6

LANGUAGE TECHNOLOGY 2020: THE META-NET PRIORITY RESEARCH THEMES

6.1 INTRODUCTION

For decades it has been obvious that one of the last remaining frontiers of IT is still separating our rapidly evolving technological world of mobile devices, computers and the internet from the most precious and powerful asset of mankind, the human mind, the only system capable of thought, knowledge and emotion. Although we use computers to write, telephones to chat and the web to search for knowledge, IT has no direct access to the meaning, purpose and sentiment behind our trillions of written and spoken words. This is why technology is unable to summarise a text, answer a question, respond to a letter and to translate reliably. In many cases it cannot even correctly pronounce a simple English sentence.

Visionaries such as Ray Kurzweil, Marvin Minsky and Bill Gates have long predicted that this border would eventually be overcome by artificial intelligence including language understanding whereas science fiction such as the Star Trek TV series suggested attractive ways in which technology would change our lives, by establishing the fantastic concept of an invisible computer that you have a conversation with and that is able to react to the most difficult commands and also of technology that can reliably translate any human and non-human language.

Many companies had started much too early to invest in language technology research and development and then lost faith after a long period without any tangible progress. During the years of apparent technological standstill, however, research continued to conquer new ground. The results were a deeper theoretical understanding of language, better machine-readable dictionaries, thesauri and grammars, specialised efficient language processing algorithms, hardware with increased computing power and storage capacities, large volumes of digitised text and speech data and new methods of statistical language processing that could exploit language data for learning hidden regularities governing our language use.

We do not yet possess the complete know-how for unleashing the full potential of language technology as essential research results are still missing. Nevertheless, the speed of research keeps increasing and even small improvements can already be exploited for innovative products and services that are commercially viable. We are witnessing a chain of new products for a variety of applications entering the market in rapid succession.

These applications tend to be built on dedicated computational models of language processing that are specialised for a certain task. People, on the other hand, apply the basic knowledge of the language they have acquired during the first few years of their socialisation, throughout their lives to many different tasks of varying complexity such as reading, writing, skimming, summarising, studying, editing, arguing, teaching. They even use this knowledge for the learning of additional languages. After people have obtained proficiency in a second language, they can already translate simple sentences more fluently than many machine translation systems, whereas truly adequate and stylistically acceptable translation is a highly skillful art gained by special training.

G. Rehm and H. Uszkoreit (eds.), *META-NET Strategic Research Agenda for Multilingual Europe 2020*,
White Paper Series, DOI: 10.1007/978-3-642-36349-8_6, © The Author(s) 2013

Today, no text technology software can translate and check for grammatical correctness and no speech technology software could recognise all the sentences it can read aloud if they were spoken by people in their normal voices. But increasingly we observe a reuse of core components and language models for a wide variety of purposes. It started with dictionaries, spell checkers and text-to-speech tools. Google Translate, Apple's Siri and IBM Watson still do not use the same technologies for analysing and producing language, because the generic processing components are simply not powerful enough to meet their respective needs. But many advanced research systems already utilise the same tools for syntactic analysis. This process is going to continue.

In ten years or less, basic language proficiency is going to be an integral component of any advanced IT. It will be available to any user interface, service and application development. Additional language skills for semantic search, knowledge discovery, human-technology communication, text analytics, language checking, e-learning, translation and other applications will employ and extend the basic proficiency. The shared basic language competence will ensure consistency and interoperability among services. Many adaptations and extensions will be derived and improved through sample data and interaction with people by powerful machine learning techniques.

In the envisaged big push toward realising this vision by massive research and innovation, the technology community is faced with three enormous challenges:

1. *Richness and diversity.* A serious challenge is the sheer number of languages, some closely related, others distantly apart. Within a language, technology has to deal with numerous dialects, sociolects, registers, professional jargons, genres and slangs.
2. *Depth and meaning.* Understanding language is a complex process. Human language is not only the key to knowledge and thought, it also cannot be interpreted without certain shared knowledge and active inference. Computational language proficiency needs semantic technologies.
3. *Multimodality and grounding.* Human language is embedded in our daily activities. It is combined with other modes and media of communication. It is affected by beliefs, desires, intentions and emotions and it affects all of these. Successful interactive language technology requires models of embodied and adaptive human interaction with people, technology and other parts of the world.

It is fortunate for research and economy that the only way to effectively tackle the three challenges involves submitting the evolving technology continuously to the growing demands and practical stress tests of real world applications. Google's Translate, Apple's Siri, Autonomy's text analytics and scores of other products demonstrate that there are plenty of commercially viable applications for imperfect technologies. Only a continuous stream of technological innovation can provide the economic pull forces and the evolutionary environments for the realisation of the grand vision.

In the remainder of the Chapter, we propose five major action lines of research and innovation:

- Three priority themes connected with powerful application scenarios that can drive research and innovation. These will demonstrate novel technologies in attractive show-case solutions of high economic and societal impact. They will open up numerous new business opportunities for European language-technology and -service providers.
- A steadily evolving system of shared, collectively maintained interoperable core technologies and resources for the languages of Europe and selected economically relevant languages of its partners. These will ensure that our languages will be sufficiently supported and represented in the next generations of IT.

- A pan-European language technology service platform for supporting research and innovation by testing and showcasing research results, integrating various services even including professional human services will allow SME providers to offer component and end-user services, and share and utilise tools, components and data resources.

The three priority research themes are:

- **Translingual Cloud** – generic and specialised federated cloud services for instantaneous reliable spoken and written translation among all European and major non-European languages.
- **Social Intelligence** – understanding and dialogue within and across communities of citizens, customers, clients and consumers to enable e-participation and more effective processes for preparing, selecting and evaluating collective decisions.
- **Socially Aware Interactive Assistants** – socially aware assistants that learn and adapt and that provide proactive and interactive support tailored to specific situations, locations and goals of the user through verbal and non-verbal multimodal communication.

These priority themes have been designed with the aim of turning our vision into reality and to letting Europe benefit from a technological revolution that will overcome barriers of understanding between people of different languages, between people and technology and between people and the knowledge of mankind. The themes connect societal needs with LT applications and roadmaps for the organisation of research, development and innovation. The priority themes cover the main functions of language: storing, sharing and using of information and knowledge, as well as improving social interaction among humans and enabling social interaction between humans and technology. As multilingualism is at the core of European culture and becoming a global norm, one theme is devoted to overcoming language barriers.

The three themes have been selected in a complex process (see Appendix C, p. 80 ff.) to ensure the needed market pull, the appropriate performance demands, the realistic testing environments and public interest. The themes represent a mix of applications with respect to the various user communities such as small businesses, large enterprises, public administration and the public at large.

6.2 PRIORITY THEME 1: TRANSLINGUAL CLOUD

6.2.1 Solutions for the EU Society

The goal is a multilingual European society, in which all citizens can use any service, access all knowledge, enjoy all media and control any technology *in their mother tongues*. This will be a world in which written and spoken communication is not hindered anymore by language barriers and in which even specialised high-quality translation will be affordable.

The citizen, the professional, the organisation, or the software application in need of cross-lingual communication will use a single, simple access point for channelling text or speech through a gateway that will instantly return the translations into the requested languages in the required quality and desired format.

Behind this access point will be a network of generic and special-purpose services combining automatic translation or interpretation, language checking, post-editing, as well as human creativity and quality assurance, where needed, for achieving the demanded quality. For high-volume base-line quality the service will be free for use but it will offer extensive business opportunities for a wide range of service and technology providers.

Selected components of this ubiquitous service are:

- use and provision platform for providers of computer-supported top-quality human translation, multilingual text authoring and quality assurance by experts

- trusted service centres: certified service providers fulfilling highest standards for privacy, confidentiality and security of source data and translations
- quality upscale models: services permitting instant quality upgrades if the results of the requested service levels do not yet fulfil the quality requirements
- domain, task and genre specialisation models
- translingual spaces: dedicated locations for ambient interpretation. Meeting rooms equipped with acoustic technology for accurate directed sound sensoring and emission

6.2.2 Novel Research Approaches and Targeted Breakthroughs

The main reason why high-quality machine translation (HQMT) has not been systematically addressed yet seems to be the Zipfian distribution of issues in MT: some improvements, the "low-hanging fruit", can be harvested with moderate effort in a limited amount of time. Yet, many more resources and a more fundamental, novel scientific approach – that eventually runs across several projects and also calls – are needed for significant and substantial improvements that cover the phenomena and problems that make up the Zipfian long tail. This is an obstacle in particular for individual research centres and SMEs given their limited resources and planning horizon. Although recent progress in MT has already led to many new applications of this technology, radically different approaches are needed to accomplish the ambitious goal of this research including a true quality breakthrough. Among these new research approaches are:

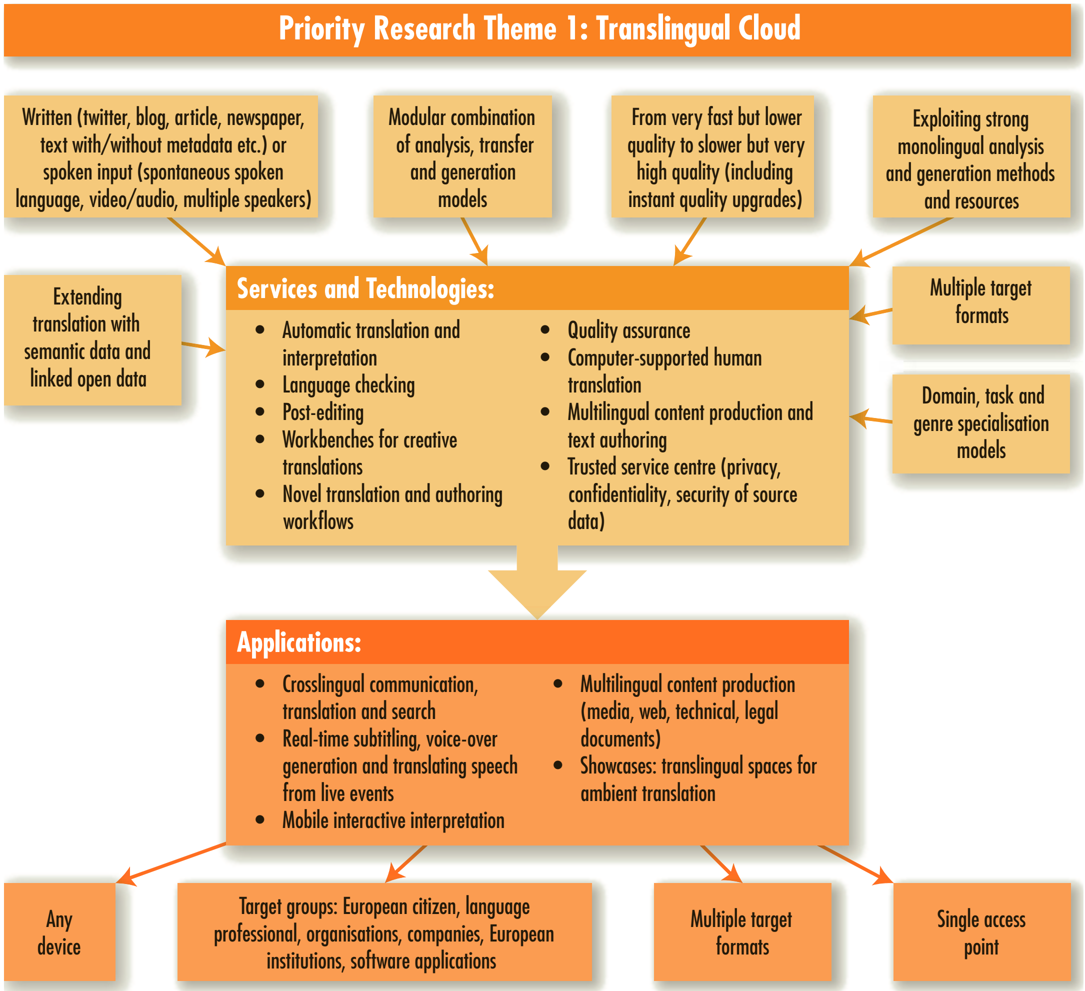
- Systematic concentration on quality barriers, i. e., on obstacles for high quality
- A unified dynamic-depth weighted-multidimensional quality assessment model with task profiling
- Strongly improved automatic quality estimation
- Inclusion of translation professionals and enterprises in the entire research and innovation process
- Improved statistical models that extract more dependencies from the data
- Ergonomic work environments for computer-supported creative top-quality human translation and multilingual text authoring
- Semantic translation paradigm by extending statistical translation with semantic data such as linked open data, ontologies including semantic models of processes and textual inference models
- Exploitation of strong monolingual analysis and generation methods and resources
- Modular combinations of specialised analysis, generation and transfer models, permitting accommodation of registers and styles (including user-generated content) and also enabling translation within a language (e. g., between specialists and laypersons).

The expected breakthroughs will include:

- High-quality text translation and reliable speech translation (including a modular analysis-transfer-generation translation technology that facilitates reuse and constant improvement of modules)
- Seemingly creative translation skills by analogy-driven transfer models
- Automatic subtitling and voice over of films
- Ambient translation

6.2.3 Solution and Realisation

The technical solutions will benefit from new trends in IT such as software as a service, cloud computing, linked open data and semantic web, social networks, crowdsourcing etc. For MT, a combination of translation brokering on a large scale and translation on demand is promising. The idea is to streamline the translation process so that it becomes simpler to use and more transpar-


Priority Research Theme 1: Translingual Cloud
Written (twitter, blog, article, newspaper, text with/without metadata etc.) or spoken input (spontaneous spoken language, video/audio, multiple speakers)
Modular combination of analysis, transfer and generation models
From very fast but lower quality to slower but very high quality (including instant quality upgrades)
Exploiting strong monolingual analysis and generation methods and resources
Extending translation with semantic data and linked open data
Services and Technologies:
Automatic translation and interpretation
Language checking
Post-editing
Workbenches for creative translations
Novel translation and authoring workflows
Quality assurance
Computer-supported human translation
Multilingual content production and text authoring
Trusted service centre (privacy, confidentiality, security of source data)
Multiple target formats
Domain, task and genre specialisation models
Applications:
Crosslingual communication, translation and search
Real-time subtitling, voice-over generation and translating speech from live events
Mobile interactive interpretation
Multilingual content production (media, web, technical, legal documents)
Showcases: translingual spaces for ambient translation
Any device
Target groups: European citizen, language professional, organisations, companies, European institutions, software applications
Multiple target formats
Single access point

9: Priority Research Theme 1: Translingual Cloud

ent for the end user, and at the same time respects important factors such as subject domain, language, style, genre, corporate requirements and user preferences. Technically, what is required is maximum interoperability of all components (corpora, processing tools, terminology, knowledge, maybe even translation models) and a cloud or server/service farm of specialised language technology services for different needs (text and media types, domains, etc.) offered by SMEs, large companies or research centres.

A platform has to be designed and implemented for the resource and evaluation demands of large-scale collaborative MT research. An initial inventory of language tools and resources as well as extensive experience in shared tasks and evaluation has been obtained in several EU-funded projects. Together with LSPs, a common service layer supporting research workflows on HQMT must be established. As third-party (customer) data is needed for realistic development and evaluation, intellectual property rights and legal issues must be taken into account from the onset. The infrastructures to be built include service clouds with trusted service centres, interfaces for services (APIs), workbenches for creative translations, novel translation workflows (and improved links to content production and authoring) and showcases such as ambient and embedded translation.

Research Priority	Phase 1: 2013-2014	Phase 2: 2015-2017	Phase 3: 2018-2020
Immediate affordable translation in any needed quality level (from sufficient to high)	Development of necessary monolingual language tools (analysis, generation) driven by MT needs; exploitation of novel ML techniques for MT purposes, using large LR and semantic resources, including Linked Open Data and other naturally occuring semantic and knowledge resources (re-purposing for MT and NLP use); experiment with novel metrics, automated, human-centered, or hybrid; use EU languages, identify remaining gaps (LR resources, tools)	Concentrate on HQMT systems using results of Phase 1; deepen development of MT-related monolingual tools; employ novel techniques aimed at HQMT, combination of systems, domain adaptation, cross-language adaptation; develop showcases for novel translation workflow; use novel metrics identified as correlated with the aims of HQMT; continue development on EU languages, identify needs for non-EU languages (MT-related) and their gaps	Deployment of MT systems in particular applications requiring HQMT, such as technology export, government and public information systems, private services, medical applications etc., using novel translation workflows where appropriate; application- and user-based evaluation driven engagement of core and supplemental technologies; coverage of EU languages and other languages important for EU business and policy
Delivering multimedia content in any language (captioning, subtitling, dubbing)	Multi-media system prototypes, combining language, speech, image and video analysis; employing novel techniques (machine learning, cross-fertilisation of features across media types); targeted evaluation metrics for system quality assessment related to MT; aimed at EU languages with sufficient resources; data collection effort to support multi-media analysis	Prototype applications in selected domains, such as public service (parliament recordings, sports events, legal proceedings) and other applications (tv archives or movie delivery, online services at content providers); continued effort at multimedia analysis, adding languages as resources become available	Deployment of large-scale applications for multi-media content delivery, public and/or private, in selected domains; development of online services for captioning, subtitling, dubbing, including on-demand translation); new languages for outside-of-the-EU delivery, continued improvement of EU languages
Cross-lingual knowledge management and linked open data	Publication of multilingual language resources as linked open data as well as linking of resources across languages; develop ontology translation components that can localise ontologies and linked datasets to different languages	Develop an ecosystem of NLP tools and services that leverage the existing multilingual resources on linked open data; develop new generation of MT technology that can profit from semantic data and linked open data	Develop methods that allow querying linked open data in different languages
Avantgarde functionalities	Consecutive interpretation and translation	Synchronous interpretation and translation	Translingual collaborative spaces

10: Priority Theme 1 – Translingual Cloud: Preliminary Roadmap

6.2.4 Impact

HQMT in the cloud will ensure and extend the value of the digital information space in which everyone can contribute in her own language and be understood by members of other language communities. It will assure that diversity will no longer be a challenge, but a welcome enrichment for Europe both socially and economically. Based on the new technology, language-transparent web and language-transparent media will help realise a truly multilingual mode of online and media interaction for every citizen regardless of age, education, profession, cultural background, language proficiency or technical skills. Showcase applications areas are:

- Multilingual content production (media, web, technical, legal documents)
- Cross-lingual communication, document translation and search
- Real-time subtitling and translating speech from live events
- Mobile interactive interpretation for business, social services, and security
- Translation workspaces for online services

6.2.5 Organisation of Research

Several very large cooperating and competing lead projects will share an infrastructure for evaluation, resources (data and base technologies), and communication. Mechanisms for reducing or terminating partner involvements and for adding new partners or subcontracted contributors should provide the needed flexibility. A number of smaller projects, including national and regional projects, will provide building blocks for particular languages, tasks, component technologies or resources. A special scheme will be designed for involving EC-funding, member states, industrial associations, and language communities.

Two major phases from 2015 to mid 2017 and from mid 2017 to 2020 are foreseen. Certain services such as multilingual access to web-information across European languages should be transferred to implementation and testing at end of phase 2017. Internet-based real-time speech translation for a smaller set of languages will also get into service at this time as well as HQMT for selected domains and tasks. A major mid-term revision with a thorough analytical evaluation will provide a possible breakpoint for replanning or termination.

A close cooperation of language technology and professional language services is planned. In order to overcome the quality boundaries we need to identify and understand the quality barriers. Professional translators and post-editors are required whose judgements and corrections will provide insights for the analytical approach and data for the bootstrapping methodology. The cooperation scheme of research, commercial services and commercial translation technology is planned as a symbiosis since language service professionals or advanced students in translation studies or related programmes working with and for the developing technology will at the same time be the first test users analytically monitored by the evaluation schemes. This symbiosis will lead to a better interplay of research and innovation.

Although the research strand will focus on advances in translation technology for innovation in the language and translation service sector, a number of other science, technology and service areas need to be integrated into the research from day one. Some technology areas such as speech technologies, language checking, authoring systems, analytics, generation and content management systems need to be represented by providers of state-of-the-art commercial products.

Supporting research and innovation in LT should be accompanied by policy making in the area of multilingualism, but also in digital accessibility. Overcoming language barriers can greatly influence the future of the EU.

Solutions for better communication and for access to content in the users' native languages would reaffirm the role of the EC to serve the needs of the EU citizens. A connection to the infrastructure programme CEF could help to speed up the transfer of research results to badly needed services for the European economy and public.

At the same time, use cases should cover areas in which the European social and societal needs massively overlap with business opportunities to achieve funding investment that pays back, ideally public-private partnerships.

Concerted activities sharing resources such as error corpora or test suites and challenges or shared tasks in carefully selected areas should be offered to accelerate innovation breakthrough and market-readiness for urgently needed technologies.

6.3 PRIORITY THEME 2: SOCIAL INTELLIGENCE AND E-PARTICIPATION

6.3.1 Solutions for the EU Society

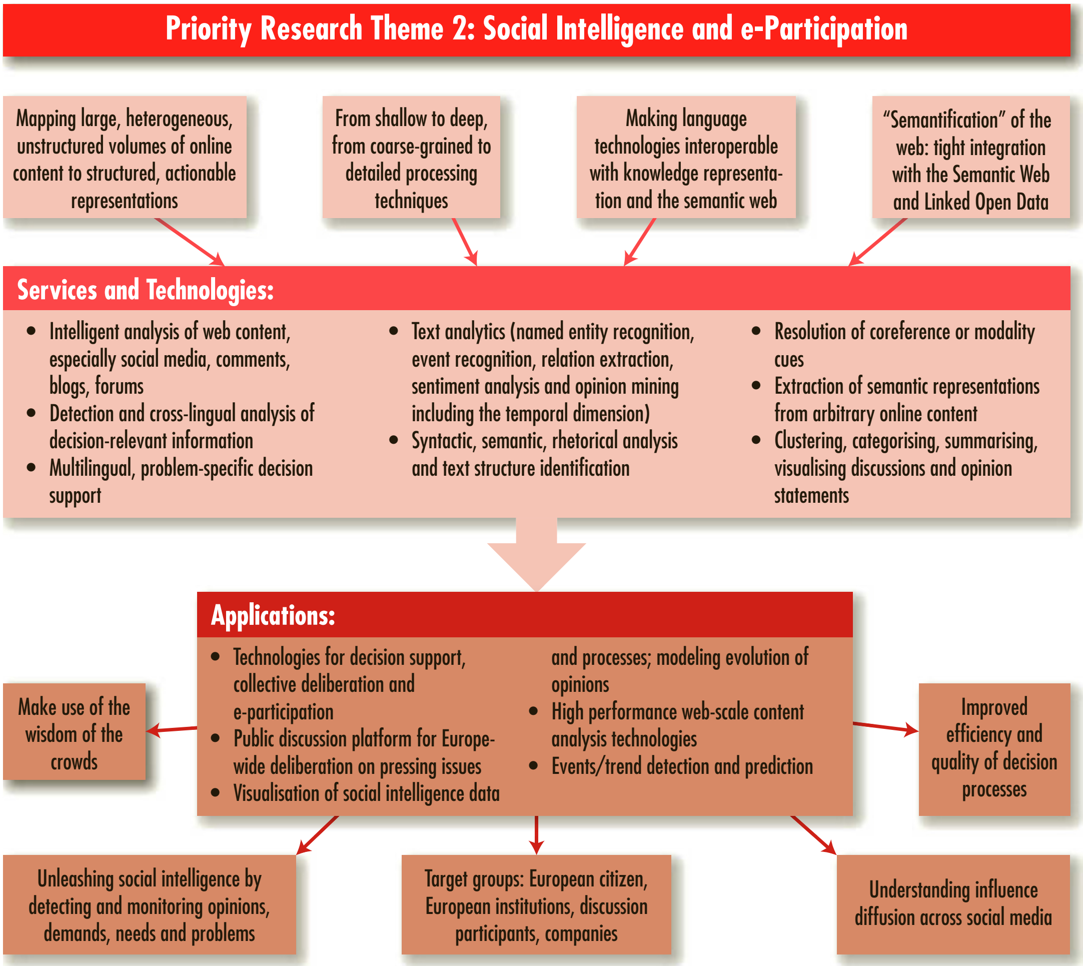
The central goal behind this theme is to use information technology and the digital content of the web for improving effectiveness and efficiency of decision-making in business and society.

The quality, speed and acceptance of individual and collective decisions is the single main factor for the success of social systems such as enterprises, public services, communities, states and supranational organisations. The growing quantity and complexity of accessible relevant information poses a serious challenge to the efficiency and quality of decision processes. IT provides a wide range of instruments for intelligence applications. Business intelligence, military intelligence or security intelligence applications collect and pre-process decision-relevant information. Analytics programmes search the data for such information and decision support systems evaluate and sort the information and apply problem-specific decision rules. Although much of the most relevant information is contained in texts, text analytics programmes today only account for less than 1% of the more than 10 billion US$ business intelligence and analytics market. Because of their limited capabilities in interpreting texts, mainly business news, reports and press releases, their findings are still neither comprehensive nor reliable enough.

Social intelligence builds on improved text analytics methodologies but goes far beyond the analysis. One central goal is the analysis of large volumes of social media, comments, communications, blogs, forum postings etc. of citizens, customers, patients, employees, consumers and other stakeholder communities. Part of the analysis is directed to the status, opinions and acceptance associated with the individual information units. As the formation of collective opinions and attitudes is highly dynamic, new developments need to be detected and trends analysed. Emotions play an important part in individual actions such as voting, buying, supporting, donating and in collective opinion formation, the analysis of sentiment is a crucial component of social intelligence.

Social intelligence can also support collective deliberation processes. Today any collective discussion processes involving large numbers of participants are bound to become intransparent and incomprehensible rather fast. By recording, grouping, aggregating and counting opinion statements, pros and cons, supporting evidence, sentiments and new questions and issues, the discussion can be summarised and focussed. Decision processes can be structured, monitored, documented and visualised, so that joining, following and benefitting from them becomes much easier. The efficiency and impact of such processes can thus be greatly enhanced.

Since many collective discussions will involve participants in several countries, e. g., EU member states or enterprise locations, cross-lingual participation needs to be supported [32]. Special support will also be provided for


Priority Research Theme 2: Social Intelligence and e-Participation
Mapping large, heterogeneous, unstructured volumes of online content to structured, actionable representations
From shallow to deep, from coarse-grained to detailed processing techniques
Making language technologies interoperable with knowledge representation and the semantic web
"Semantification" of the web: tight integration with the Semantic Web and Linked Open Data
Services and Technologies:
Intelligent analysis of web content, especially social media, comments, blogs, forums
Detection and cross-lingual analysis of decision-relevant information
Multilingual, problem-specific decision support
Text analytics (named entity recognition, event recognition, relation extraction, sentiment analysis and opinion mining including the temporal dimension)
Syntactic, semantic, rhetorical analysis and text structure identification
Resolution of coreference or modality cues
Extraction of semantic representations from arbitrary online content
Clustering, categorising, summarising, visualising discussions and opinion statements
Applications:
Technologies for decision support, collective deliberation and e-participation
Public discussion platform for Europe-wide deliberation on pressing issues
Visualisation of social intelligence data and processes; modeling evolution of opinions
High performance web-scale content analysis technologies
Events/trend detection and prediction
Make use of the wisdom of the crowds
Improved efficiency and quality of decision processes
Unleashing social intelligence by detecting and monitoring opinions, demands, needs and problems
Target groups: European citizen, European institutions, discussion participants, companies
Understanding influence diffusion across social media

11: Priority Research Theme 2: Social Intelligence and e-Participation

participants not mastering certain group-specific or expert jargons and for participants with disabilities affecting their comprehension.

6.3.2 Novel Research Approaches and Targeted Breakthroughs

A key enabler will be language technologies that can map large, heterogeneous, and, to a large extent, unstructured volumes of online content to actionable representations that support decision making and analytics tasks. Such mappings can range from the relatively shallow to the relatively deep, encompassing for example coarse-grained topic classification at the document or paragraph level or the identification of named entities, as well as in-depth syntactic, semantic and rhetorical analysis at the level of individual sentences and beyond (paragraph, chapter, text, discourse) or the resolution of co-reference or modality cues within and across sentences.

Technologies such as, e. g., information extraction, data mining, automatic linking and summarisation have to be made interoperable with knowledge representation and semantic web methods such as ontological engineering. Drawing expertise from related areas such as knowledge management, information sciences, or social sciences is a prerequisite to meet the challenge of modelling social intelligence [49]. The new research approach should target the bottleneck of knowledge engineering by:

- Semantification of the web: bridging between the semantic parts and islands of the web and the traditional web containing unstructured data;
- Merging and integrating textual data with social network and social media data, especially along the dimension of time;
- Aligning and making comparable different genres of content like mainstream-news, social media (blogs, twitter, facebook etc.), academic texts, archives etc.;
- Extracting semantic representations from social media content, i. e., creating representations for reasoning and inferencing;
- Taking metadata and multimedia data into account.

The following list contains specific targeted breakthroughs to be sought in this scenario:

- Social intelligence by detecting and monitoring opinions, demands, needs and problems;
- Detecting diversity of views, biases along different dimensions (e. g., demographic) etc. including temporal dimension (i. e., modelling evolution of opinions);
- Support for both decision makers and participants;
- Problem mining and problem solving;
- Support of collective deliberation and collective knowledge accumulation;
- Vastly improved approaches to sentiment detection and sentiment scoring (going beyond the approach that relies on a list of positive and negative keywords);
- Introducing genre-driven text and language-processing (different genres need to be processed differently);
- Personalised recommendations of e-participation topics to citizens;
- Proactive involvement in e-participation activities;
- Understanding influence diffusion across social media (identifying drivers of opinion spreading);
- More sophisticated methods for topic and event detection that are tightly integrated with the Semantic Web and Linked Open Data.
- Modelling content and opinion flows across social networks;
- Evaluation of methods by analytic/quantitative and sociological/qualitative means.

6.3.3 Solution and Realisation

Individual solutions should be assembled from a repository of generic monolingual and cross-lingual language technologies, packaging state-of-the-art techniques in robust, scalable, interoperable, and adaptable components that are deployed across sub-tasks and sub-projects, as well as across languages where applicable (e. g., when the implementation of a standard data-driven technique can be trained for individual languages). These methods need to be combined with powerful analytical approaches that can aggregate all relevant data to support analytic decision making and develop new access metaphors and task-specific visualisations.

By robust we mean technologically mature, engineered and scalable solutions that can perform high-throughput analysis of web data at different levels of depth and granularity in line with the requirements of their applications. Technology should be able to work with heterogeneous sources, ranging from unstructured (arbitrary text documents of any genre) to structured (ontologies, linked open data, databases).

To accomplish interoperability we suggest a strong semantic bias in the choice and design of interface representations: to the highest degree possible, the output (and at deeper levels of analysis also input) specifications of component technologies should be interpretable semantically, both in relation to natural language semantics (be it lexical, propositional, or referential) and extra-linguistic semantics (e. g., taxonomic world or domain knowledge). For example, grammatical analysis (which one may or may not decompose further into tagging, syntactic parsing, and semantic role labelling) should make available a sufficiently abstract, normalised, and detailed output, so that downstream processing can be accomplished without further recourse to knowledge about syntax. Likewise, event extraction or fine-grained, utterance-level opinion mining should operate in terms of formally interpretable representations that support notions of entailment and, ultimately, inference.

Finally, our adaptability requirement on component technologies addresses the inherent heterogeneity of information sources and communication channels to be processed. Even in terms of monolingual analysis only, linguistic variation across genres (ranging from carefully edited, formal publications to spontaneous and informal social media channels) and domains (as in subject matters) often calls for technology adaptation, where even relatively mature basic technologies (e. g., part-of-speech taggers) may need to be customised or re-trained to deliver satisfactory performance. Further taking into account variation across downstream tasks, web-scale language processing typically calls for different parameterisations and trade-offs (e. g., in terms of computational cost vs. breadth and depth of analysis) than an interactive self-help dialogue scenario. For these reasons, relevant trade-offs need to be documented empirically, and component technologies accompanied with methods and tools for adaptation and cost-efficient re-training, preferably in semi- and un-supervised settings.

The technical solutions needed include:

- Technologies for decision support, collective deliberation and e-participation.
- A large public discussion platform for Europe-wide deliberation on pressing issues such as energy policies, financial system, migration, natural disasters, etc.
- Visualisation of social intelligence-related data and processes for decision support (for politicians, health providers, manufacturers, or citizens).
- High-throughput, web-scale content analysis techniques that can process multiple different sources, ranging from unstructured to completely structured, at different levels of granularity and depth by allowing to trade-off depth for efficiency as required.
- Mining e-participation content for recommendations, summarisation and proactive engagement of less active parts of population.
- Detection and prediction of events and trends from content and social media networks.
- Extraction of knowledge and semantic integration of social content with sensory data and mobile devices (in near-real-time).
- Cross-lingual technology to increase the social reach and approach cross-culture understanding.

We suggest to structure the research along at least the six lines shown in Figure 12.

6.3.4 Impact

The 21st century presents us with multiple challenges including efficient energy consumption, global warming and financial crises. It is obvious that no single individual can provide answers to challenging problems such as these, nor will top-down imposed measures find social acceptance as solutions. Language technology will enable a paradigm shift in transnational public deliberation. The European Ombudsman recently realised [32] that there are problems and gaps in the way public debates and consultation are usually held in Europe – language technology can improve the situation altogether and bring about a paradigm shift in that regard.

The applications and technologies discussed in this section will change how business adapts and communicates with their customers. It will increase transparency in decision-making processes, e. g., in politics and at the same time give more power to the citizen. As a by-product, the citizens are encouraged to become better in-

Research Priority	Phase 1: 2013-2014	Phase 2: 2015-2017	Phase 3: 2018-2020
Social influence and incentives	Modelling social diversity of views across languages and cultures	Modelling social influence and incentives through game theoretic approaches using data from texts and social networks	Holistic modelling of society (or its segments) through observing a variety of data sources
Information tracking	Tracking dynamics of information diffusion across languages, cultures and media	Transforming textual and social network streams into actionable deep knowledge representations	Prediction of future events and identification of causal relationships from textual and social streams
Multimodal data processing	Joining textual data and social networks, including spatial and temporal dimensions	Joining textual and social data with unstructured sources like sensor data (smart cities), video, images, audio	Detecting inconsistencies, gaps and completeness of collected knowledge from textual and social sources
Visualisation and user interaction	Visualisation of textual and social dynamics	Adaptive human-computer interfaces boosting specific aims in interaction	Adaptive interaction systems for communication with the whole or parts of society
High-throughput analysis	Scalable processing of multimodal data (Big Data)	Real-time modelling and reasoning on massive textual and social streams	Algorithms and toolkits being able to deal with global scale analytics and reasoning with multimodal data
Knowledge-driven text analysis	Develop named-entity taggers that scale to entities described in linked open data resources; develop methods that exploit linked open data for improved disambiguation.	Develop a new generation of information extraction tools that are able to reliably extract from texts all semantic relations defined in, e. g., DBPedia	NLP systems are able to deal with linked open data and Semantic Web ontologies to analyse text at the meaning level and draw appropriate inferences

12: Priority Theme 2 – Social Intelligence and e-Participation: Preliminary Roadmap

formed in order to make use of their right to participate in a reasonable way. Powerful analytical methods will help European companies to optimise marketing strategies or foresee certain developments by extrapolating on the basis of current trends. Leveraging social intelligence for informed decision making is recognised as crucial in a wide range of contexts and scenarios:

- Organisations will better understand the needs, opinions, experiences, communication patterns, etc. of their actual and potential customers so that they can react quickly to new trends and optimise their marketing and customer communication strategies.
- Companies will get the desparately needed instruments to exploit the knowledge and expertise of their huge and diverse workforces, the wisdom of their own crowds, which are the most highly motivated and most closely affected crowds.
- Political decision makers will be able to analyse public deliberation and opinion formation processes in order to react swiftly to ongoing debates or important, sometimes unforeseen events.
- Citizens and customers get the opportunity (and necessary information) to participate and influence political, economic and strategic decisions of governments and companies, ultimately leading to more transparency of decision processes.

Thus, leveraging collective and social intelligence in developing new solutions to these 21st century challenges seems a promising approach in such domains where the complexity of the issues under discussion is beyond the purview of single individuals or groups.

The research and innovation will provide technological support for emerging new forms of issue-based, knowledge-enhanced and solution-centred participatory democracy involving large numbers of expert- and non-expert stakeholders distributed over large areas, using multiple languages. At the same time the resulting technologies will be applicable to smaller groups and also interpersonal communication as well, even though different dynamics of information exchange can be foreseen.

The research to be carried out and technologies to be developed in this priority theme will also have a big influence on the Big Data challenge and how we will make sense of huge amounts of data in the years to come. What we learn from processing language is the prime tool for processing the huge and intractable data streams that we will be confronted with in the near future.

6.3.5 Organisation of Research

Research in this area touches upon political as well as business interests and at the same time is scalable in reach from the regional to the European scale. Therefore, it is necessary to identify business opportunities and potential impact for society at different levels and to align EU level research with efforts on the national level. Furthermore, this priority theme calls for large-scale, incremental, and sustained development and innovation across multiple disciplines (especially language technology and semantic technologies) and, within each community, a certain degree of stacking and fusion of approaches. Therefore, research organisation needs to create strong incentives for early and frequent exchange of technologies among all players involved. A marketplace for generic component technologies and a service-oriented infrastructure for adaptation and composition must be created, to balance performance-based steering and self-organisation among clusters of contributing players. In this ecosystem of technology providers and integrators, component uptake and measurable contributions against the targeted breakthrough of the priority theme at large should serve as central measures of success.

6.4 PRIORITY THEME 3: SOCIALLY AWARE INTERACTIVE ASSISTANTS

6.4.1 Solutions for the EU Society

Socially aware interactive assistants are conversational agents. Their socially-aware behaviour is a result of combining analysis methods for speech, non-verbal and semantic signals.

Now is the time to develop and make operational socially aware, multilingual assistants that support people interacting with their environment, including human-computer, human-artificial agent (or robot), and computer-mediated human-human interaction. The assistants must be able to act in various environments, both indoor (such as meeting rooms, offices, appartments), outdoor (streets, cities, transportation, roads) and virtual environments (such as the web, virtual worlds, games), and also be able to communicate, exchange information and understand other agents' intentions. They must be able to adapt to the user's needs and environment and have the capacity to learn incrementally from all interactions and other sources of information.

The ideal socially aware multilingual assistant can interact naturally with humans, in any language and modality. It can adapt and be personalised to individual communication abilities, including special needs (for the visual, hearing, or motor impaired), affections, or language proficiencies. It can recognise and generate speech incrementally and fluently. It is able to assess its performance and recover from errors. It can learn, personalise itself and forget. It can assist in language training and education, and provide synthetic multimedia information analytics. It recognises people's identity, and their gender, language or accent. If the agent is embodied in a robot, it can move, manipulate objects, and interact with people.

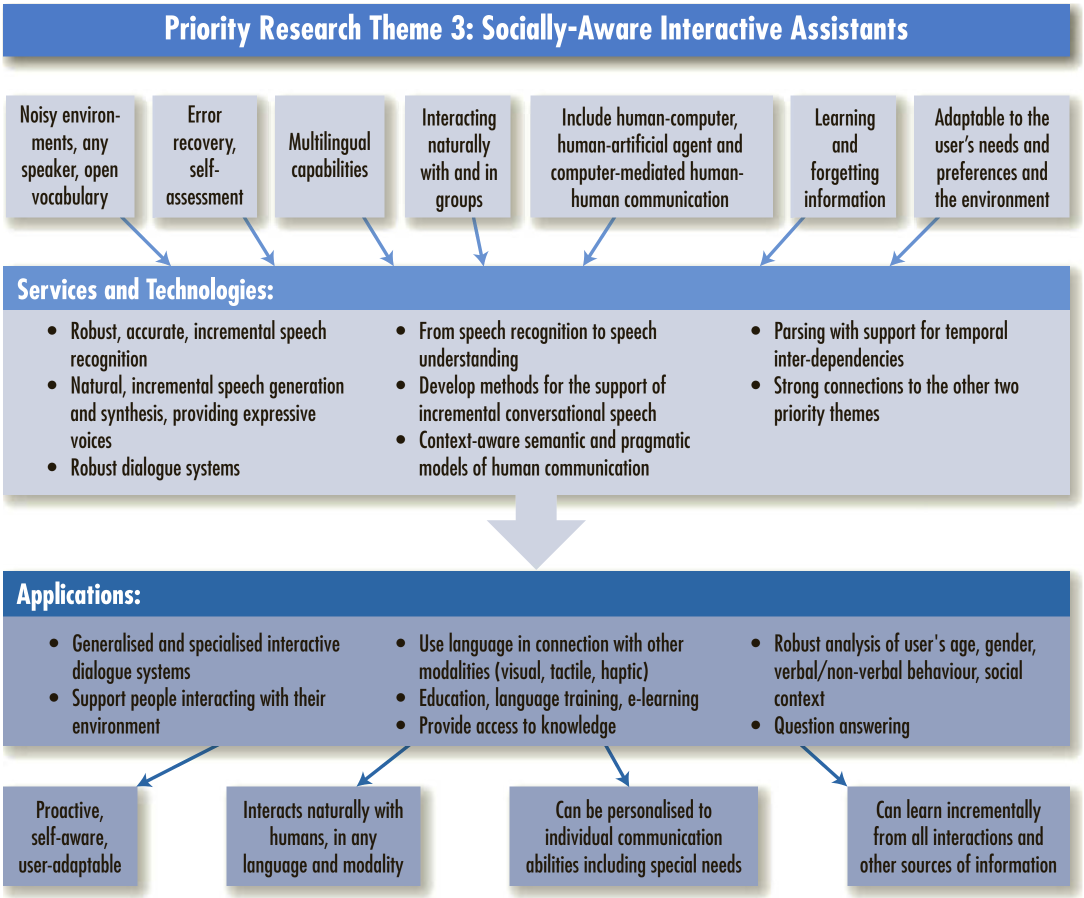
This priority theme includes several core components:

- Interacting naturally with humans (in communication, education, games, etc.) in an implicit (proactive) or explicit (spoken dialogue and/or gesticulation) manner based on robust analysis of human user identity, age, gender, verbal and nonverbal behaviour, and social context;
- Using language in connection with other communication modalities (visual, tactile, haptic);
- Conscious of its capabilities and self-learning;
- Exhibiting robust performance everywhere (indoor/outdoor, augmented reality);
- Overcoming handicap obstacles by means of suitable technologies (sign language understanding, assistive applications, etc.);
- Interacting naturally with and in groups (in social networks, with humans or artificial agents/robots);
- Exhibiting multilingual proficiency (speech-to-speech translation, interpretation in meetings and videoconferencing, cross-lingual information access);
- Referring to written support (transcription, close-captioning, reading machines, ebooks);
- Providing access to knowledge (answers to questions, shared knowledge in discussion);
- Providing personalised training (computer-assisted language learning, e-learning).

Initial steps in the right direction have already been taken – again, by US companies. Apple's intelligent assistant Siri is available on the iPhone, Google's interactive speech technologies can be used on Android and iOS devices. Recently, Microsoft announced – in a letter sent to shareholders by Microsoft CEO Steve Ballmer – that it wants to focus on the development of "new form factors that have increasingly natural ways to use them including touch, gestures and speech". Analysing this announcement, user interface expert Bill Meisel "never expected to see mentions of natural user interfaces and machine learning in a short message to shareholders by the CEO of one of the largest companies in the U.S. Their mention as focus areas suggests that areas once viewed as leading-edge technology have achieved mainstream importance, to the degree that their successful deployment can impact the future of a major company." [43]. Meisel concludes that all three companies (Apple, Google, Microsoft) are currently "developing integrated ecosystems that can tightly couple our human intelligence with computer intelligence across a range of products. And they have the budgets to make it happen." Again, Europe has to ask itself the question if we want to leave this huge field to three US companies or if the combined expertise of our continent's language technology experts is better suited to build interactive, socially aware assistants for the speakers and users of our many different languages and cultures.

6.4.2 Novel Research Approaches and Targeted Breakthroughs

In addition to significantly improving core speech and language technologies, the development of socially aware interactive assistants requires several research breakthroughs. With regard to speech recognition, accuracy (open vocabulary, any speaker) and robustness (noise, cross-talking, distant microphones) have to be improved. Methods for self-assessment, self-adaptation, personalisation, error-recovery, learning and forgetting information, and also for moving from recognition to understanding have to be developed. Concerning speech synthesis, voices have to be made more natural and expres-



Priority Research Theme 3: Socially-Aware Interactive Assistants

Noisy environments, any speaker, open vocabulary

Error recovery, self-assessment

Multilingual capabilities

Interacting naturally with and in groups

Include human-computer, human-artificial agent and computer-mediated human-human communication

Learning and forgetting information

Adaptable to the user's needs and preferences and the environment

Services and Technologies:

- Robust, accurate, incremental speech recognition
- Natural, incremental speech generation and synthesis, providing expressive voices
- Robust dialogue systems
- From speech recognition to speech understanding
- Develop methods for the support of incremental conversational speech
- Context-aware semantic and pragmatic models of human communication
- Parsing with support for temporal inter-dependencies
- Strong connections to the other two priority themes

Applications:

- Generalised and specialised interactive dialogue systems
- Support people interacting with their environment
- Use language in connection with other modalities (visual, tactile, haptic)
- Education, language training, e-learning
- Provide access to knowledge
- Robust analysis of user's age, gender, verbal/non-verbal behaviour, social context
- Question answering

Proactive, self-aware, user-adaptable

Interacts naturally with humans, in any language and modality

Can be personalised to individual communication abilities including special needs

Can learn incrementally from all interactions and other sources of information

13: Priority Research Theme 3: Socially Aware Interactive Assistants

sive, control parameters have to be included for linguistic meaning, speaking style, emotion etc. They also have to be equipped with methods for incremental conversational speech, including filled pauses and hesitations. Likewise, speech recognition, synthesis and understanding have to be integrated, including different levels of evaluation and different levels of automated annotation.

Human communication is multimodal (including speech, facial expressions, body gestures, postures, etc.), crossmodal and fleximodal: it is based on pragmatically best suited modalities. Semantic and pragmatic models of human communication have to be developed. These have to be context-aware and model situational inter-depedencies between context and modalities for arriving at robust communication analysis (multimodal content analytics, infering knowledge from multiple sensory modalities). They have to be able to detect and recover interactively from mistakes, learning continuously and incrementally. Parsing has to model temporal inter-dependencies within and between modalities in order to maximise the assistant's human-communication-prediction ability. In order to be able to design technologies, adequate semantically and pragmatically annotated language and multimodal resources have to be produced.

A common push has to be made towards more natural dialogue. This includes, among others, the recognition and production of paralinguistics (prosody, visual cues, emotion) and a better understanding of socio-emotional functions of communicative behaviour, including group dynamics, reputation and relationship management. More natural dialogue needs more advanced dialogue models that are proactive (not only reactive), that are able to detect that recognised speech is intended as a machine command, they have to be able to interpret silence as well as direct and indirect speech acts (including lies and humour). Another prerequisite for more natural dialogue is the ability of the assistant to personalise itself to the user's preferences. The digital assistant has to operate in a transparent way and be able to participate in multi-party conversations and make use of other sensory data (GPS, RFID, cameras etc.).

There is also a strong connection to the first priority theme: the multilingual assistant should be able to do speech-to-speech translation in human-human-interaction (e. g., in meetings) and to deal with different languages, accents and dialects effectively. Systems developed should also cover at least all official languages of the EU and several regional languages.

6.4.3 Solution and Realisation

The technological and scientific state-of-the-art is at a stage that finally allows tackling the development of socially aware multilingual assistants. Progress in machine learning, including adaptation, unsupervised learning from data streams, continuous learning, and transfer learning makes it possible automatically to learn certain capabilities from data. In addition, existing language and multimodal resources enable the bootstrapping of systems. Furthermore, there is interdisciplinary progress made in, e. g., social signal processing and also knowledge representation including approaches such as the Semantic Web and Linked Open Data – especially inferences and automatic reasoning on such data sets are an important prerequisite for the technologies devised here.

Technological advances are continuously being achieved in the vision-based human behaviour analysis and synthesis fields. Ubiquitous technologies are now widely available. User-centric approaches have been largely studied and crowd-sourcing is used more and more. Quantitative and objective language technology and human-behaviour understanding technology evaluations, allowing for assessing a technological readiness level (TRL), are carried out more widely, as best practice, and language resources and publicly-available annotated recordings of human spontaneous behaviour are now available.

However, there are prohibitive factors. Technology evaluation is still limited and not conducted for all languages. There is limited availability of language resources; the necessary resources do not exist yet for all languages. Publicly-available recordings of spontaneous (rather than staged) human behaviour are sparse, especially when it comes to continuous synchronised observations of multi-party interactions. Limited progress of the technology for automatic understanding of social behaviour like rapport, empathy, envy, conflict, etc., is mainly attributed to this lack of suitable resources. In addition, we still have limited knowledge of human language and human behaviour perception processes. Automated systems often face theoretical and technological complexity of modelling and handling these processes correctly.

6.4.4 Impact

The impact of this priority theme will be wide-ranging. It will impact the work environment and processes, creativity and innovation, leisure and entertainment, and the private life. Several societal and economical facts call for, but also allow for, improved and more natural interaction between humans and the real world through machines. The ageing society requests ambient intelligence. Globalisation involves the capacity to interact in many languages, and offers a huge market for new products fully addressing this multilingual necessity.

The automation of society implies more efficiency and a 24/7 availability of services and information, while green technologies, such as advanced videoconferencing, need to be prioritised. The continuously reduced costs and

Research Priority	Phase 1: 2013-2014	Phase 2: 2015-2017	Phase 3: 2018-2020
Interacting naturally with agents	Provide usable human interface, reliable speech recognition, natural and intelligible speech synthesis, limited understanding and dialogue capabilities	Provide usable dialogue interface, context and dialogue aware speech recognition and synthesis; recognise and produce emotions, understanding capabilities, context aware dialogue, using other sensors	Provide multiparty (human-agents) interface, multiple voices, mimicking, advanced understanding and advanced personalised dialogue (indirect speech acts, incl. prosodics, lies, humor)
Using language and other modalities	Multimodal interaction (speech, facial expression, gesture, body postures)	Multimodal dialogue, fusion and fission	Fleximodal dialogue, identification of best suited modalities
Conscious of its performing capacities	Confidence in hearing/understanding, recovering from mistakes	Ability to learn continuously and incrementally from mistakes by interaction	Unsupervised learning/forgetting
Multilingual proficiency	Ensure availability or portability to major EU languages; recognise which language is spoken; multilingual access to multilingual information	More languages, accents and dialects; recognise dialects, accents; exploit limited resources; cross-lingual access to information	Speech translation in human-human interactions (multiple speakers speaking multiple languages); cross-cultural support; learn new language with small effort
Resources	Install infrastructure, benchmark data, semantically annotated data (multimodal), dialogue data	Use infrastructure, more data, more languages	Use infrastructure, more data, more languages
Evaluation	Benchmark evaluation; measures and protocols for automated speech synthesis, dialogue systems, speech translation evaluation	Measure of progress; more languages	Measure of progress; more languages

14: Priority Theme 3 – Socially-Aware Interactive Assistants: Preliminary Roadmap

speed improvement of hardware allow for affordable and better technologies, that can now easily be made available online through app stores.

At the same time we still face prohibitive factors. The cultural, political and economical dimensions of language are well perceived, but its technical dimension is not. There is still a psychological barrier for communicating with machines, although this gets more and more common through the use of smartphones and applications such as Skype or Facetime.

6.4.5 Organisation of Research

In order to improve research efficiency within a public-private partnership, the preferred infrastructure were to handle the various applications in connection with the cooperative development of technologies, including the evaluation of progress, and the production of the language and human naturalistic behaviour resources which are necessary for development and testing.

To maximise impact, it is necessary to make a substantial effort in the development of integrated systems based on open architectures, and a multilingual middleware to enable the developed functionalities to be incorporated in a wide range of software. This might best be achieved through a small number of coordinating projects, attached to a federation of strategic projects with complementary goals. These projects should be objective-driven, with clear research, technology and exploitation milestones, coordinated by an on-going road-mapping effort. This includes the production of adequate language and human naturalistic behaviour corpora, semantically annotated including prosodic and non-verbal behavioural cues. This also includes the production (acquisition and annotation) of dialogue corpora from the real world, which implies an incremental system design, and either the use of synchronised continuous observations of all involved parties, or the use of similar data available online (conversations, talk shows).

Dialogue systems evaluation still needs more research on the choice of adequate metrics and protocols. The multilingual dimension that is targeted implies the availability of language resources and language technology evaluation for all languages. Handling them all together reduces the overall effort, given the possibility to use the same best practices, tools and protocols.

6.5 CORE LANGUAGE RESOURCES AND TECHNOLOGIES

The three priority research themes share a large and heterogeneous group of core technologies for language analysis and production that provide development support through basic modules and datasets (see Figure 18, p. 68). To this group belong tools and technologies such as, among others, tokenisers, part-of-speech taggers, syntactic parsers, tools for building language models, information retrieval tools, machine learning toolkits, speech recognition and speech synthesis engines, and integrated architectures such as GATE and UIMA. Many of these tools depend on specific datasets (i. e., language resources), for example, very large collections of linguistically annotated documents (monolingual or multilingual, aligned corpora), treebanks, grammars, lexicons, thesauri, terminologies, dictionaries, ontologies and language models. Both tools and resources can be rather general or highly task- or domain-specific, tools can be language-independent, datasets are, by definition, language-specific. As complements to the core technologies and resources there are several types of resources, such as error-annotated corpora for machine translation or spoken dialogue corpora, that are specific to one or more of the three priority themes.

A key component of this research agenda is to collect, develop and make available core technologies and resources through a shared infrastructure so that the research and technology development carried out in all themes can make use of them. Over time, this approach will improve the core technologies, as the specific research will have certain requirements on the software, extending their feature sets, performance, accuracy etc. through dynamic push-pull effetcs. Conceptualising these technologies as a set of shared core technologies will also have positive effects on their sustainability and interoperability. Also, many European languages other than English are heavily under-resourced, i. e., there are no or almost no resources or basic technologies available [12].

The European academic and industrial technology community is fully aware of the need for sharing resources such as language data (e. g., corpora), language descriptions (e. g., lexicons, thesauri, grammars), tools (e. g., taggers, stemmers, tokenisers) and core technology components (e. g., morphological, syntactic, semantic processing) as a basis for the successful development and implementation of the priority themes. Initiatives such as FLaReNet [50] and CLARIN have prepared the ground for a culture of sharing, META-NET's open resource exchange infrastructure, META-SHARE, is providing the technological platform as well as legal and organisational schemes. All language resources and basic technologies will be created under the core technologies umbrella. The effort will revolve around the following axes: Infrastructure; Coverage, Quality, Adequacy; Language Resources Acquisition; Openness; Interoperability.

6.5.1 Infrastructure

It is imperative to maintain and further to develop META-SHARE. Broad participation by the whole language technology community is essential in maintaining and extending the infrastructure so that acceptance is ensured. META-SHARE will be the key instrument to make language resources available, visible and accessible, to facilitate their sharing and exchange.

The following aspects are important for the next evolutionary steps of META-SHARE: definition of the basic data and software resources that should populate META-SHARE, multilingual coverage, the capacity to attract providers of useful resources or raw data sets, improvements in sharing mechanisms, and collaborative working practices between R&D and commercial users. There must also be a business-friendly framework to stimulate commercial use of resources, based on a sound licensing facility. Close cooperation with the three priority themes is of vital importance, especially for defining the set of needed core technologies and resources.

META-SHARE is not limited to data. Instead, it has to be seen as an international hub of resources and technologies for speech and language services from industries and communities. The development and proposal of tools and web services, including evaluation protocols and collaborative workbenches is deemed essential. The accumulation and sharing of resources and tools in a single place would lower the R&D costs for new applications in new language resource domains.

Sustainability covers preservation, accessibility, and operability (among other things). Collecting and preserving knowledge in the form of existing resources should be a key priority. A sustainability analysis must be part of a resource specification phase. Funding agencies should make a sustainability plan mandatory for projects concerned with the production of language resources.

6.5.2 Coverage, Quality, Adequacy

Innovation in LT crucially depends on language resources but currently there are not enough available resources to satisfy the needs of all languages, quantitatively and qualitatively. Language resources should be produced and made available for every language, every register, every domain to guarantee full coverage and high quality (see Figure 15). New methods of shared or distributed resource development can be exploited to achieve better coverage. It is important to assess the availability of existing resources with respect to their adequacy to applications and technology requirements. This involves assessing the maturity of the technologies for which new resources should be developed. Basic language resource kits should be supported and developed for all languages and, at least, key applications.

Automatic techniques should be promoted to guarantee quality through error detection and confidence assessment. The promotion of validation and evaluation can play a valuable role in fostering quality improvement. Evaluation should encompass technologies, resources, guidelines and documentation. But like the technologies it addresses, evaluation is constantly evolving, and new, more specific measures using innovative methodologies are needed to evaluate the reliability of language resources, while maximal use of existing tools should be ensured for the validation of resources.

Lists of basic language technologies should be compiled that should be either made available or researched and implemented for all languages covered by this agenda. These should include tools such as sentence boundary detection modules, tokenisers, lemmatisers, taggers, parsers, word/phrase aligners etc. as obligatory components for each language. These should also include resources needed for making the modules work for a given language. Other aspects are quality thresholds (minimum accuracy, speed, open availability, interoperability etc.) and cross-lingual evaluation campaigns. After partial attempts at these in the past (e. g., BLARK and ELARK, shared tasks such as CLEF, EuroMatrix Marathons, IWSLT, Morpho-Olympics etc.) a more coordinated, sustainable and also wider attempt is needed.

A Language Resources Impact Factor (LRIF) should be defined in order to enforce the practice of citation of resources on the model of scientific paper authoring and to calculate the actual research impact of resources. A reference model for creating resources will help address the current shortage of resources in terms of breadth (languages and applications) and depth (quality and volume).

In addition to the, putting it in general terms, unification of approaches mentioned above, a set of shared resources and technologies should be compiled for all the languages to be supported through the future initiative. The specifics of this shared set of dictionaries, text and speech corpora, terminologies, ontologies, lexicons, taggers etc. remain to be discussed and determined. It is important that they follow the same basic principles, cover not only general language but also several specific domains tailored to the priority themes, will be interlinked (for multilingual applications) and made available as free, public data sets for research and commercial purposes. The creation of such a shared set of base resources and technologies is imperative for the future European multilingual information society – currently there are many European languages that do not even have a corresponding corpus yet that fulfills certain requirements. National corpora only exist for a handful of languages, many of these corpora are not readily available for research purposes.

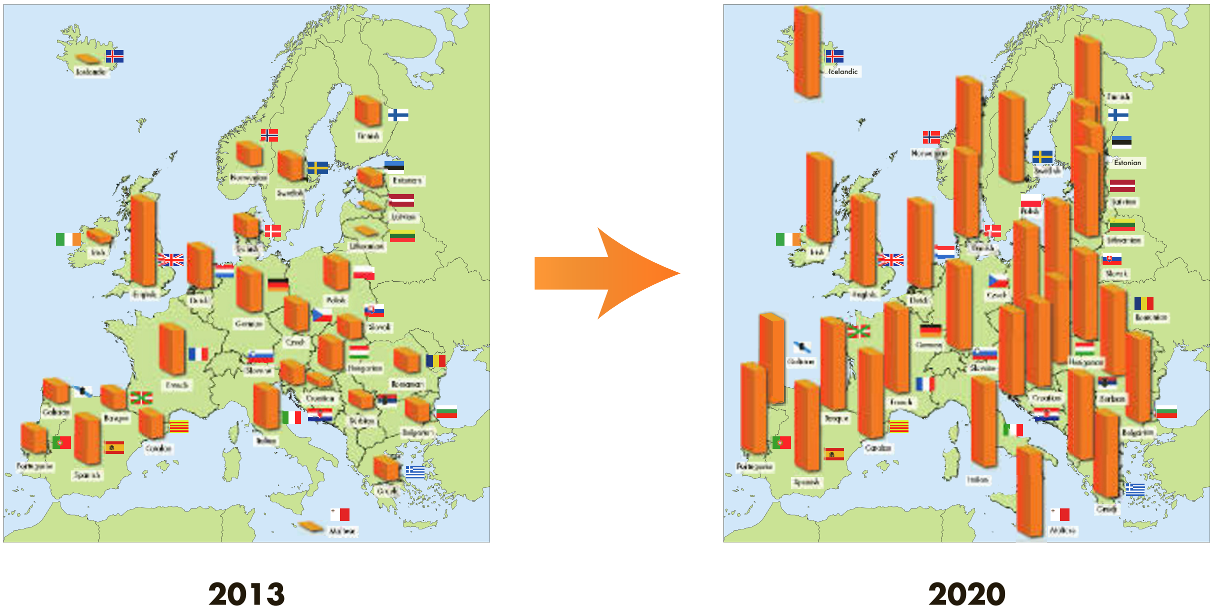
6.5.3 Language Resources Acquisition

Re-use and re-purposing should be encouraged to ensure the reuse of development methods and existing tools. With production costs constantly increasing, there is a need to invest in innovative production methods that involve automatic procedures; strategies that approach or ensure full automation for high-quality resource production should be promoted. It is worth considering the power of social media to build resources, especially for those languages where no language resources built by experts exist yet.

There are several promising experiments in crowdsourcing data collection tasks. Crowd-sourcing makes it possible to mobilise large groups of human talent around the world with just the right language skills so that we can collect what we need when we need it. For instance, it has been estimated that Mechanical Turk translation is 10 to 60 times less expensive than professional translation. A particularly sensitive case is that of less-resourced languages, where language technology should be developed rapidly to help minority-language speakers access education and the Information Society [51, 19, 20, 21].

6.5.4 Openness

There is a strong trend towards open data, i. e., data that are easily obtainable and that can be used with few, if any, restrictions. Sharing data and tools has become a viable solution towards encouraging open data [52], and the community is strongly investing in facilities such as META-SHARE for the discovery and use of resources. These facilities could represent an optimal intermediate solution to respond to the needs for data variety, ease of retrieval, better data description and community-wide access, while at the same time assisting in clearing the intricate issues associated with intellectual property rights (see Section 6.7 for more details).



15: Towards appropriate and adequate coverage of language resources and technologies for Europe

6.5.5 Interoperability

Interoperability of resources seeks to maximise the extent to which they are compatible and therefore integratable at various levels, so as to allow, for instance, the merging of data or tools coming from different sources. All stakeholders need to join forces to drive forward the use of existing and emerging standards, at least in the areas where there is some degree of consensus.

6.5.6 Organisation of Research

In order to optimise the efficiency of shared core technologies for language analysis and production as well as the further development of the infrastructure, maximise the infrastructure's impact, and ensure that requirements for research and development are met at the necessary depth for all languages in all priority themes, the organisation of this shared agenda theme should adopt the following principles: It is necessary to invest in the further development of an integrated infrastructure (i. e., META-SHARE) based on an open architecture, enabling the sharing and further development of resources. The infrastructure should support technology-specific challenges and shared tasks in order to accelerate innovation breakthrough and market-readiness for desperately needed technologies. Concerted activities and policies facilitating the sharing of resources overcoming all stumbling blocks on the way to technical, organisational and legal interoperability should be supported. EU level research must be aligned and tightly coordinated with efforts on the national levels, so that coverage and language-specific developments are efficiently achieved. An important aspect of this coordination effort is concerned with the META-NET White Paper Series [12]: in the 30 different white papers we have concrete and specific assessments of the language- and country-specific situation with regard to demands and technology gaps. The next step is to address and to fill these gaps with high-quality and robust core technologies and language resources.

Research Priority	Phase 1: 2013-2014	Phases 2 and 3: 2015-2020
Infrastructure	Maintain and extend facility(-ies) for sharing resource data and tools; promote accurate and reliable documentation of resources through metadata; cooperation between infrastructure initiatives to avoid the duplication of effort	Automatically accumulate descriptions and resources; multilingual coverage, ease of conversion into uniform formats; integrate web services (SaaS)
Coverage, quality, adequacy	Increase number of resources to address LT and application needs; address formal and content quality by promoting evaluation and validation; promote evaluation and validation activities and the dissemination of their outcomes	Increase number of resources to address LT and application needs; provide HQ resources for all European languages
Acquisition	Define and disseminate LR production best practices; enforce reusing and repurposing; towards the full automation of LR data production; methods for collaborative creation and extension of HQ resources, also to increase coverage; implement workflows of language processing services for acquisition of resources required for the implementation of the priority themes; bridge acquisition methods with linked open data and big data; share the effort for production of LRs between international bodies and individual countries	
Openness	Educate key players with basic legal know-how; elaborate specific, simple and harmonised licensing solutions for data resources; promote copyright exception for research purposes; develop legal and technical solutions for privacy protection; opt for openness of resources, especially publicly funded ones; ensure that publicly funded resources are publicly available free of charge; clear IPR at the early stages of production; try to ensure that re-use is permitted	
Interoperability	Standardisation activities, make standards operational and put them in use; establish permanent Standards Watch; promote and disseminate standards to students and young researchers; encourage/enforce use of best practices or standards in production projects; identify new mature areas for standardisation and promote joint efforts between R&D and industry	

16: Core language resources and technologies: Preliminary Roadmap

6.6 A EUROPEAN SERVICE PLATFORM FOR LANGUAGE TECHNOLOGIES

We argue for and recommend the design and implementation of an ambitious large-scale platform as a central motor for research and innovation in the next phase of IT evolution and as a ubiquitous resource for the multilingual European society. The platform will be used for testing, show casing, proof-of-concept demonstration, avant-garde adoption, experimental and operational service composition, and fast and economical service delivery to enterprises and end-users (see Figure 17).

The proposed creation of a powerful cloud or sky computing platform (see Section 3.5) for a wide range of services dealing with human language, knowledge and emotion will not only benefit the individual and corporate users of these technologies but also the providers. Large-scale ICT infrastructures and innovation clusters such as this suggested platform are also foreseen in the Digital Agenda for Europe (see [5], p. 24).

Users will be able to receive customised integrated services without having to install, combine, support and maintain the software. They will have access to specialised solutions even if they do not use these regularly.

Language technology providers will have ample opportunity to offer stand-alone or integrated services.

Providers of language services rendered by human language professionals will be able to use the platform for enhancing their services by means of appropriate technology and for providing their services stand-alone or integrated into other application services.

Researchers will have a virtual laboratory for testing, combining, and benchmarking their technologies and for exposing them in realistic trials to real tasks and users.

Providers of services that can be enabled or enhanced by text and speech processing will utilise the platform for testing the needed LT functionalities and for integrating them into their own solutions.

Citizens and corporate users will enjoy the benefits of language technology early and at no or reasonable costs through a large variety of generic and specialised services offered at a single source.

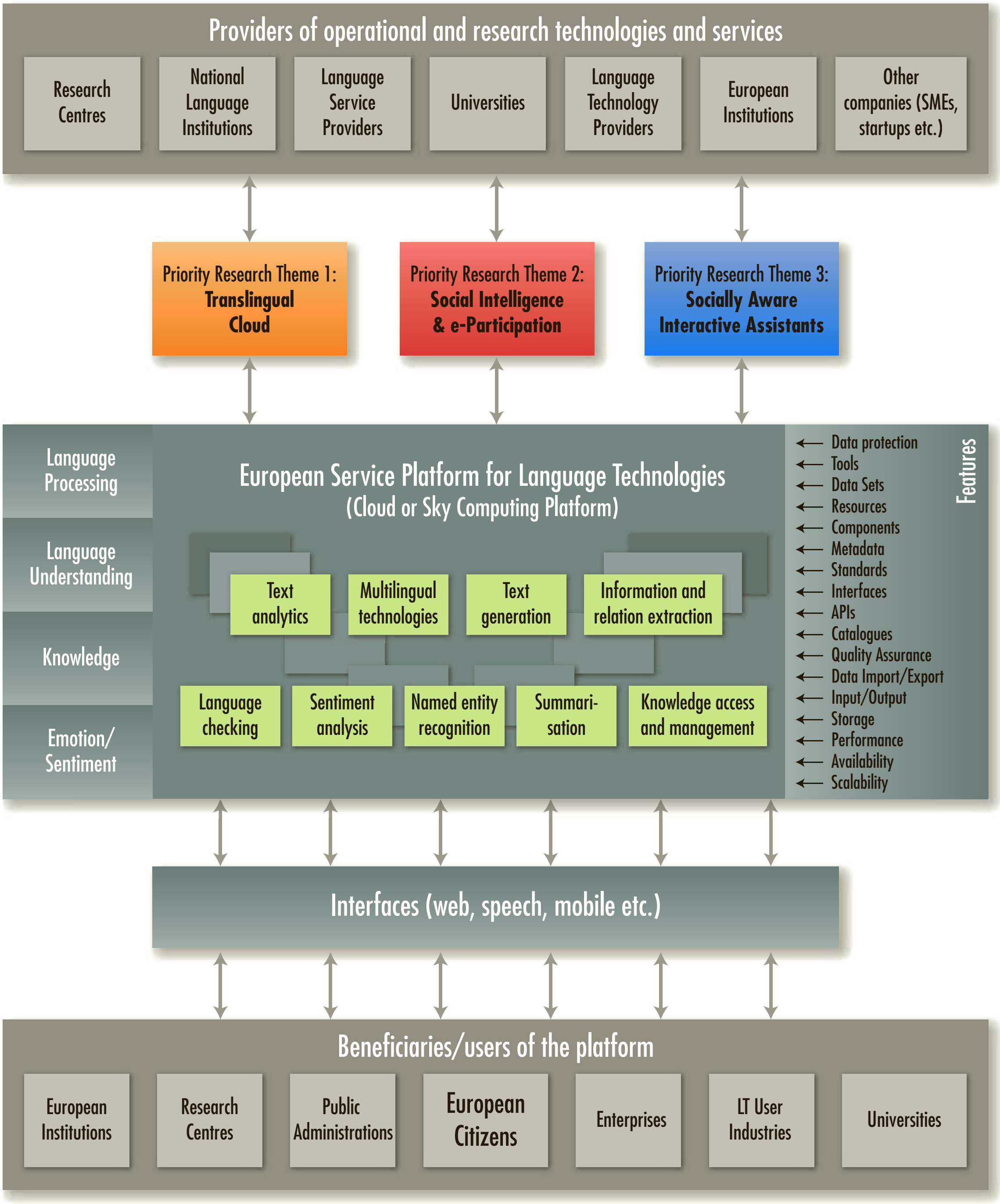
In order to allow for the gigantic range of foreseeable and currently not yet foreseeable solutions, the infrastructure will have to host all relevant simple services, including components, tools and data resources, as well as various layers or components of higher services that incorporate simpler ones. META-SHARE can play an important role in the design of the platform (see Section 6.5).

A top layer consists of **language processing** such as text filters, tokenisation, spell, grammar and style checking, hyphenation, lemmatising and parsing. At a slightly deeper level, services will be offered that realise some degree and form of **language understanding** including entity and event extraction, opinion mining and translation. Both basic language processing and understanding will be used by services that support **human communication** or realise human-machine interaction. Part of this layer are question answering and dialogue systems as well as email response applications. Another component will bring in services for processing and storing **knowledge** gained by and used for understanding and communication. This part will include repositories of linked data and ontologies, as well as services for building, using and maintaining them. These in turn permit a certain range of rational capabilities often attributed to a notion of intelligence. The goal is not to model the entire human intelligence but rather to realise selected forms of **inference** that are needed for utilising and extending knowledge, for understanding and for successful communication. These forms of inference permit better decision support, pro-active planning and autonomous adaptation. A final part of services will be dedicated to **human emotion**. Since people are largely guided by their emotions and strongly affected by the emotions of others, truly user-centred IT need facilities for detecting and interpreting emotion and even for expressing emotional states in communication.

We consider the paradigm of federated cloud services or sky computing with its emerging standards such as OCCI, OVM and CDMI and toolkits such a OpenNebula as the appropriate approach for realising the ambitious infrastructure. All three priority areas of this SRA will be able to contribute to and at the same time draw immense benefits from this platform. There are strong reasons for aiming at a single service platform for the three areas and for the different types of technologies. They share many basic components and they need to be combined for many valuable applications, including the selected showcase solutions of the three areas.

Implementation of the Platform

The creation of this platform has to be supported by public funding. Because of the high requirements concerning performance, reliability, user support, scalability, persistence as well as data protection and conformance with privacy regulation, the platform needs to be established by a consortium with strong commercial partners and also be operated by this consortium or a commercial contractor. A similar platform with slightly different desiderata and functionalities is currently built under the name Helix-Nebula for the Earth Sciences with the help of the following commercial partners: Atos, Capgemini, CloudSigma, Interoute, Logica, Orange Business Services, SAP, SixSq, Telefonica, Terradue, Thales, The Server Labs and T-Systems. Partners are also the Cloud


Providers of operational and research technologies and services
Research Centres
National Language Institutions
Language Service Providers
Universities
Language Technology Providers
European Institutions
Other companies (SMEs, startups etc.)
Priority Research Theme 1: Translingual Cloud
Priority Research Theme 2: Social Intelligence & e-Participation
Priority Research Theme 3: Socially Aware Interactive Assistants
Language Processing
Language Understanding
Knowledge
Emotion/ Sentiment
European Service Platform for Language Technologies
(Cloud or Sky Computing Platform)
Text analytics
Multilingual technologies
Text generation
Information and relation extraction
Language checking
Sentiment analysis
Named entity recognition
Summari-sation
Knowledge access and management
Features
Data protection
Tools
Data Sets
Resources
Components
Metadata
Standards
Interfaces
APIs
Catalogues
Quality Assurance
Data Import/Export
Input/Output
Storage
Performance
Availability
Scalability
Interfaces (web, speech, mobile etc.)
Beneficiaries/users of the platform
European Institutions
Research Centres
Public Administrations
European Citizens
Enterprises
LT User Industries
Universities

17: European Service Platform for Language Technologies

Security Alliance, the OpenNebula Project and the European Grid Infrastructure. These are working together with major research centres in the Earth Sciences to establish the targeted federated and secure high-performance computing cloud platform.

The intended platform for LT and neighbouring fields would be intended for a mix of commercial and non-commercial services. It would be cost-free for all providers of non-commercial services (cost-free and advertisement-free) including research systems, experimental services and freely shared resources but it would raise revenues by charging a proportional commission on all commercially provided services. In order to reduce dependence on individual companies and software products, the base technology should be supplied by open toolkits and standards such as OpenNebula and OCCI.

For each priority research theme, chances for successful showcasing and successful commercial innovation will increase tremendously if usable services of required strength and reliability could be offered on such a platform.

The platform will considerably lower the barrier for market entry for innovative technologies, especially for products and services offered by SMEs. Still, these stakeholders may not have the resources, expertise, and time to create the necessary interfaces to integrate their results into real-life services, let alone the overarching platform itself. There is still a gap between research prototypes and products that have been engineered and tested for robust applications. Moreover, many innovative developments require access to special kinds of language resources such as recordings of spoken commands to smartphones, which are difficult to get for several reasons.

Thus the service platform will be an important instrument for supporting the entire innovation chain, but, in addition, interoperability standards, interfacing tools, middle-ware, and reference service architectures need to be developed and constantly adapted. Many of these may not be generic enough to serve all application areas, so that much of the work in resource and service integration will have to take place in the respective priority theme research actions.

6.7 LEGAL CHALLENGES

Legal challenges are involved on multiple levels in our future research and technology plans as described in this agenda. One of the key challenges for our community and also for the policy makers is to push for the development of a common legal framework that would facilitate resource sharing efforts abiding by the law, benefiting from the adoption of "fair use" principles and appropriate copyright exceptions. It is of utmost importance that legislation regarding resource use and resource acquisition be harmonised, and even standardised, for all types of language resources, and that free use be allowed, at least for research or non-profit purposes (see Section 6.5).

Other areas in which we are facing or in which we expect legal challenges are the "trust" features of the European Language Technology Platform, which needs to exhibit a maximum level of data security in order to protect confidential documents (from contracts to patient data), or novel methods of acquiring written or spoken data for language resources. Any grant of access to language resources should ideally include not only the right to read the relevant content but also to allow transformative uses, dissemination and distribution of such resources and their derivatives, according to the needs and policies of language resources owners and users. Not only the acquisition but also the sharing and distribution of language resources is constantly hindered or completely disabled by legal aspects which should ideally be resolved once and for all. Legal issues such as these are severe stumbling blocks that can bring innovation to a complete standstill. In addition, content or approaches to data privacy or security that are legal in one country may be illegal in another. These aspects can be partially addressed on the software level (for example, through appropriate metadata records that reflect different legal realms) but should ideally be harmonised on the European or global level.

META-NET favours and aligns itself with the growing open data and open source movement and the idea of opening up data, resources and technologies (especially those whose development was supported through public funding) instead of locking them away. META-NET advocates the use of a model licensing scheme with a firm orientation towards the creation of an openness culture and the relevant ecosystem for language resources.

6.8 LANGUAGES TO BE SUPPORTED

The research and technology development programme specified in this agenda has a much broader scope in terms of languages to be supported than our study "Europe's Languages in the Digital Age" (Section 4, p. 27 ff.). The set of languages to be reflected with corresponding technologies include not only the currently 23 official languages of the European Union but also recognised and unrecognised regional languages and the languages of associated countries or non-member states. Equally important are the minority and immigrant languages that are in active use by a significant population in Europe (for Germany, these are, among others, Turkish and Russian; for the UK, these include Bengali, Urdu/Hindi and Punjabi). An important set of languages outside our continent are those of important political and trade partners such as, for example, Chinese, Indonesian, Japanese, Korean, Russian, and Thai. META-NET already has good working relationships with several of the respective official bodies, especially EFNIL (European Federation of National Institutions for Language), NPLD (Network to Promote Linguistic Diversity, [51]), and also the Maaya World Network for Linguistic Diversity.

The concrete composition of languages to be supported by this agenda's research programme up until the year 2020 and beyond, depends on the concrete composition of participating countries and regions and also on the specific nature of the funding instruments used and combined for realising the ambitious plan. It remains to be discussed what it means for a language to be supported through this strategic programme; most probably, the level of support will have to be determined through a concrete set of specific resources and specific base technologies that need to be researched and developed for a given language and that need to fulfill certain requirements (with regard to, among others, coverage, precision, quality, speed etc.). The next level of support would, then, be determined by including a language in one or more of the priority research themes.

Not all countries have the required expertise or human resources to take care of the technology support for their languages. For example, in Iceland there is not a single position in LT at any Icelandic university or college and there is only one company that works in this area. Those colleagues who work on LT at universities and research institutes come from either language or computer science departments; their main duties are not related to LT, still they managed to produce a few basic technologies and resources but advanced types of resources do not exist at all for Icelandic, nor do they for many other under-resourced languages. This is why we need to intensify research and establish techniques, methods and instruments for research and knowledge transfer so that colleagues in countries such as Iceland can benefit as much as possible for their own language from the research carried out in other countries for other languages. Bootstrapping the set of core language technologies and resources for all languages spoken in Europe is not a matter of a few countries joining forces but a challenge on the European scale that must be addressed accordingly to avoid digital exclusion and secure future business development.

META-NET realises that Europe is a multi-ethnic region in which many more languages than only the official ones are spoken. Therefore, it is important not only to carry out research and technology development on the official and a few additional languages but also to work on those languages that are in active use by a significant part of the population, in order to address the severe issue of linguistic ghettoisation and finally to bring about a truly multilingual European information society.

As regards funding the programme we suggest an approach that involves multiple stakeholders, especially the European Union, the Member States, Associated Countries, other countries and also regions, not only in Europe but ultimately also on other continents. Research on advanced, sophisticated monolingual technologies is to be supported by the respective countries' funding agencies primarily. Research on multilingual technologies and also research on basic technologies and resources for under-resourced languages needs to be supported by the EU along with the respective countries and regions. Specific procedures for research and knowledge transfer need to be agreed upon and put into action so that the speakers of these languages can benefit from our activities as much and as quickly as possible. In order to provide basic technology support for those languages spoken in Europe with active hubs of research outside our continent, connections to the leading research centres need to be established or intensified so that Europe can benefit from technologies that have been developed by these centres. If technologies exist, funding schemes need to be established so that they can be adopted, if necessary, to the standards that will be put into practice in Europe in the years to come, especially with regard to sharing, distribution, data formats, APIs and inclusion in the European Language Technology Platform.

6.9 RESEARCH ORGANISATION

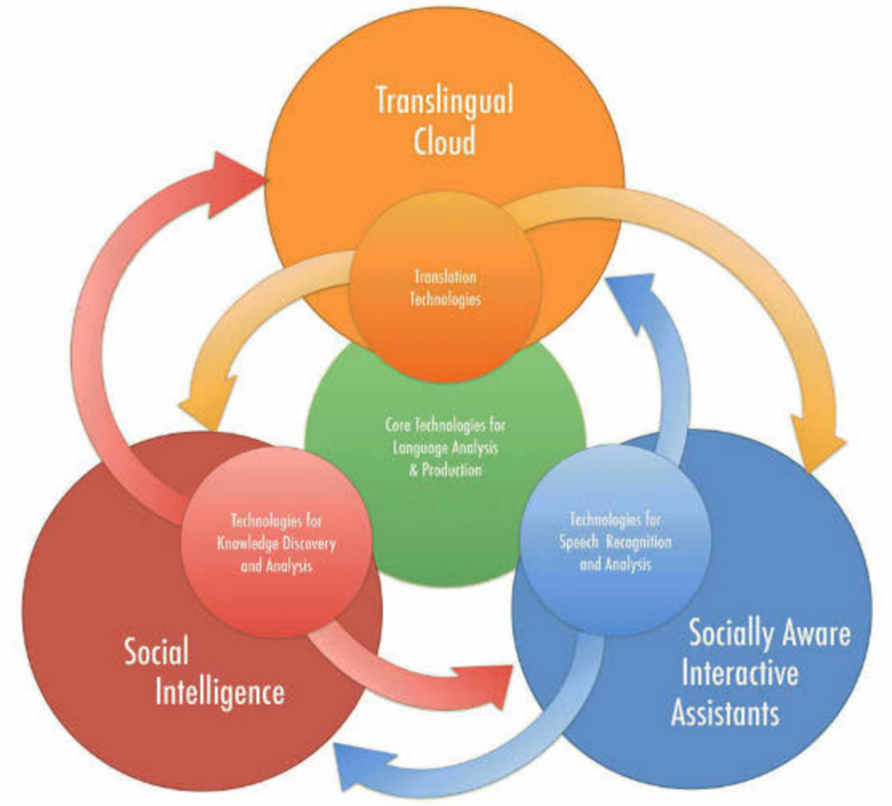
The three proposed priority research themes overlap in technologies and challenges – this is intended. The overlap reflects the coherence and maturation of the field. At the same time, the resulting division of labour and sharing of resources and results is a precondition for the realisation of this highly ambitious programme.

All three themes need to benefit from progress in core technologies of language analysis and production such as morphological, syntactic and semantic parsing and generation. But each of the three areas will concentrate on one central area of language technology: the Translingual Cloud will focus on cross-lingual technologies such as translation and interpretation; the Social Intelligence strand will take care of knowledge discovery, text analytics and related technologies; the research dedicated to the Interactive Assistants will take on technologies such as speech and multimodal interfaces (see Figure 18).

Except for a few large national projects and programmes such as Technolangue and Quaero in France, Verbmobil and Theseus in Germany and DARPA Communicator and GALE in the US, the field of language technology does not have experience with research efforts of the magnitude and scope required for the targeted advances and plans in this SRA. Nevertheless, our technology area has to follow developments in other key engineering disciplines and speed up technology evolution by massive collaboration based on competitive division of labour and sharing of resources and results. In our reflection on optimal schemes for organising we tried to draw lessons from our own field's recent history and to capitalise on experience from other fields by adopting approaches that proved successful and evading encountered pitfalls.

The final model for the organisation of collaboration will have to be guided by a thoughtful combination of the following basic approaches.

Flexible collaborative approach: For each priority theme, one or several very large cooperating and competing lead projects will share an infrastructure for evaluation, communication and resources (data and base technologies). Mechanisms for reducing or terminating partner involvements and for adding new partners or subcon-


Translingual Cloud
Translation Technologies
Core Technologies for Language Analysis & Production
Technologies for Knowledge Discovery and Analysis
Technologies for Speech Recognition and Analysis
Social Intelligence
Socially Aware Interactive Assistants

18: Scientific cooperation among the three priority research themes

tracted contributors should provide flexibility. A number of smaller projects including national and regional projects will provide building blocks for particular languages, tasks, component technologies or resources. A cooperation scheme will be designed for effectively and flexibly involving EC-funding, contributions from member states, industrial associations, and language communities, among others [53]. The choice of funding instruments will be determined in due time.

Staged approach: Two major phases are foreseen (2015-2017, 2018-2020). The major phases should be synchronised among the themes and also projects.

Evolutionary approach: Instead of banking on one selected paradigm, competing approaches will be followed in parallel with shared schemes for evaluation, merging, adopting and discontinuing research threads so that the two elements of successful evolutionary research approaches, selection and cross-fertilisation, are exploited to the maximum extent possible.

Analytical approach: Instead of the currently predominant search for an ideal one-fits-all approach, the research will focus on observed quality barriers and not shun computationally expensive dedicated solutions for overcoming particular obstacles.

Bootstrapping approach: Better systems can be derived from more and better data and through new insights. In turn, improved systems can be used to gain better data and new insights. Thus the combination of the analytical evolutionary approach with powerful machine learning techniques will be the basis for a technology bootstrapping, which has been the by far most fruitful scheme for the development of highly complex technologies.

Close cooperation with relevant areas of service and technology industries: In order to increase chances of successful commercialisation and to obtain convincing and sufficiently tested demonstrations of novel applications, the relevant industrial sectors must be strongly integrated into the entire research cycle.

Tighter research-innovation cycle: Through the collaboration between research, commercial services and commercial technology industries, especially through the shared evaluation metrics and continuous testing, the usual push-model of technology transfer will hopefully be substituted by a pull-model, in which commercial technology users can ask for specific solutions. In the envisaged research scheme, incentives will be created for competing teams each composed of researchers, commercial users and commercial developers by the participating enterprises for initiating successful innovations.

Interdisciplinary approach: A number of science, technology and service areas need to be integrated into the research from day one. Some technology areas such as speech technologies, language checking and authoring systems need to be represented by providers of state-of-the-art commercial products.

The coordination among the three research strands poses administrative challenges. Because of the described interdependencies and also because of the need to maintain and improve the obtained level of cohesion and community spirit in the European Language Technology community, a coordinating body is needed. Whether such an entity is jointly carried by the three areas or by a separate support project, needs to be determined in the upcoming discussion on the appropriate support instruments for the identified research priorities.

6.10 SUPPORTIVE POLICY MAKING

Technological progress would be even more efficient and effective if the proposed research effort could be accompanied by appropriate supportive policy making in several areas. One of these areas is multilingualism. Overcoming language barriers can greatly influence the future of the EU and the whole planet [19, 20, 21]. Solutions for better communication and for access to content in the native languages of the users would reaffirm the role of the EC to serve the needs of the EU citizens. A substantial connection to the infrastructural program CEF could help to speed up the transfer of research results to badly needed services for the European economy and public. At the same time, use cases should cover areas where the European societal needs massively overlap with business opportunities to achieve funding investment that pays back, ideally public-private partnerships.

Language policies supporting multilingualism can create a tangible boost for technology development. Some of the best results in machine translation have been achieved in Catalonia, where legislation supporting the use of the Catalan language has created an increased demand for automatic translation.

Numerous US-originating breakthroughs in IT that have subsequently led to commercially successful products of great economic impact could only be achieved by a combination of systematic long-term research support coupled with public procurement. Many types of aircraft or the autonomous land vehicle would not have seen the light of the day without massive military support, even the internet or the speech technology behind Apple Siri heavily benefited from sequences of DARPA programmes often followed by government contracts procuring earlier versions of the technology for military or civilian use by the public sector.

The greed for originality on the side of the public research funding bodies and their constant trial-and-error search for new themes that might finally help the European IT sector to be in time with their innovations has often caused the premature abortion of promising developments, whose preliminary results were more than once taken up by research centres and enterprises in the US. An example in language technology is the progress in statistical machine translation. Much of the groundwork laid in the German government-sponsored project Verbmobil (1993–2000) was later taken up by DARPA research and industrial systems including Google Translate.

In order to drive technology evolution with public funding to a stage of maturity where first sample solutions can deliver visible benefits to the European citizens and where the private sector can take up technologies to then develop a wide range of more sophisticated profitable applications, we strongly advocate a combination of

1. language policies supporting the status of European languages in the public sector,
2. long-term systematic research efforts with the goal to realise badly needed pre-competitive basic services,
3. procurement of solution development by European public administrations.

European policy making should also speed up technology evolution by helping the research community to gain affordable and less restrictive access to text and speech data repositories, especially to data that have been collected with public support for scientific and cultural purposes. Today, outdated legislation and restrictive interpretation of existing law hinder the effective use of many valuable data collections such as, for example, several so-called national corpora. The research community urgently needs the help of European and national policy makers for modes of use of these data that would boost technology development without infringing on the economic interests of authors and publishers.