Lang Resources & Evaluation (2014) 48:679–707
DOI 10.1007/s10579-014-9277-0

PROJECT NOTES

An overview of the European Union's highly multilingual parallel corpora

Ralf Steinberger · Mohamed Ebrahim · Alexandros Poulis · Manuel Carrasco-Benitez · Patrick Schlüter · Marek Przybyszewski · Signe Gilbro

Published online: 29 August 2014

Abstract Starting in 2006, the European Commission's *Joint Research Centre* and other European Union organisations have made available a number of large-scale highly-multilingual parallel language resources. In this article, we give a comparative overview of these resources and we explain the specific nature of each of them. This article provides answers to a number of question, including: What are these linguistic resources? What is the difference between them? Why were they originally created and why was the data released publicly? What can they be used

R. Steinberger (✉)
European Commission – Joint Research Centre (JRC), Ispra, Italy
e-mail: Ralf.Steinberger@jrc.ec.europa.eu

M. Ebrahim
Cognizant-SetCon GmbH, Munich, Germany
e-mail: Mohamed@alqamus.de

A. Poulis
Lionbridge Technologies, Inc, Tampere, Finland
e-mail: Alexandros.Poulis@lionbridge.com

M. Carrasco-Benitez · P. Schlüter
European Commission – Directorate General for Translation (DGT), Luxembourg, Luxembourg

M. Carrasco-Benitez
e-mail: Manuel.Carrasco-Benitez@ec.europa.eu

P. Schlüter
e-mail: Patrick.Schluter@ec.europa.eu

M. Przybyszewski
European Commission – Directorate General Education And Culture (EAC), Brussels, Belgium
e-mail: Marek.PRZYBYSZEWSKI@ec.europa.eu

S. Gilbro
European Centre for Disease Prevention and Control (ECDC), Stockholm, Sweden
e-mail: info@ecdc.europa.eu

for and what are the limitations of their usability? What are the text types, subject domains and languages covered? How to avoid overlapping document sets? How do they compare regarding the formatting and the translation alignment? What are their usage conditions? What other types of multilingual linguistic resources does the EU have? This article thus aims to clarify what the similarities and differences between the various resources are and what they can be used for. It will also serve as a reference publication for those resources, for which a more detailed description has been lacking so far (EAC-TM, ECDC-TM and DGT-Acquis).

Keywords Parallel corpora · Linguistic resources · Highly multilingual · European Union · Translation memory · JRC-Acquis · DGT-Acquis · DGT-TM · DCEP · ECDC-TM · EAC-TM · JRC EuroVoc Indexer JEX · EuroVoc · Eur-Lex

1 Introduction and motivation

In recent years, European Union organisations have released a number of large-scale multilingual linguistic parallel resources, in between 22 and 26 languages, all available via the JRC's web pages.[1] These are the full-text corpora JRC-Acquis (Steinberger et al. 2006), DGT-Acquis and *Digital Corpus of the European Parliament* (DCEP; Hajlaoui et al. 2014); the translation memories (TMs) DGT-TM (Steinberger et al. 2012b), ECDC-TM and EAC-TM; as well as the document collection accompanying the multi-label categorisation software *JRC EuroVoc Indexer* (JEX; Steinberger et al. 2012a).

This article aims at giving a structured overview over these seven resources, revealing why the EU releases this data (Sect. 2), what the specific strengths (Sect. 3) and what the limitations (Sect. 4) of these EU resources are. In Sect. 5, we list a range of usage examples for which such highly multilingual parallel corpora are particularly useful. Section 6 describes the resources from the points of view of language coverage, source language of the translations, translation quality, document types and subject domain categorisation of the individual texts, sentence splitting, as well as alignment across the languages. While Table 1 contrasts all structured features of these seven resources, Sect. 7 provides any additional information and thus more detailed background information on each of them. In Sect. 8, we give a hint on how overlap between the corpora can be avoided or reduced. Section 9 summarises the usage conditions and the historical development regarding this important aspect of text corpora. Section 10 lists a number of multilingual resources other than parallel corpora that have been made available by European Union (EU) institutions. Finally, Sect. 11 provides a brief summary.

The *Joint Research Centre* (JRC)[2] is the European Commission's in-house science service, working in a very wide range of subject areas, one of which is

[1] All EU corpora discussed here can be downloaded from https://ec.europa.eu/jrc/language-technologies.

[2] See https://ec.europa.eu/jrc/. All URLs were last visited on 7 February 2014.

Table 1 Comparative analysis of features of seven publicly accessible EU corpora

	JRC-Acquis	DGT-Acquis	JEX data	DCEP	DGT-TM	ECDC-TM	EAC-TM
Year of first release	May-2006	Dec-2012	May-2012	2014	Nov-2007	Oct-2012	Jan-2013
Updates since (last update)	Yes (Feb-2009)	No	No	No	Yes (Apr-2013)	No	No
More updates expected	No	Yes	No	Yes	Yes, annually	No	Probably yes
Documents from years (incl. updates)	1958–2006	2004–2011	1952–2011	2001–2012		Current (status 2012)	2008–2012
Alignment unit	Sentence-aligned full-text	Paragraph-aligned full-text	Full-text	Full-text	TM (sentences; TUs)	TM (sentences; TUs)	TM (sentences; TUs)
Alignment language pairs	All	All	All	All	All, via En pivot	All, via En pivot	All, via En pivot
Alignment method	Vanilla + HunAlign	In-house system (DGT + Prompsit)	Manual, document level only	Manual, currently document level only	In-house system DGT	TM (manual)	TM (manual)
Format	TEI (XML)	Muset; Formex-4 (XML); TIFF; plain text	JEX compact format (plain text)	SGML, XML and plain text	TMX (XML)	TMX (XML)	TMX (XML)
Character set	UTF-8	UTF-8	UTF-8	UTF-8	UTF-16 Little Endian	UTF-8	UTF-8
Languages	22	23	22	23	23	22 + IS + NO	22 + HR + IS + NB/NO + TR
Number of language pairs	231	253	231	253	253	300	325
Source language	Various (mostly EN)	Various (mostly EN)	Various (mostly EN)	Various (mostly EN)	Various (mostly EN)	EN	EN

Table 1 continued

	JRC-Acquis	DGT-Acquis	JEX data	DCEP	DGT-TM	ECDC-TM	EAC-TM
Size English (N° of words)	55.5 Mio	98 Mio	59 Mio	103 Mio	85.2 Mio	29 K	45 K
Size all (N° of words)	1 Bio	unknown	unknown	1.37 Bio	1 Bio	320 K	540 K
Space on disk, all (zipped)	3.7 GB	4 packages, from 3 to 81 GB	1.6 GB	6 GB	3.4 GB	3.7 MB	3.5 MB
Subject domain	Wide-coverage, all areas of public life and politics, e.g. economy, health, IT, law, agriculture, food, politics, social issues, etc.	Wide-coverage, all areas of public life and politics, e.g. economy, health, IT, law, agriculture, food, politics, social issues, etc.	Wide-coverage, all areas of public life and politics, e.g. economy, health, IT, law, agriculture, food, politics, social issues, etc.	Wide-coverage, all areas of public life and politics, e.g. economy, health, IT, law, agriculture, food, politics, social issues, etc.	Wide-coverage, all areas of public life and politics, e.g. economy, health, IT, law, agriculture, food, politics, social issues, etc.	Health-related topics (anthrax, botulism, cholera, hepatitis, etc.); organisation and activities of ECDC (e.g. job opportunities; epidemic intelligence and surveillance).	Education, training, culture, youth, administration
Text type	Legal and administrative; declarations and resolutions; agreements; acts and common objectives.	All OJ Series, i.e. Series L, ML, C, CA and CE;	OJ, L-Series, secondary legislation	Press-releases, technical announcements, meeting minutes, reports of parliamentary committees, legal and administrative, oral and written questions, meeting agendas, etc.	OJ L-Series	Web pages of ECDC	Forms and reference data for funding applications and reports related to the Life-long Learning Programme and the Youth in Action programme.

Table 1 continued

	JRC-Acquis	DGT-Acquis	JEX data	DCEP	DGT-TM	ECDC-TM	EAC-TM
Data creator	EC-JRC	EC-DGT	EC-JRC	EP-DGTRAD	EC-DGT	ECDC	EC-EAC
Translators	EC professional	EC professional	EC professional	EP professional	EC professional	EC professional	Specialised staff in national agencies
Pre-processing by	JRC + RAS + BUTE	DGT and external contractor (Prompsit), from Formex-4	JRC	EP-DGTRAD	DGT (from Formex-4)	JRC	EAC + JRC
CELEX document number	Yes	Not currently	Yes	No	Yes	No	No
EuroVoc-indexed	Yes	Not currently	Yes	No	No	No	No

See Sects. 6 to 9 for further explanations

media monitoring and—as a supporting technology—Language Technology.[3] The main product, which is freely accessible to the public, is the *Europe Media Monitor* (EMM) family of applications (Steinberger et al. 2009; Steinberger 2013).[4] As EMM's main users (EU Institutions, the 28 EU Member States, various United Nations sub-organisations, the African Union, the Organisation of American States, and many more) are highly international, it is crucial that EMM monitors and analyses the news in many languages. Driven by the need to deliver text mining applications for all (currently 24) official EU languages (but also in a variety of non-EU languages), the JRC made use of the EU's parallel text collections to create multilingual resources and to boot-strap resources and applications across languages. This need is the initial motivation behind the preparation and usage of sentence-aligned parallel corpora at the JRC. When the Publications Office of the European Institutions agreed to give their permission, in 2006, the JRC released its 22-language parallel corpus JRC-Acquis to the public. The JRC-Acquis was then the largest parallel corpus in existence, considering its language coverage (22 languages, 231 language pairs) and its size (over 1 billion words). Several other EU corpora and other language resources have been released since.

A number of further important parallel language resources need to be mentioned in this context, even though they were either the outcome of a private initiative or of EU-funded projects: (1) In 2005, Koehn (2005) released the EuroParl corpus, consisting of the verbatim reports of the speeches made in the European Parliament's plenary. EuroParl initially covered 11 languages and has since been extended to cover 21 languages with a total of about 60 million words.[5] The new *Digital Corpus of the European Parliament* (DCEP, see Sect. 7.3) excludes the verbatim reports in order to avoid overlap with EuroParl. (2) The Multext project (*Multilingual Text Tools and Corpora*; Ide and Véronis 1994) aimed at developing standards, tools, corpora and linguistic resources for a wide variety of languages. Multext initially covered six languages, but has since been extended to up to 18 languages.[6] The tools included a text editor and tools for SGML manipulation, text segmentation, morpho-lexical treatment, multilingual text alignment, a speech workbench, as well as a variety of libraries and utilities. (3) Multext-East (Erjavec and Ide 1998; Erjavec 2010), a spin-off of Multext, developed morpho-syntactic specifications and language resources for six central and eastern European languages, plus in English as a hub language. In its latest version, Multext-East covers seventeen languages.[7] The importance of these two early projects lies not only in providing language resources for a set of lesser-resourced languages, but most of all in providing *parallel and comparable* specifications and resources. This includes, for instance, the identification of a single harmonised set of morpho-syntactic features for all of these languages. (4) An important building stone in the exploitation of parallel corpora has certainly also been the Arcade evaluation

[3] For details, see https://ec.europa.eu/jrc/en/research-topic/internet-surveillance-systems.

[4] The EMM websites can be accessed publicly via http://emm.newsbrief.eu/overview.html.

[5] See http://www.statmt.org/europarl/.

[6] See http://aune.lpl.univ-aix.fr/projects/multext/.

[7] See http://nl.ijs.si/ME/.

campaign (1995–2005) for parallel text alignment systems (Chiao et al. 2006). (5) Jörg Tiedemann's OPUS open parallel corpus collection (Tiedemann and Nygaard 2004; Tiedemann 2009) is a collection of translated texts from the web which has been growing since its first release in 2003, when it consisted of 30 million words in 60 languages. By 2013, OPUS has reached a coverage of over 150 languages with altogether five billion aligned translation units.[8] The parallel EU corpora presented in this article are mostly parallel across all languages covered while the ones in OPUS contain both bilingual and multilingual corpora.

2 Why is the EU releasing this data

In a nutshell, the motivation of the EU Institutions to support the development of multilingual text analysis tools is related to four main values and objectives: (1) the development of more business potential, (2) the improvement of democracy through transparency of information, (3) the maintenance of the EU's linguistic diversity and (4) the preservation of the EU's cultural diversity. These main objectives can be derived from Directive 2003/98/EC of the European Parliament and of the Council on the re-use of public sector information,[9] as well as from Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission Documents.[10]

These four values are interlinked. Multilinguality is one of the basic principles of the EU. It is an essential part of cultural and linguistic diversity. Text mining software improves search and retrieval of relevant information and it has the propensity to make citizens more informed and to improve their democratic abilities. Machine translation and cross-lingual information access technology can give citizens also access to information across national borders, and it may thus widen their horizon and improve cross-national understanding. Language Technology applications can contribute to making the EU more transparent, egalitarian, accountable and democratic by giving the EU citizen access to legislative and policy proposals in their own and also in other languages. Language Technology may also have an impact on cross-border exploitation of other types of information and it may thus have a positive effect on an unhindered competition in the EU's internal market.

Even for the majority of the EU's 24 official languages, sufficiently efficient Language Technology applications do not yet exist. A recent study of the *Multilingual Europe Technology Alliance* META shows that 21 of 30 studied European languages (70 %) are in danger of digital extinction because the digital support for these languages is non-existent or weak at best.[11] This means that only few Language Technology software applications are available for these languages. In order to develop higher-level applications such as machine translation or

[8] See http://opus.lingfil.uu.se.

[9] See http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:EN:NOT for details and to read the full text of the regulation.

[10] See http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:330:0039:0042:EN:PDF.

[11] See http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison.

information extraction, basic resources such as corpora, dictionaries, morphological analysers and other text processing tools are needed. The simplest and most basic resource are large aligned text collections in various languages as they ease the development of monolingual and cross-lingual Language Technology applications.

The previously mentioned Directive 2003/98/EC on the re-use of public sector information recognises that public sector information such as multilingual collections of documents can be an important primary material for digital content products and services. The Commission Decision from 2011 goes beyond the 2003 Directive by encouraging EU Member States to adopt open data policies and to allow a broad use of documents held by public sector bodies. It states that the Commission has set an example that is to be followed by the Member States.

The JRC's media monitoring team has been working for many years on developing text mining software for over twenty languages and its members are very much aware of the lack of multilingual language resources. Especially *parallel* resources would be useful as their existence would massively speed up the development of highly multilingual text analysis applications. Parallel resources offer the same functionality across languages, using the same categories and the same input and output format (Steinberger 2011). Due to the JRC's awareness of the needs of the computational linguistics R&D community and due to its good contacts with the owners of large collections of parallel multilingual text data inside EU organisations, the JRC was able to push the release of several large-scale parallel corpora. Additionally, the JRC also made available a number of by-products of parts of its own media analysis engine (see Sect. 10).

EU document collections are a very rich resource, but their usefulness has its limits. Section 3 summarises the main positive features of such EU resources, Sect. 4 focuses on its limitations, and Sect. 5 lists a few concrete examples of how such parallel corpora can be used to build or improve Language Technology tools and applications.

3 What are the specific positive features of these EU resources

The most prominent feature of the resources described here is the high number of languages and the parallel nature of the corpora. While there are individual parallel corpora that exist in even more language versions, such as the *Bible* (Resnik et al. 1999),[12] the Declaration of Human Rights, or George Orwell's novel *1984* (Erjavec and Ide 1998), to our knowledge there are no other larger parallel corpora covering twenty languages. Various other large-scale corpora are available for languages with a larger population of native speakers (e.g. English, French, Spanish, Chinese) or for languages that are particularly well documented (e.g. Dutch[13]), and for language pairs involving these languages. As soon as one leaves the area of the few highly developed languages (from a Language Technology point of view), however,

[12] See http://homepages.inf.ed.ac.uk/s0787820/bible/ for the more recent distribution of an aligned Bible corpus that is much larger than that prepared by Resnik et al. (1999).

[13] See the META-NET report http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison.

resources—and especially *parallel* resources—are very rare. For lesser-used languages such as Lithuanian, Latvian, Estonian, Hungarian, Maltese and Czech, there exist much less corpora. Parallel corpora involving these languages together with English are occasionally found, but parallel corpora for pairs of lesser-used languages (e.g. Estonian-Maltese, or Latvian-Greek) are normally non-existent. EU corpora contribute to somewhat balance our English-centric field of study. Adding parallel data for lesser-used languages and language pairs is probably the biggest contribution of these EU corpora to the field. Koehn et al. (2009) have shown that building a Machine Translation (MT) system based only on EU texts can already produce quite acceptable results.

The corpus of EU documents not only talks generically about various issues, but it occasionally includes some highly specific vocabulary. One example is the *Integrated Tariff of the European Communities* (TARIC),[14] which consists of a very long multilingual list of products and their product codes that need to be used when declaring the movement of goods across borders, thus including fine-grained distinctions between product types (e.g. technology; types of fish, fruit, clothing, paper, etc.). Another example is the list of dual-use goods that are being monitored for the purpose of nuclear non-proliferation.[15]

Another useful feature of several of our corpora (see Table 1) is the fact that the documents have been manually classified according to the EuroVoc Thesaurus.[16] EuroVoc is a wide-coverage thesaurus with over six thousand classes that covers the interests of the European Institutions, and thus a wide range of fields including trade and finance, industry, agriculture, health, nuclear science, fishery, social rights, religion and the working environment, to name just a few. This not only has a positive impact on the wide range of vocabulary occurring in these texts, but the subject domain codes for each document also allow filtering the documents, e.g. to extract the specialist terminology for each of the fields.

4 What are the limitations of these EU resources

EU parallel resources also have a number of restrictions. One of them is the fact that they cover almost only EU languages (small amounts of text in Norwegian, Icelandic, Croatian and Turkish are included, see Table 1), thus excluding many of the world's biggest languages. Another restriction is linked to the text domain, which for the bulk of the corpus is legal and administrative. ECDC-TM deals with public health and EAC-TM is concerned with education and culture, but these translation memories are small compared to the other data. With DCEP, for the first time, larger numbers of press releases have become available, which is a useful extension to the dominant administrative sub-language.

It goes without saying that legislation covers almost all parts of human life and thus includes science, finance, social and personal matters, agriculture, transport, as

[14] See http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2003:103:0001:0031:EN:PDF.

[15] See http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:134:0001:0269:en:PDF.

[16] See http://eurovoc.europa.eu/.

well as any other part of society that needs regulating, meaning that vocabulary of all these fields will of course be included in the corpora. However, verbs will occur almost exclusively in third person, and the register of the language and the choice of words are rather different from other text types, such as news, fiction, personal letters and social media texts. Even the syntax and the lexical choice of the EU corpora is likely to be different from that found in other text types: Legal texts have strict requirements regarding their form, and vocabulary often has the status of terminology, which needs to be used consistently in order to avoid misunderstandings, while the requirement in fiction would typically be to use a varied lexicon and colourful phrases.

While the EU corpora are useful for the automatic generation of multilingual dictionaries and for the training of at least a basic statistical machine translation system (Koehn et al. 2009), a generic and better MT system would require more varied training data. Genre studies would be restricted to the few text types included. Named entity recognition work would be biased by the entities found in these text types (less persons and locations, more state organisations). Sentiment analysis or opinion mining work would probably not make much sense on EU corpora. Finally, summarisation software for legal and administrative documents would face entirely different challenges compared to those for news clusters or other less structured document types.

These and more restrictions clearly apply to the EU corpora, making them just one component among many linguistic resources needed to develop an encompassing set of text analysis software applications. They do of course also have their very own uses not shared by other text types, such as legal text mining and knowledge representation, or studies on translation consistency. To summarise, one can say that the released EU corpora are a very good stepping stone that allows developing a first basic version of text mining tools for a range of languages and language pairs for which otherwise no or very little resources would be available, but the EU's resources need to be complemented by many others.

5 What can the EU's parallel corpus data be used for (usage examples)

One area that crucially depends on parallel data is the creation of models for Statistical Machine Translation (SMT). The initial work on SMT made use of proceedings of Canadian parliament debates (Hansard) available in English and French. Since 2001, SMT work funded by DARPA focused on translation from Chinese and Arabic into English, for which models were trained using large parallel corpora such as those made up of United Nations publications (Eisele and Chen 2010). The EuroParl corpus with its originally 11 languages allowed the creation of SMT systems for up to 110 language pairs (Koehn 2005) and provided a crucial precondition for work in projects like *EuroMatrix* and *EuroMatrix Plus*.[17] The publication of the JRC-Acquis in 2006 enabled the creation of SMT systems for 462 European language pairs (Koehn et al. 2009). Parallel corpora involving more than

[17] See http://www.euromatrix.net/ and http://www.euromatrixplus.net/.

two languages have also been used to improve MT results by exploiting triangulation, either through the unions of multiple translation correspondences (e.g. Cohn and Lapata 2007) or through their intersections (e.g. Chen et al. 2009).

The value of not only multi-monolingual, but *parallel* resources (corpora, dictionaries, tools) cannot be estimated high enough because it makes the effort of developing, training and testing multilingual text mining tools more efficient and comparable (Steinberger 2011). Various scientific events have therefore focused on building or exploiting parallel corpora.[18] Apart from for SMT, parallel collections of sentences have been used concretely for the following tasks, and probably more:

- Producing multilingual lexical and semantic resources such as dictionaries and ontologies;
- Training and testing information extraction software;
- Annotation projection across languages for Named Entity Recognition (Ehrmann et al. 2011), sentiment analysis (Steinberger et al. 2011a, b), multi-document summarisation (Turchi et al. 2010), semantic role labelling (Padó and Lapata 2009), part-of-speech annotation, word sense disambiguation, and more (Yarowsky et al. 2001); Annotation projection allows saving annotation time and it creates more comparable resources for many languages;
- Improving monolingual text analysis by exploiting patterns in various other languages: Naseem et al. (2009) report massive improvements in unsupervised part-of-speech tagging by analysing parallel texts in up to eight languages;
- Automatic creation of parallel tree banks (e.g. Zhechev and Way 2008);
- Cross-lingual word sense disambiguation (e.g. Lefever and Hoste 2010);
- Cross-lingual textual entailment (e.g. Mehdad et al. 2010);
- Cross-lingual plagiarism detection (Potthast et al. 2011);
- Checking translation consistency automatically (e.g. Tufiş 2004);
- Multilingual and cross-lingual clustering and classification (e.g. Wei et al. 2008);
- Creation of multilingual semantic space, e.g. using Lexical Semantic Analysis (Landauer and Littman 1991) or Kernel Canonical Correlation Analysis (Vinokourov et al. 2003), for the purpose of cross-lingual information retrieval, multilingual clustering and classification, or any other cross-lingual purposes.

When the full text (i.e. information on the ordering of the sentences in the document) is available, further uses are possible:

[18] Events dedicated to building and exploiting parallel corpora are, for instance, the workshop series on 'Annotation and exploitation of parallel corpora' (e.g. http://www.bultreebank.org/AEPC2/); 'Slavic parallel corpora' (http://www.slavistik.uni-mainz.de/606.php); 'Parallel corpora and linguistic theory' (http://paralleltext.info/sle2013/); 'Annotation and Alignment of parallel corpora for linguistic research' (http://www.dagstuhl.de/13043); 'ATA-AMTA Workshop on users and uses for parallel corpora' (http://permalink.gmane.org/gmane.science.linguistics.corpora/11156); and 'Workshop on building and using parallel texts: data-driven machine translation and beyond' (http://www.statmt.org/wpt05/). The CLEF Initiative and its evaluation labs are also highly relevant for this field (http://www.clef-initiative.eu/).

- Annotation projection for co-reference and discourse structure;
- Translation studies and comparative language studies;
- Making use of full-text information to improve SMT;
- Testing and benchmarking alignment software (for sentences, words, etc.).

To summarise, multilingual parallel corpora are useful (a) as monolingual corpora for each of the languages involved; (b) as resources to project linguistic annotation from one highly resourced language to various others and to thus produce multilingual comparable annotated corpora while saving annotation time; (c) to improve monolingual text analysis tools by exploiting features and semantic distinctions found in other languages; (d) to produce bilingual or even multilingual cross-language resources such as dictionaries and multilingual vector space representations or to produce applications such as MT systems, cross-lingual word sense disambiguation, cross-lingual information retrieval, etc.

6 Communalities and differences regarding each of the seven linguistic EU resources

In this section, we present the seven linguistic EU resources from the points of view language coverage, source language of the translations, translation quality, usage of document identifiers, subject domain categorisation and alignment granularity. In Sect. 7, we will then provide further detail on each corpus resource. Sections 6 and 7 are complemented by Table 1, which presents contrastive corpus features in tabular form. Together, they provide background information on each of the multilingual linguistic resources, which will help the reader get a better understanding of their usefulness and purpose.

As the various corpora have been produced by different organisations and for different purposes, their format and the information available for each of them are unfortunately not consistent. Asking for more and for more structured information was not an option because several of the corpus providers have no personal interest in releasing this data. They rather made it available as a favour when they were told how useful this information could be to the R&D community.

All the corpus resources discussed here (and more) can be downloaded via the JRC's Language Technology Resources webpage (See Footnote 1).

Up to June 2013, there were 23 official **EU languages**. Croatian (ISO-Code: HR) was added on 1 July 2013 as a 24th language.[19] As the corpora described in this article were produced before that date, there are no Croatian documents in the corpora described here, with the exception of the EAC-Translation Memory, which also includes other non-EU languages. Croatian EU documents will become available with future releases (e.g. as part of the yearly updates of DGT-TM). In order to join the EU, Croatia needed to adopt the existing body of EU law (the *Acquis Communautaire*), so

[19] The 24 official EU languages as of January 2014 are Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish. We use the two-digit ISO codes to represent the languages.

at least the currently valid EU legislation already exists in Croatian translation. The University of Zagreb has sentence-aligned this document collection as part of the CESAR project[20] with the equivalent English documents, following the specifications and the format of the JRC-Acquis resource. This alignment is available for download via the META-Share servers.[21] For most of the EU's corpus resources, no documents are furthermore available in the Irish language (Gaeilge, GA). Irish became an official EU language in 2007, but the EU Institutions are currently exempt from the obligation to draft all acts in Irish.[22] The number of Irish documents was thus often too small to be included. In Table 1, the numbers 23 and 22 (languages) respectively refer to either all official EU languages except Croatian, or to all official EU languages except Croatian and Irish. Note that many more languages are spoken in the EU without being official EU languages, including regional languages such as Catalan, Basque and Ladin, minority languages such as Romani and Russian and immigrant languages such as Arabic and Chinese. There are even two national languages that are not official EU languages: Luxembourgish (Luxembourg) and Turkish (Cyprus).

The **source language** for most documents produced by the EU institutions is no longer known. This information is not part of the explicit meta-information available for the documents. However, at least for the DGT-Acquis corpus, the source language may be inferable by comparing the document dates of the various language versions in the Formex-4 format[23] (XML). Generally speaking, it is known that, in 2008, 72 % of all EC documents were drafted in English and 11 % in French. It is likely that at least some documents were translated via an intermediate language, i.e. that there are translations of translations. The source language for ECDC-TM and EAC-TM is known to be English.

The **translation quality** of EU documents is typically very good, especially for legal documents, because translations are carried out by highly trained professional translators. Documents are checked from both a legal and a language point of view during a multi-step revision process: the quality is controlled at the level of each translation service, by legal services and by the EC's Publications Office (PO). There is much focus on translation consistency, which includes involvement of the public administrations of the EU Member States. EU translators work in a cutting-edge IT environment, with many custom-built enhancements aimed at streamlining the work and ensuring quality and consistency. While it is in the nature of translation that its quality is always arguable, it can be assumed that the translation quality is on average of a very high standard.

Many EU documents are labelled using the **CELEX document number,** which is a unique document identifier containing also information on the document type and the publication year.[24] This document number can be used to separate the corpora into more homogeneous sub-corpora, such as treaties, international

[20] See http://www.cesar-project.net/.

[21] The META-Share download page is http://meta-share.ffzg.hr/repository/browse/croatian-translations-of-acquis/547866326c1811e28a985ef2e4e6c59e6758e8d15e7a445e9471e185a758b50c/.

[22] http://publications.europa.eu/code/pdf/370000en.htm#fn4-2.

[23] See http://formex.publications.europa.eu/formex-4/formex-4.htm.

[24] The structure of CELEX document numbers is explained at http://eur-lex.europa.eu/en/tools/faq.htm#1.12.

agreements, preparatory acts, case law, national implementing measures, parliamentary questions, etc. This CELEX number is only available for the JRC-Acquis and the JEX data. For Translation Memories, it is not useful to keep information on the originating document because many translation units are found in more than one document. The European Parliament uses its own document categories and DCEP is categorised according to these, but they are not compatible with the European Commission's document categories.

Most EU documents have been manually multi-label categorised according to the wide-coverage **EuroVoc Thesaurus**, which is a hierarchically organised list of over six thousand subject domains reflecting the interests of the European Union and its Member States (See Footnote 16). Knowing the EuroVoc descriptor codes of documents allows, for example, compiling subject domain-specific sub-corpora, e.g. to derive subject domain-specific vocabulary or to train document classifiers (Steinberger et al. 2012a). At the moment, only JRC-Acquis and the JEX data are categorised according to EuroVoc. In the future, EuroVoc information may also become available for the DGT-Acquis corpus.

Translation Memories consist of individual '**Translation Units**' (TU) and their translations. TUs are typically full sentences, but they can also be headings or full paragraphs. The latter is due to the fact that many legal texts are written in long sentences, sometimes spanning over several paragraphs. In such cases, each paragraph (and thus less than a sentence) is one TU. Similarly, in the 'sentence'-aligned full-text corpus JRC-Acquis, it is best to talk about TUs or of alignment units because not all aligned units are sentences. JRC-Acquis and DGT-TM were produced by automatically splitting larger documents into sentences (and the same will be the case for DCEP). The Translation Memories ECDC-TM and EAC-TM were created by translating sentence by sentence, meaning that both the sentence borders and the cross-lingual alignment are very reliable. All TMs distributed via the JRC's webpages are in the same format, i.e. in TMX, which is an XML standard for TMs.

The **alignment** granularity of the various corpora varies between document level alignment (i.e. we only know that a full document is the translation of another) and the alignment of *translation units*, which often are sentences (see the previous paragraph): The JEX corpus data is only aligned at document level as it was compiled for the purpose of document classification. The DCEP data is currently also only aligned at document level, but efforts are underway to align translation units using the HunAlign sentence alignment tool (Varga et al. 2005; see Sect. 7.3). The JRC-Acquis was sentence-aligned using two different freely available aligners so that users can choose the one they prefer (HunAlign and Vanilla; see Sect. 7.1). Developers may also want to use the two alignments as training or test data when developing their own aligners. DGT-Acquis was aligned using an in-house aligner. By definition, the three Translation Memories are aligned at translation unit level, but while DGT-TM was aligned automatically using an in-house system, ECDC-TM and EAC-TM were produced directly when translating the English source language texts into the other languages. Automatic alignments obviously leave more space for alignment errors, but at least the DGT-TM alignments were thoroughly tested by human translators, who use the alignment output for their manual or

interactive translation efforts. JRC-Acquis has been aligned separately for each language pair (the same plan exists for DCEP), while DGT-Acquis and DGT-TM alignments were performed for language pairs involving English and the alignments of other language pairs were inferred with English as a pivot language. For selected language pairs, third-party developers have produced word alignments for some of the resources.[25] It is furthermore possible to produce word or word-n-gram alignments for any language pair on the basis of the aligned sentences by using tools such as *Giza++* (Och and Ney 2003) or *Anymalign* (Lardilleux and Lepage 2009). The latter allows to align any number of languages simultaneously.

7 Further details about each of the linguistic resources

Table 1 shows the various features of the seven corpora in comparison, including their size, the languages and language pairs included, details on method and granularity of the alignment, formatting information, subject domains covered, text types included, information on updates, who created and who translated the corpora (expected translation quality), who processed the data to convert it into a machine-readable format, and more. In this section, we provide some additional background information on each of the multilingual linguistic resources, such as special features and information on why and how they were produced. This section thus contains information that might be useful for a better understanding of each of the datasets and that could not be described by addressing the general issues discussed in Sect. 6. By definition, Sect. 7 is thus rather heterogeneous.

7.1 JRC-Acquis

JRC-Acquis (Steinberger et al. 2006) was the first of the sentence-aligned and pre-processed corpora distributed by the European Commission. In its latest version (version 3.0), it comprised 22 languages, i.e. all of nowadays' 24 official EU languages except Irish and Croatian. The internet portal Eur-Lex (then still called CELEX), which contains the full text of many EU documents in HTML or PDF format, had already been freely accessible, but the data was not prepared for computational linguistics usage, i.e. the documents were not explicitly aligned, the text was not split into sentences, no sentence or alignment information was available, etc. EuroParl (Koehn 2005) had already been released in 2005 by Philipp Koehn from Edinburgh University, containing verbatim reports of the European Parliament's plenary sessions. Like DGT-Acquis and DCEP, JRC-Acquis contains the full text of the documents, which allows additional uses of the data compared to the translation memory collections, where the context of the sentences is not accessible (see Sect. 5 for details).

The JRC downloaded many documents from the Eur-Lex website, combined them with other EU documents already available in-house, selected all documents

[25] At http://pelcra.pl/res/parallel/word-aligned/, for instance, Polish-English word alignments can be found.

that were available in at least ten languages (of which at least three had to be from the countries that accessed the EU in 2004 or later), cleaned them, sentence-aligned them for all possible language pair combinations using the aligners *Vanilla*[26] and HunAlign (Varga et al. 2005), and made the corpus available in TEI-compliant XML format.[27] The cleaning and pre-processing of the 22-language corpus was necessary to convert the various data formats into UTF8-encoded XML format; verify the language of the documents (and discard those that were not in the expected language); split the text into numbered paragraph chunks; identify and label the signature at the end of the document (consisting of place, date, names and lists of addresses); and identify and label the annex of the document (half the documents contain annexes, typically consisting of lists of names, goods, addresses, etc.; annexes are not always attached to each language version of the document).

Due to the purely language-based criteria used to select the documents of the JRC-Acquis, its document types are mixed, including all or many sub-types of the documents available on Eur-Lex. Eur-Lex contains all documents that are part of the Official Journal (OJ) of the European Union.[28]

The JRC-Acquis is the only of the EU corpora that was sentence-aligned using two different types of alignment software. This allows comparing alignment performance and users can choose the results they prefer. To reduce the storage size, cross-lingual alignments are stored as meta-information and a provided software tool allows users to produce bilingual aligned corpora for any language pair.

7.2 DGT-Acquis

The DGT-Acquis is a family of four multilingual parallel corpora in up to 23 languages, i.e. all official EU languages except Croatian. The DGT-Acquis was produced by the European Commission's (EC) *Directorate General for Translation* (DGT). It includes all the series of the *Official Journal of the European Union* (OJ; see Footnote 28), i.e. the series L, LM, C, CA and CE. The collection contains documents from May 2004 to December 2011 in the XML format Formex-4 (see Footnote 23). Older OJ documents would have been harder to process because they only exist as OCR files or, since about 1994, in SGML format. The intention is to update this data collection with documents from the years 2012 and 2013 and to use proper in-house XML processing techniques rather than the string processing used for the first data release.

DGT-Acquis could be considered an update of the JRC-Acquis in that the motivation was to provide the R&D community with a large aligned full-text parallel corpus that contains several document types. However, DGT-Acquis also differs a lot from JRC-Acquis: (a) It was built in a more systematic way (selection

[26] The Vanilla software used implements the Gale and Church (1993) alignment algorithm.

[27] While the Vanilla alignment was performed at the JRC, the separate HunAlign alignment was carried out by the *Media Research Centre* at *Budapest University of Technology and Economics*. The Romanian documents were collected and pre-processed by the *Research Institute for Artificial Intelligence* at the *Romanian Academy of Sciences*.

[28] See http://publications.europa.eu/official/index_en.htm for more information on the Official Journal.

of all documents of all years since 2004 in all OJ series); (b) the data was not processed (selected, cleaned, aligned, etc.) at the JRC, but by DGT and the external firm Prompsit[29]; (c) the full-text documents were paragraph-aligned using in-house software rather than being sentence-aligned using publicly accessible software tools; (d) the same data is available in four packages with different levels of alignment (original data; file level alignment in Formex-4; file level alignment in plain text; and paragraph level alignment in plain text), allowing the users to access the data with the most appropriate processing level for their own needs and to re-process the data; and (e) the data is encoded in a very different container format called the *Multilingual Dataset Format*[30] (*muset*; Carrasco-Benitez 2008).

The motivation to use the muset format was to bridge the gap between large quantities of multilingual parallel data (big data) and the general movement towards Linked Open Data, which makes the automatic usage of data easier. How to handle *Big Multilingual Linked Data* is in the middle of evolving and muset would evolve accordingly.

The original data (Formex-4 in XML) is also included to invite the community at large to improve the results. The fact that both the TIFF image files of each document and the Formex-4 XML version of the document are available opens up an entirely new usage, i.e. that of training or testing optical character recognition (OCR) software, or the like.

7.3 DCEP (Digital Corpus of the European Parliament)

The *Digital Corpus of the European Parliament* (DCEP; Hajlaoui et al. 2014) is the latest EU corpus. At the time of writing, it is about to be released. It covers all official EU languages except Croatian. It has been prepared by the European Parliament's (EP) *Directorate General for Translation* and it contains the majority of the documents published on the European Parliament's official website.[31] EP's Directorate General for Translation has created and made publicly accessible this corpus to contribute to the European Parliament's policy of multilingualism, designed to ensure the equal treatment of languages. To avoid overlapping with the *EuroParl* corpus (Koehn 2005), DCEP does not contain the verbatim reports of the European Parliament's plenary sessions (CRE documents).

DCEP contains a variety of document types, including the following (in brackets: Number of English words for each category, out of a total of 103 Mio English words): reports (29 Mio), adopted texts (19 Mio), written answers to questions (15 Mio), written questions (12 Mio), national or EU-wide press releases (12 Mio), motions (7 Mio) and minutes of plenary meetings (3 Mio). As the parliamentary decision process involves proposing ideas, discussing them and voting about them (with each step being documented), there will be groups of topically related documents with different formulations and levels of detail, depending on the stage of the decision process.

[29] See http://www.prompsit.com/.

[30] See http://dragoman.org/muset/ for details.

[31] See http://www.europarl.europa.eu/.

Most DCEP documents are available in several languages, but this first version of DCEP is not an entirely parallel corpus and some documents exist only in one language version. DCEP is currently aligned at document level and it does not yet offer ready-made sentence alignment. However, work is underway to sentence-align the DCEP for all language pairs, using the HunAlign tool, and to also distribute this alignment data.

The corpus comes in two versions: the *source* directory contains the corpus in its original format (SGML or XML), while the *strip* directory contains the same document in plain text format, i.e. without SGML and XML tags. The corpus is further subdivided by language and then by document type. An *index* folder contains one file that links the different language versions of the same document, allowing users to compile bilingual or multilingual corpora. Like DGT-Acquis, DCEP offers a rather wide range of document types (and thus writing styles), compared to the more legislation-centred DGT-TM corpus and to the JEX data. Future versions of DCEP will also contain documents which are currently only available in MS-Word or in PDF format.

7.4 DGT-TM

The first version of DGT-TM was released in the year 2007, including EU documents up to the year 2006. There have been three updates since (releases 2011, 2012 and 2013) and it is planned to release new data every year. The data up to the year 2013 includes 23 languages (all official EU languages except Croatian), but the 23rd language, Irish, is as usual much under-represented. The next release of DGT-TM (release 2014, including the data for the year 2013) is expected to include 30,000 Irish and about 200,000 Croatian Translation Units (TUs). While the alignment of TUs in the first version of DGT-TM were manually produced or verified, the later versions were produced by automatically sentence-aligning full-text documents, using DGT's in-house alignment software *Euramis*. The TM is used by DGT's (human) translators, who give feedback in case they encounter wrong alignments, so that the alignment quality is good. DGT also uses the TM (and other parallel sources) to train their own in-house statistical machine translation (SMT) system MT@EC, which is based on Moses (Koehn et al. 2007) and which—as of January 2014—covers 24 languages and 552 language pairs (of which 58 direct). Since DGT-TM has been released, it became a resource that is used a lot by human translators. For that same reason, the number of downloads of DGT-TM is higher than for any of the other EU resources. DGT-TM was built exclusively on the basis of legislative documents (L-Series of the OJ), meaning that it is equivalent to part of the Acquis Communautaire (the body of EU law). The choice of using the OJ's L-Series to produce TMs is motivated by the fact that the L-Series are considered to be most useful for EU translators. The C-Series of the OJ Journal may be added in the future.

While processing the OJ data, DGT performed a number of changes to the original sentences. These include omitting TUs that are of low value for the translators (short sentences, long sentences, obvious mismatches, etc.); delete

sentence enumerators; re-insert diacritics where these had been replaced by transcriptions, etc. Details can be found on the download page for DGT-TM.[32] For these reasons, the DGT-TM TUs may not exactly match the equivalent sentences in the full-text documents that might be included in the full-text corpora JRC-Acquis and DGT-Acquis. The documents were aligned in accordance with the segmentation rules used at DGT. The extraction did keep the Eur-Lex document number, from which other information (e.g. year and document type) can be derived, but as many repeating TUs were omitted while compiling the corpus, the full set of sentences for each document cannot be reconstructed.

DGT-TM is accompanied by software that allows producing bilingual TMs for any language pair of choice by directly accessing the downloadable zip files. Bilingual alignments for language pairs not involving English are produced by going via the pivot language English.

For a detailed description of DGT-TM, see Steinberger et al. (2012b).

7.5 ECDC-TM

The Translation Memory ECDC-TM was provided by the *European Centre for Disease Prevention and Control* (ECDC), which is an EU agency in Stockholm focusing on public health issues.[33] The TM was produced when translating the organisation's English web pages into the 24 languages of the *European Economic Area* (EEA), which includes the EU Member States plus three countries of the *European Free Trade Area* (EFTA), i.e. Iceland, Liechtenstein and Norway). Croatia was not yet part of the EU when the translation memory was produced. ECDC-TM is much smaller than the previously mentioned resources, but it has the advantage that it covers a rather different subject domain.

ECDC-TM is accompanied by software that allows producing bilingual TMs for any language pair of choice, via the pivot language English. ECDC-TM was released with the motivation in mind that the data might help software providers produce better machine translation tools, which will in turn benefit the readers of the ECDC's website.

7.6 EAC-TM

The Translation Memory EAC-TM was provided by the EC *Directorate General for Education and Culture*[34] (EAC). It was created from translation files used for translating electronic forms such as project or funding applications and report forms for decentralised actions of two EU programmes: EAC's *Life-long Learning Programme* (LLP) and the *Youth in Action Programme*. The contents in the electronic forms are technically split into two types: (a) the labels and contents of electronic forms (referred to as 'Forms' Data) and (b) checkboxes and drop-down contents (referred to as 'Reference Data'). Due to the different types of data, the two

[32] http://ipsc.jrc.ec.europa.eu/index.php?id=197.

[33] For details on ECDC, see http://www.ecdc.europa.eu.

[34] For details on DG EAC, see http://ec.europa.eu/dgs/education_culture/.

collections are kept separate. For example, labels can be 'Country', 'Please specify your home country' etc., while examples for reference data are 'Germany', 'Basic/general programmes', 'Education and Culture' etc. EAC-TM is much smaller than most other resources discussed here, but it has the advantage that it covers a rather different subject domain, namely that of education, training, culture, youth and sports. Furthermore, it covers more languages than any other corpus described in this article: in addition to the usual 22 languages, it includes documents in Croatian, Icelandic, Norwegian and Turkish. EAC-TM exists in these languages because the respective countries are eligible for participation in the EAC programmes. EAC-TM should be enriched with new data every year as the tools that are being translated are evolving and also the number of tools translated keeps increasing. It is hoped that it will be possible to prepare this new data to release an update of the EAC Translation Memories.

EAC-TM is accompanied by software that allows producing bilingual TMs for any language pair of choice, via the pivot language English. EAC did not initially have the intention of releasing their Translation Memories publicly, but—once asked by the JRC—they fully supported the idea of making their data available because it was expected to be in the public interest.

7.7 Date accompanying the JRC EuroVoc Indexer (JEX)

JEX is software that automatically multi-label-classifies documents according to the categories of the EuroVoc Thesaurus (Steinberger et al. 2012a). This software performs profile-based category ranking by first learning the profiles and by comparing the profiles to the new document to which EuroVoc categories need to be assigned.

We describe JEX in this article because the software release includes tens of thousands of parallel documents in 22 languages that were used to train the categorisation software for each of these languages. These documents are exclusively of CELEX type 3 (secondary legislation), i.e. they are the legal documents that get published in the L-Series of the EC's Official Journal (OJ). Most JEX documents will thus also be included in JRC-Acquis, DGT-Acquis and/or DGT-TM. The major advantage of this corpus over the other ones mentioned in this article is that it consists of a homogeneous set of manually EuroVoc-categorised documents that can be used to produce automatic EuroVoc multi-label classification software for thousands of classes trained on a parallel document collection. Furthermore, the corpus is accompanied by software against which the results of other systems can be compared (see also Sects. 6 and 10). This data is very suitable to test or train multilingual and cross-lingual clustering and categorisation software. The plain text JEX documents are aligned at document level, but not at sentence level.

8 Overlap between the resources

Several of the mentioned resources overlap, which creates a problem for users who want to make use of the entire collection of parallel corpora made available via the

JRC's website. This section gives an indication of which corpora are unique, which ones are potentially overlapping, and how users can possibly identify the overlapping documents or TUs.

The resources ECDC-TM, EAC-TM and DCEP should each be made up of completely unique document sets so that they can simply be merged without any risk of overlap. Note, however, that individual sentences or possibly larger sections of documents may nevertheless be repeated verbatim (e.g. headers like "Article 1" or phrases such as "see the Footnote for details"). Having verbatim repetitions across different documents is normal and the reason of existence of translation memories is indeed to detect in a new document such previously translated sections in order to avoid translating them again manually. This helps saving the translators' time and effort. To give an idea of the amount of overlap that can be expected, we calculated some statistics on overlap regarding the DGT-TM releases dated 2007 and 2011, which do not have any documents in common: An analysis of the English-Danish sets of TUs in both collections revealed that slightly more than 3.5 % of the English-Danish sentence pairs are exact duplicates. There are even duplicate sentences within the 2011 release of DGT-TM. These are simply sentences or headers that occur repeatedly, across documents. While the majority of these repeated TUs have been excluded when creating this TM, others (especially from the earlier years) are included: There are 2,275 TUs that are identical across all 22 languages, 13,345 that are identical for ten or more languages, and 330,158 TUs that are identical for any language pair. It was a conscious decision by DGT that at least some of the repeated TUs should be part of the release in order to give some indication of the frequency of these TUs. For that purpose, repetitions were collected up to a certain point, after which only new TUs were added.

Whether or not these repeated text fragments should be used when training SMT systems or when using the corpora for other purposes depends on a design decision of the system developers: repetitive segments are redundant and, for instance, no new vocabulary can be learnt from them; on the other hand, repeated sections may help to strengthen certain word combinations in the language models and give a bigger weight to word co-occurrences and word sequences that are used more frequently.

The overlap between JRC-Acquis, DGT-Acquis, DGT-TM and the JEX corpus are of a different nature. Each of these corpora was collected by different persons and for a different purpose. Apart from the repetitions of identical *text segments* across different documents, discussed in the previous paragraphs, these four corpora partially include entire identical *documents*. The best way to exclude repeated documents is to use the document type, the document creation year, as well as the CELEX document identifiers. These document identifiers are unique for each document and they provide information on the document type and its year of creation. Documents created in different years, even for similar corpora such as JRC-Acquis, DGT-Acquis or the JEX corpus, are non-identical documents. Table 1 provides information on the document types used and on the years covered by each collection. Unfortunately, the current version of the DGT-Acquis does not contain information on the CELEX document identifier. We hope that a future version will include this important piece of information.

Two additional details need to be mentioned: (1) As the document pre-processing differs between the collections, the same document or TU in different collections may be represented differently. (2) In DGT-TM, TUs are accompanied by the CELEX identifier of the document in which they have been found. However, when TUs were found in several documents, not all their CELEX numbers are listed.

9 Usage rights/licences

Several of the seven resource collections described here have different usage rights. It is important to stress that the only legally binding usage rights are those distributed with the corpora, that the section you are currently reading has no legal value, that it does not necessarily represent the views of the EU institutions and that in any case it omits many details. Instead, we summarise here—in our own words—our own understanding of what the main usage restrictions are and we provide some historical background information on the usage rights.

OJ documents have been available online since 1998,[35] but in the first years, users needed to purchase the documents in order to get access to them. Due to the insight that the public will benefit from freely accessible EU documents (see also Directive 2003/98/EC of the European Parliament and of the Council on the re-use of public sector information; Footnote 9), OJ documents later became freely accessible. However, when the JRC inquired with the Publications Office of the EU institutions whether it could release the data sets it had collected and it was using, it did not initially get permission. It was only in 2006 that this permission was explicitly granted, resulting in the release of the JRC-Acquis corpus. A major worry of EU lawyers was then that the documents might be used for *legal* purposes rather than for the development of computational linguistics applications. For that reason, the usage conditions of the JRC-Acquis (which was the first parallel corpus released by an EU institution) stated that (a) the European Communities consider the OJ (and more) to be in the public domain, that (b) users need to state prominently that only EU legislation printed in the paper edition of the OJ is deemed authentic and that (c) translations of this documentation should only be made on the basis of the authentic version printed in the OJ. It is thus obvious that any usage restrictions are *not* due to copyright reasons—which are the major reason for usage restrictions regarding most other corpora.

With its release in 2007, DGT-TM was the second multilingual parallel resource available via the JRC's web pages. Like the JRC-Acquis, DGT-TM could be freely downloaded, but its usage was initially limited to research use, while the current version (release 2012 of the DGT-TM) explicitly allows "all kinds of use which comply with the conditions laid down in the Commission Decision of 12 December 2011 on the re-use of Commission documents, published in Official Journal of the European Union L330 of 14 December 2011, pages 39 to 42" (See Footnote 10). For further details on this Commission Decision, see also Sect. 2. The conditions also say that the data users are under an obligation to state the source of the documents used and that the EC retains ownership of the data. An important feature

[35] See http://eur-lex.europa.eu/en/tools/faq.htm#1.2.

of the conditions is also the limitation of liability of the data providers: the data and the accompanying software (to extract bilingual TMs from the zip files) are made available without any guarantee, e.g. regarding the accuracy of the data and regarding potential consequences of errors in the software.

The more recently published linguistic resources EAC-TM, ECDC-TM and DGT-Acquis all make reference—directly or indirectly—to the same 2011 re-use policy document as DGT-TM.

The usage conditions for DCEP, to be released via the JRC by the European Parliament, have not yet been finalised to date, but they are likely to be along the lines of the Commission Decision on the re-use of Commission documents, as well.

Several of the parallel text corpora are accompanied by software that allows users to quickly and easily extract parallel data for the language pairs of interest. The usage conditions for these pieces of software are mostly concerned with stating that it cannot be guaranteed that the software works flawlessly, and with limiting the liability of the software providers regarding potential damages that may occur when the software is used. In the case of the JRC EuroVoc Indexer software JEX, the software is the major part of the release while the accompanying text documents (which are probably mostly included also in the other resources) are mostly there to allow testing and re-training the software. For that purpose, a separate end-user licence agreement (EULA) was formulated for JEX.

To summarise our own understanding of the usage conditions: While the creators of many artistic and commercial data collections (such as prose, news texts, dictionaries, etc.) are to a large extent concerned with the creative or commercial aspects of copyright, the EU organisations mostly try to limit their liability and to avoid that the text corpora are used as the basis of legal decisions. Their usage by translators or to train SMT systems is encouraged, be it for commercial or for R&D purposes. However, users strictly have to comply to the usage conditions distributed with the individual resources they are using and any doubts must be clarified with the respective contact point for the resources.

10 Further EU resources besides parallel corpora

The release of the first large-scale parallel corpora EuroParl (Koehn 2005) and JRC-Acquis (Steinberger et al. 2006) became possible because the EU institutions decided to make their document collections freely accessible on the Eur-Lex[36] web site so that they could be harvested. Prior to that, the same EU documents had to be purchased. As we observed in Sect. 2, this opening up was a consequence of the generic insight that public data has a variety of uses and that the general public would benefit if these resources were available and also used by research and development organisations. The same insight led to the release of the database of EU terminology IATE,[37] the multilingual wide-coverage thesaurus EuroVoc,[38] plus

[36] See http://eur-lex.europa.eu/.

[37] See http://iate.europa.eu/.

[38] See http://eurovoc.europa.eu/.

other resources for translators[39] (EC&DGT 2008). **Eur-Lex** provides free access to European Union law and other documents considered to be public, written in all official EU languages. The **IATE** website (*Inter-Active Terminology for Europe*) gives access to a database of EU inter-institutional terminology. IATE has been used in the EU institutions and agencies since 2004 for the collection, dissemination and shared management of EU-specific terminology. **EuroVoc** is a multilingual thesaurus originally built specifically for the manual indexing and retrieval of multilingual documentary information of the EU institutions, but it is now much more widely used, e.g. by the libraries of many national governments in the EU. It is a multi-disciplinary thesaurus covering fields that are sufficiently wide-ranging to encompass both Community and national points of view, with a certain emphasis on parliamentary activities. EuroVoc is a controlled set of vocabulary which can also be used outside the EU institutions, particularly by parliaments.

JRC has publicly released its *JRC EuroVoc indexer* software **JEX** (Steinberger et al. 2012a), which multi-label classifies documents according to the multilingual EuroVoc thesaurus and thus allows establishing links between documents written in different languages. As many EU documents have been classified according to EuroVoc, the EU corpora can be used to train JEX or other multi-label classification software, or they can be used to generate subject domain-specific dictionaries.

An entirely different resource produced by the JRC is **JRC-Names** (Steinberger et al. 2011a, b). As a by-product of the large-scale news analysis in over 20 languages since 2004, combined with targeted Wikipedia mining, large collections of named entities and their many spelling variants (including across languages and writing scripts) were released to support R&D organisations in improving search and retrieval, but also to develop named entity recognition and other software for a wide range of languages. JRC-Names was produced by applying Named Entity Recognition (NER) to huge collections of multilingual news articles and by deciding automatically for each newly identified name whether it was a new name or a spelling variant of a previously known name. Variants such as *Witali Klitschko*, *Vitali Klitsjko* and *Виталий Кличко* are recognised as belonging to the same named entity[40] by transliterating names into the Roman script (where applicable), by applying empirically derived spelling normalisation rules and by then using string distance metrics, resulting in up to hundreds of different spelling variants for the same entity. For details, see Steinberger et al. (2011a, b). JRC-Names is automatically updated daily to include recently identified names and name variants. As of February 2014, it includes 582,000 names and name variants in 27 scripts. A 2012 snapshot of JRC-Names has also been prepared in Linked Open Data (LOD) format.[41] It is planned to complete and extend the LOD representation of JRC-Names.

As the same group within the JRC works on many other text analysis applications, often covering over twenty languages, a **number of further smaller**

[39] See http://ec.europa.eu/dgs/translation/publications/.

[40] i.e. the Ukrainian boxer and politician, see http://emm.newsexplorer.eu/NewsExplorer/entities/en/19011.html.

[41] For download and more information, see http://datahub.io/dataset/jrc-names.

resources have been produced and could be released. These include a set of sentiment-annotated quotations (Balahur et al. 2010) and a set of multilingual document clusters annotated for multi-document summarisation purposes (Turchi et al. 2010). More resources will become available via the JRC's Language Technology Resources site (see Footnote 1).

Following the same philosophy that EU public data can be reused for the benefit of the population, EU organisations now make many more types of data available via the ***European Union Open Data Portal***, where information on trade, transport, waste, employment, telecommunication, health, geography and much more can be downloaded, including as linked data.[42] The beta version of the portal was publicly launched in April 2013 and the non-beta version came live in December 2013. The portal's front page states: "Data are free to use, reuse, link and redistribute for commercial or non-commercial purposes", showing that the open data policy of the EU institutions has now been realised to its full extent.

11 Summary

Already in 2003, EU legislators turned their insight regarding the usefulness of EU public data (including raw language data) into a directive that encouraged the re-use of public sector information (see Footnote 9). They recognised that such data can be used to support the development of automatic tools to analyse language and that such tools may lead to higher transparency because citizens would get better access to information (e.g. legislation), including across languages (see Sect. 2). Since the release of the JRC-Acquis parallel corpus in 2006, six more EU corpora have been made available.

The major advantages of the EU corpora are the number of languages (there are currently 24 official EU languages), their parallelism (which is particularly important for lesser-used languages and language pairs), and the fact that they speak about almost any subject of human life that needs regulating (wide vocabulary). However, at the same time, these EU document collections also have limited use because—with few exceptions—they only cover EU languages, and because of the restricted language register of EU documents (mostly legal and administrative).

These parallel EU resources can be used for many purposes, including the development, training and testing of a wide range of Language Technology applications. Due to their parallelism, they can be used to develop multilingual tools quicker, e.g. by making use of annotation projection, and they can be used to develop cross-lingual natural language processing applications such as Machine Translation and more. Exploiting the composite document identifiers used for most EU documents, corpus users can separate out different text types or publication years. The subject domain classification of the documents allows selecting sub-corpora covering different semantic fields and thus identifying specialised

[42] See https://open-data.europa.eu/. Quote extracted on 7 February 2014.

vocabulary. The uniform presentation of the corpora allows saving pre-processing efforts and makes them easy to use.

Each of the seven corpora discussed in this article has its own features. The corpora differ regarding the covered subject domain (e.g. public health, or education and culture), their size (some with up to over one billion words), and the homogeneity (some are very homogeneous and allow exact comparison, while others are more heterogeneous and offer more different language registers). The Translation Memories serve not only computational linguists, but are intensely used by a large number of human translators world-wide.

As the corpora were put together by different organisations and for different purposes, they partially overlap, but with the help of the unique document identifiers, the information on the text type and the creation year of the documents, it should be possible to at least reduce the number of overlapping documents.

In the course of the years, the usage conditions continued to become more relaxed, so that the latest resources can be used for almost any purpose.

Besides parallel corpora, the EU has made accessible a number of other multilingual resources, including terminology databases, resources for translators, thesauri, multilingual lists of spelling variants for entities, and multi-label classification software. Several of the resources that already exist will be updated regularly by adding the latest data and thus expanding the resources continuously.

12 List of abbreviations

Acquis	Acquis Communautaire (EU body of common rights and obligations)
Bio	Billion
BUTE	Budapest University of Technology and Economics
CELEX	Previous name of Eur-Lex; name of document identifiers
DARPA	Defense Advanced Research Projects Agency
DCEP	Digital Corpus of the European Parliament
DGT	Directorate General for Translation of the EC
DG-TRAD	EP Directorate General for Translation
EAC	Directorate General for Education and Culture of the EC
EC	European Commission
ECDC	European Centre for Disease Prevention and Control (EU Agency)
EMM	Europe Media Monitor
EP	European Parliament
EU	European Union
EULA	End-user license agreement
Eur-Lex	Service providing access to the legal texts of the EU
EuroVoc	Multilingual thesaurus maintained by the Publications office of the EU
HTML	HyperText Markup Language
IATE	InterActive Terminology for Europe
IT	Information Technology

JEX	JRC EuroVoc Indexer
JRC	Joint Research Centre (Directorate General of the European Commission)
K	Thousand
LOD	Linked Open Data
META	Multilingual Europe Meta Alliance
Mio	Million
MT	Machine Translation
MULTEXT	Multilingual text tools and corpora
Muset	Multilingual Dataset Format
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
OJ	Official Journal of the European Union
OJ C-Series	OJ series including reports, minutes, statements and announcements including the judgments of the European Court of Justice and the General Court, but also calls for expressions of interest, public contracts, and more (see Footnote 28)
OJ L-Series	OJ series 'Legislation', including regulations, directives, decisions, recommendations and opinions
PDF	Portable Document Format
PO	Publications Office
R&D	Research & Development
RAS	Romanian Academy of Sciences
SMT	Statistical Machine Translation
TARIC	Integrated Tariff of the European Communities
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format
TM	Translation Memory
TMX	Translation Memory eXchange (an XML format)
TU	Translation Unit
URL	Uniform Resource Locator (web address)
UTF8	UCS Transformation Format-8-bit (character encoding)
XML	Extensible Markup Language

References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., et al. (2010). Sentiment analysis in the news. In *Proceedings of the 7th international conference on language resources and evaluation (LREC'2010)*, Valletta, Malta, 19–21 May 2010, pp. 2216–2220.

Carrasco-Benitez, M. T. (2008). Open architecture for multilingual parallel texts. http://arxiv.org/ftp/arxiv/papers/0808/0808.3889.pdf.

Chen, Y., Kay, M., & Eisele A. (2009). Intersecting multilingual data for faster and better statistical translations. In *Proceedings of human language technologies: The 2009 annual conference of the*

North American chapter of the association for computational linguistics, Boulder, Colorado, pp. 128–136.

Chiao, Y.-C., Kraif, O., Laurent, D., Nguyen, T. M. H., Semmar, N., Stuck, F., et al. (2006). Evaluation of multilingual text alignment systems: The ARCADE II project. In *Proceedings of the 5th international conference on language resources and evaluation (LREC'2006)*, Genoa, Italy, pp. 1975–1978.

Cohn T., & Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th annual meeting of the association for computational linguistics*, Prague, Czech Republic, pp. 728–735.

EC&DGT (2008). European Commission & Directorate General for Translation—Translation tools and workflow. Office for Official Publications, Brussels, Belgium.

Ehrmann, M., Turchi, M., & Steinberger, R. (2011). Building a multilingual named entity-annotated corpus. In *Proceedings of the 8th international conference recent advances in natural language processing (RANLP'2011)*, Hissar, Bulgaria.

Eisele A., & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the international conference on language resources and evaluation (LREC 2010)*, Valletta, Malta, pp. 2868–2872.

Erjavec, T. (2010). MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*, Valletta, Malta, pp. 2544–2547.

Erjavec T., & Ide, N. (1998). The MULTEXT-East corpus. In *Proceedings of the first international conference on language resources and evaluation (LREC)*, Granada, Spain.

Gale, W., & Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19*(1), 75–102.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Varga, D., & Steinberger, R. (2014). DCEP—Digital Corpus of the European Parliament. In *Proceedings of the 9th edition of its language resources and evaluation conference*, Reykjavik, Iceland.

Ide N., & Véronis, J. (1994). MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 15th international conference on computational linguistics (CoLing)*, Kyoto, Japan, pp. 588–592.

Koehn, P. (2005). EuroParl: A parallel corpus for statistical machine translation. In *Proceedings of the machine translation summit*, Phuket, Thailand, pp. 79–86.

Koehn, P., Birch, A., & Steinberger, R. (2009). 462 Machine Translation Systems for Europe. In L. Gerber, P. Isabelle, R. Kuhn, N. Bemish, M. Dillinger, & M.-J. Goulet (Eds.), *Proceedings of the twelfth machine translation summit (MT-Summit XII)*, Ottawa, Canada, August 2009, pp. 65–72.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*, demonstration session, Prague, Czech Republic, June 2007.

Lardilleux, A., & Lepage, Y. (2009). Sampling-based multilingual alignment. In *International conference on recent advances in natural language processing (RANLP'2009)*, Borovets, Bulgaria, pp. 214–218.

Lefever, E., & Hoste, V. (2010). SemEval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval'2010)*, Uppsala, Sweden, pp. 15–20.

Landauer T., & Littman, M. (1991). A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the 11th international conference 'Expert Systems and Their Applications'*, Vol. 8, pp. 77–85.

Mehdad Y., Negri, M., & Federico, M. (2010). Towards cross-lingual textual entailment. In *Proceedings of human language technologies*, Los Angeles, CA, USA, pp. 321–324.

Naseem, T., Snyder, B., Eisenstein, J., & Barzilay, R. (2009). Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence, 36*, 341–385.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research, 36*, 307–340.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. In *Language resources and evaluation. Special issue on plagiarism and authorship analysis*, Vol. 45, no. 1, pp. 45–62.

Resnik, P., Olsen, M. B., & Diab, M. (1999). The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities, 33*(1–2), 129–153.

Steinberger, R. (2011). A survey of methods to ease the development of highly multilingual Text Mining applications. *Language Resources and Evaluation Journal, 46*(2), 155–176.

Steinberger, R. (2013). Multilingual and cross-lingual news analysis in the Europe Media Monitor (EMM). In M. Lupu, E. Kanoulas, & F. Loizides (Eds.), *Multidisciplinary information retrieval. 6th information retrieval facility conference (IRFC'2013), Limassol, Cyprus*. Springer Lecture Notes in Computer Science, Vol. 8201, pp. 1–4.

Steinberger, R., Ebrahim, M., & Turchi, M. (2012a). JRC EuroVoc Indexer JEX—A freely available multi-label categorisation tool. In *Proceedings of the 8th international conference on language resources and evaluation (LREC'2012)*, Istanbul, 21–27 May 2012.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012b). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on language resources and evaluation (LREC'2012)*, Istanbul, 21–27 May 2012, pp. 454–459.

Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R., & van der Goot, E. (2011a). Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the 8th international conference recent advances in natural language processing (RANLP'2011)*, Hissar, Bulgaria, 12–14 September 2011.

Steinberger, R., Pouliquen, B., Kabadjov, M., & van der Goot, E. (2011b). JRC-Names: A freely available, highly multilingual named entity resource. In *Proceedings of the 8th international conference recent advances in natural language processing (RANLP'2011)*, Hissar, Bulgaria, 12–14 September 2011, pp. 104–110.

Steinberger, R., Pouliquen, B., & van der Goot, E. (2009). An Introduction to the Europe media monitor family of applications. In F. Gey, N. Kando, & J. Karlgren (Eds.), *Information access in a multilingual world—proceedings of the SIGIR 2009 workshop (SIGIR-CLIR'2009)*, Boston, USA, 23 July 2009, pp. 1–8.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20 + languages. In *Proceedings of the 5th international conference on language resources and evaluation (LREC'2006)*, Genoa, Italy, 24–26 May 2006, pp. 2142–2147.

Tiedemann, J. (2009). News from OPUS—A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing* (Vol. V, pp. 237–248). John Benjamins, Amsterdam/Philadelphia.

Tiedemann, J., & Nygaard, L. (2004). The OPUS corpus—Parallel and free. In *Proceedings of the 4th international conference on language resources an evaluation (LREC)*, Lisbon, Portugal, pp. 1183–1186.

Tufiş, D. (2004). Term translations in parallel corpora: Discovery and consistency check. In *Proceedings of the 4th international conference on language resources an evaluation (LREC)*, Lisbon, Portugal, pp. 1981–1984.

Turchi, M., Steinberger, J., Kabadjov, M. & Steinberger, R. (2010). Using parallel corpora for multilingual (multi-document) summarisation evaluation. Multilingual and multimodal information access evaluation. Springer Lecture Notes for Computer Science, LNCS 6360/2010, pp. 52–63.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, Borovets, Bulgaria, pp. 590–596.

Vinokourov, A., Shawe-Taylor, J., & Cristianini, N. (2003). Inferring a semantic representation of text via cross-language correlation analysis: Advances in neural information processing systems 15. In S. Becker, S. Thrun & K. Obermayer (Eds), (pp. 1473–1480). Cambridge, MA: MIT Press.

Wei, C.-P., Yang, C. C., & Lin, C.-M. (2008). A Latent Semantic Indexing-based approach to multilingual document clustering. *Decision Support Systems, 45*(2008), 606–620.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001) Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT'01*, San Diego.

Zhechev, V., & Way, A. (2008). Automatic generation of parallel treebanks. In *Proceedings of the 22nd international conference on computational linguistics (CoLing'2008)*, Manchester, UK, Vol. 1, pp. 1105–1112.