a folha

Boletim da língua portuguesa nas instituições europeias

http://ec.europa.eu/translation/portuguese/magazine

English version of an article published in No. 45 — Summer 2014

European Union multilingual corpora for reuse in translation

Hilário Leal Fontes
Directorate-General for Translation — European Commission

[click here for Portuguese version]

Bi- or multilingual parallel corpora provide the basis for today's state-of-the-art statistical machine translation. They are also necessary for searching translated terminology in context.

To have an idea of the possibilities of machine translation based on parallel corpora, it only takes using machine translation engines freely available on the Web, like Google Translate[(1)] or Bing Translator[(2)]. To search translated terminology in context, the concordancing services provided by Linguee[(3)], Glosbe[(4)] or MyMemory[(5)] are good examples of what can be obtained.

Although these services have incorporated many other bi- and multilingual sources, they started with European Union (EU) parallel corpora and, for Chinese, Russian and Arabic, with parallel corpora from the United Nations.

In this paper I would like to give a brief overview of the parallel corpora built from the multilingual texts of the European Union that can be downloaded and used directly with computer-aided translation tools and/or used for building statistical machine translation engines. But let's first have an idea of where these corpora come from and of the still untapped potential of the multilingual texts published by the EU.

The EU translates millions of pages per year. The Directorate-General for Translation (DGT) of the European Commission alone translates around two million pages per year[(6)], including about 80,000 pages into Portuguese.

Considering that the Commission employs just over a third of all EU translators, it can be estimated that the EU as a whole generates more than 200,000 pages per year of translations from other languages into Portuguese.

While part of these pages is not for publication (communication with the Portuguese Government, letters to citizens and/or organisations who contact the Commission in Portuguese, etc.), the vast majority of the pages produced by the Commission in Portuguese are published in the L and C series of the Official Journal of the European Union (OJEU)[(7)]), on EU Bookshop[(8)] and on the *europa.eu* portal[(9)].

The overwhelming majority of pages produced by DGT into Portuguese are also translated into all or almost all other official EU languages. The situation in the other EU translation services is probably not fundamentally different[10].

With a view to re-use translations, DGT translations and publications are aligned at sentence level and stored in a database called Euramis, and from this database they are re-used with our computer assisted translation (CAT) tools. It is also this large memory that provides the corpora for building the engines that power the Commission's machine translation service, MT@EC.

The Euramis database is not publicly available but DGT regularly publishes the content of the L (Legislation) series of the OJEU aligned at sentence level, for the benefit of translators, researchers and anyone who may find aligned parallel corpora useful. DGT has not been alone in this. Indeed, the first large-scale aligned parallel corpora based on EU texts were published by Prof. Philipp Koehn[11] — Europarl — and the Commission's Joint Research Centre (JRC) — JRC-Acquis.

Let us look now at the aligned parallel corpora built with EU texts which are publicly available.

Europarl[12]

It is probably the multilingual parallel corpus most widely used in research and for building MT engines with European languages. The successive versions of this corpus added the more recent texts to earlier texts and corrected the odd technical issue.

The texts at the source of the Europarl corpus are the transcriptions of the European Parliament debates and were compiled by Prof. Philipp Koehn, who needed a multilingual corpus for research in the domain of statistical machine translation, in particular for the development of the Moses project. Currently, it features release 7 and contains more than 2 million aligned sentences between English and Portuguese. The language is quite colourful, typical of parliamentary debates.

JRC-Acquis[13]

The JRC published in 2006 the '*Acquis communautaire*' corpus — the body of all EU legislation in force at a given time — in 20 official languages, with approximately 9 million words per language. The latest published version, of January 2014, covers 22 languages and contains, on average, nearly 48 million words per language.

DGT-Acquis[14]

Strictly speaking, «*Acquis*» is a misnomer. While the aim of the JRC-Acquis project was to compile the body of legislation in force on 1 May 2004 (this legislation had to be translated in all accession languages), DGT-Acquis is the result of the compilation of the C (Information and Notices) and L (Legislation) series of the OJEU published from May 2004 onwards. Currently, DGT-Acquis covers texts published up to the end of 2011.

In this corpus, the alignment has been done down to paragraph level and it is possible to obtain, upon request, the unprocessed original XML files. It contains about 5 million paragraphs in every language and covers 23 languages (Croatian not included yet). Irish has a very reduced coverage and Maltese has a somewhat reduced coverage.

Using this corpus directly with CAT tools or for building machine translation engines without a prior finer segmentation may not yield the best results. Content-wise, there is a significant overlapping with DGT Memory (cf. below) because DGT Memory also contains the OJ L series, but the OJ C series is not included in any other corpus. The C series is of particular interest because its language style, while being still administrative, is less legislative. A new version featuring the post-2011 texts should be published later this year. This collection provides an excellent basis for a university project (Minho, are you listening?) to align and publish it with a finer granularity.

DGT Memory[15]

The Directorate-General for Translation of the European Commission (DGT) published in 2007[16] a series of legal texts aligned at sentence level and checked manually by its own language departments over the previous ten years.

In 2011 a new collection was published with the automatic unchecked alignment of the L series of the Official Journal of the European Union from May 2004 until the end of 2010.

In 2012 and 2013 it received an annual increment corresponding to acts published in the previous year.

This collection has been published in 22 languages (all official languages except Irish and Croatian) and comes with tools that allow the extraction of bilingual corpora in any pair of the 22 languages. The language style is partly the typical language of legislative acts, but many texts have technical content.

Other JRC collections

The JRC has also published small translation memories with texts from the Directorate-General for Education and Culture (DG EAC)[17] and from the European Centre for Disease Prevention and Control (ECDC)[18].

Per-Fide project[19]

This is a commendable initiative by the University of Minho, which has processed with their tools several versions of corpora published by DGT and the JRC in six languages (DE, EN, ES, FR, IT and PT) and has compiled a brand new collection of texts from the European Central Bank and several other parallel corpora which do not deal directly with EU affairs.

Within the same project, there is also a concordancing interface[20] which allows searches in the above-mentioned corpora and in other corpora which are only available for online consultation.

Opus website[21]

This site contains many collections of aligned parallel corpora in a large number of languages, among which the following aligned parallel corpora from EU texts with Portuguese:

a) *EMEA*, texts of the European Medicines Agency; (b) European Central Bank (the corpus compiled by the University of Minho); c) Europarl, releases 3 and 7; d) European Constitution; e) EUBookshop. This latter corpus is a compilation of texts published in the EU virtual library.

hilario.fontes@ec.europa.eu

[1] Google Translator, https://translate.google.com/.
[2] Bing Translator, http://www.bing.com/translator/.
[3] Linguee, http://www.linguee.com/.
[4] Glosbe, http://glosbe.com/
[5] MyMemory, http://mymemory.translated.net/.
[6] 1 page = 1500 characters without spaces.
[7] EUR-Lex: *Official Journal of the European Union*, http://eur-lex.europa.eu/oj/direct-access.html?locale=en.
[8] EU Bookshop, https://bookshop.europa.eu/en/home/.
[9] Europa, http://europa.eu/index_en.htm.
[10] The translations produced by the translation service of the Court of Justice are published in the European Court reports,

http://bookshop.europa.eu/en/bundles/reports-of-cases-cbbpqep2IxiOAAAAE1GKQ16Jon/ and in EUR-Lex (*EU law and related documents - EU case-law*), http://eur-lex.europa.eu/collection/eu-law/eu-case-law.html?locale=en.

[11] Philipp Koehn, http://homepages.inf.ed.ac.uk/pkoehn/.

[12] Statistical Machine Translation, *European Parliament Proceedings Parallel Corpus 1996-2011*, http://www.statmt.org/europarl/.

[13] Joint Research Centre, *Language Technology Resources: JRC-Acquis*, http://ipsc.jrc.ec.europa.eu/index.php?id=198.

[14] Joint Research Centre, *Language Technology Resources: DGT-Acquis*, http://ipsc.jrc.ec.europa.eu/index.php?id=783.

[15] Joint Research Centre, *Language Technology Resources: DGT-Translation Memory* https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory.

[16] This early collection was taken offline in 2013.

[17] Joint Research Centre, *Language Technology Resources: EAC-TM*, http://ipsc.jrc.ec.europa.eu/index.php?id=784.

[18] Joint Research Centre, *Language Technology Resources:ECDC-TM*, http://ipsc.jrc.ec.europa.eu/index.php?id=782.

[19] Per-Fide, *Resources*, http://per-fide.ilch.uminho.pt/site.pl/resources.en.

[20] Per-Fide, *Query: Type: bilingual + Select language: PT-EN*, http://per-fide.ilch.uminho.pt/query/bilingual/PT-EN.

[21] Opus, http://opus.lingfil.uu.se/.