Title	Overview of Parallel Corpora
Version	2.5
Author(s)	DF, JL
Date	21-11-2017
Status	For distribution
Distribution	NCF
ID	CE-2017-1095

Table of contents

1. Background and approach

In the following survey, our aim is to provide an overview of parallel corpora that contain languages relevant for all the countries that are members or observers of CLARIN ERIC. Our motivation was to identify to what extent these resources exist and are easily available, and check which information about the resources is available, thereby highlighting the aspects in which accessibility of these corpora as well as the presentation of the relevant information can be optimised from a User Involvement perspective.

The corpora were identified in five steps:

(i) through the VLO,
(ii) on the repositories or websites of the national consortia,
(iii) through META-SHARE,
(iv) through the LRE Map, and
(v) through the national UI coordinators

Since we have been able to identify a large number of corpora, we have split the overview into two major subsections. Section (2) focuses on corpora that are already part of the CLARIN infrastructure in that they are either provided by the national consortia or can be found through the VLO. By contrast, section (3) focuses on corpora that are not yet part of the infrastructure and cannot be found through the VLO. Each subsection first presents each corpus in terms of its most relevant information and then provides a recap of the identification, availability and the status of the metadata for all the corpora.

2. Parallel corpora in the CLARIN infrastructure

Table (1) shows that 81 corpora are part of the CLARIN infrastructure. Each corpus is described in terms of the following information:

(i) the languages of the compiled data,
(ii) the directionality of the translations,
(iii) the sources of the data,
(iv) the size of the corpus,
(v) annotation with a focus on alignment, and
(vi) the availability and license of the corpus.

Subsections (2.1.)-(2.4.) provide an overview of these points. The hyperlinks were last accessed 15 November 2017.

Table 1: Overview of corpora in the CLARIN infrastructure

Corpus name	Description
The CLUVI parallel corpus 23 million tokens annotation unknown downloadable	This is a multilingual corpus that contains the following six language combinations: *English-Galician; Galician-Spanish; French-Galician; English-Galician-French-Spanish; Spanish-Catalan-Basque* from the period between 2003 and 2012 and the following domains: fiction, computing, popular science, law, and administration. The corpus contains 23 million tokens. It is unclear how the corpus is annotated. A 6-million-token subcorpus can be downloaded through the corpus webpage. The data are available under the CC-BY NC-SA 3.0 license.
The English-Slovak Parallel corpus size unknown Automatic morphological annotation downloadable	This is a bilingual English-Slovak corpus, with mixed/unclear directionality. The texts are from the legal domain (Acquis), Europarl, the Official Journal of the European Union and some data are taken from the OPUS corpus. The size and period of the data in the corpus are unknown. The corpus is automatically morphologically annotated. It can be downloaded here through the LINDAT repository and is available under the CC-BY NC-SA 3.0 license.
The French-Croatian Parallel corpus 263 million tokens sentence-aligned unavailable	This is a bilingual French-Croatian corpus of fictional texts with mixed/unclear directionality. The corpus consists of 263 million tokens from various periods. The corpus is aligned at the sentence level. Though the description in the VLO says the corpus is available, the relevant website that is linked to appears to be broken. It is also unclear under which license the corpus is available.
The Catalan-Spanish Parallel Corpus 100 million tokens sentence-aligned unavailable	This is a bilingual corpus of Catalan newspaper articles and their Spanish translations. The corpus consists of 100 million tokens from unknown periods. The corpus is aligned at the sentence level. The corpus is unavailable for download or through a concordancer and the license is unclear.
The Croatian-Slovenian Parallel Corpus size unknown annotation unknown unavailable	This is a bilingual Croatian-Slovenian corpus with unclear directionality. All the relevant metadata are unclear – the types of data compiled, the size, the periods covered, and the annotation. Though the description in the VLO says the corpus is available, the relevant website that is linked to appears to be broken. It is also unclear under which license the corpus is available.
The English-Nepali Parallel Corpus 1.2 million tokens partially sentence-aligned unavailable	This is a bilingual English-Nepali corpus texts on national development with unclear directionality. The corpus consists of 1.2 million tokens from an unknown period. A subset of the corpus is aligned at the sentence level. The corpus is unavailable for download or through a concordancer and the license is unclear.

The Polish-Lithuanian Parallel Corpus size unknown annotation unknown downloadable	This is a bilingual Polish-Lithuanian corpus with unclear directionality. All the relevant metadata are unclear – the types of data compiled, the size, the periods covered, and the annotation. The corpus is available for download in the CLARIN-PL repository under the IS PAS corpora license.
CzEng 1.6 206.4 million tokens sentence-aligned available downloadable	This is a bilingual English-Czech corpus. The corpus is bidirectional, with original texts in English and Czech, and accompanying translations. It is unclear which domain and period the texts come from. The corpus consists of 206.4 million tokens. The corpus is sentenced aligned. The corpus is available for download on the corpus website under the CC-BY NC-SA 3.0 license. For the relevant publication, see Bojar et al. (2016).
Kacenka 3.3 million tokens annotation unknown downloadable	This is a bilingual corpus of English fictional texts from various periods and their Czech translations. The corpus consists of 3.3 million tokens. Though the documentation says the corpus is aligned, it is unclear at which levels. The corpus is unavailable and the license is unclear.
The KOTUS Finnish-Swedish Parallel Corpus 4.3 million tokens sentence-aligned downloadable	This is a bilingual corpus of Finnish corporate press releases, surveys, reports, laws and regulations, as well as Government proposals to Parliament from the period between 1993 and 2004 and their Swedish translations. The corpus consists of 4.3 million tokens. The corpus is sentence-aligned. It is available for download through the Finnish repository Kielipankki and through the concordancer *Korp* under the CC-BY license.
COMPARA size unknown sentence-aligned searchable on-line	This is a bilingual corpus of Portuguese texts from fiction, academic prose, journalism and tourism from an unknown period and their English translations. The size of the corpus is unknown, although it is growing. The corpus is sentence-aligned. It is available through a concordancer on the corpus website under the CC-BY license. For the relevant publication, see Frankenberg Garcia and Santos (2003).
The DPC – Dutch Parallel Corpus 10.8 million tokens sentence-aligned unavailable	This is a multilingual Dutch-English and Dutch-French corpus of fiction, journalism, instructive texts and administration from an unknown period. The corpus consists of 10.8 million tokens. The corpus is sentence-aligned. It is still under construction and not yet available. For the relevant publication, see Macken et al. (2007).
Parallel corpus newsletters IFT FR-GR size unknown	This is a bilingual corpus of French IFT newsletters and their Greek translations. Most of the relevant metadata are unclear – the size, the period and the annotation.

annotation unknown unavailable	It is still under construction and not yet available. The data will be available under the CC-BY license.
ACCURAT balanced test corpus for under resourced languages 4,608 sentences annotation unknown for download	This is a multilingual corpus of Greek, Slovenian, Romanian, Latvian, Estonian, Croatian and Lithuanian. It consists of 4,608 sentences. It is unclear how this corpus is aligned. It is available for download through the CLARIN:el repository under the CC-BY license.
UP/TAP 31,849 sentences sentence-aligned for download	This is a bilingual corpus of Greek and Portuguese from the travel domain. It consists of 31,849 sentences. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under the CC-BY license.
European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English 1.2 million sentences sentence-aligned for download	This is a bilingual corpus of Greek and English from the European Parliament from 1996 to 2011. The corpus consists of 1.2 million sentences. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under the CC-ZERO license.
EMEA Corpus size unknown sentence aligned for download	This is a multilingual corpus of data from roughly 20 languages taken from the European Medicines Agency document. The size of the corpus is unclear. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution) license.
ECDC Translation Memory 320,000 tokens sentence-aligned for download	This is a multilingual corpus of data from roughly 20 languages taken from the public health domain. There are 320,000 tokens in the corpus. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution) license.
DGT-Translation Memory 10.1 million tokens unclear alignment for download	This is a multilingual corpus of data from roughly 20 languages taken from the European Legislation. The corpus consists of 10.1 million tokens. The corpus is aligned, though it is unclear how. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution) license.
DGT-Acquis size unknown sentence-aligned for download	This is a multilingual corpus of data from 23 languages taken from the Official Journal of the European Union from the period between 2004 and 2011. The size in terms of tokens or sentences is unclear. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution)

	license.
EAC Translation Memory 320,000 tokens sentence aligned for download	This is a multilingual corpus of data from more than 50 languages taken from the law, education and culture domains. The corpus consists of 320,000 tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution) license.
A parallel corpus collected from the European Constitution 3 million tokens sentence aligned for download	This is a multilingual corpus of data from 21 languages taken from European Constitution documents. The corpus consists of 3.01 million tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution) license.
A parallel corpus of KDE4 localization files (v.2) 60 million tokens sentence aligned for download	This is a multilingual corpus of data from 92 languages taken from KDE4 localization files. The corpus consists of 60 million tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under the CC-BY license.
European Central Bank parallel corpus 757 million tokens sentence aligned for download	This is a multilingual corpus of data from 19 languages taken from the website and documentation of the European Central Bank. The corpus consists of 757 million tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under Open For Reuse With Restrictions (Attribution) license.
OpenSubtitles2011 8.31G tokens sentence and word aligned for download	This is a multilingual corpus of data from 54 languages taken from the OpenSubtitles website. The corpus consists of 8.31G tokens The corpus is sentence and word aligned. It is available for download through the CLARIN:el repository under the Open For Reuse With Restrictions (Attribution) license.
SPC - Stockholm Parallel Corpora 1.32 million tokens sentence aligned for download	This is a multilingual corpus of data from English, Afrikaans, Chinese and Greek from the law domain. The corpus consists of 1.32 million tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under the Open For Reuse With Restrictions (Attribution) license.
Tatoeba 12 million tokens sentence aligned for download	This is a multilingual corpus of 117 languages—it is a collection of sentenced from Tatoeba. The corpus consists of 12 million tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under CC-BY license.
DGT-TM-2016 373 million tokens sentence aligned for download	This is a multilingual corpus of data from around 30 languages taken from the European Legislation. The corpus consists of approx. 373 million tokens. The corpus is sentence aligned. It is available for download through the

	CLARIN:el repository under the Open For Reuse With Restrictions (Attribution) license.
QTLP English-Greek Corpus for the MEDICAL domain 62,452 sentence pairs sentence aligned unavailable	This is a bilingual corpus of English and Greek with data taken from the medical domain. The corpus consists of 62,452 pairs of sentences. The corpus is sentence aligned. Though the corpus is listed in the CLARIN:el repository, it isn't available for download due to the MS-NC-NoReD license.
QTLP German-Greek Corpus for the MEDICAL domain 2,752 pairs of sentences sentence aligned unavailable	This is a bilingual corpus of German and Greek with data taken from the medical domain. The corpus consists of 2,752 pairs of sentences. The corpus is sentence aligned. Though the corpus is listed in the CLARIN:el repository, it isn't available for download due to the MS-NC-NoReD license.
QTLP Portuguese-Greek Corpus for the MEDICAL domain 62,608 sentence pairs sentence aligned unavailable	This is a bilingual corpus of Portuguese and Greek with data taken from the medical domain. The corpus consists of 62,608 pairs of sentences. The corpus is sentence aligned. Though the corpus is listed in the CLARIN:el repository, it isn't available for download due to the MS-NC-NoReD license.
QTLP English-Greek Corpus for the AUTOMOTIVE domain 2,946 sentence pairs sentence aligned unavailable	This is a bilingual corpus of English and Greek with data taken from the automotive domain. The corpus consists of 2,946 pairs of sentences. The corpus is sentence aligned. Though the corpus is listed in the CLARIN:el repository, it isn't available for download due to the MS-NC-NoReD license.
QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain 59,297 sentence pairs sentence aligned unavailable	This is a bilingual corpus of Portuguese and Greek with data taken from the automotive domain. The corpus consists of 59,297 pairs of sentences. The corpus is sentence aligned. Though the corpus is listed in the CLARIN:el repository, it isn't available for download due to the MS-NC-NoReD license.
Text Corpus - EMEL 43,000 tokens annotation unclear for download	This is a bilingual corpus of English and French data taken from NLP conference papers. The corpus consists of 43,000 tokens. It is unclear how the corpus is aligned. It is available for download through the CLARIN:el repository under the CC-BY license.
FREL 701,401 tokens annotation unclear unavailable	This is a bilingual corpus of French and Greek with data taken from literature. The corpus consists of 701,401 tokens. It is unclear how the corpus is annotated. The corpus is unavailable (though listed in the CLARIN:el repository) as it is still under negotiation.
Inlerlingual Perspectives 18 articles annotation unclear for download	This is a bilingual corpus of English and Greek with data taken from research articles. The corpus consists of 18 articles. It is unclear how the corpus is annotated. It is available for download through the CLARIN:el repository under the CC-BY license.

aformes 376,250 tokens annotation unclear for download	This is a bilingual corpus of English and Greek with data taken from a journal of undergraduate creative writing of the Faculty of English Language and Literature. The corpus consists of 376,250 tokens. It is unclear how the corpus is annotated. It is available for download through the CLARIN:el repository under the CC-BY license.
GLOSSOLOGIA size unknown annotation unclear for download	This is a multilingual corpus of French, Greek, English, and German data taken from a journal of general and historical Greek linguistics. The size of the corpus is unclear in terms of sentences and tokens. It is unclear how the corpus is annotated. It is available for download through the CLARIN:el repository under the CC-BY license.
Civitas Gentium size unknown annotation unclear for download	This is a multilingual corpus of English, Greek and French data from scientific papers, book re-views and opinions in the fields of the Geographic Analytic. The size of the corpus is unclear in terms of sentences or tokens. It is unclear how the corpus is annotated. It is available for download through the CLARIN:el repository under the CC-BY license.
Official Journal of the European Union size unknown partial paragraph and sentence alignment unavailable	This is a multilingual corpus of data from 23 languages taken from the Official Journal of the European Union. The size of the corpus (or rather, its 4 subcorpora) is unclear. Certain parts of the corpus are paragraph aligned, certain parts are sentence aligned. Though the corpus (or rather its 4 subcorpora) is listed in the CLARIN:el repository, the corpus can't be downloaded due to its "other" license.
INTERA Corpus - the Greek-English part 4 million tokens sentence aligned for download	This is a bilingual corpus of Greek and English data taken from the law, education, environment, tourism and health domains. It consists of 4 million tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under CC-BY license.
Greek-Bulgarian Bul-TM parallel corpus 10 million tokens sentence aligned for download	This is a bilingual corpus of Greek and Bulgarian data taken from domains related to society and politics. It consists of 10,000,000 tokens. The corpus is sentence aligned. It is available for download through the CLARIN:el repository under CC-BY license.
OPUS corpus size unknown sentence-aligned downloadable and searchable on-line	This is a multilingual corpus that compiles texts from all world languages. It consists of various subcorpora that compile data from literature, EU Translation Memories, political documents, open subtitles, UN documents, texts from the EUROPARL corpus, etc. The size of the corpus is unclear, though it is growing. The corpus is sentence-aligned. It is available for download and through a dedicated concordancer under the CC-BY license. For the relevant publication, see Tiedemann (2009).

Opus, Helsinki Korp Version 2.7 billion tokens sentence-aligned searchable on-line	This is a multilingual variant of the OPUS corpus that compiles data from the following sixteen languages: Czech, Danish, Dutch, English, Estonian, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Russian, Swedish, Spanish, and Turkish. It consists of 2.7 billion tokens from texts from various domains and unknown periods. The corpus is available through the concordancer *Korp* under the CC-BY license.
SzegedParalell: angol-magyar párhuzamos korpusz size unknown annotation unknown unavailable	This is an English-Hungarian corpus. The directionality of the translation is unclear. All the other metadata are unclear as well. The corpus is available for download.
Tourism English-Croatian Parallel Corpus 2.0 140,000 tokens annotation unknown downloadable	This is an English-Croatian corpus of texts from tourist websites from an unknown period. The directionality of the translation is unclear. The corpus consists of 140,000 tokens. It is unclear how the corpus is annotated. It is available for download through the CLARIN.SI repository under the CLARIN.SI User Licence for Internet Corpora.
LOGON parallel tourist corpus of Norwegian-English texts 500,000 tokens annotation unknown searchable on-line	This is a bilingual corpus of Norwegian tourist texts from an unknown period and their English translations. The corpus consists of approximately 500,000 tokens. It is unclear how the corpus is annotated. The corpus is available through a dedicated concordancer on the corpus webpage. The license under which it is available is unknown.
Serbian-English parallel corpus srenWaC 1.0 23.1 million tokens annotation unknown downloadable	This is a bilingual corpus of Serbian web texts from an unknown and their English translations. The corpus consists of 23.1 million tokens. It is unclear how the corpus is annotated. It is available for download through the CLARIN.SI repository under the CLARIN.SI User License for Internet Corpora.
Parallel Bible Corpus size unknown annotation unknown unavailable	This is a multilingual Bible corpus from many languages (possibly more than 100). It contains various contemporary and historic translations. The size of the corpus is unknown (save for the fact that it consists of “1169 unique translations”. Apart from Unicode normalization, it is unclear how the corpus is annotated. It is unavailable and the license is unclear.
Croatian-English parallel corpus hrenWaC 2.0 99,001 sentence pairs sentence-aligned	This is a bilingual Croatian-English corpus. Though the directionality of translations is not explicitly reported, the data are crawled from top-level Croatian .hr domains. The corpus consists of 99,001 sentence pairs.

downloadable	The corpus is sentence-aligned and available for download from the CLARIN.SI repository under the CLARIN.SI User License for Internet Corpora.
MULTEXT-East "1984" annotated corpus 4.0 1,064,424 tokens sentence-aligned downloadable	This is a multilingual corpus of George Orwell's *1984* original novel in English and its translations into 11 languages (Bulgarian, Czech, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak and Slovenian). The corpus consists of 1,064,424 tokens. The corpus is sentence-aligned. The corpus is available for download on the CLARIN.SI repository under the CC-BY license. For the relevant publication, see Erjavec (2012).
Slovene-English parallel corpus slenWaC 1.0 718,315 tokens sentenced-aligned downloadable	This is a bilingual Slovene-English corpus. Though the directionality of translations is not explicitly stated, the texts from an unknown period are crawled from top-level Slovenian .si domains. The corpus consists of 718,315 tokens. The corpus is sentenced-aligned and available for download from the CLARIN.SI repository under the CLARIN.SI User License for Internet Corpora.
Finnish-English parallel corpus fienWaC 1.0 2.9 million tokens sentenced-aligned downloadable	This is a bilingual Finnish-English corpus. Though the directionality of translations is not explicitly stated, the texts from an unknown period are crawled from top-level Finnish .fi domains. The corpus consists of 2.9 million tokens. The corpus is sentenced-aligned and available for download from the CLARIN.SI repository under the CLARIN.SI User License for Internet Corpora.
Amharic-English bilingual corpus 500 million tokens annotation unknown unavailable	This is a bilingual corpus compiling data for Amharic and English—it is unidirectional in that the original texts are in Amharic and the translations in English. The data compiled are from the legal and news domains in Amharic script. The corpus consists of approx. 500 million tokens from an unknown period. It is unknown how the corpus is annotated. The corpus is not publicly available due to the ELRA_VAR license.
CRATER 2 Corpus 4 million tokens morphosyntactically tagged unavailable	This is a multilingual English, French and Spanish corpus. The directionality of the translation is unclear. Furthermore, it is unclear which texts the corpus contains and from which period. The corpus consists of approx. 4, million tokens. The corpus is marked with morphosyntactic tagging. The corpus is not publicly available due to the ELRA_VAR license.
Croatian-English parallel corpus size unknown annotation unknown unavailable	This is a bilingual corpus of Croatian weekly newspapers from the period between 1998 and 2000 and their English translations. The size of the corpus is unknown. It is unclear how the corpus is annotated. Though it should be available for download, the link to the landing page seems to be broken.
Estonian-English parallel	This is a bilingual corpus of Estonian laws and their translations into English

corpus 307,000 sentences sentence-aligned downloadable	and EU legislation translated into Estonian. The corpus consists of 307,000 sentences. The corpus is sentence-aligned. The corpus is available for download through the corpus webpage under the CLARIN ACA licence.
Estonian-French parallel corpus 65 million tokens annotation unknown searchable on-line	This is a bilingual Estonian-French corpus. The directionality of the translation is unclear, as are the source and period of the texts. The corpus consists of 65 million tokens. It is not clear how the corpus is annotated. The corpus is available through a dedicated concordancer under the CLARIN ACA licence.
Europarl Parallel Corpus 650,000 tokens sentence-aligned downloadable	This is a multilingual corpus of 21 languages. It consists of proceedings from the European parliament from the period between 1996 and 2011. The corpus consists of 650,000 tokens. The corpus is sentence-aligned. It is available for download through the corpus webpage under the CC-ZERO license.
European Parliament Interpretation Corpus (EPIC) 177,295 tokens PoS tagged and lemmatised unavailable	This is a multilingual corpus of original texts and translations in Italian, English and Spanish in all possible combinations. The texts are taken from the European parliament from an unknown period. The corpus consists of 177,295 tokens. The data in the corpus are PoS-tagged and lemmatised. The corpus is publicly unavailable due to the ELRA_VAR license.
GeFRePaC - German French Reciprocal Parallel Corpus 30 million tokens sentence- and partially word-aligned unavailable	This is a bilingual corpus of German and French. The corpus is bidirectional. The texts are taken from the European Union CELEX Database from an unknown period. The corpus consists of 30 million tokens. The corpus is sentence-aligned as well as partially word aligned. The corpus is not publicly available due to the ELRA_VAR license.
JRC-Acquis Multilingual Parallel Corpus 1 billion tokens sentence-aligned downloadable	This is a multilingual corpus consisting of data from 22 languages. The texts are from the Acquis Communautaire from various periods beginning in the 1950s. The corpus consists of 1 billion tokens. The corpus is sentence-aligned and available for download from the webpage of the European Commission, although the license is not clear. For the relevant publication, see Steinberger et al. (2014).
MLCC Multilingual and Parallel Corpora 10.2 million tokens annotation unclear unavailable	This is a multilingual corpus of data from the following 9 languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish. The data are taken from the *Official Journal of the European Communities* from the period between 1986 and 1994. The corpus consists of 10.2 million tokens. It is not clear how the corpus is annotated. The corpus is not publicly available due to the ELRA_VAR license.

MULCOLD - Multilingual Corpus of Legal Documents 1.2 million tokens annotation unclear searchable on-line	This is a multilingual corpus of Russian, English, Swedish and Finnish. The directionality of the translations is unclear. The texts are from international conventions and treaties from unknown periods. There are 1.2 million tokens in the corpus. It is unclear how the corpus is annotated. The corpus is available through the concordancer *Korp* under the CC-BY-ND license.
MUSA Multilingual Multimodal Corpus 1.3 million tokens aligned with transcripts unavailable	This is a multilingual corpus of Greek, English and French. The directionality of the translations is unclear. The data are taken from movie subtitles from unknown periods. There are 1.3 million tokens in the corpus. The data are aligned with "transcripts and scripts". The corpus is not available and the license is unclear.
PELCRA Polish-English parallel corpora 262 million tokens annotation unclear searchable on-line	This is a bilingual corpus of Polish and English. The corpus is bidirectional from an unknown period and of unknown text types. It consists of 262 million tokens. The corpus is "manually aligned and annotated for equivalence types". It is available through a dedicated concordancer.
PANACEA English-French and English-Greek parallel corpus size unclear annotation unclear unavailable	This is a multilingual corpus of English texts taken from the "environment" and "labour legislation" domain and their French and Greek translations. It is unclear how many tokens there are in the corpus. It is unclear how the corpus is annotated. The corpus is not publicly available due to the ELRA_VAR license.
ParFin 359,494 tokens sentence searchable on-line	This corpus contains Finnish literary texts of 1990-2010 and their Russian translations. There are 5,360,000 tokens in the corpora. ParFin is aligned at the sentence level. The corpus is available through the concordancer *Korp* under the CLARIN_RES licence.
ParRus 5. million tokens paragraph searchable on-line	This corpus contains Russian classical and 20th century literature and their translations into Finnish. There are 5.9 million tokens in the corpora. ParRus is aligned the paragraph level. The corpus is available through the concordancer *Korp* under the CLARIN_RES licence.
Czech-Slovak Parallel Corpus size unclear morphologically annotated downloadable	This is a bilingual corpus of Czech and Slovak. The directionality of the translations is unclear. The texts are from the domain of law (Acquis), Europarl, the Official Journal of the European Union and some are taken from the OPUS corpus. The size of the corpus is unknown, as is the period of the texts. Apart from "automatic morphological annotation", it is unclear how the corpus is annotated. It is available for download through the LINDAT repository under the CC-BY license.
CsEnVi Pairwise Parallel Corpora 31 million tokens	This is a multilingual corpus of Vietnamese, Czech and English data. The directionality of the translations is unclear. The texts are from TED talks and subtitles from the CLUVI corpus from unknown periods. The corpus consists

annotation unclear downloadable	of 31 million tokens. It is unclear how the data are annotated. The corpus is available for download through the LINDAT repository under the CC-BY license.
EnTam: An English-Tamil Parallel Corpus (EnTam v2.0) 169,871 sentences sentence-aligned downloadable	This is a bilingual corpus of the English Bible and various texts from the news and cinema domains from unknown periods and their Tamil translations. The corpus consists of 169,871 sentences. The corpus is sentence-aligned. The corpus is available for download through the LINDAT repository under the CC-BY license.
HindEnCorp 0.5 132,300 sentences annotation unclear downloadable	This is a bilingual English-Hindi corpus of texts from TED talks, news articles, Wikipedia, etc. from unknown periods. The directionality of the translations is unclear. The corpus consists of 132,300 sentences. It is unclear how the data are annotated. The corpus is available for download through the LINDAT repository under the CC-BY license.
English-Hindi Parallel Corpus size unknown sentence-aligned downloadable	This is a bilingual English-Hindi corpus. The directionality of the translations is unclear, as are the sources of the data, the size of the corpus and the time periods of the data. The corpus is sentence-aligned and available for download through the LINDAT repository under the CC-BY license.
Czech-English Manual Word Alignment 112,765 tokens word-aligned downloadable	This is a bilingual corpus of English and Czech data. The directionality of the translations is unclear. The data are taken from e-books, *Reader's Digest*, the *Kačenka* magazine, Acquis Communautaire, the Project Syndicate and PCEDT from various periods. The corpus consists of 112,765 tokens. The corpus is aligned at the word level and is available for download through the LINDAT repository under the CC-BY license.
Czech and English abstracts of ÚFAL papers 200,000 tokens document-aligned downloadable	This is a bilingual corpus of English and Czech abstracts of ÚFAL papers. The corpus consists of approx. 200,000 tokens. The corpus is aligned at the document level and is available for download through the LINDAT repository under the CC-BY license.
English-Czech Corpus from Wikipedia 7.5 million tokens annotation unknown downloadable	This is a bilingual corpus of English Wikipedia articles from an unknown period and their Czech translations. The corpus consists of 7.5 million tokens. It is unclear how the corpus is annotated. The corpus is available for download through the LINDAT repository under the CC-BY license.
Europarl QTLeap WSD/NED corpus 52 million tokens WSD, NER, CR downloadable	This is a multilingual corpus for the following language pairs: Bulgarian-English, Czech-English, Portuguese-English, Spanish-English and Basque-English. The corpus is unidirectional in that the target language is English. The texts are taken from the European Parliament from an unknown period. The corpus consists of 52 million tokens. The data in the corpus is annotated as follows: (i) word sense

	disambiguation, (ii) named-entity disambiguation, (iii) coreference resolution, though the alignment is not clear. The corpus is available for download through the LINDAT repository under the CC-BY license.
UMC 0.1: Czech-Russian-English Multilingual Corpus 1.8 million tokens sentence-aligned downloadable	This is a multilingual corpus of Czech, Russian and English. The directionality of the translations is unknown. The texts are taken from the Project Syndicate website from the period between 1995 and 2008, which focuses on news articles and commentaries. The corpus consists of 1.8 million tokens. The corpus is sentence-aligned. The corpus is available for download through the LINDAT repository under the CC-BY license.
FTA/Eng and FTA/Spa 3 million tokens alignment unclear concordancer	This is a bidirectional bilingual corpus of English and Spanish texts related to the Free Trade Agreement. The corpus consists of roughly 3 million tokens. It is unclear how the corpus is annotated. The corpus is available through the concordancer *Corpuscule* under the CLARIN_ACA license.
The Norwegian-Spanish Parallel Corpus 6 million tokens alignment unclear concordancer and download	This is a unidirectional bilingual corpus of Norwegian and Spanish fictional and non-fictional texts from between 2000 and 2009 with Spanish being the target language. The corpus consists of roughly 6 million tokens. It is unclear how the corpus is annotated. The corpus is available through the concordancer *Corpuscule* and downloadable in the CLARINO repository under the CLARIN_ACA license

2.1. Identification

All of the corpora in table 1 can be found through the VLO except for the following 35:

(i) *Parallel corpus newsletters IFT FR-GR*, which is otherwise available in the CLARIN:el repository;
(ii) *ACCURAT balanced test corpus for under resourced languages*, which is otherwise available in the CLARIN:el repository;
(iii) *UP/TAP*, which is otherwise available in the CLARIN:el repository;
(iv) *European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English*, which is otherwise available in the CLARIN:el repository;
(v) *EMEA Corpus*, which is otherwise available in the CLARIN:el repository;
(vi) *ECDC Translation Memory,* which is otherwise available in the CLARIN:el repository;
(vii) *DGT-Translation Memory,* which is otherwise available in the CLARIN:el repository;
(viii) *DGT-Acquis,* which is otherwise available in the CLARIN:el repository;
(ix) *EAC Translation Memory,* which is otherwise available in the CLARIN:el repository;
(x) *A parallel corpus collected from the European Constitution,* which is otherwise available in the CLARIN:el repository;
(xi) *A parallel corpus of KDE4 localization files (v.2) ,* which is otherwise available in the CLARIN:el repository;

(xii) *European Central Bank parallel corpus,* which is otherwise available in the CLARIN:el repository;
(xiii) *OpenSubtitles2011,* which is otherwise available in the CLARIN:el repository;
(xiv) *SPC - Stockholm Parallel Corpora,* which is otherwise available in the CLARIN:el repository;
(xv) *Tatoeba,* which is otherwise available in the CLARIN:el repository;
(xvi) *DGT-TM-2016,* which is otherwise available in the CLARIN:el repository;
(xvii) *QTLP English-Greek Corpus for the MEDICAL domain,* which is otherwise available in the CLARIN:el repository;
(xviii) *QTLP German-Greek Corpus for the MEDICAL domain,* which is otherwise available in the CLARIN:el repository;
(xix) *QTLP Portuguese-Greek Corpus for the MEDICAL domain,* which is otherwise available in the CLARIN:el repository;
(xx) *QTLP English-Greek Corpus for the AUTOMOTIVE domain,* which is otherwise available in the CLARIN:el repository;
(xxi) *QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain,* which is otherwise available in the CLARIN:el repository;
(xxii) *Text Corpus – EMEL,* which is otherwise available in the CLARIN:el repository;
(xxiii) *FREL,* which is otherwise available in the CLARIN:el repository;
(xxiv) *Interlingual Perspectives,* which is otherwise available in the CLARIN:el repository;
(xxv) *aformes,* which is otherwise available in the CLARIN:el repository;
(xxvi) *GLOSSOLOGIA,* which is otherwise available in the CLARIN:el repository;
(xxvii) *Civitas Gentium,* which is otherwise available in the CLARIN:el repository;
(xxviii) *Official Journal of the European Union,* which is otherwise available in the CLARIN:el repository;
(xxix) *INTERA Corpus - the Greek-English part,* which is otherwise available in the CLARIN:el repository;
(xxx) *Greek-Bulgarian Bul-TM parallel corpus,* which is otherwise available in the CLARIN:el repository;
(xxxi) *Opus, Helsinki Korp Version*, which is otherwise available in the Finnish repository *Kielipankki*;
(xxxii) *SzegedParalell: angol-magyar párhuzamos korpusz*, which is referred to on the webpage of the Hungarian consortium;
(xxxiii) *LOGON parallel tourist corpus of Norwegian-English texts*, which is available through CLARINO;
(xxxiv) *PELCRA Polish-English parallel corpora*, which can be found on the webpage of the Polish consortium;
(xxxv) *FTA/Eng and FTA/Spa*, which is available on the Corpuscule website of CLARINO.

The OPUS corpus is an exception in that the VLO only finds four of its subcorpora, but not the entire corpus:

(i) Opus ECB Corpus;

(ii) Opus Subtitles Corpus;
(iii) Opus EU Corpus;
(iv) OPUS Localization Corpus.

Additionally, an out-of-date variant of *JRC-Acquis Multilingual Parallel Corpus* is listed under the VLO, which links to the LINDAT repository – there, the link to the landing page is broken.

2.2. Availability

2.2.1. Through concordancers

In terms of availability, 10 corpora are exclusively available through concordancers:

(i) *The KOTUS Finnish-Swedish Parallel Corpus*, which is available through *Korp*;
(ii) *Opus, Helsinki Korp Version*, which is also available through *Korp*;
(iii) *LOGON parallel tourist corpus of Norwegian-English texts*, which is available through a dedicated concordancer;
(iv) *MULCOLD - Multilingual Corpus of Legal Documents,* which is available through *Korp*;
(v) *COMPARA*, which is available through a dedicated concordancer;
(vi) *Estonian-French parallel corpus*, which is available through a dedicated concordancer;
(vii) *PELCRA Polish-English parallel corpora*, which is also available through a dedicated concordancer;
(viii) *ParFin*, which is available through Korp;
(ix) *ParRus,* which is available through Korp;
(x) *FTA/Eng* and *FTA/Spa*.

2.2.2. For download

In total, 45 corpora are available for download, 40 through CLARIN repositories and 6 on dedicated webpages.

The following 11 corpora are available through the LINDAT repository:

(i) *The English-Slovak Parallel corpus*;
(ii) *Czech-Slovak Parallel Corpus*;
(iii) *CsEnVi Pairwise Parallel Corpora*;
(iv) *Parallel Corpus (EnTam v2.0)*;
(v) *HindEnCorp 0.5*;
(vi) *English-Hindi Parallel Corpus*;
(vii) *Czech-English Manual Word Alignment*;
(viii) *Czech and English abstracts of ÚFAL papers*;
(ix) *English-Czech Corpus from Wikipedia*;
(x) *Europarl QTLeap WSD/NED corpus*;
(xi) *UMC 0.1: Czech-Russian-English Multilingual Corpus*.

The following 6 corpora are available through the CLARIN.SI repository:

(i) *Tourism English-Croatian Parallel Corpus 2.0*;
(ii) *Serbian-English parallel corpus srenWaC 1.0*;
(iii) *Croatian-English parallel corpus hrenWaC 2.0*;
(iv) *MULTEXT-East "1984" annotated corpus 4.0*;
(v) *Slovene-English parallel corpus slenWaC 1.0*;
(vi) *Finnish-English parallel corpus fienWaC 1.0*.

The following 22 corpora are available through the CLARIN:el repository:

(i) *ACCURAT balanced test corpus for under resourced languages;*
(ii) *European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English;*
(iii) *UP/TAP;*
(iv) *EMEA Corpus;*
(v) *ECDC Translation Memory;*
(vi) *DGT-Translation Memory;*
(vii) *DGT-Acquis;*
(viii) *EAC Translation Memory;*
(ix) *A parallel corpus collected from the European Constitution;*
(x) *A parallel corpus of KDE4 localization files (v.2);*
(xi) *European Central Bank parallel corpus;*
(xii) *OpenSubtitles2011;*
(xiii) *SPC - Stockholm Parallel Corpora;*
(xiv) *Tatoeba;*
(xv) *DGT-TM-2016;*
(xvi) *Text Corpus – EMEL;*
(xvii) *Interlingual Perspectives;*
(xviii) *aformes;*
(xix) *GLOSSOLOGIA;*
(xx) *Civitas Gentium;*
(xxi) *INTERA Corpus - the Greek-English part;*
(xxii) *Greek-Bulgarian Bul-TM parallel corpus*.

The Polish-Lithuanian Parallel Corpus is available for download through the CLARIN-PL repository.

The following 5 corpora are available for download on their own dedicated pages:

(i) *The CLUVI parallel corpus*;
(ii) *CzEng 1.6*;
(iii) *Estonian-English parallel corpus*;
(iv) *EuroParl Parallel Corpus*;
(v) *JRC-Acquis Multilingual Parallel Corpus*;

2.2.3. Through a concordancer and for download

The *OPUS* corpus, the *KOTUS Finnish-Swedish Parallel Corpus* and *The Norwegian-Spanish Parallel Corpus* are available both for download and through a concordancer.

2.2.4. Unavailable

The following 23 corpora are completely unavailable. It must be emphasised that, in the case of corpora (vi)-(vii), the links to their download pages are broken; it is possible then that the respective authors may want these corpora to be available.

(i) *The Catalan-Spanish Parallel Corpus*;
(ii) *The English-Nepali Parallel Corpus*;
(iii) *Parallel corpus newsletters IFT FR-GR* (still in preparation);
(iv) *European Parliament Interpretation Corpus (EPIC)*;
(v) *The Croatian-Slovenian Parallel Corpus*;
(vi) *Croatian-English parallel corpus;*
(vii) *The French-Croatian Parallel corpus;*
(viii) *QTLP German-Greek Corpus for the MEDICAL domain;*
(ix) *QTLP English-Greek Corpus for the MEDICAL domain*
(x) *QTLP Portuguese-Greek Corpus for the MEDICAL domain;*
(xi) *QTLP English-Greek Corpus for the AUTOMOTIVE domain;*
(xii) *QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain;*
(xiii) *FREL;*
(xiv) *Parallel Bible Corpus*
(xv) *Official Journal of the European Union;*
(xvi) *The DPC – Dutch Parallel Corpus*;
(xvii) *Amharic-English bilingual corpus*;
(xviii) *CRATER 2 Corpus*;
(xix) *GeFRePaC - German French Reciprocal Parallel Corpus*;
(xx) *MLCC Multilingual and Parallel Corpora*;
(xxi) *Kacenka*;
(xxii) *SzegedParalell: angol-magyar párhuzamos korpusz*.
(xxiii) *PANACEA English-French and English-Greek parallel corpus*.

2.3. Metadata

The metadata on the corpora are generally poor.

2.3.1. Languages

As to the languages that are represented, 32 of the corpora are multilingual. Except for the *Parallel Bible Corpus*, which represents most of the world's languages, the other corpora are focussed on European languages, with *EnTam: An English-Tamil Parallel Corpus (EnTam v2.0)*, English-Hindi Parallel Corpus,

HindEnCorp 0.5 and *CsEnVi Pairwise Parallel Corpora* being the exceptions, as they include Hindi, Tamil and Vietnamese as well.

(i) *Parallel Bible Corpus* – more than 150 languages;
(ii) *Tatoeba* – 117 languages;
(iii) *A parallel corpus of KDE4 localization files (v.2)* – 92 languages;
(iv) *OpenSubtitles2011* – 54 languages;
(v) *OPUS corpus* – more than 50 languages;
(vi) *EAC Translation Memory* – more than 50 languages;
(vii) *DGT-TM-2016* – around 30 languages;
(viii) *DGT-Acquis* – 23 languages;
(ix) *Official Journal of the European Union* – 23 languages;
(x) *JRC-Acquis Multilingual Parallel Corpus* – 22 languages;
(xi) *Europarl Parallel Corpus* – 21 languages;
(xii) *A parallel corpus collected from the European Constitution* – 21 languages;
(xiii) *EMEA Corpus* – roughly 20 languages;
(xiv) *ECDC Translation Memory* – roughly 20 languages;
(xv) *DGT-Translation Memory* – roughly 20 languages;
(xvi) *European Central Bank parallel corpus* – 19 languages;
(xvii) *Opus, Helsinki Korp Version* – 16 languages;
(xviii) *MULTEXT-East "1984" annotated corpus 4.0* – 12 languages;
(xix) *MLCC Multilingual and Parallel Corpora* – 9 languages;
(xx) *ACCURAT balanced test corpus for under resourced languages* – 7 languages;
(xxi) *The CLUVI parallel corpus* – 6 languages;
(xxii) Europarl QTLeap WSD/NED corpus – 6 languages;
(xxiii) MULCOLD - Multilingual Corpus of Legal Documents – 4 languages;
(xxiv) *SPC - Stockholm Parallel Corpora* – 4 languages;
(xxv) *GLOSSOLOGIA* – 4 languages;
(xxvi) *The DPC – Dutch Parallel Corpus* – 3 languages;
(xxvii) *CRATER 2 Corpus* – 3 languages;
(xxviii) *Civitas Gentium* – 3 languages;
(xxix) *European Parliament Interpretation Corpus (EPIC)* – 3 languages;
(xxx) *MUSA Multilingual Multimodal Corpus* – 3 languages;
(xxxi) *PANACEA English-French and English-Greek parallel corpus* – 3 languages;
(xxxii) *CsEnVi Pairwise Parallel Corpora* – 3 languages.

2.3.2. Size

The largest identified corpus in the parallel corpora family is likely *Opus*—the Helsinki Corp Version consists of approx. 2.7 billion tokens, whereas the smallest corpus is *Text Corpus - EMEL*, as it consists of approx. 43,000 tokens. The size is unknown for the following 16 corpora:

(i) *The English-Slovak Parallel corpus*;

(ii) *The Croatian-Slovenian Parallel Corpus*;
(iii) *The Polish-Lithuanian Parallel Corpus*;
(iv) *COMPARA;*
(v) *Parallel corpus newsletters IFT FR-GR;*
(vi) *OPUS corpus;*
(vii) *SzegedParalell: angol-magyar párhuzamos korpusz;*
(viii) *Parallel Bible Corpus*;
(ix) *Croatian-English parallel corpus*;
(x) *PANACEA English-French and English-Greek parallel corpus*;
(xi) *EMEA Corpus;*
(xii) *DGT-Acquis;*
(xiii) *GLOSSOLOGIA;*
(xiv) *Civitas Gentium;*
(xv) *Official Journal of the European Union;*
(xvi) *Czech-Slovak Parallel Corpus*.

2.3.3. Alignment

Explicit information regarding the level of alignment is available for the following 43 corpora, so a little less than half. All of the corpora listed below are aligned at the sentence level except for (xii) and (xxxiii), which are partially aligned at the word level as well, (xv), which is paragraph aligned, (xviii), which is word aligned, and (xix), which is aligned at the document level. It is unclear how corpora (xxi) and (xxviii) are aligned.

(i) *The French-Croatian Parallel corpus*;
(ii) *The Catalan-Spanish Parallel Corpus*;
(iii) *The English-Nepali Parallel Corpus;*
(iv) *The KOTUS Finnish-Swedish Parallel Corpus;*
(v) *COMPARA;*
(vi) *The DPC – Dutch Parallel Corpus;*
(vii) *OPUS corpus;*
(viii) *Croatian-English parallel corpus hrenWaC 2.0;*
(ix) *Finnish-English parallel corpus fienWaC 1.0;*
(x) *Estonian-English parallel corpus;*
(xi) *Europarl Parallel Corpus;*
(xii) GeFRePaC - German French Reciprocal Parallel Corpus;
(xiii) *JRC-Acquis Multilingual Parallel Corpus;*
(xiv) ParFin;
(xv) ParRus;
(xvi) *Parallel Corpus (EnTam v2.0);*
(xvii) *English-Hindi Parallel Corpus;*
(xviii) *Czech-English Manual Word Alignment;*
(xix) *Czech and English abstracts of ÚFAL papers;*

(xx) *UMC 0.1: Czech-Russian-English Multilingual Corpus;*
(xxi) Kacenka;
(xxii) *OPUS corpus;*
(xxiii) *Europarl Parallel Corpus;*
(xxiv) *UP/TAP;*
(xxv) *European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English;*
(xxvi) *EMEA Corpus;*
(xxvii) *ECDC Translation Memory;*
(xxviii) *DGT-Translation Memory;*
(xxix) *DGT-Acquis;*
(xxx) *EAC Translation Memory;*
(xxxi) *A parallel corpus collected from the European Constitution;*
(xxxii) *A parallel corpus of KDE4 localization files (v.2);*
(xxxiii) *European Central Bank parallel corpus;*
(xxxiv) *OpenSubtitles2011;*
(xxxv) *SPC - Stockholm Parallel Corpora;*
(xxxvi) *Tatoeba;*
(xxxvii) *DGT-TM-2016;*
(xxxviii) *QTLP English-Greek Corpus for the MEDICAL domain;*
(xxxix) *QTLP German-Greek Corpus for the MEDICAL domain;*
(xl) *QTLP Portuguese-Greek Corpus for the MEDICAL domain;*
(xli) *QTLP English-Greek Corpus for the AUTOMOTIVE domain;*
(xlii) *QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain;*
(xliii) *INTERA Corpus - the Greek-English part;*
(xliv) *Greek-Bulgarian Bul-TM parallel corpus.*

2.3.4. License

Most corpora (32) are available under the CC-BY license, 10 corpora (all of which in CLARIN:el) are licensed under the Open for Reuse with Restrictions license, 6 corpora are licensed under the restrictive ELRA_VAR license and are thus unavailable, 5 corpora are licensed under the MS-NC-NoReD, so are unavailable, 2 corpora are available under CC-ZERO, 5 corpora are available under CLARIN.SI User License, 4 corpora under the CLARIN_ACA license, 2 corpora are available under the CLARIN_RES licence, 1 corpus is available under an unknown “other” license, and 1 corpus under the IS PAS license. Information regarding the license is unavailable for 13 corpora.

3. Parallel corpora not in the CLARIN infrastructure

Table 2 lists 25 corpora that are not part of the CLARIN infrastructure. We again first provide a summary of each individual corpus and then an overview of the metadata.

Table 2: Overview of the corpora not in the CLARIN infrastructure

Corpus name	Description
The TRIS corpus 1,758,419 tokens sentence-aligned unavailable	This is a bilingual corpus of German original texts and their Spanish translations from the European Commission from the period between 1997 and 2010. The corpus consists of 1,758,419 tokens. The corpus is sentence-aligned. Though the corpus appears available for download, the link is broken. For the relevant publication, see Escartín (2012).
TED-Parallel-Corpus 300,000 sentences annotation unclear downloadable	This is a multilingual corpus of TED talks in English from unknown periods and translations into 11 languages – Arabic, Simplified Chinese, Traditional Chinese, Dutch, French, German, Hebrew, Italian, Japanese, Korean and Russian. The corpus consists of 300,000 sentences. It is unclear how the corpus is annotated. It is available for download; the license is unclear.
The United Nations Parallel Corpus 334,953,817 tokens annotation unclear downloadable	This is a multilingual corpus of official records and other parliamentary documents of the United Nations that are in the public domain. The data are in seven languages: English, Russian, Spanish, French, Chinese and Arabic. The corpus consists of 334,953,817 tokens. It is unclear how the corpus is annotated. The corpus is available for download on the corpus webpage. The license under which it is available is unknown. For the relevant publication, see Ziemski et al. (2016).
SPOOK corpus	This is a multilingual corpus of English-Slovene, German-Slovene, French-Slovene and Italian-Slovene; the corpus is unidirectional in that the target language is Slovene. The data are taken from literature, news items, and technical fields from unknown periods. The size of the corpus is unknown. The corpus us unavailable. The licence is unknown.
ParCor - A Parallel Pronoun-Coreference Corpus Size unknown pronoun coreference downloadable	This is a bilingual corpus of English TED talks and EU Bookshop publications from unknown periods and their German translations. The size and of the corpus is unknown. The corpus is manually-annotated for pronoun coreference and can be dowloaded from the OPUS webpage. For the relevant publication, see Guillou et al. (2014).
MultiUN: Multilingual UN Parallel Text 2000—2009 1 billion tokens sentence-aligned downloadable	This is a multilingual corpus of texts from the period between 2000 and 2009 taken from the United Nations website in the following language combinations: *Spanish-Chinese, Chinese-Spanish, French-Chinese, Chinese-French*. The corpus consists of approx. 1 billion tokens. The corpus is sentence-aligned. The corpus is available for download on the

	corpus webpage, although it is unclear under which license the data are available. For the relevant publication, see Eisele and Chen (2010).
Bulgarian-X language Parallel Corpus 1,202,209,147 tokens annotation unclear searchable on-line	This is a multilingual corpus of texts in 50 languages. We were unable to identify the domains from which the data were extracted. The corpus consists of 1,202,209,147 tokens. It is unclear how the corpus is annotated. It is available through a dedicated concordancer under the CC-BY license.
EUbookshop 3.5 billion tokens sentence-aligned downloadable	This is a multilingual corpus of texts from EU law books and related publications from unknown periods in 48 languages. The corpus consists of approx. 3.5 billion tokens. The corpus is sentence-aligned. The corpus is available for download on the OPUS webpage, although it is unclear under which license. For the relevant publication, see Skadiņš et al. (2014)
Parallel Global Voices Size unknown sentence-aligned downloadable	This is a multilingual of data crawled from the Global Voices webpage in approx. 50 languages. The corpus consists of 174,629 documents, token size and period unknown. The corpus is sentence-aligned. The corpus is available for download on the webpage of the corpus under the CC-BY license.
Scielo corpus Size unknown annotation unknown unavailable	This is a multilingual corpus that consists of texts from the Scielo database of scientific publications in the following language pairs: *English-French, English-Spanish, English-Portuguese*. All the other metadata are unknown – size, period, and annotation. The corpus is unavailable.
REVISTA PESQUISA FAPESP PARALLEL CORPORA 150,000 sentences sentence and word aligned downloadable	This is a multilingual corpus that consists of data from the Brazilian magazine *REVISTA PESQUISA FAPESP* in the following two language pairs: Portuguese-English and Portuguese-Spanish. The corpus is unidirectional with Portuguese as the original language. The corpus consists of 150,000 aligned sentences. The corpus is sentence- and word-aligned. The corpus is available for download on the corpus webpage.
Estonian Open Parallel Corpus 2012. Estonian-English 2,500,000 tokens annotation unclear downloadable	This is a bilingual corpus of Biblical and judicial texts from unknown periods in English and Estonian—the translations are in English. The corpus consists of 2,500,000 tokens. It is unclear how the corpus is annotated. It is available for download through META-SHARE under the CC-BY license.
Linguatools Webcrawl Parallel Corpus German-English 2015	This is a bilingual corpus of German and English texts from between 2013 and 2015. All the other metadata are unknown—directionality, sources and size.

size unknown annotation unclear unavailable	It is not clear how the corpus is annotated and is not publicly available due to the ELRA_VAR license.
Manually aligned CES Polish-English parallel corpus 1,445,000 tokens sentence aligned unavailable	This is a bilingual corpus of Polish CES reports from unknown periods and their English translations. The corpus consists of 1,445,000 tokens. The corpus is sentence-aligned. Though the corpus is said to be available for download under the CC-BY license but the link appears to be broken.
Parallel Wiki size unknown annotation unclear unavailable	This is a multilingual corpus of Wikipedia texts in the following language pairs: English-German, English-Romanian, and English-Spanish. The original texts are in English. All the other metadata are unclear – size, period, and annotation. Though the corpus should be available for download under the CC-BY-NC license, the link seems to be broken.
PELCRA multilingual parallel corpora 143,000,000 tokens sentence aligned unavailable	This is a multilingual corpus of texts from the CORDIC and RAPID websites, as well as from press releases of the European Parliament and the European Southern Observatory in 25 languages; the directionality is not clear. There are 143,000,000 tokens in the corpus. The corpus is sentence-aligned. Though the corpus is supposed to be for download, the link appears to be broken. The data should be available under the CC-BY license.
SETimes 43,142,458 tokens partially sentence aligned unavailable	This is a multilingual corpus of texts from the *setimes.com* website from unknown periods in 10 languages; the directionality is not clear. There are 43,142,458 tokens in the corpus. The corpus is partially aligned at the sentence level. Though the corpus is supposed to be for download but the link appears to be broken. The data should be available under the CC-BY license. For the relevant publication, see Tyers and Alperen (2010).
The NAACL 2003 English-Romanian corpus size unknown annotation unclear unavailable	This is a bilingual corpus of English and Romanian. All the metadata are unclear – directionality, sources of the data, size, periods covered, and annotation. Though the corpus is supposed to be for download but the link appears to be broken. It is unclear under which license the data should be available.
Slovak-English Parallel Corpus 556 million tokens sentence aligned searchable on-line	This is a bilingual corpus of Slovak and English; the corpus is bidirectional. The sources of the data and time periods are unknown. The corpus consists of approx. 556 million tokens. The corpus is sentence-aligned. The corpus is available through a dedicated

	concordancer, though the license is unknown.
REVEAL-THIS Corpus size unknown annotation unclear unavailable	This is a multilingual corpus of English, French and Greek. All the metadata are unclear – directionality, sources of the data, size, periods covered, and annotation. Though the corpus is supposed to be for download but the link appears to be broken. It is unclear under which license the data should be available.
µtopia 1,500,000 tokens annotation unclear downloadable	This is a multilingual corpus of texts from Twitter and Microblogs from unknown periods in the following language pairs: English-Mandarin, English-Arabic, English-Russian, English-Korean, English-Japanese. The directionality of the translations is unclear. The corpus consists of approx. 1,500,000 tokens. It is unclear how the corpus is annotated. The corpus is available for download on the corpus webpage, although the license is unknown.
Parallel English-Irish corpus of legal texts size unknown sentence aligned downloadable and searchable on-line	This is a bilingual corpus of legal texts in English and Irish with unknown directionality. The size of the corpus is unknown. The corpus is sentence aligned and is available both for download and through a dedicated concordancer on the corpus webpage, though the license is unclear.
The English-Swedish Parallel Corpus 3,500,000 tokens paragraph aligned unavailable	This is a bilingual corpus of fictional and non-fictional sources from various periods in English and Swedish—the corpus is bidirectional with original texts both in English and Swedish. The corpus consists of approx. 3,500,000 tokens. The corpus is paragraph-aligned and does not seem to be available.
LILA parallel corpus 8 million tokens 8 million tokens sentence aligned searchable on-line	This is a bilingual corpus of fictional and non-fictional sources from between 1991 and 2012 in Lithuanian and Latvian—the corpus is bidirectional with original texts in both languages. The corpus consists of approx. 8 million tokens. The corpus is sentence-aligned. The corpus is available through a dedicated concordancer, though the license is unclear. For the relevant publication, see Utka et al. (2012).
QTLeap Corpus V1.2 140,000 tokens sentence aligned download	This is a multilingual corpus of texts related to computer and IT troubleshooting for the following language pairs: Bulgarian-English, Czech-English, Portuguese-English, Spanish-English and Basque-English—the corpus is unidirectional in that the target language is English. The corpus consists of 140,000 tokens. The corpus is sentence-aligned and is available for download through META-SHARE under the CC-BY license.
QTLeap News Corpus 1,104 sentences sentence aligned	This is a multilingual corpus of news articles from various periods in English, Czech, German and Spanish data. It is unidirectional in that the source language is English. The corpus consists of 1,104 sentences.

unavailable	The corpus is sentence –aligned and appears to be unavailable.

3.1. Identification

The following 12 corpora were found through META-SHARE:

(i) *The TRIS corpus*;
(ii) *Bulgarian-X Language Parallel Corpus;*
(iii) *Estonian Open Parallel Corpus 2012. Estonian-English;*
(iv) *Linguatools Webcrawl Parallel Corpus German-English 2015;*
(v) *Manually aligned CES Polish-English parallel corpus;*
(vi) *Parallel Wiki;*
(vii) *SETimes*
(viii) *The NAACL 2003 English-Romanian corpus;*
(ix) *Slovak-English Parallel Corpus;*
(x) *REVEAL-THIS Corpus;*
(xi) *QTLeap Corpus V1.2;*
(xii) *QTLeap News Corpus*.

The following 6 corpora were found through the LRE Map:

(i) *ParCor - A Parallel Pronoun-Coreference Corpus*;
(ii) *MultiUN: Multilingual UN Parallel Text 2000—2009*;
(iii) *EUbookshop*;
(iv) *Parallel Global Voices*;
(v) *Scielo corpus*;
(vi) *REVISTA PESQUISA FAPESP PARALLEL CORPORA*.

The following 6 corpora were found on Google:

(i) *TED-Parallel-Corpus;*
(ii) *The United Nations Parallel Corpus;*
(iii) *SPOOK corpus;*
(iv) *µtopia;*
(v) *Parallel English-Irish corpus of legal texts;*
(vi) *The English-Swedish Parallel Corpus*.

Finally, the hyperlink to the *LILA parallel corpus* was obtained through personal correspondence.

3.2. Availability

3.2.1. Through concordancers

The following 3 corpora are available exclusively through concordancers. In all cases, the concordancers are dedicated – i.e. they serve as components integrated as parts of the relevant corpus websites.

(i) *Bulgarian-X language Parallel Corpus*;
(ii) *Slovak-English Parallel Corpus;*
(iii) *LILA parallel corpus*.

3.2.2. For download

The following 10 corpora are available as downloads exclusively. The corpora *ParCor - A Parallel Pronoun-Coreference Corpus* and *EUbookshop* are also subcorpora of *OPUS*, which was described in table 1.

(i) *TED-Parallel-Corpus*;
(ii) *The United Nations Parallel Corpus;*
(iii) *ParCor - A Parallel Pronoun-Coreference Corpus;*
(iv) *MultiUN: Multilingual UN Parallel Text 2000—2009;*
(v) *EUbookshop;*
(vi) *Parallel Global Voices;*
(vii) *REVISTA PESQUISA FAPESP PARALLEL CORPORA;*
(viii) *Estonian Open Parallel Corpus 2012. Estonian-English;*
(ix) *µtopia;*
(x) *QTLeap Corpus V1.2*.

3.2.3. Unavailable

The following 9 corpora are unavailable. In the case of corpora (iii)-(ix), downloads links are offered but are broken, so it is possible that their respective authors want them to be available.

(i) *Scielo corpus;*
(ii) *QTLeap News Corpus;*
(iii) *The TRIS corpus;*
(iv) *Manually aligned CES Polish-English parallel corpus;*
(v) *Parallel Wiki;*
(vi) *PELCRA multilingual parallel corpora;*
(vii) *SETimes;*
(viii) *The NAACL 2003 English-Romanian corpus;*
(ix) *REVEAL-THIS Corpus*.

3.3. Metadata

As in table 1, the metadata on annotation of the corpora not available in the CLARIN infrastructure are generally poor. It must be emphasised that the state of the documentation is relatively consistent both in the case of the infrastructure and of the corpora that are not part of it. As the following subsections will show, metadata on annotation are missing for roughly half of the corpora, whereas the size of the corpus is relatively well-documented, as was the case of the corpora in the CLARIN infrastructure.

3.3.1. Languages

17 out of the 26 corpora are multilingual. As in table (1), the corpora mostly represent European languages, with the bigger exceptions being the *Parallel Global Voices* corpus*,* which consists not only of data from European languages like Danish, German, Greek, and English, but also of Japanese, Indonesian, Arabic, Amharic and Bangla data; *µtopia*, which comprises of Japanese and Korean data, and the *TED-Parallel-Corpus*, which in part consists of Arabic, Simplified Chinese, Traditional Chinese, Japanese, Hebrew and Korean.

(i) *Bulgarian-X language Parallel Corpus* – approx. 50 languages;
(ii) *Parallel Global Voices* – approx. 50 languages;
(iii) *EUbookshop* – 48 languages;
(iv) *PELCRA multilingual parallel corpora* – 25 languages;
(v) *TED-Parallel-Corpus* – 11 languages;
(vi) *SETimes* – 10 languages;
(vii) *The United Nations Parallel Corpus* – 6 languages;
(viii) *µtopia* – 6 languages;
(ix) *QTLeap Corpus V1.2* – 6 languages;
(x) *Spook corpus* – 5 languages;
(xi) *Scielo corpus* – 4 languages;
(xii) *Parallel Wiki* – 4 languages;
(xiii) *QTLeap News Corpus* – 4 languages;
(xiv) *MultiUN: Multilingual UN Parallel Text 2000—2009* – 3 languages;
(xv) *REVISTA PESQUISA FAPESP PARALLEL CORPORA* – 3 languages;
(xvi) *REVEAL-THIS Corpus* – 3 languages.

3.3.2. Size

Information regarding size is relatively well documented – it is available for the following 18 (out of 25) corpora. However, corpus size is not always given in terms of tokens—in the case of corpora (ii), (ix) and (ixx), the size is given in sentences, while in the case of corpus (viii), the size is given in terms of documents. The largest corpora are *MultiUN: Multilingual UN Parallel Text 2000—2009* (approx. 1 billion tokens), *Bulgarian-X language Parallel Corpus* (approx. 1.2 billion tokens) and *EUbookshop* (approx. 3 billion tokens); the smallest corpus is *QTLeap Corpus V1.2* (approx. 140,000 tokens).

(i) *The TRIS corpus*;

(ii) *TED-Parallel-Corpus*;
(iii) *The United Nations Parallel Corpus*;
(iv) *MultiUN: Multilingual UN Parallel Text 2000—2009*;
(v) *MultiUN: Multilingual UN Parallel Text 2000—2009*;
(vi) *Bulgarian-X language Parallel Corpus*;
(vii) *EUbookshop*;
(viii) *Parallel Global Voices*;
(ix) *REVISTA PESQUISA FAPESP PARALLEL CORPORA*;
(x) *Estonian Open Parallel Corpus 2012. Estonian-English*;
(xi) *Manually aligned CES Polish-English parallel corpus*;
(xii) *SETimes*;
(xiii) *Slovak-English Parallel Corpus*;
(xiv) *µtopia*;
(xv) *The English-Swedish Parallel Corpus*;
(xvi) *LILA parallel corpus*;
(xvii) *QTLeap Corpus V1.2*;
(xviii) *QTLeap News Corpus*.

3.3.3. Alignment

Information regarding text alignment is available only for the following 11 corpora, so less than half:

(i) *The TRIS corpus*;
(ii) *Parallel Global Voices*;
(iii) *Manually aligned CES Polish-English parallel corpus*;
(iv) *PELCRA multilingual parallel corpora;*
(v) *SETimes*;
(vi) *Slovak-English Parallel Corpus;*
(vii) *Parallel English-Irish corpus of legal texts*;
(viii) *The English-Swedish Parallel Corpus*;
(ix) *LILA parallel corpus*;
(x) *QTLeap News Corpus*;
(xi) *QTLeap News Corpus*.

3.3.4. License

Information on the license is available only for 9 corpora. The 8 corpora listed below are available under the CC-BY license. By contrast, the data in the *Linguatools Webcrawl Parallel Corpus German-English 2015* are licensed under the restrictive ELRA_VAR license.

(i) *Bulgarian-X language Parallel Corpus;*
(ii) *Parallel Global Voices;*
(iii) *Estonian Open Parallel Corpus 2012. Estonian-English;*

(iv) Manually aligned CES Polish-English parallel corpus;

(v) Parallel Wiki;

(vi) PELCRA multilingual parallel corpora;

(vii) SETimes;

(viii) QTLeap Corpus V1.2.