow the intentions underpinning data catalogues established in the course of the open data movement since the early 2000s to allow the sharing, exchange and trading of data. Data spaces and markets enable the availability of and allow accessibility to high-quality data, which follow standards (thus providing data interoperability) accompanied by clear licensing conditions. The Gaia-X[12] initiative defines a "data space" as "refer[ring] to a type of data relationship between trusted partners, each of whom apply the same high standards and rules to the storage and sharing of their data. However, of key importance to the concept of a data space is that data are not stored centrally but at source and are therefore only shared (via semantic interoperability) when necessary. A data space is the sum of all its participants – which may be data providers, users and intermediaries. Data spaces can be nested and overlapping, so that a data provider, for example, can participate in several data spaces all at once. Data sovereignty and trust are essential for the working of data spaces and the relationships between participants."

Data at scale – Practical LT solutions require high-quality data at scale and for a broad range of domains and available in various languages, with clear licences and fair conditions attached. Data infrastructures, data spaces and data markets provide

[11] https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy

[12] https://gaia-x.eu/what-is-gaia-x/

powerful means to discover, evaluate and access relevant data as well as related data-driven services, that are required for LT solutions.

Knowledge Graphs – A Knowledge Graph is a knowledge base that uses a graph-structured data model or topology to integrate data. KGs are used to store interlinked descriptions of entities – objects, events, situations or concepts – while also encoding the semantics underlying the terminology used.[13] Since the development of the Semantic Web, KGs have often been associated with Linked Open Data (LOD) projects, focusing on the connections between concepts and entities (Soylu et al. 2020; Auer et al. 2018). They are prominently associated with and used by search engines such as Google or Bing; knowledge-engines and personal assistants such as Wolfram Alpha, Apple's Siri, and Amazon Alexa; and social networks such as LinkedIn and Facebook. LT solutions require not only targeted datasets but also high-quality, interlinked, meaningful and contextualised data that can easily be used, quickly expanded and efficiently maintained with reasonable effort. KGs provide these characteristics and contribute to the data and knowledge backbone for LT.

Semantic AI – Modern approaches tend to combine statistical AI (ML) and symbolic AI (models like ontologies, knowledge bases for common sense knowledge, and cultural resources, among others). In October 2020, Agarwal defined semantic AI[14] as "provid[ing] a framework to perform end to end complex tasks automatically. It uses many different machine learning and logic-based approaches, and also utilizes the background knowledge often stored in knowledge graphs."

Innovative data and metadata management tools – Innovative data and metadata management tools enable the availability and accessibility of high-quality data and data interoperability (using relevant standards), that provide powerful data governance mechanisms (following relevant regulations), that enable mechanisms for the assessment of ethics in data, and that allow improvements in data literacy. In addition such tools should support (perhaps in combination with) secure data sharing mechanisms (data spaces), provide strong capability for interlinking data, support meaning and context (KGs) and provide semantic AI capability.

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

From the start of the open data movement in 2007 with its eight principles of open government data, the requirements of industry data as well as organisation-based data-sharing and collaboration have found their feet and culminated in the next era of data sharing: data catalogues and data portals, as well as, more recently, data spaces and data markets. In the area of LT, data availability, accessibility, aggregation, shar-

[13] https://ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph

[14] https://medium.com/@dr.puneet.a/what-is-semantic-ai-is-it-a-step-towards-strong-ai-5f0355be3597

ing and reuse have received attention since the early 1990s, with associations and organisations providing LR catalogues, like the European Language Resources Association[15] or the Linguistic Data Consortium.[16] Since the early 2010s, several research and innovation projects have contributed to the field including FLaReNet and META-NET with META-SHARE[17] (Piperidis 2012). They provided recommendations, specifications and implementations of platforms promoting and facilitating data discovery, sharing and reuse. At the same time, CLARIN[18] (Hinrichs and Krauwer 2014) has been established as a research infrastructure providing access to digital language data for scholars in the social sciences and humanities, and beyond. CLARIN is associated with the EUDAT Collaborative Data Infrastructure (EUDAT CDI),[19] and contributes to the European Open Science Cloud (EOSC)[20] with the EOSC-related project Social Sciences and Humanities Open Cloud (SSHOC)[21] and its data market for social sciences and humanities.[22]

Another example of research, development and infrastructure activities supported by the implementation of the Public Sector Information Directive[23] is the ELRC-SHARE repository[24] (Piperidis et al. 2018) that is used for documenting, storing, browsing and accessing LRs that are collected through the European Language Resource Coordination[25] initiative (Lösch et al. 2018) and considered useful for feeding the CEF Automated Translation (CEF.AT) platform.

In 2022, the European Language Grid (ELG)[26] (Rehm et al. 2020a; Rehm 2023) released the ELG platform providing access to LT resources and services from all over Europe, enabling users to try out the services or use the ELG APIs. ELG built bridges to a wide range of language data platforms including the European AI on Demand Platform (Labropoulou et al. 2023).

Turning to the LT industry, there are products like the TAUS Marketplace,[27] as well as APIs for lexicographical information or Natural Language Processing (NLP) APIs giving access to services from part-of-speech tagging and dependency parsing to MT, summarisation and question answering. Finally, there are active industry as-

[15] http://www.elra.info
[16] https://www.ldc.upenn.edu
[17] http://www.meta-share.org
[18] https://www.clarin.eu
[19] https://www.eudat.eu
[20] https://eosc-portal.eu
[21] https://sshopencloud.eu
[22] https://marketplace.sshopencloud.eu
[23] https://digital-strategy.ec.europa.eu/en/policies/public-sector-information-directive
[24] https://elrc-share.eu
[25] https://lr-coordination.eu
[26] https://www.european-language-grid.eu
[27] https://datamarketplace.taus.net

sociations and networks like LT-Innovate[28] or BDVA/DAIRO[29] that support the idea of data collection and provision and sharing to support better LT in the future.

Most if not all of the above platforms and initiatives have now endorsed the FAIR principles, adopting them as a de facto standard. In this context, data interoperability has been an important factor, related (mostly but not exclusively) to efficient data use and processing, as well as data exchange and sharing. There are dozens of standards regarding data in place worldwide, set up by several standardisation bodies in a range of industry domains. This diversity of data-related standards reinforces the problem as there is relatively little mapping between such standards and approaches. In the context of ELG and with regard to the wider area of AI/LT platform interoperability, initial attempts have been made at cross-platform search and discovery of resources and services, on the one hand, and composition of cross-platform service workflows, on the other (Rehm et al. 2020b).

Since the open data and data sharing movement began, every digital asset has needed to be accompanied by a clear and dedicated licence. While this issue has become more and more important, there are quite a lot of possible licences to choose from, inevitably reinforcing legal interoperability problems. While there are multiple commercial licensing options not centrally registered, a good source for open licences is the Open Definition of the Open Knowledge Foundation.[30]

Several data regulations and directives have been developed by the European Union over the last decade. They are an important foundation of the data economy, as well as the realisation of a working, sustainable data infrastructure across Europe. Some of the most important ones include, among others: GDPR,[31] European Strategy for Data,[32] European Data Governance (Data Governance Act),[33] EU Open Data Strategy and PSI Directive,[34] European Approach to Artificial Intelligence, including the EC AI Strategy,[35] Digital Single Market Strategy for Europe,[36] and Digital Action Education Plan.[37] As far as LT for DLE in Europe is concerned, all of these regulations have a clear impact. In terms of this deep dive, the Data Governance Act has a strong implication for data, LRs and KGs, as it lays the groundwork for the development of common data spaces in strategic sectors.

Setting technical issues to one side, data and ethics is a topic in which regulators and standards (such as those mentioned above) play a crucial role. After many years' discussion about data and ethics but also about AI and ethics, a standard has been published: *IEEE P7000 Engineering Methodologies for Ethical Life-cycle Concerns*

[28] https://www.lt-innovate.org

[29] https://www.bdva.eu

[30] https://opendefinition.org/guide/data/

[31] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[32] https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf

[33] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767

[34] https://digital-strategy.ec.europa.eu/en/policies/open-data

[35] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[36] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192

[37] https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en

Working Group. It establishes a process model by which engineers and technologists can address ethical considerations throughout the various stages of system initiation, analysis and design. Expected process requirements include management and engineering views of new IT product development, computer ethics and IT system design, value-sensitive design, and stakeholder involvement in ethical IT system design.[38]

Data literacy is an underlying component of digital dexterity: an employee's ability and desire to use existing and emerging technology to drive better business outcomes. The European Union supports data literacy and beyond in the Digital Action Education Plan,[39] and globally programmes like the World Bank's Data Use and Literacy Programme[40] support the awareness, education and implementation of data literacy. Nevertheless, compared to data and data-related technologies available, the issue of data literacy lags far behind and needs more action and effort to be applied.

The idea of a KG follows the basic principles of the semantic web and linked data. For LTs, the KG principles have great potential for modelling common-sense knowledge and domain-specific knowledge, as well as provisioning rich context and meaning in monolingual, bilingual, multilingual and cross-lingual applications. KGs are often assembled from numerous sources, and as a result, can be highly diverse in terms of structure and granularity.

KGs aim to serve as an ever-evolving shared substrate of knowledge within an organisation or community (Noy and McGuinness 2001). We distinguish two types of KGs: open KGs and enterprise KGs. Open KGs are published online, making their content accessible for the public good. Enterprise KGs are internal to a company and applied to commercial use-cases. Applications based on KGs include search, recommender systems, personal agents, advertising, business analytics, risk assessment, and automation. Useful further reading includes Blumauer and Nagy (2020), Abu-Salih (2021), Colon-Hernandez et al. (2021), Ji et al. (2022), and Li et al. (2021).

The technological leaps in LT and AI in the past few years and the widely recognised importance of data and knowledge resources for their accomplishment have called for new concepts and instruments in the area of data technologies and naturally so also in AI and LTs. In Europe, data spaces are a (relatively) new concept and solution to stimulate the data economy by providing secure and trustworthy mechanisms and platforms for data sharing and data trading. The European Commission lists a number of data spaces in its Data Strategy as of February 2020[41] that is strongly interconnected with the EU Data Governance Act.[42] EU Member States have supported research on data spaces in recent years, as for example Gaia-X[43] and the International Data Spaces initiative (Germany) that channeled into the establishment of the International Data Spaces Association (IDSA) and the publication of several standards and recommendations in the field (IDS Information Model or the

[38] https://sagroups.ieee.org/7000/

[39] https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en

[40] https://www.worldbank.org/en/programs/data-use-and-literacy-program

[41] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy

[42] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[43] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

Reference Architecture Model),[44] or the Data Market Austria (DMA)[45] prototype for a public marketplace for data trading. In January 2023, the European Commission launched the Common European Language Data Space which aims to focus on language data and models discoverability, sharing and trading covering all EU languages and aiming to support a wide range of LT applications in different modalities, domains and contexts.

2.2 Main Gaps

The following observations have been formulated, collected and further analysed together with researchers and practitioners in the field and reflect our joint understanding of the current gaps in the components of this deep dive.

There is untapped potential when it comes to data available in archives as well as old data files. There is a real need for open AI models in LT that are provided to interested parties with open licences. Not only ready-to-use models are required, but also the raw data needs to be made available in order for developers to create their own models. Annotated corpora are often available mainly in English, and it is often the case that they are not available in other languages, let alone all those required for different technologies and applications. The ELG dashboard[46] offers a visual overview of the current standing of Europe's languages (and beyond) with respect to available language data, tools and services. Through such availability counts the dashboard approximates the technological readiness of each language (see Chapter 3). There is an urgent need for monolingual, bilingual and multilingual domain-specific corpora. Such data can only rarely be found via available resources, mostly because it simply is not there, but also because of incorrect or missing documentation of data and metadata. Manually annotated data is lacking; although the quality of automated and semi-automated annotations is increasing, manual annotation by human experts in a certain field is still the best means of acquiring high-quality data.

Overall, there are missing open LRs. Domain-specific LRs are required to be available for scientific purposes with open licences. If the FAIR principles were systematically applied, this would be a huge benefit where data and metadata is concerned, but they are not really being rolled out properly. Although Europe has benefited from a strong open data movement for about 15 years now, there is still a gap in the provision of clearly specified licences for data. At the moment, benchmark approaches are not harmonised or standardised, and benchmarks on domain-specific vocabularies and annotated data and corpora are often missing. Metadata provides only very limited data provenance. Overall data quality is weak and so it often happens that use-cases cannot be realised as specified as labeled data is not available for the use-case at hand. Non-existing policies around data and metadata management

[44] https://internationaldataspaces.org/use/reference-architecture/

[45] https://datamarket.at

[46] https://live.european-language-grid.eu/catalogue/dashboard

that should be part of a data governance model often result in low data and metadata quality. There are increasingly many data silos in place that are neither connected nor interoperable, and there are more and more data infrastructures available that are simply not interoperable either, as the harmonisation of relevant standards in the field is missing. This is a clear problem and gap in the combination of research data (e. g., via EOSC)[47] and industry data (e. g., via industry data markets) as well as data from public administration or government data catalogues and portals (e. g., the European Open Data Portal).[48] More and clearer directives and regulations in the field should be developed to overcome these gaps in relation to data, LRs, and KGs. The effect of regulations on data-related topics should be evaluated continuously and regulations and directives adapted for identified gaps and changing environments. For example, GDPR has a strong effect on data collection.

Guidelines and policies are not available for each language in order to achieve DLE in Europe. Data for non-EU languages and beyond are not sufficiently in place, and so services for such languages cannot be developed with sufficient quality for them to be useful. National crowd-sourcing platforms that facilitate data collection for low-resource languages are not available hampering DLE in Europe.

There is a strong need for education that can deliver improved understanding of better data management processes in science, academia, as well as in business and industry. This should lead to better understanding of the value of data, and so improve data management principles and techniques. There is a need to inform educational bodies of the importance of sharing data; for example, if more learner corpora were made asvailable, this would lead to improved computer-assisted language learning and adaptive educational technologies. More senior staff and experts in AI need to work on data-related topics and deliver AI and deep learning mechanisms.

As an overall gap, there is a strong difference with regard to the level of digitisation in Europe. Data catalogues and portals often provide metadata only with links to the listed data that is provided by the data publishers and data owners themselves, with only a small amount also providing the data itself. The resulting issues and gaps relate to 1. *the availability of and access to the data itself*, as information in catalogues as to whether such data continues to be provided by publishers and owners is insufficient; 2. *lack of interoperability in metadata but mainly in the data itself*. The metadata often provides data interoperability (e. g., by using the same catalogue software CKAN),[49] and at least in Europe (but also beyond) we are making use of the de facto metadata standard for open data and data portals DCAT-AP (Data Catalogue Vocabulary (DCAT) expanded for Application Profiles);[50] and 3. *a fragmentation of data catalogues and data portals*.

[47] https://eosc-portal.eu

[48] https://data.europa.eu

[49] https://ckan.org

[50] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

Regarding data spaces and data markets, the TRUSTS project (Trusted Secure Data Sharing Space)[51] has carried out a study[52] on the definition and analysis of the EU and worldwide data market trends and industrial needs for growth, that includes a section on data market challenges, which includes a good summary of the gaps and challenges in this area (Figure 1). All these gaps and issues can only be addressed by working business models in the area of data sharing and trading in a working and successful data economy. IDSA published a relevant report in May 2021.[53]

Fig. 1 Challenges of data marketplaces

For a KG to become useful for a downstream application there is a need for it to contain a certain amount of application- and domain-specific knowledge. Often, openly available resources are not suitable for a particular task, so to reduce the entry barriers there is a need to be able to generate a suitable ontology or schema for said task and then to populate the schema with instances.

Currently, a KG is mainly developed based on textual and numerical data as an input format, with other formats like video and audio only very rarely taken into account. Working LT in the required languages using mechanisms like speech-to-text could support the creation of KGs.

Finally, there is a gap in the availability of comprehensive KGs. While there are some common-knowledge KGs even freely available (DBpedia, or Yago being just two examples), there is a clear lack of bigger KGs in specific domains and industries, that can act as a kind of foundation model, but also as training input for AI algorithms. Even if such specific graphs were available, there is a clear gap in the availability of multilingual domain-specific KGs that can be used for LT applications.

---

[51] https://www.trusts-data.eu

[52] https://www.trusts-data.eu/wp-content/uploads/2021/07/D2.1-Definition-and-analysis-of-the-EU-and-worldwide-data-market-trends-....pdf

[53] https://internationaldataspaces.org/the-ecosystem-effect-of-business-models-driven-by-data-sovereignty/

Regarding the gaps in semantic AI, we see that the fields of statistical and symbolic AI are still not fully combined; the two fields often exist in isolation beside one another and so cannot provide their full potential to the solution of a problem. This is largely the case overall in the machine learning and semantic web communities, but it is also the case in areas like LT, or domains like health or energy.

Finally, the following gaps regarding innovative data and metadata management tools have been identified: 1. *The need for user-friendly, flexible, open-source corpus annotation tools* that can easily be used by linguists as well as domain experts in-house and with fair costs and conditions. 2. *The need for user-friendly visualisation tools* in order to be able to understand the content of datasets at hand quickly and properly without the need for significant efforts in data integration and data wrangling. 3. *Better detection techniques for harmful content* are required to avoid bias, and identify and filter toxic content, or fake news and fake data, etc. In a time where AI and ML are being used more and more, even small portions of toxic data and content can influence an algorithm during training and so needs to be identified and filtered out. 4. *Better techniques for corpus filtering* are required regarding domain filtering, noise cleaning (see above, also) as well as the filtering and removal of bias. 5. *A clear lack of preservation technologies and tools* have been identified that are required to ensure that lesser-spoken languages can be archived for the long term (e. g., that are available on tape only) and made available as data that is easy to use, including the provision of proper data documentation in the form of rich metadata. 6. *Intelligent data analytics of small content nuggets* is needed, as, at the moment, often only huge corpora are being analysed by the available technologies and tools, but there is an increasing trend towards smarter data analytics that can be applied on ever smaller datasets, including for instance to just one paragraph or section, rather than the whole text. 7. *Add-on business models* are needed as gaps have been identified in the area of business models around data creation and provision, and so the development of tools and technologies is often limited to small experiments in funded research projects. Having clearly defined and successfully working business models in place would improve the industrial development in the field and stimulate the availability of the required innovative data and metadata management tools.

3 The Future of the Area

3.1 Contribution to Digital Language Equality

The major issue is the lack of available relevant and required data and LRs, as well as KGs in all European languages, official or not. At the moment sufficient data is available mainly in English, and to a lesser degree in German, French and Spanish. However, even in these languages data gaps exist that hinder LT development.

Looking further into this area it is easy to identify an even greater gap in the availability of data regarding dialects of European languages as well as regional languages. Dialects and regional languages exist, they are actively used and form

part of a country's or a region's identity and culture. Language diversity is so strong that sometimes in a small region several different languages or dialects are used.

In addition, there is very little data available for sign languages which is a clear issue for the inclusion of those with disabilities, as well as there being little to no respective data available for non-EU languages that are widely spoken in the EU, like Turkish or Arabic, for example.

DLE is a fundamental aspect of a functioning European society, in which diversity and inclusion are valued in every single EU Member State and across Europe with its colourful regional cultures and identities. The lack of DLE in Europe carries the risk of dividing society as it fosters misunderstanding, and may even support the promotion of toxic content, fake news, or lead to wrong interpretations of regional policies and regulations or the misinterpretation of research results in times of crisis.

We have identified the following three approaches: 1. *Digital Language Equality Strategy*: more funding and support by regional and national governments and the European Union to support the development of DLE in Europe for years to come for EU languages as well as regional languages and dialects (and for non-EU languages, too), aided by a data and LR matrix that shows which data and LRs should be available when and for which languages (see Chapter 45); 2. *Crowdsourcing and citizen science*: the creation of the required data needs the support of native speakers as well as linguists with the respective language experience and skills, and the support by data experts providing guidance with regard to the creation of useful and high-quality data; and 3. *Data-related business models*: these are required in the field to foster data creation and acquisition for minor languages and dialects by industry and the private sector.

In addition, and to allow DLE for certain domains like health, for instance, there is a strong need for the continuous development and maintenance of monolingual, bilingual and multilingual domain-specific vocabularies and KGs, to enable the multilingual and cross-lingual development of innovative domain-specific applications that provide value to the economy and society as a whole.

3.2 Breakthroughs Needed

Based on the identified components, the state-of-the-art analysis and the gap analysis, the areas of data infrastructures, data spaces and data markets are major issues where future technology visions and breakthroughs are needed in the field, as this area provides the overall umbrella for the availability and accessibility of the required data for powerful LTs that can help bring about DLE in Europe.

The main breakthroughs needed in terms of *data infrastructures, data spaces and data markets* include: 1. designing working architectures and ensuring effective workflows for compliant data provision and consumption; 2. developing specifications and building blocks to enable data and metadata interoperability; 3. developing and deploying technologies that embed data sovereignty and build trust among data providers and consumers; 4. developing specifications and building blocks that en-

able data value creation including data publishing and discovery mechanisms as well as accounting and billing; and 5. specifying and developing data governance models with clear roles, rules and policies for all stakeholders.

A recent study by the European Commission (Cattaneo et al. 2020) examines trends in data markets. The study measures *the value of a data market*, i. e., "the marketplace where digital data is exchanged as products or services as a result of the elaboration of raw data", and the *value of the data economy*, i. e., "[by] measur[ing] the overall impacts of the Data Market on the economy as a whole". The study compares the value of the data market and data economy from 2018 to 2019. It also projects the facts and figures for the year 2025 based on three scenarios.

Growth in data markets and the data economy brings with it several implications. According to the European Commission,[54] the total number of data professionals (i. e., those who deal with data endeavours as their primary task) will continue to rise consistently. Many opportunities will open in data-related jobs, and more knowledge workers are needed. Despite these positive trends, there is still a potential lack of supply of data professionals in high-growth scenarios. Companies taking a role as data providers and data buyers will also grow in number and market share.

KGs and semantic AI combined and provided as part of a data infrastructure can bring clear value, and should be part of any data infrastructure in the future. Gartner Research states that from 2021 onwards, graphs will form the foundation of modern data analytics with the capabilities to enhance and improve user collaboration, ML models and explainable AI. Although graph technologies are not new to data analytics, there has been a shift in thinking about them as organisations identify an increasing number of use-cases where they could play an important role. In fact, as many as 50% of Gartner client inquiries around the topic of AI involve a discussion around the use of graph technology.[55] In 2020, it was estimated that by 2023, graph technologies would facilitate rapid contextualisation for decision making in 30% of organisations worldwide.[56]

The main breakthroughs needed in the area of *KGs and semantic AI* include: 1. developing KG principles and technology from the current status of a "rising star" to a natural part of any data infrastructure and any data-related organisational infrastructure; 2. fostering the development of multilingual KGs under fair conditions and costs for use and re-use; 3. fostering the development of domain-specific KGs under fair conditions and costs for use and re-use; 4. KGs need a higher level of automation in their creation and maintenance, and more consideration needs to be given to the format of data beside textual data, such as audio and video; 5. a high level of deep and continuous learning will enable KGs to maintain themselves over time regarding new domain-specific and language-specific terminology. This means that new terms will be identified, analysed and inserted into the graph in the correct position, as well as being applied to the applications used by the KG; 6. bringing together the two main AI communities of statistical AI and symbolic AI to work together on

[54] https://op.europa.eu/s/vbSA

[55] https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021

[56] https://info.tigergraph.com/gartner-graph-steps-onto-the-main-stage-of-data-and-analytics

future semantic AI approaches; and 7. developing the areas of responsible AI and explainable AI by making use of semantic AI in multilingual environments to provide AI-based applications that deliver the correct results with benefits for research, industry and society.

The global enterprise metadata management market is forecast to grow at a rate of 20.3% from USD 7.45 Billion in 2019 to USD 27.24 Billion by 2027. Enterprise metadata management (EMM) provides the control and clarity needed to manage the change that often accompanies a complex enterprise data ecosystem. EMM and the various pieces of management software created for it provide administration for data integration, and allow users to inspect the metadata's links and roles.[57]

The main breakthroughs needed in the area of *innovative data and metadata management tools* include: 1. the development of tools that can be easily integrated with data infrastructures, data spaces and data markets; 2. the development of technologies and tools that can identify and remove bias, toxic content and fake data from data and content; 3. the provision of tools in the field of semantic AI, thus combining statistical and symbolic AI, that provide out-of-the-box responsible and explainable AI capability; 4. the development of a landscape where models and algorithms based on semantic AI can be created, ultimately with smaller amounts of data; 5. tools for data and metadata management that work not only in major languages like English but which can be easily adapted with low cost to smaller languages or dialects; 6. tools that allow deeper modelling of cultural aspects, gender aspects, etc. to avoid bias in data; 7. tools that are able to combine input from various types of data like text, images, audio and video but also gestures; and 8. tools along the whole data life cycle for all languages and all relevant use-cases are required to ensure powerful LT which can help enable DLE.

3.3 Technology Visions and Development Goals

We identified several technology visions and development goals for the area of data, LRs and KGs regarding DLE as a result of a comprehensive list of use-cases in the field, highlighting the related requirements. The majority of use-cases for LT involve human-to-machine and human-to-human communication and interaction via digital tools. To a large extent, these can be categorised using the concepts of *conversational AI* and *platforms and insight engines* that are covered by the other deep dive chapters in this book. In summary, the following excerpts represent identified data and technology development goals:

- LRs (speech, text) for official EU languages as well as for other European and non-European languages, for languages of minorities and dialects;
- pre-trained and fine-tuned language models for general and vertical domains for at least all EU-24 languages;
- speech models addressing at least the EU-24 languages;

[57] https://www.reportsanddata.com/report-detail/enterprise-metadata-management-market

- NLP pipelines of tokenisers, taggers, parsers etc., which require labelled linguistic datasets (e. g., treebanks) and evaluation sets;
- interfaces and content should be available in *all* languages via the web, i. e., the information available on a specific object, person or event provides the same amount of information in all languages;
- knowledge and content available in the form of audio files should be available in all languages so that it can be easily consumed;
- appropriate data required to train and develop monolingual, bilingual and multilingual models that cover the type of knowledge (domain-specific) and the type of language required for MT, (multi)document summarisation and speech-to-text technologies;
- efficient APIs required to integrate organisation-specific data and systems with social media platforms;
- pseudonymised or anonymised data for all EU languages, as well as domain-specific annotated corpora;
- data and models which address gender bias or minority bias etc.;
- data and technologies for identifying and ideally also removing toxic content, hate speech, fake news;
- comprehensive multilingual ontologies in vertical domains;
- KGs for common concepts, event descriptions for daily activities, and patterns for frequent questions;
- text-to-speech resources for common vocabularies and terminologies, as well as computer vision technologies for sign languages;
- data and technologies for modelling culture specific phenomena;
- better designed crowdsourcing platforms to enable more citizen science efforts towards building speech and language systems.

Some of these points are already available and in use in different data infrastructures. Beyond investing in the design and development of the missing parts, it is the integrated combination of all of them that could, from a technology perspective, be the main breakthrough and technology vision for the future management of metadata and data, as well as of LRs, that can act as the backbone for powerful LTs to realise DLE in Europe. Existing LT data infrastructure providers, such as ELG, ELRC-SHARE, CLARIN, META-SHARE, and ELRA as well as industrial and national initiatives can provide the seeds for a kind of federated data infrastructure, i. e., a data space that enables seamless and trusted interactions between data providers and data consumers, and enables cross-fertilisation by means of interoperability, aided among others by semantic KG technologies. Interoperability challenges can be broadly classified in four different layers:

- *technical interoperability*, enabling technical components (i. e., data space connectors) to communicate with each other;
- *semantic interoperability*, ensuring that attributes and policies have the same meaning;
- *organisational interoperability*, ensuring that the different (business) procedures and operations are compatible;

- *legal interoperability*, ensuring that contractual statements are legally equivalent.

Different federation architectures can be designed for building data spaces ranging from architectures with some central components (e. g., a data space catalogue) to fully decentralised ones. Whatever the architectural choice, data spaces will promote data sovereignty, enhance data exchange and trading, and enable the creation of value from data. The Language Data Space, coupled with the data space-inherent data integration capabilities, and developments in machine learning, deep learning, transfer learning and federated learning is expected to help fill in the gaps. Of key importance in the development of language data spaces is the compliance of data and operations with the rules, regulations and values of the European Union. LTs themselves are expected to play a crucial role in ensuring such compliance. Privacy preservation technologies, such as data anonymisation technologies and ethics compliance (through bias detection technologies, say), will be important tools in the hands of data providers, data consumers and data space operators. By its nature, the Language Data Space is conceived of as one of the horizontal data spaces in the data space ecosystem designed by the European Commission. In addition to the intra-data space interoperability, the Language Data Space will have to ensure interoperability with vertical data spaces (e. g., health, manufacturing, skills, mobility, etc), enabling cross-fertilisation, data discovery, exchange and trading at the inter-data space level.

Zooming out of the data spaces discourse and moving to technology visions regarding data access and sharing in general, one of the top-10 data and analytics technology trends identified by Gartner Research is the notion of a *Semantic Data Fabric*.[58] Although the notion was already identified in 2019, they predicted that the first real-world implementations would not be available before 2023. According to Gartner Research,[59] a data fabric enables frictionless access and sharing of data in a distributed data environment. It enables a single consistent data management framework, which allows seamless data access and processing by design across otherwise siloed storage. In the coming years, bespoke data fabric designs will be deployed primarily as a static infrastructure, forcing organisations into a new wave of costs to completely redesign their infrastructures for more dynamic data-mesh approaches. A data fabric must have the ability to collect and analyse all forms of metadata, and analyse and convert passive metadata to active metadata. It must have the ability to create a KG that can operationalise the data fabric design, and enable users to enrich data models with semantics. Extreme levels of distribution, scale and diversity of data assets add complexity to data integration rendering necessary a strong data integration backbone to enable versatile data sharing.

[58] https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo

[59] https://www.gartner.com/en/documents/3978267/data-fabrics-add-augmented-intelligence-to-modernize-you

3.4 Towards Deep Natural Language Understanding

Several areas of this deep dive on data, LRs, and KGs have already provided an overview of the state-of-the-art, a gap analysis and an outlook towards deep NLU. The way to help achieve deep NLU is once again by enabling the previously listed components for data, i. e., availability and accessibility of data and metadata; quality of data; interoperability; licences and data-related regulations; data and ethics; and data literacy. Related to these components, where data and metadata are concerned, data infrastructures, data spaces and data markets, integrating KGs, semantic AI and innovative data and metadata management tools need to be built. Furthermore, the following areas are of great importance: the ability to model emotions and culture-specific phenomena to facilitate cross-cultural understanding; the availability of world- and situation-specific knowledge in as many languages as possible; and of course tools that allow the modelling as well as the continuous learning of such attributes need to be built.

Continuous adaptation of LRs in all languages via automated and handcrafted mechanisms is key for deep NLU, to ensure new concepts and terminology are immediately taken into account and provided in monolingual, bilingual, and multilingual formats to ensure that new topics (like the COVID-19 pandemic) can be handled properly, but also so that the impact can be fully understood by a broad population to avoid bias, for example. Issues in digital language inequality will clearly support the division of societies, which needs to be avoided at all cost given the precarious times we live in, and the global nature of the problems we all face.

4 Summary and Conclusions

Data, LRs, and KGs form the basis and backbone for LTs. We identified the following main components: availability and accessibility of data and metadata; quality of data; data interoperability; licensing and data-related regulations; data and ethics; and data literacy. All of these need to be tackled in the future to allow data collection and provision with fair conditions and costs for all relevant stakeholders to help bring about DLE in Europe. Related to these components, where data and metadata are concerned, we identified the following technology concepts, methodologies and tools, that are currently on the rise and that are also part of our technology vision for 2030: data infrastructures, data spaces and data markets; KGs; semantic AI; and innovative data and metadata management tools.

As an add-on component, we tackled the topic of data-related business models, as we identified the importance of sustainable data-related business models as a prerequisite for a working data economy and ecosystem that stimulate and foster the above-listed data-related components in a well-functioning LT landscape.

Besides technology, interoperability and data-related aspects, there must be a strong focus on applying all these mechanisms and methodologies to the widest range of languages possible, at least to the EU-24 languages but also regional and

minority languages and also local dialects, as well as to non-European languages that are widespread across Europe. Without such data and LRs in place, DLE simply cannot be achieved.

To fill the identified gaps in data, LRs and KGs, we recommend a future path for Europe towards comprehensive and interlinked data infrastructures, which provide interoperability out-of-the-box by following harmonised and well-tested standards, regarding 1. (semantic) data interoperability as well as 2. services and 3. innovative data and metadata management tools available in all phases of the data lifecycle.

Metadata, data, data-driven tools and services need to be easily integratable into these data infrastructures, without today's huge efforts in data cleaning and integration, or service and tool integration. This future technology vision of integrated and interoperable data spaces follows the approach of federated architectures interlinking data providers and consumer spaces in a trusted framework. Existing data platforms and infrastructures as well as newly developed ones should be integrated where appropriate and possible.

In such a federated ecosystem, data regarding a domain or language can easily be identified, used, re-used, and evaluated for specific use-cases. Data-driven services can be delivered to meet an end-user's requirements. Crowdsourcing and citizen science mechanisms will allow human-machine interaction to foster data acquisition, cleaning and enrichment (e. g., annotation, classification, quality validation and repair, domain-specific model creation, etc.). Raw data can be loaded into available tools to build models for specific use-cases, but also existing algorithms, models or vocabularies will be available for easy loading and re-use to avoid unnecessary energy consumption/computing power to deliver energy-efficient data management.

A high level of importance needs to be placed on privacy protection (related to personal identifiable information, PII, and beyond) and the avoidance of bias (e. g., on gender), and the respective privacy preservation and ethics compliance technologies should be available to all stakeholders.

Data infrastructures require working and sustainable business models that promote data sovereignty, enable data trading, sharing and collaboration. Policies and sustainable data governance models around data creation, data provision and data sharing will be needed. Targeted publicly funded programmes and activities in the area of data literacy are needed from early education onward, to ensure that sufficient human resources in the field are available in the future.

In addition, we need to invest in the collection and development of data and LRs that are relevant for LT to ensure the availability of sufficient data in all EU languages. We make recommendations in three areas: 1. targeted national and European funding along a matrix of relevant resources and languages, combined with 2. more measures in the fields of crowdsourcing and citizen science, and 3. the development of functioning data-related business models, all of which are of critical importance (see Chapter 45).

Europe has a number of difficult problems to solve if DLE is to be achieved, including 1. the specifics of the European language space with EU official languages, a broad range of dialects and regional languages, as well as a high number of non-EU languages in use by a growing number of citizens across the continent, 2. the

European societal characteristics with a rich variety and diversity in culture and society, and 3. the overall challenging requirements of the continuous digitisation in a more and more globalised world, and the related critical need for an efficient, working (language) data infrastructure, that provides a rich, easy-to-use and sustainable backbone for European LT. Despite these challenges, there is a huge potential to become a world leader in LT and a role model for DLE if they can be overcome.

The availability of high-quality data, LRs and KGs in as many languages as possible, that are easily accessible with fair conditions and costs in a clearly specified legal environment providing transparent rules and regulations, has clear benefits and brings with it a competitive advantage for all stakeholders. For the European research community to foster innovations in the field, for the European industry to successfully compete in a global market, and for the benefit of European citizens and society, data, LRs, and KGs are crucial if European DLE is to be achieved.

References

Abu-Salih, Bilal (2021). "Domain-Specific Knowledge Graphs: A Survey". In: *Journal of Network and Computer Applications* 185, p. 103076.

Auer, Sören, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal (2018). "Towards a Knowledge Graph for Science". In: *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. WIMS '18. Novi Sad, Serbia: Association for Computing Machinery. https://doi.org/10.1145/3227609.3227689.

Blumauer, Andreas and Helmut Nagy (2020). *Knowledge Graph Cookbook: Recipes for Knowledge Graphs that work*. https://www.poolparty.biz/the-knowledge-graph-cookbook/.

Cattaneo, Gabriella, Giorgio Micheletti, Mike Glennon, Carla La Croce, and Chrysoula Mitta (2020). *The European Data Market Monitoring Tool: Key Facts & Figures, First Policy Conclusions, Data Landscape and Quantified Stories: D2.9 Final Study Report*. Publications Office. DOI: 10.2759/72084.

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal (2021). "Combining Pre-Trained Language Models and Structured Knowledge". In: *arXiv preprint arXiv:2101.12294*.

Elliot, Bern, Anthony Mullen, Adrian Lee, and Stephen Emmott (2021). *Gartner Research: Hype Cycle for Natural Language Technologies*.

Hinrichs, Erhard and Steven Krauwer (2014). "The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: ELRA, pp. 1525–1531. http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.

Ji, Shaoxiong, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu (2022). "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2, pp. 494–514. DOI: 10.1109/TNNLS.2021.3070843.

Kaltenböck, Martin, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg (2022). *Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs*. European Language Equality (ELE); EU project no. LC-01641480 – 1010-18166. https://european-language-equality.eu/reports/data-knowledge-deep-dive.pdf.

Labropoulou, Penny, Stelios Piperidis, Miltos Deligiannis, Leon Voukoutis, Maria Giagkou, Ondřej Košarko, Jan Hajič, and Georg Rehm (2023). "Interoperable Metadata Bridges to the wider Language Technology Ecosystem". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 107–127.

Li, Xinyu, Mengtao Lyu, Zuoxu Wang, Chun-Hsien Chen, and Pai Zheng (2021). "Exploiting Knowledge Graphs in Industrial Products and Services: A Survey of Key Aspects, Challenges, and Future Perspectives". In: *Computers in Industry* 129, p. 103449.

Lösch, Andrea, Valerie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith (2018). "European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management". In: *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), pp. 1339–1343.

Noy, Natalya and Deborah L. McGuinness (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Stanford Knowledge Systems Laboratory. http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html.

Piperidis, Stelios (2012). "The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: ELRA.

Piperidis, Stelios, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou (2018). "Managing Public Sector Data for Multilingual Applications Development". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA. http://www.lrec-conf.org/proceedings/lrec2018/pdf/648.pdf.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch (2020a). "European Language Grid: An Overview". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3359–3373. https://www.aclweb.org/anthology/2020.lrec-1.413/.

Rehm, Georg, Dimitrios Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julián Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John P. McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš (2020b). "Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability". In: *Proc. of the 1st Int. Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020)*. Ed. by Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs. Marseille, France, pp. 96–107. https://www.aclweb.org/anthology/2020.iwltp-1.15.pdf.

Sebastian-Coleman, Laura (2012). *Measuring Data Quality for Ongoing Improvement: a Data Quality Assessment Framework*. Newnes.

Soylu, Ahmet, Oscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Francisco Yedro Martínez, Matej Kovacic, Matej Posinkovic, Ian Makgill, Chris Taggart, Elena Simperl, Till C. Lech, and Dumitru Roman (2020). "Enhancing Public Procurement in the European Union Through Constructing and Exploiting an Integrated Knowledge Graph". In: *The Semantic Web – ISWC 2020 – 19th International Semantic Web Conference, 2020, Proceedings*. LNCS. Germany: Springer, pp. 430–446.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.18. http://www.nature.com/articles/sdata201618.

Chapter 44
Strategic Plans and Projects in Language Technology and Artificial Intelligence

Itziar Aldabe, Aritz Farwell, German Rigau, Georg Rehm, and Andy Way

Abstract This chapter on existing strategic plans and projects in Language Technology and Artificial Intelligence is based on an analysis of around 200 documents and is divided into three sections. The first provides a synopsis of international and European reports on Language Technology. The second constitutes a review of existing European Strategic Research Agendas, initiatives, and national plans related to Language Technology. The third contains a SWOT analysis designed to identify the factors that will need to be addressed to help solve the challenge of digital language inequality in Europe. Among the principal conclusions presented is the contention that our continent requires sophisticated multilingual, cross-lingual and monolingual LT for all European languages: LT *for* Europe that is *made in* Europe.[1]

1 Introduction

In varietate concordia (united in diversity) is the official Latin motto of the European Union (EU), adopted in 2000. According to the European Commission, "the motto means that, via the EU, Europeans are united in working together for peace and prosperity, and that the many different cultures, traditions and *languages in Europe* are a positive asset for the continent" [emphasis added].[2] All 24 official EU languages are granted equal status by the EU Charter and the Treaty on the EU. The EU is also home to over 60 regional and minority languages which are protected and promoted under the European Charter for Regional or Minority Languages (ECRML)

Itziar Aldabe · Aritz Farwell · German Rigau
University of the Basque Country, Spain, itziar.aldabe@ehu.eus, aritz.farwell@ehu.eus, german.rigau@ehu.eus

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

Andy Way
Dublin City University, ADAPT Centre, Ireland, andy.way@adaptcentre.ie

[1] This chapter is an abridged version of Aldabe et al. (2022).

[2] http://europa.eu/abc/symbols/motto/index_en.htm

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_44

since 1992,[3] in addition to migrant languages and various sign languages, spoken by some 50 million people. The Charter of Fundamental Rights of the EU under Article 21[4] states that "any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, *language*, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited" [emphasis added].

Multilingualism is a cultural cornerstone of Europe and signifies part of what it means to be and to feel European. However, not only do language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across languages, a dilemma for which no common EU policy has been proposed, many languages themselves are also endangered or on the edge of extinction (even more so from a digital perspective). This is illustrated in the *UNESCO Atlas of the World's Languages in Danger* (Moseley 2010),[5] where a map of Europe shows threatened languages, including black flags that correspond to already extinct languages.

Without a concerted effort to prevent the further deterioration of Europe's linguistic ecosystem, this current snapshot is likely to worsen. And while it may well be that no silver bullet exists to remedy the situation, one approach offers a means to provide immediate support and address the issue of linguistic barriers: Language Technology (LT) and language-centric Artificial Intelligence (AI).

Because natural language is at the heart of human intelligence, it is and must be at the heart of our efforts to develop AI technologies.[6] By the same token, all sophisticated and effective AI-powered tools are impossible without mastery of language.[7] This is why language and LT represent the next great frontier in AI.[8] Already arguably the hottest field in AI, LT also represents one of its fastest growing application areas.[9] In fact, together with vision and robotics, several recent international reports place LT as one of the three core application areas within AI. Its rise to prominence is due to the various methods LT has developed over the years to make the information contained in written and spoken language explicit or to generate written and spoken language. For this reason, it has become the nerve centre of the software that processes unstructured information and exploits the vast amount of data contained in text, audio and video files, including those from the web and social media. Despite the inherent difficulties in many of the tasks performed, current LT support allows for many advanced applications which would have been unthinkable only a few years ago. Among these may be counted speech recognition, speech synthesis, text analytics and machine translation (MT), used by hundreds of millions of people on a daily basis.[10] It is now common to utilise search engines, recommender

[3] https://en.m.wikipedia.org/wiki/European_Charter_for_Regional_or_Minority_Languages

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

[5] http://www.unesco.org/languages-atlas/

[6] https://hbr.org/2022/04/the-power-of-natural-language-processing

[7] https://www.nytimes.com/2022/04/15/magazine/ai-language.html

[8] https://www.forbes.com/sites/robtoews/2022/02/13/language-is-the-next-great-frontier-in-ai

[9] https://analyticsindiamag.com/is-nlp-innovating-faster-than-other-domains-of-ai/

[10] https://www.nimdzi.com/nimdzi-language-technology-atlas-2020/

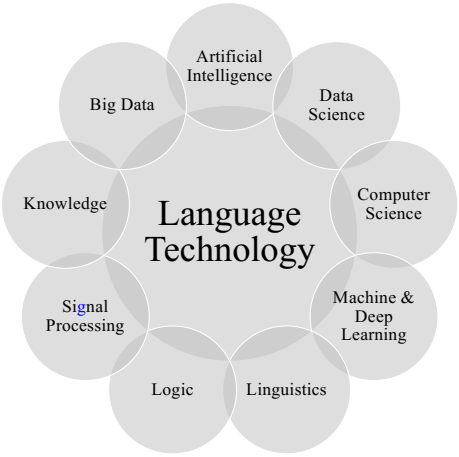

Fig. 1 Language Technology as a multidisciplinary field

systems, virtual assistants, chatbots, text editors, text predictors, MT systems, automatic subtitling, automatic summaries, and inclusive technology, all made possible thanks to LT. The field's rapid development promises even more encouraging results in the near future and its increasing social relevance has been highlighted in national and regional AI and LT strategies both inside and outside of Europe, as well as in prioritised strategic areas for research, development and innovation (R&D&I).

With this in mind, it should not be forgotten that LT is also multidisciplinary in nature, combining knowledge in computer science (and specifically in AI), mathematics, linguistics and psychology, among others. Figure 1 depicts some of the most important disciplines involved in LT. This uniqueness must be weighed in any public or private AI initiative that includes LT, especially given that funding for LT start-ups is booming and only the proper application of LT will allow the enormous volumes of multilingual written and spoken data in sectors as diverse as health, justice, education, or finance to be adequately processed and understood.[11]

Early-stage funding in 2021 amounts to just over USD 1 billion for companies offering solutions that make significant use of NLP, providing a picture of what funders think is innovative.[12] This belief is only reinforced by technology advances such as ChatGPT, whose creator, OpenAI, projects USD 1 billion in revenue by 2024.[13] Similarly, reports from analysts and consulting firms forecast enormous growth in the global LT market based on the explosion of applications observed in recent years and the expected exponential growth in unstructured digital data. For instance, ac-

[11] https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming

[12] https://towardsdatascience.com/nlp-how-to-spend-a-billion-dollars-e0dcdf82ea9f

[13] https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/

cording to an industry report from 2019,[14] the global NLP market size is expected to grow from USD 10.2 billion in 2019 to USD 26.4 billion by 2024, at a CAGR of 21% is set during the forecast period 2019-2024.[15] A recent report from 2021 estimates that the global NLP market is predicted to grow from USD 20.98 billion in 2021 to USD 127.26 billion in 2028 at a CAGR of 29.4% in the forecasted period.[16] NLP in Europe will witness market growth of 19.7% CAGR and is expected to reach USD 35.1 billion by 2026.[17] As a final example, according to Global Newswire the global NLP market is estimated to reach an expected value of USD 341.7 billion by 2030, growing at a CAGR of 27.6% during the forecast period.[18] These numbers indicate that the return on investment (ROI) will be massive so it is imperative that Europe is at the heart of this growth in future.

The attention paid to AI and LT in the social, political, and economic spheres reflect the significance of the technology for today's world. This chapter on the existing strategic plans and projects in LT and AI touches on all three of these areas. It is based on an analysis of close to 200 documents (Aldabe et al. 2022) and is divided into three sections. Section 2 provides a synopsis of international and European reports on LT. In addition to trends in innovation, many of these discuss the socioeconomic and political impact of AI and LT from a policy perspective. Section 3 constitutes a review of the existing European Strategic Research Agendas (SRAs), initiatives, and national plans related to LT. A main focus of these is the question of multilingualism and equal technological support for Europe's languages through the application of LT. Section 4 contains a SWOT analysis of the strategic documents and projects, which is designed to identify the factors that will need to be addressed to help solve the pressing issue of digital language inequality in Europe.

2 International Reports on Language Technology

AI capabilities are rapidly evolving and it has become one of the 21st century's most transformative technologies.[19] The growing interest in AI at a global political, scientific and social level has led several international organisations to draft a number of reports and initiatives in recent years. These often focus on the socioeconomic impact of AI technologies and applications with respect to policy.

[14] https://www.businesswire.com/news/home/20191230005197/en/Global-Natural-Language-Processing-NLP-Market-Size

[15] https://www.analyticsinsight.net/potentials-of-nlp-techniques-industry-implementation-and-global-market-outline/

[16] https://www.analyticsinsight.net/the-global-nlp-market-is-predicted-to-reach-us127-26-billion-by-2028/

[17] https://www.analyticsinsight.net/nlp-in-europe-is-expected-to-reach-us35-1-billion-by-2026/

[18] https://www.globenewswire.com/en/news-release/2022/09/29/2525379/0/en/Natural-Language-Processing-NLP-Market-Worth-USD-341-7-Billion-with-a-27-6-CAGR-by-2030-Report-by-Market-Research-Future-MRFR.html

[19] https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape/

2.1 Reports from International Organisations

The Organisation for Economic Co-operation and Development (OECD),[20] a frequent contributor to this discourse, has helped coordinate dialogue on the subject at international fora (notably the G7, G20, EU and UN), offered practical advice to governments on how to actualise AI policy, and stressed the potential that digital technologies demonstrate in responding to societal challenges.[21] Its 2021 report, *State of the implementation of the OECD AI Principles: Insights from national AI policies*, identifies challenges and best practices for the implementation of the five policy recommendations to national governments contained in its OECD AI Principles. These are: 1. invest in AI R&D; 2. foster a digital ecosystem for AI; 3. shape an enabling policy environment for AI; 4. build human capacity and preparation for labour market transformation; and 5. foment international co-operation for trustworthy AI. The report comes on the heels of the OECD's *The Digitalisation of Science, Technology and Innovation*, which emphasises that cutting-edge NLP techniques are opening new analytical possibilities. Among those listed is the ability to recognise victims of sexual exploitation on the internet based on facial detection and social network analysis (Chui et al. 2018). Advances such as this have caught the attention of researchers and policy makers in various countries, who have begun to experiment with NLP to track emerging research topics and technologies. As the report underscores, policy makers use these results to formulate science and innovation policy initiatives, support investments in R&D&I, and evaluate public programmes.[22]

Similar policy guidance and assessments appear elsewhere.[23] The Inter-American Development Bank[24] (IDB), for instance, suggests constructing a shared understanding of AI in order to take better advantage of its opportunities and applications while simultaneously coming to grips with its risks.[25] The World Economic Forum,[26] which provides a framework for governments that wish to develop national AI strategies, assists those responsible for crafting policy in how to ask pertinent questions, follow best practices, identify and involve stakeholders, and create a set of outcome

[20] https://www.oecd.org

[21] See, e. g., *Artificial Intelligence in Society* (https://doi.org/10.1787/eedfee77-en); *State of the implementation of the OECD AI Principles: Insights from national AI policies* (https://doi.org/10.1787/1cd40c44-e); *The Digitalisation of Science, Technology and Innovation* (https://doi.org/10.1787/b9e4a2c0-en).

[22] To help policy makers, regulators, legislators and others characterise AI systems deployed in specific contexts, the OECD has developed a user-friendly tool to evaluate AI and LT systems from a policy perspective (https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm).

[23] See, e. g., the *Helsinki Initiative on Multilingualism in Scholarly Communication* (https://www.helsinki-initiative.org/en).

[24] https://www.iadb.org

[25] https://publications.iadb.org/en/artificial-intelligence-for-social-good-in-latin-america-and-the-caribbean-the-regional-landscape-and-12-country-snapshots

[26] https://www.weforum.org

indicators.[27] UNESCO[28] extends these considerations to the educational sphere, recommending that governments and other stakeholders, in accordance with their legislation and public policies, respond to education-related opportunities and challenges presented by AI. The *Beijing Consensus on Artificial Intelligence and Education*, an outcome document issued by UNESCO in 2019, stresses the multidisciplinary nature of AI and urges readers to consider the role of AI tools in teaching and learning, highlighting its effectiveness in aiding students with learning impairments or who study in a language other than their mother tongue.[29] In the area of library science, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*, a position paper from OCLC,[30] notes structural inequalities are perpetuated by data-driven policies (Padilla 2020) and sets an agenda for tackling positive and negative impacts of data science, machine learning, and AI on libraries.[31]

Finally, in early 2022, based on the report *Facilitating the implementation of the European Charter for Regional or Minority Languages through artificial intelligence*, first published in 2020[32] and updated in 2022,[33] the Committee of Experts of the European Charter for Regional or Minority Languages of the Council of Europe (CoE) adopted a statement on the promotion of regional or minority languages through AI.[34] The Committee of Experts encourages states to promote the inclusion of regional or minority languages into research and study on AI with a view to supporting the development of relevant applications as well as to establish, in cooperation with the users of such languages and the private sector, a structured approach to the use of AI applications in the different fields covered by the Charter.

The attention paid to AI and LT in policy reports reflects the social, political, and economic importance that the technology has garnered in today's world; and the same holds true for organisations that trace trends in innovation. In its report, *Technology Trends 2019 Artificial Intelligence*,[35] the World Intellectual Property Organization[36] found that 50% of all AI patents have been published in just the past five years, a striking illustration of how rapidly innovation is advancing in the field. The report, which classifies AI technology trends into techniques, functional applications, and application fields, furthermore points to LT as one of AI's most significant functional applications, attributing over a quarter of all AI-related patents to NLP and speech processing. The number is unsurprising given the current levels of excite-

[27] https://www3.weforum.org/docs/WEF_National_AI_Strategy.pdf

[28] https://en.unesco.org

[29] https://unesdoc.unesco.org/ark:/48223/pf0000368303 See also, UNESCO's *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*, a 2019 report which, among other breakthroughs, noted a Chinese AI system that is able to correct student essays as a milestone in LT for education (https://unesdoc.unesco.org/ark:/48223/pf0000366994).

[30] https://www.oclc.org/en/about.html

[31] https://doi.org/10.25333/xk7z-9g97

[32] https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

[33] https://rm.coe.int/min-lang-2022-4-ai-and-ecrml-en/1680a657c5

[34] https://rm.coe.int/declaration-ai-en/1680a657ff

[35] https://www.wipo.int/publications/en/details.jsp?id=4386

[36] https://www.wipo.int

ment associated with NLP within AI, where the rising star is turning many heads. A case in point is the *State of AI Report* for 2021,[37] issued by UK AI investors with an eye toward stimulating informed conversation on AI and its implications going forward. The report, which considers research, talent, industry, and politics, discusses the emergence of large language models and notes that the latest generation are unlocking new NLP use-cases. Indeed, the arrival of Transformers as a general purpose architecture for ML has been a revelation, beating the state-of-the-art in domains as disparate as computer vision and protein structure prediction.

2.2 Reports from the United States

Reports from the US tell an analogous story to their international counterparts. In its 2021 and 2022 *AI Index Reports*,[38] for example, the Institute for Human-Centered AI (HAI) at Stanford University reviews the growth of research papers and conferences over time and by region, tracks AI accuracy on several benchmarks, focuses on trends in jobs and investment, and examines various national AI strategies. The reports also devote space to data and analysis concerning AI with respect to education, diversity, and ethics. Key takeaways include the observation that 65% of the new PhDs in the US chose jobs in industry over academia compared to 44% the previous year, that there is still little data available on the ethical challenges surrounding AI, and that the AI workforce remains predominantly male. The 2022 report also highlights that while current language models are setting records on technical benchmarks, they are also increasingly reflecting biases from their training data. These findings are accompanied by HAI's Global Vibrancy Tool,[39] which measures performance on various economic, inclusion, and R&D factors across several countries. The tool can create an overall index for the full list of 26 countries and it is of note that none of the top ten is an EU member state. The worrisome nature of the latter data point is compounded in an examination of the global balance and flow of top AI scientists provided by the Paulson Institute's Macro Polo think tank in its Global AI Talent Tracker report.[40] According to this analysis, the US lead in AI is built on attracting international talent, with more than two-thirds of the top-tier AI researchers working in the US having received undergraduate degrees in other countries. Although 18% of the top-tier AI researchers are European, only 10% of them work in Europe.

These final details should sound alarm bells in Europe. As demonstrated in *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence*, released by the AI100 project in 2021, remarkable progress has been made in AI over the past five years and we may anticipate that its effects will ripple out

[37] https://www.stateof.ai

[38] https://aiindex.stanford.edu/report/

[39] https://aiindex.stanford.edu/vibrancy/

[40] https://macropolo.org/digital-projects/the-global-ai-talent-tracker/

for many years to come. Prepared by a panel of experts from around the globe, the report makes clear that the ability of computers to perform sophisticated language- and image-processing tasks has advanced significantly and that more investment of time and resources are required to meet the challenges posed by AI's rapidly evolving technologies. On the one hand, this includes greater government involvement in the areas of regulation and digital education. In an AI-enabled world, citizens young and old must be literate in these new digital technologies. On the other, this means addressing fears that AI technologies will contribute to unemployment in some sectors. A Blumberg Capital survey of 1,000 American adults found that about half are concerned that AI threatens their livelihood. Indeed, despite the fact that 72% agreed that AI would help remove tedious tasks and free up time to concentrate on more creative ones, 81% were reluctant to surrender these tasks to an algorithm for fear of being supplanted.[41] As the authors of *Gathering Strength, Gathering Storms* indicate, AI is leaving the laboratory and entering our lives, having a "real-world impact on people, institutions, and culture."[42]

This perspective is shared by the National Security Commission. In addition to raising concerns that the United States risks falling behind China and other countries in the AI race, its recent 750-page report encourages the federal government to step up investment in the area.[43] Specifically, the commission calls for a *modest down payment* of $40 billion, along with hundreds of billions more in the coming years to galvanise future breakthroughs and help democratise AI research. Moreover, the report provides policy makers with a guide to ensure the US is prepared to defend against AI threats, promote AI innovation, and make responsible use of AI for national security. It is also worth mentioning that the report lists Natural Language Understanding as one of the six uses for deployed AI today. This view, which coincides with the general consensus on LT expressed above, is further reinforced by the Future Today Institute[44] in its 2021 *Tech Trends Report* on AI.[45] The group not only identifies NLP as an area that is experiencing high interest, investment, and growth, but also forecasts that NLP algorithms will do more in the future, including, for example, aid in interpreting genetic changes in viruses.

2.3 Reports from the European Union

Reports from the EU paint an equally upbeat picture about present and future expectations regarding science and technology. A recent Eurobarometer survey on European citizens' knowledge and attitudes towards these shows that 86% believe the overall

[41] https://blumbergcapital.com/ai-in-2019/

[42] https://ai100.stanford.edu

[43] https://www.nscai.gov/2021-final-report

[44] https://futuretodayinstitute.com

[45] https://2021techtrends.com/AI-Trends

influence of science and technology is positive.[46] EU citizens expect a range of technologies currently under development, including AI (61%), to improve their way of life over the next 20 years. The case for AI and LT is further laid out by various European Institutions in reports and policy initiatives that highlight their extensive impact on society and what must be done to shepherd this influence. These include, among others, *European Artificial Intelligence (AI) leadership, the path for an integrated vision*;[47] *Strategy on AI*;[48] *Ethics Guidelines for Trustworthy AI*;[49] *Liability for AI and other emerging technologies*;[50] *On Artificial Intelligence: A European approach to excellence and trust*;[51] and *Coordinated Plan on AI*.[52] All agree that AI is an area of strategic importance, a key driver of economic development, and a means to provide solutions to many societal challenges. As such, they concur that the socioeconomic, legal and ethical impact of AI must be carefully measured. For instance, the Joint Research Center (JRC) Science for Policy report, *The Changing Nature of Work and Skills in the Digital Age*,[53] observes that employment opportunities related to the development and maintenance of AI technologies and Big Data infrastructures are expected to grow, whereas jobs that are most vulnerable to automation appear to be those that require relatively low levels of formal education, do not involve complex social interaction, or demand routine manual tasks. Keeping this range in mind is a reminder that digital technologies may not only create or destroy some lines of work, but also fundamentally change what people do on the job and how they do it.

The European Commission's new *Coordinated Plan on AI*, which affirms that NLP is one of the most rapidly advancing fields in AI, is designed to address such potential turbulence.[54] The 2021 plan, in conjunction with the first-ever legal framework for AI,[55] will guarantee the safety and rights of people and businesses, while strengthening AI uptake, investment and innovation across the EU. It is also seen as the EU's next step in fostering global leadership in trustworthy AI, deemed necessary if European AI is to be globally competitive while respecting European values. This is of particular concern given that the EC's 2021 Strategic Foresight Report, *The EU's capacity and freedom to act*,[56] stresses the EU's capabilities in AI, Big

[46] https://europa.eu/eurobarometer/surveys/detail/2237

[47] https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2018)626074

[48] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence#Building-Trust-in-Human-Centric-Artificial-Intelligence

[49] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[50] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199

[51] https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[52] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence

[53] https://publications.jrc.ec.europa.eu/repository/handle/JRC117505

[54] https://digital-strategy.ec.europa.eu/en/library/new-coordinated-plan-artificial-intelligence

[55] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

[56] https://ec.europa.eu/info/strategy/strategic-planning/strategic-foresight/2021-strategic-foresight-report_en

Data and Robotics lag behind the world's leaders, the US and China. To strengthen digital sovereignty and European AI, the report encourages stakeholders to promote values via the finance, development and production of next generation tech.

One important area of focus must be high-value data, a key factor in improving performance and building robust AI models. The EC wants to ensure legal clarity in AI-based applications, especially regarding data. Its proposed regulation on data governance will help by boosting data sharing across sectors and member states, while the General Data Protection Regulation (GDPR) is a major step towards building trust.[57] The member states also recently agreed to a negotiating mandate on a proposal for a Data Governance Act (DGA).[58] The DGA is part of a wider policy to give the EU a competitive edge in the increasingly data-driven economy. The aim is to promote the availability of data that can be utilised to power applications and advanced solutions in AI, personalised medicine, green mobility, smart manufacturing and numerous other areas. While these regulations support the privacy and rights of European citizens, it should be pointed out that significant barriers to the access and re-use of language resources remain, especially with regard to competition with countries that have adopted the "fair use" doctrine, such as the US, Japan or Korea.

Research infrastructures play a role in this regard, including the Common Language Resources and Technology Infrastructure (CLARIN), an ESFRI Landmark and ERIC which offers access to LRs and LTs for researchers in the humanities and social sciences.[59] Not every EU Member State is officially affiliated with it, while others participate only as observers (Belgium joined CLARIN in 2021 and Spain will join in 2023). Additionally, because research funding agencies provide unbalanced resources to the different Member States, European languages are not equally supported by CLARIN (de Jong et al. 2020). This problem has received more attention in the EU project European Language Grid (ELG), which started in 2019 and concluded in June 2022. The ELG cloud platform contains more than 14,000 running services and resources for all European languages (Rehm et al. 2021; Rehm 2023).[60]

Experience with infrastructures such as these has demonstrated that the EU's approach to data infrastructures must be crafted with Big Data technology and LT in mind. The ESFRI roadmap includes a "Landscape Analysis" that provides an advanced analysis of the scientific needs and existing research infrastructure gaps as well as directions for strategic investments in the future that would help maintain Europe's leadership in the global context. According to its findings, research infrastructures in LT are indispensable in breaking new ground because they represent a core aspect of Big Data technology due to the volume and variety of data generated by the accumulation of unstructured text. And as the main task in AI's communication domain, NLP encompasses applications such as text generation, text mining, text classification, MT and speech recognition. Put differently, LT's ability to analyse, understand and generate information expressed in natural language is crucial for im-

[57] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[58] https://www.consilium.europa.eu/en/press/press-releases/2021/10/01/eu-looks-to-make-data-sharing-easier-council-agrees-position-on-data-governance-act/

[59] https://www.clarin.eu

[60] https://www.european-language-grid.eu

proving human-computer interaction. This view is confirmed by AI Watch, the EC's knowledge service responsible for monitoring the development, uptake and impact of AI, in three recent reports, *Defining Artificial Intelligence*, *Artificial Intelligence in public services* and *AI Watch, road to the adoption of Artificial Intelligence by the public sector*.[61] By way of example, the latter identified and employed 230 cases of AI usage in public services in order to extract emerging trends in AI, revealing that well over half of the cases are closely related to LT.

Relatedly, the EC's Directorate-General for Communications Networks, Content and Technology (DG CNECT), in collaboration with the Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (DG GROW), opened a consultation in 2021 that examined use-cases for website translation at small and medium-sized enterprises (SMEs) and surveyed multilingual websites in an effort to analyse language barriers across EU Member States.[62] The inquiry identified specific market needs that could be addressed through public solutions, such as eTranslation,[63] and by European language service providers. Of the over 1,000 SMEs that responded, 75% expressed interest in participating in the EC's subsequent pilot programme to make their website automatically multilingual. When the *European Language Industry Survey* (ELIS)[64] – then known as the *EUATC survey* – was run for the first time in 2013, MT was still primarily seen as a threat and a challenge. Only a few language companies saw it as an opportunity. Today, 65% of language company respondents see the improved quality of neural MT as an opportunity rather than a threat. According to the 2022 survey, 58% of those companies have implemented the technology and an additional 20% are planning to do so. This potential willingness to incorporate LT and AI corresponds with a separate study conducted by Eurostat[65] in 2020. It found that 7% of EU enterprises with at least ten employees used AI applications, 2% utilised ML to analyse big data internally, and 1% evaluated big data internally with the help of LT. Moreover, 2% provided a chat service, where a chatbot or virtual agent generated natural language replies to customers.

3 Major Language Technology Initiatives in Europe

First, we take a closer look at European initiatives (Section 3.1) and then examine national and also regional initiatives (Section 3.2).

---

[61] https://knowledge4policy.ec.europa.eu/ai-watch_en; https://publications.jrc.ec.europa.eu/repository/handle/JRC118163; https://publications.jrc.ec.europa.eu/repository/handle/JRC120399; https://joinup.ec.europa.eu/collection/innovative-public-services/news/ai-watch-road-adoption-artificial-intelligence

[62] https://digital-strategy.ec.europa.eu/en/library/report-sme-survey-multilingual-websites

[63] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation; https://ec.europa.eu/education/knowledge-centre-interpretation/eu-initiatives-language-technologies_en

[64] https://elis-survey.org

[65] https://ec.europa.eu/eurostat/

3.1 European Initiatives

The European Parliament recently emphasised that "multilingualism presents one of the greatest assets of cultural diversity in Europe and, at the same time, [is] one of the most significant challenges for the creation of a truly integrated EU." (European Parliament 2018). The belief is reflected in the EU's promotion of multilingualism, which falls within the scope of a variety of EU policy areas. While many of the multifaceted efforts to support Europe's languages are bearing fruit, still greater attention must be paid to removing barriers to intercultural and inter-linguistic dialogue as a means to stimulate mutual understanding. One means to achieve this is through language technology. Nonetheless, although official EU languages are granted equal status politically, they are far from equally supported from a technological perspective (see, e. g., Rehm and Uszkoreit 2012; Rehm et al. 2014; Rehm and Hegele 2018; Rehm et al. 2020b, as well as the chapters in Part I of this book).

Several strategic documents have contributed to the European debate on this subject in the past decade, including *The FLaReNet Strategic Language Resource Agenda* (Soria et al. 2014), *META-NET Strategic Research Agenda for Multilingual Europe 2020* (Rehm and Uszkoreit 2013; Rehm et al. 2016), *Language Technologies for Multilingual Europe: Towards a Human Language Project* (Rehm 2017), and the STOA report, *Language Equality in the digital age: Towards a Human Language Project* (STOA 2018). The latter helped pave the way for the preparation of the European Parliament's joint ITRE/CULT resolution, *Language equality in the digital age* (European Parliament 2018),[66] adopted in a plenary meeting in September 2018 with an overwhelming majority of 592 votes in favour, 45 against and 44 abstentions.

Approval of the resolution by such a wide margin demonstrates the importance and relevance of the issue. It includes more than 40 recommendations, structured into four sections: "Improving the institutional framework for language technology policies at EU level", "Recommendations for EU research policies", "Education policies to improve the future of language technologies in Europe" and "Language technologies: benefits for both private companies and public bodies". Among the most salient items are the following (emphases added; some items abbreviated):

- The report "recommends that in order to raise the profile of language technologies in Europe, the Commission *should allocate the area of 'multilingualism and language technology' to the portfolio of a Commissioner*; considers that the Commissioner responsible *should be tasked with promoting linguistic diversity and equality at EU level*, given the importance of linguistic diversity for the future of Europe;" (item 14)
- "suggests *ensuring comprehensive EU-level legal protection for the 60 regional and minority languages*, recognition of the collective rights of national and linguistic minorities in the digital world, and mother-tongue teaching for speakers of official and non-official languages of the EU;" (item 15)

[66] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html

- "calls on *the Member States to develop comprehensive language-related policies and to allocate resources and use appropriate tools in order to promote and facilitate linguistic diversity and multilingualism in the digital sphere*; stresses the *shared responsibility of the EU and the Member States* and in developing databases and translation technologies for all EU languages, including languages that are less widely spoken; calls for *coordination between research and industry* with a common objective of enhancing the digital possibilities for language translation and with open access to the data required for technological advancement;" (item 17)
- "calls on the Commission *to establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels*, tailored specifically to Europe's needs and demands; emphasises that the programme should seek to tackle *deep natural language understanding* and increase efficiency by sharing knowledge, infrastructures and resources, with a view to developing innovative technologies and services, in order *to achieve the next scientific breakthrough* in this area and help to reduce the technology gap between European languages; stresses that this should be done with the participation of research centres, academic, enterprises […] and other relevant stakeholders;" (item 25)
- "believes that […] *European education policies should be aimed at retaining talent in Europe*, should analyse the current educational needs related to language technology […] and, based on this, *provide guidelines for the implementation of cohesive joint action at European level* […] including the language-centric artificial intelligence industry;" (item 34)
- "points to the need *to promote the ever-greater participation of women in the field of European studies on language technologies*, as a decisive factor in the development of research and innovation" (item 36)

To these recommendations may be added the remarks made by EC Commissioner Corina Crețu in her closing statement at the hearing on the resolution:

> Ensuring appropriate technological support for all European languages will […] create jobs, growth and opportunities in the DSM [(Digital Single Market)]. It will enhance the quality of public services, and reinforce a stronger sense of unity and belonging throughout Europe. […] [U]nder the next Multiannual Financial Framework (MFF), we will need to reinforce funding, research and education actions. […] [O]vercoming language barriers in the digital environment is essential for an inclusive society, a vibrant DSM and for unity in diversity.

Crețu's statement is in line with previous public appeals voiced in 2016 by former EC Vice President Andrus Ansip and in 2017 by Director General Roberto Viola (DG CNECT) for the need to strengthen multilingualism through technologies.[67]

More recently, the EP's CULT Committee adopted a resolution on AI in the cultural, creative and educational sector in which multilingual and linguistic diversity

[67] See *How multilingual is Europe's Digital Single Market?* (https://ec.europa.eu/commission/commissioners/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en); *Multilingualism in the Digital Age: a barrier or an opportunity* (https://ec.europa.eu/digital-single-market/en/blog/multilingualism-digital-age-barrier-or-opportunity).

is also taken into account.[68] Regarding the latter, the resolution calls for: 1. AI technologies to be regulated and trained in order to ensure non-discrimination, gender equality, pluralism, as well as cultural and linguistic diversity; 2. specific indicators to measure diversity in order to promote European ventures and prevent algorithm-based recommendations that negatively affect the EU's cultural and linguistic diversity; and 3. an ethical framework for the use of AI technologies in EU media that guarantees access to culturally and linguistically diverse content. Such a framework would also address the misuse of AI to disseminate fake news and disinformation.[69] The resolution goes hand in hand with a study commissioned by the EC that explores the possibilities of applying AI technologies in ten domains that also belong to the cultural, creative and educational sector. The study aims to inspire creative entrepreneurs as well as policy-makers with concrete use cases and recommendations for the application of AI,[70] focusing partly on language-centric AI (NLP, NLU, speech technologies). The resolution also reflects the conclusions of the Education, Youth, Culture and Sport Council held on 4-5 April 2022, which called for the development of an ambitious digital policy for language technologies, translation, and lifelong language learning and teaching. This objective fits with the EU's desire to take advantage of new technologies to foster multilingualism, which it hopes will facilitate access to culture and nurture cultural exchange.[71]

A key commonality in these documents and initiatives is the idea that LT must be *made in* Europe *for* Europe. This approach will not only strengthen Europe's place at the pole position of research excellence, but also contribute to future European cross-border and cross-language communication, economic growth and social stability. The past few years have witnessed a flurry of white papers and SRAs offering roadmaps and recommendations for how best to attain the goal. In 2019, the European Language Resource Coordination (ELRC) white paper, *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters*, underscored that the main challenge is a lack of appreciation for the value of language data.[72] To help overcome this perception, the group issued several recommendations aimed at the European and national policy level, including:

- Updating the Open Data Directive (2019/1024/EU) so that it references language data as a high-value data category.[73]
- Conducting a study on language data to identify and quantify the value of language data for citizens, public administrations and businesses.

---

[68] https://www.europarl.europa.eu/news/en/press-room/20210311IPR99709/ai-technologies-must-prevent-discrimination-and-protect-diversity; https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1663438&t=e&l=en

[69] https://op.europa.eu/en/publication-detail/-/publication/b8722bec-81be-11e9-9f05-01aa75ed71a1

[70] https://digital-strategy.ec.europa.eu/en/library/study-opportunities-and-challenges-artificial-intelligence-ai-technologies-cultural-and-creative

[71] https://www.consilium.europa.eu/en/meetings/eycs/2022/04/04-05/

[72] https://lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf

[73] https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

- Updating national policies (e. g., Open Data policies, digital agenda or strategies for AI) to explicitly support the sharing of language data and LT.
- Including obligatory (language) data management plans in all relevant national funding policies and calls for proposals if not yet included.
- Conducting national surveys to assess translation practices in public administrations at all levels.

These steps will contribute to the development of an inclusive European digital society, a task for which European LT is essential. However, still others are required. The *Report on the Joint Stakeholder Consultation on Research and Innovation in Web Accessibility and Language Technologies*, for instance, highlights that greater work must be done to develop systems capable of adapting and personalising digital content according to individual needs, particularly in terms of accessibility and language.[74] Research into sign languages represents one avenue that merits greater attention, given that sign languages are increasingly becoming recognised as official national languages. Another relevant is the accessibility of information in multimodal contexts with respect to formatting and the understanding of content.

Fortunately, it is evident that the EU is not blind to LT's crucial role in building Europe's digital society and has already begun to dedicate funding and launch initiatives to advance LT and AI. Research, industry, and the public sector have benefitted from these actions. Two prominent examples include the Horizon 2020 Programme and the Connecting Europe Facility (CEF).[75] LT was embedded in the former within research and innovation in the field of information technologies, content technologies, multilingual internet and AI. Through the latter, MT tools (eTranslation) and tools for the management of thesauri and glossaries have been developed (VocBench).[76] There is, however, much left to be done. The *Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem* provides an analysis of the EU's LT market (including Norway and Iceland) and the adoption of LT by public administrations, both at the EU and national levels.[77] The report underscores that EU industry is fragmented and that many small players struggle to compete with the global giants that dominate the market. It further notes that European businesses and the public sector have become dependent on these non-European global companies, which have massive amounts of data at their disposal due to both copyright disparities between the EU (explicit permission required by European entities) and the US (fair use copyright exception), as well as intensive use of their popular systems.

---

[74] https://ec.europa.eu/digital-single-market/en/news/report-joint-stakeholder-consultation-research-and-innovation-web-accessibility-and-language-0. See also the New European Media's SRIA: https://nem-initiative.org/wp-content/uploads/2020/06/nem-strategic-research-and-innovation-agenda-2020.pdf?x98588

[75] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/information-and-communication-technologies; https://ec.europa.eu/digital-single-market/en/connecting-europe-facility; https://ec.europa.eu/digital-single-market/en/language-technologies

[76] https://ec.europa.eu/isa2/solutions/vocbench3_en

[77] https://op.europa.eu/en/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en/format-PDF/source-106906783

Nonetheless, the dependency on American or Chinese systems and the torrent of data flowing out of Europe mask areas in which European initiatives may make real the ideal of LT made *in* Europe *for* Europe. Several large international tech companies, by way of example, provide MT services free of charge. EU industry, by contrast, is experienced in navigating through Europe's many languages and European MT developers have successfully deployed services for the public sector through the support of EU-funded programmes. LT made for Europe means harnessing this know-how to support MT for all its languages and create domain-specific and application-specific MT while simultaneously being attentive to security and privacy issues. Moveover, as stated in *My Europe. My language: With language technologies made in the EU*,[78] LT offers opportunities to reduce language barriers across Europe and in the DSM at the intersection of Big Data, AI and HPC. Indeed, the European High Performance Computing Joint Undertaking[79] (EuroHPC JU), a joint initiative between the EU, European countries and private partners, is developing a world-class supercomputing ecosystem in Europe.[80] The Language Data Space EU project, a platform and marketplace for the collection, creation, sharing and re-use of multilingual and multimodal language data, was launched in January 2023.[81]

The EC has also established public-private partnerships (PPPs) in the area of AI.[82] As detailed by Curry et al. (2021), the Big Data Value PPP, created by the EC and the BDVA in 2014, represented a substantial collective effort on the part of the European data community to formulate a set of technical research priorities for Big Data. According to the report, Europe's multilingualism presents a particular challenge when it comes to data:

> Large amounts of data are being made available in a variety of formats ranging from unstructured to semi-structured to structured formats […] A great deal of this data is created or converted and further processed as text. Algorithms or machines are not able to process the data sources due to the lack of explicit semantics. In Europe, text-based data resources occur in many different languages, since customers and citizens create content in their local language. This multilingualism of data sources means that it is often impossible to align them using existing tools because they are generally available only in the English language. Thus, the seamless aligning of data sources for data analysis or business intelligence applications is hindered by the lack of language support and gaps in the availability of appropriate resources.[83]

[78] https://digital-strategy.ec.europa.eu/en/library/my-europe-my-language-language-technologies-made-eu-brochure

[79] https://eurohpc-ju.europa.eu

[80] https://digital-strategy.ec.europa.eu/en/activities/work-programmes-digital

[81] https://digital-strategy.ec.europa.eu/en/funding/language-data-space-call-tenders

[82] https://adr-association.eu

[83] https://elements-of-big-data-value.eu/research-priorities-for-big-data-value

The Big Data Value PPP's successor, the AI, Data and Robotics Partnership (formed in 2020 along with BDVA,[84] euRobotics,[85] ELLIS,[86] CLAIRE,[87] and EurAI[88]) expanded on this issue and zeroed in on NLP's importance in its Strategic Research, Innovation and Deployment Agenda,[89] "Natural Language Processing has particular resonance within Europe's multi-lingual landscape and offers the potential to harmonise human interaction." Unfortunately, although the PPP includes LT experts, research groups and companies via some the groups involved, currently no European LT association or network is represented in the PPP.

The initiative, however, complements the Coordinated Plan on Artificial Intelligence (CPAI) proposed by the European Commission for the period 2021-2027. The plan, which considers AI an area of strategic importance and aims to propel Europe to the forefront in terms of developing and exploiting AI technologies, calls for the EU to provide a minimum one billion euro annual investment in Horizon Europe and Digital Europe, although the objective is to reach twenty billion euros a year between public and private investments.[90] The focus is on four key areas: increasing investment in AI; the availability of data; the promotion of talent; and ensuring security, ethics and trust in AI. Success in these domains leans on the belief that member states must develop and coordinate their own national AI strategies, of which an analysis and comparison is provided in the report *AI Watch: National strategies on Artificial Intelligence: A European perspective in 2019*.[91]

3.2 National and Regional Initiatives

The perspective that the EU Member States should be responsible for their individual AI strategies stems partly from the observation that each country or region is best placed to address their own particular needs. The response by European countries to the CPAI has been largely positive and the number of states with an AI strategy (29 out of 30; only Croatia has no official strategy as of yet) demonstrates its success. Moreover, it is in the national plans that currently exist where many of the initiatives concerning LT and language-centric AI reside, although this is not to say that dedicated LT programmes are widespread in Europe. And in comparison to non-EU national AI initiatives, Europe's member states lag behind when LT is taken into account. Since Canada published the world's first national AI strategy in 2017, more

[84] https://www.bdva.eu

[85] https://www.eu-robotics.net

[86] https://ellis.eu

[87] https://claire-ai.org

[88] https://eurai.org

[89] https://adr-association.eu/wp-content/uploads/2020/09/AI-Data-Robotics-Partnership-SRIDA-V3.0-1.pdf

[90] https://knowledge4policy.ec.europa.eu/ai-watch/coordinated-action-plan-ai_en

[91] https://ec.europa.eu/jrc/en/publication/ai-watch-national-strategies-artificial-intelligence-european-perspective-2019

than 30 other countries and regions have published similar documents as of December 2020.[92] Several non-EU nations merit brief consideration here due to the explicit inclusion of NLP in their plans. China's AI strategy, one of the most comprehensive in the world, singles out NLU technology as a decisive area to promote university AI curricula and in its pursuit of AI talent (Zhang et al. 2021). The UK, which emphasises a strong partnership between business, academia, and government, created a pilot programme for under-18-year-olds to encourage careers in the AI sector, explicitly mentioning NLP. India's approach to AI considers the multilingual reality of the country a means to achieve technological leadership in AI and cites the development of an advanced NLP infrastructure for its languages as a stepping stone in that direction.[93] Finally, the US emphasises the crucial role LT plays in AI and NLU appears as one of the six "Uses for Deployed AI Today" in the National Security Commission on Artificial Intelligence's *Final Report*, published in 2021.[94]

In Europe, only a handful of dedicated national programmes funded projects related to LT before 2018.[95] Instead, financial support for the development of LT was generally provided through generic R&D&I calls in most member states. The Spanish case is one of those notable exceptions. The Spanish government has recently announced a new strategic plan for economic recovery and transformation (PERTE) called "The New Economics of Language."[96] The PERTE is presented as an opportunity to take advantage of the potential of Spanish and co-official languages for economic growth and international competitiveness in areas such as AI, translation, learning, cultural dissemination, audiovisual production, research and science. It has a budget of 1.1 billion euros in public funds and aims to mobilise another billion in private investment. Additionally, following the lines of the Spanish Plan for the Advancement of LT,[97] several regional governments have also launched LT initiatives, including AINA (Catalonia),[98] Nós (Galicia)[99] and GAITU (the Basque Country).[100]

At the European level, LT received better support through calls in various programmes: FP7, Horizon 2020, CEF Telecom, CIP ICT-PSP, EUREKA and EU-

[92] https://aiindex.stanford.edu/report/

[93] *AI in India: A Policy Agenda*. The report also highlights natural language voice recognition as a way to to account for the diversity in languages and digital skills in the Indian context and recommends the creation of annotated data sets for their languages to add incremental value to existing services ranging from e-commerce to agriculture.

[94] https://www.nscai.gov/2021-final-report. See also, the *American AI Initiative*.

[95] Spanish *Plan for the Advancement of Language Technology*: https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/tecnologias-del-lenguaje-en-Europa.aspx

[96] https://planderecuperacion.gob.es/como-acceder-a-los-fondos/pertes/perte-nueva-economia-de-la-lengua

[97] https://plantl.mineco.gob.es/Paginas/index.aspx

[98] https://politiquesdigitals.gencat.cat/ca/tic/aina-el-projecte-per-garantir-el-catala-en-lera-digital/

[99] https://www.xunta.gal/hemeroteca/-/nova/134792/xunta-usc-ponen-marcha-lsquo-proxecto-nosrsquo-que-permitira-incorporar-galego

[100] https://www.irekia.euskadi.eus/es/news/76846-gobierno-vasco-presentado-gaitu-plan-accion-las-tecnologias-lengua-2021-2024-cual-tiene-objetivo-integrar-euskera-las-tecnologias-linguisticas

	LT-related funding			Artificial Intelligence	
	None at all	Some funding	Dedicated LT programme	AI strategy	LT funding through AI
Austria	●			●	
Belgium		●		D	●
Bulgaria		●		●	
Croatia	●				
Cyprus				●	
Czechia		●		●	
Denmark			●	●	●
Estonia			●	●	●
Finland		●		●	
France		●		●	●
Germany		●		●	●
Greece		●		D	
Hungary		●		●	
Iceland			●	●	
Ireland		●		●	
Italy		●		●	
Latvia		●		●	
Lithuania		●		●	
Luxembourg		●		●	
Malta		●		●	●
Netherlands		●		●	
Norway		●		●	
Poland		●		●	
Portugal		●		●	
Romania		●		D	
Serbia	●			●	
Slovakia	●			●	
Slovenia		●		●	
Spain			●	●	
Sweden		●		●	

Table 1 The Language Technology funding situation in Europe (2019/2021), extracted from Rehm et al. (2020b) and updated with the newest AI strategies (D: Draft)

ROSTARS, among others. However, in these most funding for LT projects gradually reduced as well. If these findings are compared to those presented by Rehm et al. (2020b), we observe a slight increase in the number of language-centric AI initiatives over the next couple of years (see Table 1 and Figure 2).[101] It is noteworthy that only 12 European countries out of the 30 studied explicitly consider LT within their national policy initiatives. This is significant because the successful development of the next generation of innovative AI technology relies on setting aside funding

[101] According to Rehm et al. (2020b), only four of the 30 surveyed countries do not have some level of LT funding. Four countries have programmes dedicated to LT (Denmark, Estonia, Iceland, Spain), six provide funding for LT-related topics through AI (Belgium, Denmark, Estonia, France, Germany, Malta) and two (Ireland, Latvia) that do not have LT programmes, but rather a language strategy defined by their governments. See also Rehm et al. (2016, 2020a, 2021).

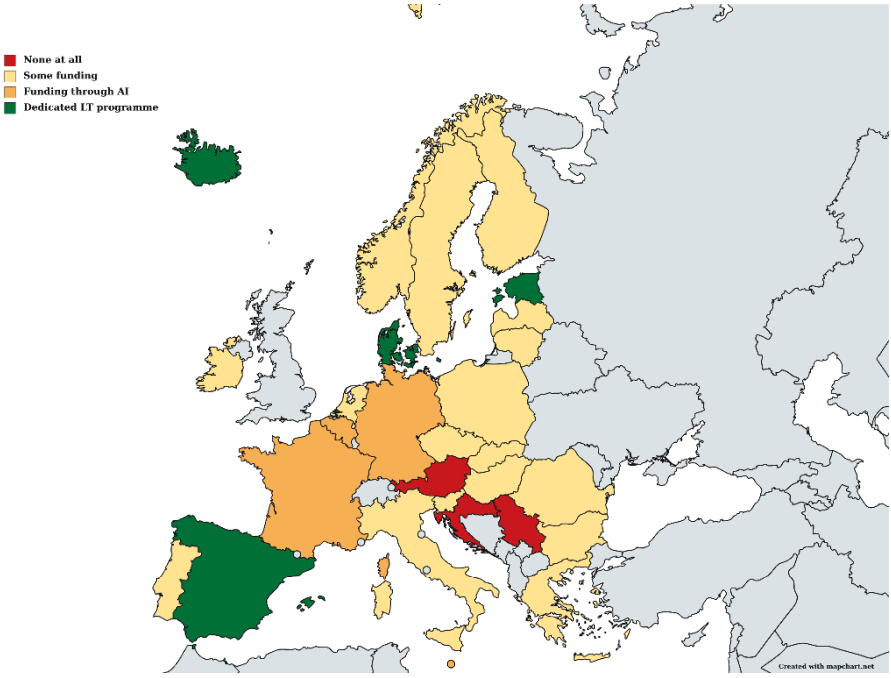

Fig. 2 The Language Technology funding situation in Europe

exclusively for LT. The same holds true for European countries that hope to incorporate LT-based AI applications, such as interactive dialogue systems and personal virtual assistants, into public services.[102]

4 SWOT Analysis

This section summarises, in the form of a SWOT analysis, the most relevant findings of the reports, documents and initiatives that were reviewed for this chapter. It attempts to identify the most significant favourable and unfavourable factors that must be addressed to make digital language equality a reality in Europe by 2030.

[102] https://digital-strategy.ec.europa.eu/en/news/new-report-looks-ai-national-strategies-progress-and-future-steps

4.1 Strengths

- Emergence of powerful new deep learning techniques, tools that are revolutionising LT.
- Important basic LT has been developed and applications that are used on a daily basis by hundreds of millions of users for speech recognition, speech synthesis, text analytics and MT are available.
- Existence of multiple national and European LT research networks, associations, communities and other relevant stakeholders whose objective is to promote all kinds of activities related to research, development, education and industry in the field of LT, both nationally and internationally.
- Existence of unique, valuable and potentially extremely useful data resources that can be exploited by current LT. An enormous amount of data is expressed in human language.
- Increasing number of companies in LT and good level of readiness for the implementation of LT in production environments.
- LT contributes to the development of inclusive digital societies, and is critical for responding to social challenges (accessibility, transparency, equity).

4.2 Weaknesses

- Deep learning LT and large pre-trained language models have shortcomings. Language models have limited real-world knowledge, can generate biased and factually incorrect text, may contain personal information, etc. They are also expensive to train and have a heavy carbon footprint. It is important to understand the limitations of large pre-trained language models and put their success in context.
- The LT markets are currently dominated by large non-EU actors, which do not address the specific needs of a multilingual Europe; Europe remains far behind due to market fragmentation, insufficient funding and legal barriers, thus hindering online commerce and communication. Europe does not fully exploit its enormous potential in LT.
- LT currently only plays a rather subordinate role in the political agenda and public debate of the EU and most of its Member States.
- There is a general misconception and over-hyping of actual AI and LT capabilities. AI is often perceived in a polarised fashion as either “magical” technology that can solve any problem or as a threat to jobs and workers, who will be replaced by machines.
- No EU policy has been proposed to address the problem of language barriers.
- GDPR/Copyright is a major barrier to the access and re-use of language resources, in competition with countries that adopt the “fair use” doctrine.

- The Open Data Directive (2019/1024/EU) does not include language data as a "high-value data category". Most of the data require extensive IPR clearing (to address Copyright and GDPR).
- There is a lack of adequate LT policies and sustainability plans at the European and national levels to properly support European languages through LT. Only four of the 30 European countries studied have a dedicated LT national programme, only six have included LT funding through national AI strategies.
- There is scarce and limited LT support for non-official EU languages.
- No European LT association is represented in the new Data, AI and Robotics public-private partnership.
- There is a lack of necessary resources (experts, HPC capabilities, etc.) compared to large US and Chinese enterprises that lead the development of new LT systems. In particular, the "computing divide" between large firms and non-elite universities increases concerns around bias and fairness within AI technology, and presents an obstacle towards democratising AI.
- Compared to English, there are far fewer LT resources and tools including language resources, annotated corpora, pre-trained language models, benchmark datasets, software libraries, etc.
- There is an uneven distribution of resources (funding, open data, language resources, scientists, experts, computing facilities, IT companies, etc.) by country, region and language.
- There is a weak open data sharing culture for many public stakeholders and SMEs.
- The investment in AI does not reflect the real importance of LT.
- There is a fragmented European market with an extremely large and varied base of about 1,000 SME companies that develop LT. Small to medium national technology companies have little capital and investment in LT capabilities. The markets are small for low-resource language speakers.
- In many countries, there are weak links between academia and industry and insufficient effective mechanisms for knowledge transfer.
- There is weak internationalisation of R&D&I and innovation.

4.3 Opportunities

- Many new powerful monolingual, multilingual and cross-lingual deep learning LT capabilities are available.
- LT is key for the realisation and support of European multilingualism.
- LT is used in practically all everyday digital products and services, since most use language to some extent, especially all internet-related products such as search engines, social networks and e-commerce services.
- LT can impact on sectors of fundamental importance to the well-being of all European citizens, such as health, administration, justice, education, culture, tourism, etc.

- LT offers effective solutions to facilitate monolingual and multilingual communication, including for the deaf and hard of hearing, the blind and visually impaired and those with language-related disabilities or impairments.
- LT is one of the most important AI application areas with a fast growing economic impact. Enormous growth is expected in the global LT market based on the explosion of applications observed in recent years and the expected exponential growth in unstructured digital data.
- Europe can play an economic leading role with its neighbouring countries through good partnerships based on the use of LT customised to other languages.
- Growing trend for the LT market and industry in Europe regarding the exploitation of digital resources and data of linguistic interest. Digitisation is one of the key means to generate new economic growth.
- Consolidation of a competitive LT industry that harnesses the potential of research and academia both in educating well-trained LT professionals and in transferring research results to industry and public administrations.
- Increasing awareness about the possibilities of AI and LT and the necessity to invest and coordinate efforts.
- Substantial breakthroughs and fast development of LT offer new opportunities for digital communication; current multilingual and cross-lingual deep learning LT allows for the creation of new multilingual pre-trained language models and systems that can leverage and balance LT across all European languages.
- Ensure openness of infrastructures for data and technologies.

4.4 Threats

- In comparison to 2012, the results of the European Language Equality project in 2022 show that the gap between English and all other languages appears to be getting *bigger* instead of smaller.
- Development of non-explainable techniques and deep learning models without any commonsense or up-to-date knowledge, with social biases, containing personal and private data, with a heavy impact on carbon footprint, etc.
- AI is a broad area, which overshadows and dwarfs the importance, benefits and contributions of LT, especially in Europe.
- Loss of LT skills and human capital trained in Europe due to the lack of sufficient research, transfer and funding opportunities.
- Inability to retain in, or attract to, the EU researchers and workers skilled in LT and AI.
- Growing development of the sector in US and China that will eventually penetrate the European application market, limiting the Digital Language Equality opportunities as described in this report.
- The complexity of copyright, GDPR, Open Data directives etc. makes access to language resources too costly, unclear and risky.

- Fear of many jobs becoming redundant due to the deployment of AI-powered technologies.

5 Conclusions

Europe's multilingual nature is also one of the main obstacles to a truly connected, cross-lingual communication and information space. Moreover, while language diversity is at the core of European identity, many of our languages are in danger of digital extinction because they are not sufficiently supported through Language Technologies (Moseley 2010; Rehm and Uszkoreit 2012; STOA 2018; European Parliament 2018).[103] Sophisticated multilingual, cross-lingual and monolingual LT for all European languages would future-proof our languages as cornerstones of our cultural heritage and richness. In recent years, European research in LT has faced increased competition from other continents, especially with respect to breakthroughs in AI. These scientific advancements have led to global commercial successes, from which the respective regions benefit especially. As a consequence, many European scientists, including young high-potential researchers, are leaving Europe to continue their work abroad. Europe must invest in retaining and attracting these researchers. Our continent is in need of powerful LT made *in* Europe *for* all European citizens, tailored to our unique cultures, societies and economic requirements so that a linguistically fragmented Europe may become a truly unified and inclusive one. This ambitious but worthy effort involves supporting its rich and diverse linguistic cultural heritage, from broadly spoken languages to minority and regional languages, as well as the languages of immigrants and important trade partners, benefiting European citizens, European industry and European society.

References

Aldabe, Itziar, Georg Rehm, German Rigau, and Andy Way (2022). *Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (second revision)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-strategic-documents-v3.pdf.

Chui, Michael, Martin Harryson, James Manyika, Roger Roberts, Rita Chung, Ashley van Heteren, and Pieter Nel (2018). "Notes from the AI frontier: Applying AI for social good". In: *McKinsey Global Institute*.

Curry, Edward, Andreas Metzger, Sonja Zillner, Jean-Christophe Pazzaglia, and Ana García Robles, eds. (2021). *The Elements of Big Data Value: Foundations of the Research and Innovation Ecosystem*. Cham: Springer.

de Jong, Franciska, Bente Maegaard, Darja Fišer, Dieter van Uytvanck, and Andreas Witt (2020). "Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Mar-

[103] http://www.unesco.org/languages-atlas/index.php?hl=en&page=atlasmap

seille, France: European Language Resources Association, pp. 3406–3413. https://aclanthology.org/2020.lrec-1.417.

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Moseley, Christopher (2010). *Atlas of the World's Languages in Danger*. http://www.unesco.org/culture/en/endangeredlanguages/atlas.

Padilla, Thomas (2020). "Responsible Operations: Data Science, Machine Learning, and AI in Libraries". In: *American Archivist* 83, pp. 483–487.

Rehm, Georg, ed. (2017). *Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda*. CRACKER and Cracking the Language Barrier federation. http://cracker-project.eu/sria/.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch (2020a). "European Language Grid: An Overview". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3359–3373. https://www.aclweb.org/anthology/2020.lrec-1.413/.

Rehm, Georg and Stefanie Hegele (2018). "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs". In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: ELRA, pp. 3282–3289. https://aclanthology.org/L18-1519.pdf.

Rehm, Georg, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020b). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: ELRA, pp. 3315–3325. https://www.aclweb.org/anthology/2020.lrec-1.407/.

Rehm, Georg, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres

Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals (2021). "European Language Grid: A Joint Platform for the European Language Technology Community". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*. Kyiv, Ukraine: ACL, pp. 221–230. https://www.aclweb.org/anthology/2021.eacl-demos.26.pdf.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Rehm, Georg and Hans Uszkoreit, eds. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg etc.: Springer. http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf.

Rehm, Georg, Hans Uszkoreit, Sophia Ananiadou, Núria Bel, Audronė Bielevičienė, Lars Borin, António Branco, Gerhard Budin, Nicoletta Calzolari, Walter Daelemans, Radovan Garabík, Marko Grobelnik, Carmen García-Mateo, Josef van Genabith, Jan Hajič, Inma Hernáez, John Judge, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Joseph Mariani, John McNaught, Maite Melero, Monica Monachini, Asunción Moreno, Jan Odjik, Maciej Ogrodniczuk, Piotr Pęzik, Stelios Piperidis, Adam Przepiórkowski, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Sandford Pedersen, Inguna Skadiņa, Koenraad De Smedt, Marko Tadić, Paul Thompson, Dan Tufiş, Tamás Váradi, Andrejs Vasiļjevs, Kadri Vider, and Jolanta Zabarskaite (2016). "The Strategic Impact of META-NET on the Regional, National and International Level". In: *Language Resources and Evaluation* 50.2, pp. 351–374. DOI: 10.1007/s10579-015-9333-4. http://link.springer.com/article/10.1007/s10579-015-9333-4.

Rehm, Georg, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Váradi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, and Sigve Gramstad (2014). "An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"". In: *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*. Ed. by Laurette Pretorius, Claudia Soria, and Paola Baroni. Reykjavik, Iceland, pp. 30–37. http://georg-re.hm/pdf/CCURL-2014-META-NET.pdf.

Soria, Claudia, Nicoletta Calzolari, Monica Monachini, Valeria Quochi, Núria Bel, Khalid Choukri, Joseph Mariani, Jan Odijk, and Stelios Piperidis (2014). "The language resource Strategic Agenda: the FLaReNet synthesis of community recommendations". In: *Language Resources and Evaluation* 48, pp. 753–775. https://doi.org/10.1007/s10579-014-9279-y.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. https://data.europa.eu/doi/10.2861/136527.

Zhang, Daniel, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault (2021). "The AI Index 2021 Annual Report". In: DOI: 10.48550/ARXIV.2103.06312. https://arxiv.org/abs/2103.06312.

Chapter 45
Strategic Research, Innovation and Implementation Agenda for Digital Language Equality in Europe by 2030

Georg Rehm and Andy Way

Abstract This chapter presents the ELE Programme (ELE Consortium 2022). Reacting to the landmark resolution (European Parliament 2018), its vision is to achieve digital language equality in Europe by 2030. The programme was prepared jointly with many stakeholders from the European Language Technology, Natural Language Processing, Computational Linguistics and language-centric AI communities, as well as with representatives of relevant initiatives and associations, and language communities. Europe still suffers from strong inequalities in terms of technology support of its languages. English is still by far the language with the best technological support, followed by a cluster of three languages (German, Spanish, French) that already have only half the technological support of English. More than half of the around 90 languages surveyed have either weak or no technological support at all. The ELE Programme is foreseen to be a shared, long-term funding programme tailored to Europe's needs, demands and values. For the EU we foresee the role of providing resources for coordinating the programme, for providing shared infrastructures, for maintaining the scientific goals and programme principles, etc. The participating countries have the role of providing resources for the development of technologies and datasets for their own languages. Key goals are to reduce the technology gap between English and all other European languages and to address the lack of available language data. The ELE Programme tackles the following overarching themes: *Language Modelling*, *Data and Knowledge*, *Machine Translation*, *Text Understanding* and *Speech*. These interconnected themes focus upon the socio-political goal of establishing DLE in Europe and on the scientific goal of Deep Natural Language Understanding, both by 2030.[1]

Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Germany, georg.rehm@dfki.de

Andy Way
Dublin City University, ADAPT Centre, Ireland, andy.way@adaptcentre.ie

on behalf of the whole European Language Equality consortium and all contributors.
European Language Equality EU Project, coordinator@european-language-equality.eu

[1] This chapter is a revised version of the ELE *Strategic, Research, Innovation and Implementation Agenda for Digital Language Equality* (ELE Consortium 2022), which is also available online: https://european-language-equality.eu/agenda/.

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_45

1 Executive Summary

The overall vision of the ELE Programme is to achieve complete digital language equality (DLE) in Europe by 2030. The programme was prepared jointly with many relevant stakeholders from the European Language Technology (LT), Natural Language Processing (NLP), Computational Linguistics and language-centric Artificial Intelligence (AI) communities, as well as with representatives of relevant initiatives and associations, and language communities. The ELE Programme responds to the call "to establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe's needs and demands", as specified by the European Parliament Resolution *Language equality in the digital age* (European Parliament 2018). The results of the ELE project show that English is still by far the language with the best and most thorough technological support, followed by a cluster of three languages (German, Spanish, French) that have only half the technological support of English. After yet another gap, the long tail of languages with fragmentary support starts with Finnish, Italian and Portuguese. More than half of the around 90 languages surveyed have either weak or no technological support at all. In comparison to previous results from 2012 (Rehm and Uszkoreit 2012), the gap between English and the other languages appears to be getting *bigger* instead of smaller. With the exceptions of English, German, French and Spanish, all languages we investigated exist in socio-political and economic ecosystems that do *not* incentivise, encourage or foster the development of technologies for these languages. While all 30 European countries we surveyed have put in place national AI strategies, almost all of these national strategies seem to have either ignored or left out the topic of languages and language-centric AI.[2]

The ELE Programme is foreseen to be a shared, long-term, coordinated and collaborative LT funding programme tailored to Europe's needs, demands and values, including multilingualism and language equality in general. For the EU we foresee the role of providing resources for coordinating the programme, for providing shared infrastructures, for maintaining the scientific goals and programme principles, etc. The participating countries have the role of providing resources for the development of technologies and datasets for their own languages. Key goals are to reduce the technology gap between English and all other European languages and to address the lack of available language data: this is true for all European languages except English. The ELE Programme focuses upon *openness*: open source, open access and open standards as well as interoperability and standardisation. A key emphasis is on the creation of large open access language models for all European languages, including

[2] Despite our original findings, in the interim, Spain has funded the 1.1B€ PERTE New Economy of Language programme to "maximize the value of Spanish and co-official languages in the new digital economy and artificial intelligence", see https://planderecuperacion.gob.es/como-acceder-a-los-fondos/pertes/perte-nueva-economia-de-la-lengua. Accordingly, rather than be seen as a laggard in this space, Spain now represents what could and should be done to support European languages and associated technology, and the PERTE programme stands as a template for other nations to adapt to their particular situation.

the creation of datasets and multilingual models, symbolic knowledge, models that include discourse capabilities as well as grounding and other sophisticated features currently out of reach for existing state-of-the-art technologies. The ELE Programme is expected to have a runtime of nine years. In addition to overall coordination, the ELE Programme tackles the following overarching themes: *Language Modelling*, *Data and Knowledge*, *Machine Translation*, *Text Understanding* and *Speech*. These interconnected themes focus upon the socio-political goal of achieving DLE in Europe and on the scientific goal of Deep Natural Language Understanding, both by 2030. The ELE Programme strengthens and makes optimal use of infrastructures, data spaces and services provided by other European initiatives.

The global NLP market is estimated to reach 341.7B$ by 2030. In contrast, the modest investment needed to implement the ELE Programme will not only bring about DLE in Europe but it will also move European research and industry in this field into a dominant position for years to come.

2 Multilingual Europe and Digital Language Equality

Languages are the most common and versatile way for humans to convey and access information. We use language, our most natural means of communication, to encode, store, transmit, share and manipulate information. We use language in everyday life to interact with others and our environment and as social glue, to express and to explain ourselves, to convince, agree with and rebut others. Our laws and constitutions are written in language. We use it in science, commerce, teaching and passing on knowledge to the next generations, for pleasure, creativity and aesthetic enjoyment in puns, jokes and literature. History and culture are recorded, interpreted and enjoyed through language. Our languages are a core part of our identities.

Human languages are incredibly complex: a single word (phrase, sentence, text) can have many meanings, a single meaning can be expressed by many different words (but meaning depends on linguistic and situational context), we can use language literally and metaphorically, language and knowledge are highly intertwined, we do not articulate important parts of a message if these parts are presumed shared knowledge by the community (this includes situational knowledge), important parts of meaning reside in what can be inferred from what has been said. At the same time, language changes. New words are invented, some old ones are dropped, even the structure (syntax and morphology) of languages and the meaning of words change over time. These aspects make human languages fundamentally different from the formal languages of mathematics, logic and computer science. This is also what makes human languages so efficient, elegant, flexible and enjoyable. Finally, there are many human languages (6,000+), not even counting regional and dialectical variants. All these aspects are at the core of human languages and they make it hard for computers to "fully understand" human language and to "properly" process human language in the context of "full and deep understanding".

Languages are at the heart of every aspect of life and their role is crucial to the future of European countries, citizens, businesses, and of the European Union as a whole. Full Digital Language Equality (see Chapter 3) in Europe can deliver an impact in the following four high-priority areas.

Digital Language Equality will have a positive and unprecedented impact on all European languages. We must ensure that no European languages remain under-resourced (see Chapter 4 for an overview and Chapters 5 to 37 for in-depth analyses), but that they are equipped with the same high level of technological support already enjoyed by very few of them (Chapter 2). This, in itself, will deliver a major impact on all European citizens and businesses: supporting all languages in the interest of equality and fairness empowers and brings advantages to their speakers, while reflecting the democratic and inclusive spirit of the EU.

Digital Language Equality will make a contribution to establishing a fair, inclusive and sustainable multilingual Digital Single Market: this will be achieved by helping to future-proof all European languages through digital technologies, and especially preventing the threat of digital extinction for those that suffer from weak support. By fostering a more inclusive and cooperative business and social environment, companies and citizens will benefit from sharing knowledge, digital services and products on an equal footing, overcoming the fragmentation that is caused by many European languages lagging behind, which severely penalises their speakers as well as regional and local communities. Action in this vital area is particularly urgent due to the increasing range of economic, educational and social opportunities that are afforded online and delivered remotely, from e-commerce to online shopping, to web-based recruitment services, online teaching programmes and professional training courses, among others.

Digital Language Equality will help research in Europe, mobilising and leveraging their full potential to start reclaiming scientific and industrial leadership from US-based and Asian competitors, particularly large tech enterprises as well as academic institutions and research centres, that pose fierce competition in several fields. The ELE Programme will instigate regional, national and EU-wide collaboration among scientists from academia and industry covering a broad range of disciplines, ensuring the mix of competencies that is required to deliver substantial and lasting impact at the forefront of scientific and technological progress.

Digital Language Equality will act as a multiplier of opportunities. It will help to aggregate the players that are required to unlock the full potential of an EU-wide effort to exchange and share widely-agreed methodologies, resources and technologies with a focus on promoting the digital equality of European languages: this will benefit the use and promotion of all European languages, encouraging in particular those that have traditionally lagged behind.

3 What is Language Technology and How Can it Help?

Language Technology (LT) is concerned with studying and developing systems capable of processing human languages. Over the years, the field has developed different methods to make explicit the information contained in written and spoken language – and increasingly for other modalities such as sign language, for example – or to generate or synthesise written or spoken language (see Chapter 2 for more details). Despite the inherent difficulty of many of the tasks performed, current LT support allows many advanced applications which would have been unthinkable only a few years ago. LT is present in our daily lives, for example, through search engines, recommendation systems, virtual assistants, chatbots, authoring assistants, text predictors, automatic translation systems, automatic subtitling, automatic summarisation tools, etc. Its rapid development in recent years predicts even more encouraging and also exciting results in the near future. LT is providing solutions for the following main application areas: Machine Translation, Speech Processing, Text Analysis, Information Extraction and Information Retrieval, Natural Language Generation, Human-Computer Interaction (see Chapter 2 as well as Chapters 40 to 43 for in-depth analyses of the state-of-the-art).

4 A Shared European Programme for Language Technology and Digital Language Equality in Europe by 2030

Fully in line with the recommendations of the European Parliament resolution *Language equality in the digital age* (European Parliament 2018), our recommendations, as analysed in the chapters of the present book can be summarised as follows.

The vision described in this book is fully compatible with current EU policy, needs and demands; in fact, they are mission-critical. Missing investment in the underdeveloped areas of LT and language-centric AI will result in the digital extinction of languages, i. e., only global languages spoken by large numbers of speakers, including, crucially, outside the EU, will prevail and the global LT/NLP market will continue to be dominated by the US and China, while the European LT community will be pushed aside even further.

The main concept of the ELE Programme is a collaboration between the EU, and in particular the European Commission, and all participating countries and regions since funding and further investment are needed on all levels. Funding on the level of the EU should enable overarching coordination and EU-wide technological infrastructure. It should cover the topics which require pan-European coordination such as shared tasks, protocols, multilingual dataset creation based on the same principles in line with European values and priorities, etc. Coordination on the European level is needed because language communities are still too fragmented and mostly too small. Further effort should be invested into adequate policy-making, distributed research infrastructures and technological platforms like ELG (Rehm 2023) and the

Common European Language Data Space, with flexible access to sufficient High Performance Computing (HPC) facilities. Additionally, national and regional funding should complement the European funding with regard to language-specific research and development. The main gaps to be filled in these respects and the most important anticipated developments are described, among others, in the language reports (see Chapters 5 to 37).

This section summarises our main recommendations for this shared programme (more detailed recommendations are contained in the previous chapters of this book). First, we outline the possible cornerstones for suitable policy and infrastructure recommendations, as well as ideas for the realisation of a governance model. Second, we revise the technology and data recommendations suggested by the ELE consortium (derived from Chapters 39 to 44), which are closely related to those discussed in the *Language equality in the digital age* resolution (European Parliament 2018).

Further, in terms of our research recommendations, the ELE consortium together with the wider LT community has developed a clear vision for the different areas of LT. We see an urgent need to refocus and massively strengthen European LT/NLP research through a large-scale initiative as a shared, collaborative pan-European effort between the EU and those countries and regions that participate in the initiative, i. e., the *ELE Programme*. This endeavour should include the participation of research centres, academia, companies (particularly SMEs and startups), and other relevant stakeholders. As LT is aggregated and applied to more complex settings, interdisciplinary research and activities are becoming more relevant in order to further boost developments and allow synergies to become apparent. To achieve *Deep Natural Language Understanding*, we need to finance and investigate fields such as cognitive, neural and symbolic AI further.

The ELE Programme should boost pan-European long-term basic research as well as knowledge and technology transfer between research labs and industry. Frequently mentioned areas and tasks for basic and applied research where further investigation is needed include, among others, systematic language data collection (text, dialogue, vision, sign language and other forms of interactions), speech analysis, AI, human-computer interaction, machine learning, robotics, natural language understanding and processing tasks such as machine reading, text analysis, machine translation, chatbots, virtual assistants and summarisation.

4.1 Policy Recommendations

- Reinforce European leadership in LT by establishing the ELE Programme as a large-scale, long-term coordinated funding programme for research, development, innovation and education with the *societal goal* of achieving Digital Language Equality in Europe and the *scientific goal* of Deep Natural Language Understanding, both by 2030.
- Ensure comprehensive EU-level legal protection for the more than 60 regional and minority languages spoken in Europe.

- Empower recognition of the collective rights of national and linguistic minorities in the digital world (including sign languages).
- Encourage mother-tongue teaching for speakers of official and non-official languages of the EU.
- Safeguard sufficient funding to support new technological approaches, based on increased computational power and better access to sizeable amounts of data.
- Develop specific initiatives within current funding schemes, especially Horizon Europe and Digital Europe (including the Recovery Plan for Europe), to boost long-term basic research as well as knowledge and technology transfer between countries and regions, and between academia and industry.
- Support the coordination between research and industry to enhance the digital possibilities for LT and Open Access to language data.
- Define and develop a minimum set of language resources and capacities that all European languages should possess (see Krauwer 2003) .
- Develop common policy actions and protocols for language data sharing by public administrations at all levels. Language data should be included as a high-value data category in the Open Data Directive (2019/1024/EU).
- Enable and empower European SMEs and startups to easily access and use LT in order to grow their businesses independent of language barriers, also thanks to e-commerce and online marketplaces.
- Create the necessary appealing conditions to attract and retain qualified and diverse international LT personnel in Europe.
- Encourage all EU-funded projects to have a language diversity plan and to include direct or associated partners from a less-widely spoken language.
- Empower and encourage administrations at all levels to improve access to online services and information in different languages.
- Create a European network of centres of excellence in LT to increase industry visibility and to design national research agendas.
- Implement and maintain long-term an overall EU-wide policy framework to achieve European LT sovereignty.
- Facilitate EU Member States' acquisition of LT for their local industries without depending on non-European technology providers.

4.2 Governance Model

- Structure the ELE Programme as a shared, collaborative and coordinated programme between the EU and all countries and regions that participate.
- Allocate the area of multilingualism, linguistic diversity and language technology to the portfolio of a EU Commissioner.
- Set up a large lobby for EU regional and minority languages.
- Create a pan-European network of research centres to facilitate the coordination and also implementation of the ELE Programme at all levels.

- Promote a distributed centre for linguistic diversity that will strengthen awareness of the importance of lesser-used, regional and minority languages.
- Design and apply new forms of research funding and organisation to ease the transition from application-oriented basic research to commercially-focused technology development.

4.3 Technology and Data Recommendations

- Develop large open-source language models that work for all European languages, optimised in terms of compute time and cost.
- Address the lack of available data and define the minimum amount of language resources and capabilities that all European languages should possess.
- Add more focus on systematic and comprehensive language data collection (text, dialogue, multimodal) and exploit automatic data generation (synthetic data), crowd-sourcing and translation of high-quality data.
- Develop new methodologies for transfer and adaptation of resources and technologies to other domains and languages.
- Develop high-performance applications (in terms of speed and quality) for all languages that respect safety, security and privacy.
- Ensure efficient adaptations to applications, both in terms of language, domain, efficiency, power consumption, ease of maintenance, and quality assurance.
- Develop methods to overcome the unequal data availability, by focusing on, e. g., annotation transfer, multilingual models preserving quality, few-shot or zero-shot learning.
- Unleash the power of monolingual and multilingual public sector data, data from broadcasters, social media, publishers, etc.
- Enforce open ecosystems, open standards and interoperability (including Open Source and Open Access).
- Focus on research on bias for strengthening inclusiveness and accessibility, to respect and promote European values and principles.
- Focus upon Green LT with a small compute and carbon footprint (e. g., model compression).
- Foster publicly available resources that facilitate innovation and research for both commercial and non-commercial actors.
- Construct a multilingual LT benchmark, a European "SuperGLUE"-style (Wang et al. 2019) shared benchmark, that tracks progress.
- Define the minimum language resources that all European languages should possess in order to prevent digital extinction.

4.4 Infrastructure Recommendations

- Strengthen existing and create new research infrastructures and LT platforms that support research and development activities, including collaboration, knowledge sharing, and Open Access to data, tools and technologies.
- Fill the identified gaps in data, language resources and knowledge graphs and create a future path for Europe towards comprehensive and interlinked data infrastructures.
- Develop clear and robust protocols to ensure flexible access to sufficient GPU-based HPC infrastructure and robust protocols to process sensitive data.
- Ensure sufficient operational capacity, especially for Large Language Models (LLMs) and flexible access to GPU-based HPC facilities.
- Follow the idea of a Semantic Data Fabric including rich semantics for the development of an integrated and interoperable data infrastructure.

4.5 Research Recommendations

4.5.1 Recommendations for all Research Areas

- Gather and make available the critical mass of resources in terms of data, HPC facilities, and expertise from pan-European LT research labs and centres, with support from the EC as well as national and regional administrations.
- Create sufficient multilingual and multimodal data of quality (responsible, legal, diverse, unbiased, ethical, representative, etc.), in all European languages and domains (media, health, legal, education, etc.).
- Provide flexible access to HPC facilities for LT research and industry. HPC facilities should provide clear and robust protocols to process sensitive data.
- Develop better benchmarks and datasets (ethical, responsible, legal, etc.) for all languages, domains, tasks and modalities.
- Combine interactive LT (conversational AI) with text, knowledge, and multimedia technologies for a new generation of applications that can address the deeper questions of communication, common sense and reasoning.
- Encourage trustworthy, unbiased, inclusive, non-discriminatory LT/AI, making interpretability and explainability of AI models a priority.
- Develop further the areas of responsible AI by combining statistical and symbolic AI in multilingual environments to provide AI-based applications that deliver accurate results and benefits for research, industry, and society.
- Focus on methods and learning architectures to overcome the highly unequal data availability, such as annotation transfer, synthetic data and their proper use in machine learning, multilingual models preserving quality and coverage and few-shot or zero-shot learning.

- Focus on Green LT and investigate new efficient methods to extend, reuse and adapt existing pre-trained language models or develop new ones with much reduced carbon footprint.
- Develop language- and culture-specific technologies that cover more linguistic phenomena and text types, focusing on accessibility, through sign language, avatar technology, etc.
- Provide transparency of AI models with regard to accuracy and fairness.
- Reframe LT/NLP as a quantum computing problem.

4.5.2 Machine Translation

- Develop near-real-time MT across all modalities (speech, text, signs, etc.) and adaptive MT, where the system learns from interaction with users.
- Move towards context-aware methodologies that go beyond text data and include images, videos, tables, etc. by developing multimodal MT systems.
- Develop low-resource MT by deepening research on projection and structural organisation of embeddings to comprehend how structurally different languages and their respective embedding spaces can be mapped to one another.

4.5.3 Speech Processing

- Enhance speech resources and create acoustic models to cover all European languages, including non-standard varieties and dialects.
- Improve the handling of audio conditions currently perceived as difficult (e. g., multiple simultaneous speakers in noisy environments speaking spontaneously and highly emotionally in a mix of languages).
- Develop high-quality, natural synthetic voices, allowing users to obtain content in the language of their choice.
- Improve context modelling to handle the translation of speech models across larger volumes of text.
- Support research in the direction of combining speech, NLU and NLP with other modalities, such as image and vision.
- Address privacy and security threats in areas of speech synthesis, voice cloning and speaker recognition.

4.5.4 Text Analytics and Natural Language Understanding

- Create large Open-Access language models for all European languages (for fine-tuning and downstream tasks), datasets (for training and testing), multilingual models, models that include symbolic knowledge and discourse features.
- Increase the adoption of approaches based on self-supervised, zero-shot, and few-shot learning.

- Support research in NLU which integrates speech, NLP, and contextual information as well as additional modes of perception.
- Strengthen basic research in neurosymbolic approaches to NLP/NLU, including grounding and the use of human-understandable databases and sources.
- Strengthen progress in reinforcement-based learning, novel dialogue management strategies, and situation-aware natural language generation.
- Strengthen interdisciplinary research and enable better modelling of multimodal environments.

4.6 Implementation Recommendations

- Structure the ELE Programme into three phases of similar duration.
- Facilitate discussions between the EU, the European Commission in particular, and all participating countries to define the goals and the financial setup.
- Encourage participating countries to invest into the development of large language models, data sets, technologies, and tools for their own languages.
- Encourage the EU to establish legislation to promote participation.
- Encourage the EU to invest in the pan-European coordination of all language-specific projects and initiatives, support mechanisms, infrastructures, data procedures, cross-cutting projects, etc. and provide flex funds for bootstrapping poorly supported languages.
- Structure the ELE Programme into six themes covering: Language Modelling, Data and Knowledge, Machine Translation, Text Understanding, Speech, and Infrastructure and support each theme by coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs).

5 Roadmap towards Digital Language Equality in Europe

5.1 Main Components

Language Technologies have the potential to overcome the linguistic divide in the digital age. However, we need to define actions, tools, processes and actors that need to be involved. The ELE SRIA includes a roadmap with concrete steps for the implementation that carry tangible and measurable outputs.

The main scientific goal of the ELE Programme is *Deep Natural Language Understanding in Europe by 2030*. Efficiency will be increased by sharing knowledge, infrastructures and resources, with a view to developing innovative technologies and services, in order to achieve the next scientific breakthrough in this area and help reduce the technology gap between Europe's languages with the collaboration of research centres, academic experts, industry and other relevant stakeholders. Crucially,

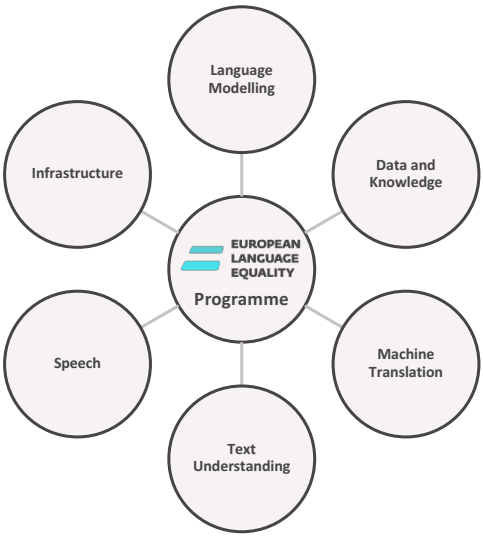

Fig. 1 The six main themes of the ELE Programme

the long-term ELE Programme will involve significantly intensified coordination between the participating countries and languages.

The main societal and economic goal of the ELE Programme is *Digital Language Equality in Europe in 2030*. The focus is on language equality and the provisioning of technologies, services and resources outside the often-preferred languages to achieve and maintain long-term technological sovereignty in this crucial application area. For regional, minority and lesser spoken languages, we need to find a (technological) way to consider Deep Natural Language Understanding within a common approach, to create synergies and increase efficiency of the solutions and their design and development. To narrow the digital divide, there is a pressing urgency for novel techniques that would bring less-resourced languages to a level comparable to state-of-the-art results for resource-rich languages. This includes the leveraging of multimodal and multilingual resources to support the development of applications for languages and varieties with scarce resources.

This roadmap towards Digital Language Equality in Europe by 2030 provides a path and the means to ensure that the two goals outlined above are met. To tackle this challenge, the ELE Programme combines the following six themes (see Figure 1).

Language Modelling This theme includes research, development and deployment activities regarding LLMs, especially multilingual and multimodal, generative LLMs that include text, speech, image, video, etc. Time and resources need to be invested for experiments, developing novel approaches, shared tasks, etc. For novel research approaches we need to combine national projects and data sets with international consortia. With regard to innovation and deployment, LLMs will be applied in industrial sectors and use-cases.

Data and Knowledge The Data and Knowledge theme is focused on the collection, production, annotation, curation, quality assessment, standardisation, etc. of text data, spoken data, video data, and other multimodal data, primarily with regard to their application as data for pre-training different sorts of LLMs.

Machine Translation The MT theme is focused on improving the automated translation from one natural language into another (including sign languages and other modalities). While Europe has a strong foundation in this field, research needs to combine novel, groundbreaking approaches with results of the Data and Knowledge as well as Language Modelling themes (see above). The results need to be applied in different industrial sectors and use-cases. Deployment needs to be fast, agile and driven by excellent teams.

Text Understanding The Text Understanding theme aims to improve the identification and labelling of information regarding all levels of linguistic analysis underlying any natural language text (or other modalities). This requires exploring new strands of research and building on synergies with the other themes. An equally important aspect is applicability in industry.

Speech The Speech theme addresses one big challenge of the European LT community, i. e., the shift from text-to-speech and multimodal processing (including research towards grounding). While progress in the area of speech applications has been made in the last decade, we also need novel research paradigms. This theme will benefit from the themes Data and Knowledge as well as Language Modelling. The development of relevant industry applications is another goal.

Infrastructure The Infrastructure theme involves the extension, maintenance and interoperability of platforms such as European Language Grid (ELG) and Language Data Space (LDS). ELG has the potential of functioning as one of the primary platforms to support the activities of the ELE Programme. Moreover, ELG can be further developed into the focal point for best practices and the development of bridges to other relevant platforms. New features and functionalities need to be implemented for a higher adaptability. Other important factors are the provisioning of GPUs and standardisation.

5.2 Actions, Budget, Timeline, Collaborations

The *Language equality in the digital age* resolution (European Parliament 2018) strongly encourages the "establish[ment of] a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, […] tailored specifically to Europe's needs and demands".

As a direct response, the ELE project (Rehm et al. 2022a) has developed the DLE Metric (Gaspari et al. 2022a; Grützner-Zahn and Rehm 2022) as a measure to assess and track the advancement towards DLE in Europe empirically (Chapter 3) and, in parallel, an outline of necessary actions. These have been informed by 66 project

reports[3] that comprise more than 2400 pages with condensed findings, summarised in the form of the present book. A total of 92 languages have been taken into account. We have included voices from research, industry and civil society. In terms of research on Europe's languages, we prepared over 30 reports on the situation of individual languages (Chapters 5 to 37, Chapter 4 contains an overview analysis). In addition, we collected input through various surveys and more than 60 expert interviews (Chapters 4, 38 and 39). To cover the industry angle, our industry partners produced four technical deep dives and collected feedback in a number of surveys for further information (Chapters 40 to 43). Civil society was represented by the European citizen survey with about 20,000 responses (Chapters 4, 38 and 39).

The ELE Programme has a foreseen runtime of nine years, divided into three phases of three years each. Implementing the ELE Programme will significantly improve the state-of-the-art of LT and NLP and language-centric AI research (Chapter 2), create DLE in Europe and put Europe back into the global pole position of research and industrial applications of this type of technology (Chapter 44).

5.2.1 Actions

We foresee different types of projects, implemented using the different EC project types: coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs), see Table 1.

Coordination and Support Actions (CSAs) are needed to support research activities and policies (networking, exchange, access to research infrastructures, conferences, etc.). The ELE Programme envisages three CSAs for the overall programme coordination. These include, among others, the maintenance of the ELE principles, quality assurance approaches, shared tasks, etc. Additional CSAs are needed for the themes *Data and Knowledge* as well as *Language Modelling* as these are fundamental for all other themes as well. Another CSA is needed for supporting and further developing shared infrastructures.

[3] See Gaspari et al. (2021), Agerri et al. (2021), Gaspari et al. (2022b), Sarasola et al. (2022), Koeva and Stefanova (2022), Melero et al. (2022a), Tadić (2022), Hlavacova (2022), Pedersen et al. (2022), Steurs et al. (2022), Maynard et al. (2022), Muischnek (2022), Lindén and Dyster (2022), Adda et al. (2022), Sánchez and García-Mateo (2022), Hegele et al. (2022a), Gavriilidou et al. (2022), Jelencsik-Mátyus et al. (2022), Rögnvaldsson (2022), Lynn (2022), Magnini et al. (2022), Skadiņa et al. (2022), Gaidienė and Tamulionienė (2022), Anastasiou (2022), Rosner and Borg (2022), Eide et al. (2022), Ogrodniczuk et al. (2022), Branco et al. (2022), Păiș and Tufiș (2022), Garabík (2022), Krek (2022), Melero et al. (2022b), Borin et al. (2022), Prys et al. (2022), Krstev and Stanković (2022), Ćušić (2022), Giagkou et al. (2022), Moshagen et al. (2022), Robinson-Jones and Scarse (2022), Hajič et al. (2021), Thönnissen (2022), Eskevich and Jong (2022), Rufener and Wacker (2022), Hajič et al. (2022), Hegele et al. (2022b), Gísladóttir (2022), Kirchmeier (2022), Hicks (2022), Blake (2022), Hrasnica (2022), Heuschkel (2022), Bērziņš et al. (2022), Backfried et al. (2022), Gomez-Perez et al. (2022), Kaltenböck et al. (2022), Way et al. (2022b), Way et al. (2022a), Aldabe et al. (2022b), Aldabe et al. (2022a), ELE Consortium (2022), Hegele et al. (2021a), Hegele et al. (2021b), Rehm et al. (2022b), Marheinecke et al. (2022) and Rehm et al. (2022c).

Research and Innovation Actions (RIA) are collaborative projects funding research activities that allow the exploration of new technologies, new methods, new products, or improvements of existing ones. Research is the fundamental prerequisite for DLE. Over the last decade, the community has developed a clear vision of the work needed in the different areas of LT. To achieve Deep NLU, we need to invest in and further research the areas of language modelling, machine translation, text understanding and speech.

Innovation Actions (IAs) consist of activities directly aiming at producing improved products, processes or services. They may include prototyping, testing, demonstrating, piloting, large-scale product validation and market replication.

	Type	Number
ELE Programme – overall coordination	CSA	3
Theme Data and Knowledge – coordination	CSA	3
Theme Language Modelling – coordination	CSA	3
Theme Language Modelling – research	RIA	15
Theme Language Modelling – innovation and deployment	IA	15
Theme Machine Translation – research	RIA	12
Theme Machine Translation – innovation and deployment	IA	12
Theme Text Understanding – research	RIA	12
Theme Text Understanding – innovation and deployment	IA	12
Theme Speech – research	RIA	12
Theme Speech – innovation and deployment	IA	12
Theme Infrastructure – support	CSA	3

Table 1 Different types and number of projects foreseen in the ELE Programme

5.2.2 Budget

As a shared programme between the EU and the participating countries, the final financial setup needs to be discussed between all involved parties. For the EU part of the budget, we suggest the breakdown shown in Table 2. In addition to these investments, which relate to the overarching coordination, research and innovation projects, the participating countries and regions are expected to invest in their languages themselves, while the languages with fragmentary, weak or no technical support can request funding from the European Union (*flexible funds*, see below).

In addition to the sum of 690M€ for the actions implementing the theme-related projects of the ELE Programme, we envisage investing an additional 150M€ as *flexible funds* for languages with fragmentary, weak or no technical support since we anticipate that a number of participating countries will require complementary fund-

ELE Programme (overall coordination)	60M€
Theme Data and Knowledge	45M€
Theme Language Modelling	195M€
Theme Machine Translation	120M€
Theme Text Understanding	120M€
Theme Speech	120M€
Theme Infrastructure	30M€
Sum	**690M€**
Flexible funds	150M€
Total	**840M€**

Table 2 Budget breakdown of the ELE Programme (EU contribution only; numbers are indicative)

ing from the EU. A more detailed breakdown of the different themes with their associated project types and runtime is shown in Table 4.

The complementary national/regional investments required on the individual language level are difficult to predict. We group the languages into three clusters (see Table 3) and provide indicative investments, which relate to the whole duration of the ELE Programme. Other factors (e. g., number of speakers, etc.) can be taken into account to arrive at more precise numbers.

Languages with *weak or no support*	40-50M€ each
Languages with *fragmentary support*	30-40M€ each
Languages with *moderate support*	20-30M€ each

Table 3 Indicative investments required by language, provided by the participating countries

This language-specific funding is foreseen to be provided by the participating countries. However, the EU should help bootstrap the development of technologies for languages that are not doing well digitally, using the suggested flexible funds.

5.2.3 Timeline

The ELE Programme is foreseen to have a runtime of nine years, divided into three phases of three years each (Table 4). The CSA and RIA projects are expected to run for three years each while the IA projects have a runtime of two years so that they can focus on the innovation and deployment aspects.

Phase 1: 2024-2026 Phase 1 lays a strong foundation for the overall ELE Programme. All projects start in Phase 1, except for the Innovation Actions.

Phase 2: 2027-2029 Phase 2 drives forward all projects of all types while continuing the Coordination Actions.

Phase 3: 2030-2032 Phase 3 continues the Coordination Actions and finishes off all projects in 2032.

	Type	Num.	Phase 1 2024	2025	2026	Phase 2 2027	2028	2029	Phase 3 2030	2031	2032	Budget Each	Sum
ELE Programme – overall coordination	CSA	3										20M€	60M€
Theme Data and Knowledge – coordination	CSA	3										15M€	45M€
Theme Language Modelling – coordination	CSA	3										15M€	45M€
Theme Language Modelling – research	RIA	15										5M€	75M€
Theme Language Modelling – innovation and deployment	IA	15										5M€	75M€
Theme Machine Translation – research	RIA	12										5M€	60M€
Theme Machine Translation – innovation and deployment	IA	12										5M€	60M€
Theme Text Understanding – research	RIA	12										5M€	60M€
Theme Text Understanding – innovation and deployment	IA	12										5M€	60M€
Theme Speech – research	RIA	12										5M€	60M€
Theme Speech – innovation and deployment	IA	12										5M€	60M€
Theme Infrastructure – support	CSA	3										10M€	30M€
													690M€
Flexible funds for languages with fragmentary, weak or no technological support.													**150M€**
													840M€

Table 4 Project types, timeline and indicative budget breakdown of the ELE Programme (EU)

5.2.4 Collaborations with Related Initiatives

The ELE Programme complements related initiatives, projects and organisations and it will make use of the services, resources and infrastructures provided by these initiatives. We can group these different stakeholders into several broader categories:

Data spaces and data infrastructures:
Various EU/EC Data Spaces including the Common European Language Data Space (LDS), Media Data Space and others; Big Data Value Association (BDVA) and Data, AI and Robotics (DAIRO);[4] Gaia-X;[5] International Data Spaces Association (IDSA);[6] etc.

Research and research data infrastructures:
European Open Science Cloud (EOSC);[7] German National Research Data Infrastructure (NFDI);[8] CLARIN ERIC;[9] Research Data Alliance (RDA);[10] etc.

Various AI initiatives:
ADRA;[11] CLAIRE;[12] LEAM;[13] HumanE-AI;[14] OpenGPT-X;[15] etc.

AI on Demand Platform:
AI-on-Demand Platform;[16] European Language Grid (ELG);[17] etc.

High performance computing:
EuroHPC Joint Undertaking;[18] etc.

Standardisation:
World Wide Web Consortium (W3C);[19] DIN;[20] etc.

[4] https://www.bdva.eu, https://www.bdva.eu/DAIRO
[5] https://gaia-x.eu
[6] https://internationaldataspaces.org
[7] https://eosc.eu
[8] https://www.nfdi.de
[9] https://www.clarin.eu
[10] https://www.rd-alliance.org
[11] https://adr-association.eu
[12] https://claire-ai.org
[13] https://leam.ai
[14] https://www.humane-ai.eu
[15] https://opengpt-x.de
[16] https://www.ai4europe.eu
[17] https://www.european-language-grid.eu
[18] https://eurohpc-ju.europa.eu
[19] https://www.w3.org
[20] https://www.din.de

6 Concluding Remarks

Large-scale studies such as the META-NET White Paper Series (Rehm and Uszkoreit 2012), the STOA study (STOA 2018) and the ELE language reports (see Chapter 4 for an overview and Chapters 5 to 37 for in-depth analyses) have shown that many languages are in danger of digital extinction because they are not sufficiently supported through Language Technologies. Digital Language Equality is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age (Chapter 3). In alignment with what the Language Technology community has promoted for more than a decade, the European Parliament adopted a resolution on *Language equality in the digital age* that suggested initiating a large-scale European LT research, development and innovation programme and to intensify research and funding to achieve Deep Natural Language Understanding and also Digital Language Equality (European Parliament 2018).

Languages are at the heart of every aspect of life. Understanding language is key for building intelligent systems. Over the coming years, AI is expected to transform every industry and society as a whole. There are trends and megatrends that bear closely on digital technologies. Among others, these include accelerating hyperconnectivity, shifts in the nature of work, increasing digitalisation, new modes of learning, expanding consumerism, novel approaches to politics and governance, changes in healthcare, etc. LT and NLP are, by now, considered important driving forces. Language Technology will play a deciding role in how these unfold.

Language tools and resources have increased and improved since the end of the last century, a process further catalysed by the advent of deep learning and neural networks over the past decade. We find ourselves today in the midst of a significant paradigm shift in LT and language-centric AI. This revolution has brought noteworthy advances to the field along with the promise of substantial breakthroughs in the coming years. However, this transformative technology poses problems from a research advancement, environmental, and ethical perspective. Furthermore, it has also laid bare the acute digital inequality that exists between languages. In fact, many sophisticated NLP systems are unintentionally exacerbating this imbalance due to their reliance on vast quantities of data derived mostly from English-language sources. Other languages lag far behind English in terms of digital presence and even the latter would benefit from greater support. Moreover, the striking asymmetry between official and non-official European languages with respect to available digital resources is very worrisome. The unfortunate truth is that European Language Technology is failing to keep pace with the newfound and rapidly evolving changes in the field.

One need look no further than what is happening today across the diverse topography of state-of-the-art LT and language-centric AI for confirmation of the current linguistic unevenness. The paradox at the heart of recent LT advances is evident in almost every LT discipline. Our ability to reproduce ever better synthetic voices has improved sharply for well-resourced languages, but dependence on large volumes of high-quality recordings effectively undermines attempts to do the same for low-resource languages. Multilingual NMT systems return demonstrably improved

results for low- and zero-resource language pairs, but insufficient model capacity continues to haunt transfer learning because large multilingual datasets are required, forcing researchers to rely on English as the best-resourced language. A similar language discrepancy is also found in several of the domain sectors: medical corpora, models and knowledge bases suffer from this disparity, as do users of under-resourced languages in education, where access to language-related tools is limited for most smaller language communities.

However, this time of transition also represents an opportunity to right the ship. Now is the moment to seek balance between European languages in the digital realm. There are ample reasons for optimism. Although there is more work that can and must be done, Europe's leading language resource repositories, platforms, libraries, models and benchmarks have begun to make inroads in this regard.

Over the last decade, the community has developed a clear vision of the work needed in the different areas of LT. The ELE project has devised an outline of necessary actions in the form of concrete recommendations. The ELE Programme, specified in the form of the SRIA and roadmap presented in this chapter, will serve as the blueprint for achieving DLE in Europe. While the political and societal goal is reaching full *Digital Language Equality across all European languages* (and, at the same, preventing digital extinction of many of our languages in Europe), the scientific goal envisioned to be reached by 2030 is *Deep Natural Language Understanding*.

Deep Natural Language Understanding is still an open research problem far from being solved since all current approaches have severe limitations. The development of new LT systems would not be possible without sufficient resources (data, experts, compute facilities, etc.). Creation of carefully designed evaluation benchmarks and annotated data sets for every language and domain of application is needed to foster technological progress, while encouraging deeper understanding of the mechanisms by which they are achieved. All these efforts will then lead to long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication, to create transparent digital language equality in Europe in all aspects of society, from government to businesses to the citizens.

We foresee an ELE Programme of nine years (2024-2032). This period will be divided into three phases of three years each, combining coordination actions (CSAs), research actions (RIAs) as well as actions for innovation and deployment (IAs). The whole community, meaning all relevant scientific and industrial stakeholders from all Member States and Associated Countries, need to be involved. The ELE Programme will tackle the following central themes: Language Modelling, Data and Knowledge, Machine Translation, Text Understanding, and Speech.

As a shared programme between the EU and the participating countries, we suggest an EU budget of 690M€, plus 150M€ of flexible funds to help bootstrap the development of technologies for languages with fragmentary, weak or no technical support. This will be supplemented by national and regional funding.

The ELE Programme is meant to develop into the focal point in which all coordinated developments come together. In this regard, the European Institutions and national as well as regional governments and language institutes must be involved in creating resources, tools and technologies for their own languages. It is exactly

the large scale of the effort that will accelerate the developments and advance the state-of-the-art that will make it possible to join forces that have so far never been joined. This will make it possible to address all European and other relevant languages, all cultures with their particular background and framing of the world, all relevant domains, and all stakeholders by means of a substantial number of use cases. We are convinced that this initiative, built around a coordinated giant pool of shared data sets, open evaluations, open competitions, shared tasks, standardisation efforts, etc. in the literal sense of Open Science, will have a much-needed, substantial and lasting impact in terms of interoperability, development costs, quality and, thus, uptake of the truly game-changing technologies developed in the ELE Programme. Research in Europe must focus on creating the new paradigm of Language Technology, fully harnessing the power of current and emerging AI methods that are based on vast data sets and knowledge bases. With a concerted effort and significant funding, digital language equality will be achieved, for the benefit of *all* Europeans.

References

Adda, Gilles, Annelies Braffort, Ioana Vasilescu, and François Yvon (2022). *Deliverable D1.14 Report on the French Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-french.pdf.

Agerri, Rodrigo, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa (2021). *Deliverable D1.2 Report on the State of the Art in Language Technology and Language-centric AI*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-state-of-the-art.pdf.

Aldabe, Itziar, Aritz Farwell, and German Rigau (2022a). *Deliverable D3.3 Report on the final round of feedback collection*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/feedback-collection.pdf.

Aldabe, Itziar, Georg Rehm, German Rigau, and Andy Way (2022b). *Deliverable D3.1 Report on existing strategic documents and projects in LT/AI (second revision)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-strategic-documents-v3.pdf.

Anastasiou, Dimitra (2022). *Deliverable D1.24 Report on the Luxembourgish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-luxembourgish.pdf.

Backfried, Gerhard, Marcin Skowron, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernaez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz (2022). *Deliverable D2.14 Technology Deep Dive – Speech Technologies*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/speech-deep-dive.pdf.

Bērziņš, Aivars, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari,

Rudolf Rosa, Riccardo Superbo, and Andy Way (2022). *Deliverable D2.13 Technology Deep Dive – Machine Translation*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/MT-deep-dive.pdf.

Blake, Oliver (2022). *Deliverable D2.10 Report from LIBER*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LIBER.pdf.

Borin, Lars, Rickard Domeij, Jens Edlund, and Markus Forsberg (2022). *Deliverable D1.33 Report on the Swedish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166 ELE. https://european-language-equality.eu/reports/language-report-swedish.pdf.

Branco, António, Sara Grilo, and João Silva (2022). *Deliverable D1.28 Report on the Portuguese Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-portuguese.pdf.

Ćušić, Tarik (2022). *Deliverable D1.36 Report on the Bosnian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-bosnian.pdf.

Eide, Kristine, Andre Kåsen, and Ingerid Løyning Dale (2022). *Deliverable D1.26 Report on the Norwegian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-norwegian.pdf.

ELE Consortium (2022). *Deliverable D3.4 Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/SRIA-and-roadmap.pdf.

Eskevich, Maria and Franciska de Jong (2022). *Deliverable D2.3 Report from CLARIN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLARIN.pdf.

European Parliament (2018). *Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI)*. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Gaidienė, Anželika and Aurelija Tamulionienė (2022). *Deliverable D1.23 Report on the Lithuanian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-lithuanian.pdf.

Garabík, Radovan (2022). *Deliverable D1.30 Report on the Slovak Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-slovak.pdf.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way (2022a). "Introducing the Digital Language Equality Metric: Technological Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 1–12. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.1.pdf.

Gaspari, Federico, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way (2022b). *Deliverable D1.3 Digital Language Equality (full specification)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166 ELE. https://european-language-equality.eu/reports/DLE-definition.pdf.

Gaspari, Federico, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou (2021). *Deliverable D1.1 Digital Language Equality (preliminary definition)*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/DLE-preliminary-definition.pdf.

Gavriilidou, Maria, Maria Giagkou, Dora Loizidou, and Stelios Piperidis (2022). *Deliverable D1.17 Report on the Greek Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-greek.pdf.

Giagkou, Maria, Penny Labropoulou, Stelios Piperidis, Miltos Deligiannis, Athanasia Kolovou, and Leon Voukoutis (2022). *Deliverable D1.37 Database and Dashboard*. European Language

Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/DLE-dashboard.pdf.

Gísladóttir, Guðrún (2022). *Deliverable D2.7 Report from ECSPM*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ECSPM.pdf.

Gomez-Perez, Jose Manuel, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn (2022). *Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/text-analytics-deep-dive.pdf.

Grützner-Zahn, Annika and Georg Rehm (2022). "Introducing the Digital Language Equality Metric: Contextual Factors". In: *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*. Ed. by Itziar Aldabe, Begoña Altuna, Aritz Farwell, and German Rigau. Marseille, France, pp. 13–26. http://www.lrec-conf.org/proceedings/lrec2022/workshops/TDLE/pdf/2022.tdle-1.2.pdf.

Hajič, Jan, Maria Giagkou, Stelios Piperidis, Georg Rehm, and Natalia Resende (2021). *Deliverable D2.1 Specification of the consultation process*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-process.pdf.

Hajič, Jan, Tea Vojtěchová, and Maria Giagkou (2022). *Deliverable D2.5 Report from META-NET*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-META-NET.pdf.

Hegele, Stefanie, Rémi Calizzano, Annika Grützner-Zahn, Katrin Marheinecke, and Georg Rehm (2021a). *Deliverable D4.1 Promotional materials and PR package*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/promotional-materials.pdf.

Hegele, Stefanie, Barbara Heinisch, Antonia Popp, Katrin Marheinecke, Annette Rios, Dagmar Gromann, Martin Volk, and Georg Rehm (2022a). *Deliverable D1.16 Report on the German Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-german.pdf.

Hegele, Stefanie, Katrin Marheinecke, Jens-Peter Kückens, and Georg Rehm (2021b). *Deliverable D4.2 Communication and dissemination plan*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/communication-dissemination-plan.pdf.

Hegele, Stefanie, Katrin Marheinecke, and Georg Rehm (2022b). *Deliverable D2.6 Report from ELG*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELG.pdf.

Heuschkel, Maria (2022). *Deliverable D2.12 Report from Wikipedia*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-Wikipedia.pdf.

Hicks, Davyth (2022). *Deliverable D2.9 Report from ELEN*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-ELEN.pdf.

Hlavacova, Jaroslava (2022). *Deliverable D1.8 Report on the Czech Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-czech.pdf.

Hrasnica, Halid (2022). *Deliverable D2.11 Report from NEM*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-NEM.pdf.

Jelencsik-Mátyus, Kinga, Enikő Héja, Zsófia Varga, Tamás Váradi, László János Laki, and Győző Yang Zijian (2022). *Deliverable D1.18 Report on the Hungarian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-hungarian.pdf.

Kaltenböck, Martin, Artem Revenko, Khalid Choukri, Svetla Boytcheva, Christian Lieske, Teresa Lynn, German Rigau, Maria Heuschkel, Aritz Farwell, Gareth Jones, Itziar Aldabe, Ainara Estarrona, Katrin Marheinecke, Stelios Piperidis, Victoria Arranz, Vincent Vandeghinste, and Claudia Borg (2022). *Deliverable D2.16 Technology Deep Dive – Data, Language Resources, Knowledge Graphs*. European Language Equality (ELE); EU project no. LC-01641480 – 1010-18166. https://european-language-equality.eu/reports/data-knowledge-deep-dive.pdf.

Kirchmeier, Sabine (2022). *Deliverable D2.8 Report from EFNIL*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-EFNIL.pdf.

Koeva, Svetla and Valentina Stefanova (2022). *Deliverable D1.5 Report on the Bulgarian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-bulgarian.pdf.

Krauwer, Steven (2003). "The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap". In: *Proceedings of the International Workshop Speech and Computer (SPECOM 2003)*. Moscow, Russia.

Krek, Simon (2022). *Deliverable D1.31 Report on the Slovenian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-slovenian.pdf.

Krstev, Cvetana and Ranka Stanković (2022). *Deliverable D1.35 Report on the Serbian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-serbian.pdf.

Lindén, Krister and Wilhelmina Dyster (2022). *Deliverable D1.13 Report on the Finnish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-finnish.pdf.

Lynn, Teresa (2022). *Deliverable D1.20 Report on the Irish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-irish.pdf.

Magnini, Bernardo, Alberto Lavelli, and Manuela Speranza (2022). *Deliverable D1.21 Report on the Italian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-italian.pdf.

Marheinecke, Katrin, Annika Grützner-Zahn, and Georg Rehm (2022). *Deliverable D4.4 Report on ELE Conference*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/conference.pdf.

Maynard, Diana, Joanna Wright, Mark A. Greenwood, and Kalina Bontcheva (2022). *Deliverable D1.11 Report on the English Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-english.pdf.

Melero, Maite, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas (2022a). *Deliverable D1.6 Report on the Catalan Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-catalan.pdf.

Melero, Maite, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas (2022b). *Deliverable D1.32 Report on the Spanish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-spanish.pdf.

Moshagen, Sjur Nørstebø, Rickard Domeij, Kristine Eide, Peter Juel Henrichsen, and Per Langgård (2022). *Deliverable D1.38 Report on the Nordic Minority Languages*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-nordic-languages.pdf.

Muischnek, Kadri (2022). *Deliverable D1.12 Report on the Estonian Language*. Reports on European Language Equality (ELE) | Coordinator: Prof. Dr. Andy Way, Co-Coordinator: Prof. Dr. Georg Rehm, received funding from the European Union (EU project no. LC-01641480 – 101018166). https://european-language-equality.eu/reports/language-report-estonian.pdf.

Ogrodniczuk, Maciej, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski (2022). *Deliverable D1.27 Report on the Polish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-polish.pdf.

Păiș, Vasile and Dan Tufiș (2022). *Deliverable D1.29 Report on the Romanian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-romanian.pdf.

Pedersen, Bolette Sandford, Sussi Olsen, and Lina Henriksen (2022). *Deliverable D1.9 Report on the Danish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-danish.pdf.

Prys, Delyth, Gareth Watkins, and Stefano Ghazzali (2022). *Deliverable D1.34 Report on the Welsh Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-welsh.pdf.

Rehm, Georg, ed. (2023). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Cham, Switzerland: Springer.

Rehm, Georg, Federico Gaspari, German Rigau, Maria Giagkou, Stelios Piperidis, Annika Grützner-Zahn, Natalia Resende, Jan Hajic, and Andy Way (2022a). "The European Language Equality Project: Enabling digital language equality for all European languages by 2030". In: *The Role of National Language Institutions in the Digital Age – Contributions to the EFNIL Conference 2021 in Cavtat*. Ed. by Željko Jozić and Sabine Kirchmeier. Budapest, Hungary: Nyelvtudományi Kutatóközpont, Hungarian Research Centre for Linguistics, pp. 17–47.

Rehm, Georg, Stefanie Hegele, and Katrin Marheinecke (2022b). *Deliverable D4.3 Report on EP/EC Workshop*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/EC-workshop.pdf.

Rehm, Georg, Stefanie Hegele, and Katrin Marheinecke (2022c). *Deliverable D4.6 ELE Book Publication*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/book-publication.pdf.

Rehm, Georg and Hans Uszkoreit, eds. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg etc.: Springer.

Robinson-Jones, Charlie and Ydwine R. Scarse (2022). *Deliverable D1.39 Report on the West Frisian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-frisian.pdf.

Rögnvaldsson, Eiríkur (2022). *Deliverable D1.19 Report on the Icelandic Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-icelandic.pdf.

Rosner, Mike and Claudia Borg (2022). *Deliverable D1.25 Report on the Maltese Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-maltese.pdf.

Rufener, Andrew and Philippe Wacker (2022). *Deliverable D2.4 Report from LT-innovate*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-LTInnovate.pdf.

Sánchez, José Manuel Ramírez and Carmen García-Mateo (2022). *Deliverable D1.15 Report on the Galician Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-galician.pdf.

Sarasola, Kepa, Itziar Aldabe, Arantza Diaz de Ilarraza, Ainara Estarrona, Aritz Farwell, Inma Hernaez, and Eva Navas (2022). *Deliverable D1.4 Report on the Basque Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-basque.pdf.

Skadiņa, Inguna, Ilze Auziņa, Baiba Valkovska, and Normunds Grūzītis (2022). *Deliverable D1.22 Report on the Latvian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-latvian.pdf.

Steurs, Frieda, Vincent Vandeghinste, and Walter Daelemans (2022). *Deliverable D1.10 Report on the Dutch Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-dutch.pdf.

STOA (2018). *Language equality in the digital age – Towards a Human Language Project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2. https://data.europa.eu/doi/10.2861/136527.

Tadić, Marko (2022). *Deliverable D1.7 Report on the Croatian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-croatian.pdf.

Thönnissen, Marlies (2022). *Deliverable D2.2 Report from CLAIRE*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/consultation-CLAIRE.pdf.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, pp. 3261–3275. https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Way, Andy, Georg Rehm, Jane Dunne, Maria Giagkou, Jose Manuel Gomez-Perez, Jan Hajič, Stefanie Hegele, Martin Kaltenböck, Teresa Lynn, Katrin Marheinecke, Natalia Resende, Inguna Skadina, Marcin Skowron, Tereza Vojtěchová, and Annika Grützner-Zahn (2022a). *Deliverable D2.18 Report on the state of Language Technology in 2030*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/LT-in-2030.pdf.

Way, Andy, Georg Rehm, Jane Dunne, Jan Hajič, Teresa Lynn, Maria Giagkou, Natalia Resende, Tereza Vojtěchová, Stelios Piperidis, Andrejs Vasiljevs, Aivars Berzins, Gerhard Backfried, Marcin Skowron, Jose Manuel Gomez-Perez, Andres Garcia-Silva, Martin Kaltenböck, and Artem Revenko (2022b). *Deliverable D2.17 Report on all external consultations and surveys*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/external-consultations.pdf.