European Parliament Proceedings Parallel Corpus 1996-2011

For a detailed description of this corpus, please read:

Europarl: A Parallel Corpus for Statistical Machine Translation, *Philipp Koehn*, MT Summit 2005, pdf.

Please cite the paper, if you use this corpus in your work. See also the extended (but earlier) version of the report (ps, pdf).

The Europarl parallel corpus is extracted from the proceedings of the European Parliament. It includes versions in 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek.

The goal of the extraction and processing was to generate sentence aligned text for statistical machine translation systems. For this purpose we extracted matching items and labeled them with corresponding document IDs. Using a preprocessor we identified sentence boundaries. We sentence aligned the data using a tool based on the Church and Gale algorithm.

Release v7

On 15 May 2012 we released a further expanded and improved version of the corpus. Previous versions are available here. The corpus is released as a source release with the document files and a sentence aligner, and parallel corpora of language pairs that include English.

Changes since v6

- added 01/2011 - 11/2011 data, now up to around 60 million words per language
- further refined preprocessing, cleaning

All formats contain document (<CHAPTER id>), speaker (<SPEAKER id name language>), and paragraph (<P>) mark-up on a separate line. The data is stored in one file per day, and in smaller units for newer data.

Some documents have the SPEAKER tag attribute LANGUAGE which indicates what language the original speaker was using.

To use the parallel corpora with tools like GIZA++, you want to:

- tokenize the text (required)
- lowercase the text (recommended)
- strip empty lines and their correspondences (required)
- remove lines with XML-Tags (starting with "<") (required)

Download

- source release (text files), 1.5 GB
- tools (preprocessing tools and sentence aligner only), 8.6 KB
- parallel corpus Bulgarian-English, 41 MB, 01/2007-11/2011
- parallel corpus Czech-English, 60 MB, 01/2007-11/2011
- parallel corpus Danish-English, 179 MB, 04/1996-11/2011
- parallel corpus German-English, 189 MB, 04/1996-11/2011
- parallel corpus Greek-English, 145 MB, 04/1996-11/2011
- parallel corpus Spanish-English, 187 MB, 04/1996-11/2011
- parallel corpus Estonian-English, 57 MB, 01/2007-11/2011
- parallel corpus Finnish-English, 179 MB, 01/1997-11/2011
- parallel corpus French-English, 194 MB, 04/1996-11/2011
- parallel corpus Hungarian-English, 59 MB, 01/2007-11/2011
- parallel corpus Italian-English, 188 MB, 04/1996-11/2011
- parallel corpus Lithuanian-English, 57 MB, 01/2007-11/2011
- parallel corpus Latvian-English, 57 MB, 01/2007-11/2011
- parallel corpus Dutch-English, 190 MB, 04/1996-11/2011
- parallel corpus Polish-English, 59 MB, 01/2007-11/2011
- parallel corpus Portuguese-English, 189 MB, 04/1996-11/2011
- parallel corpus Romanian-English, 37 MB, 01/2007-11/2011
- parallel corpus Slovak-English, 59 MB, 01/2007-11/2011
- parallel corpus Slovene-English, 54 MB, 01/2007-11/2011
- parallel corpus Swedish-English, 171 MB, 01/1997-11/2011

Size of the Corpus

Sizes for single-language data after removing XML.

Language	Sentences	Words
Bulgarian	411,636	-
Czech	668,595	13,195,311
Danish	2,323,099	47,761,381
German	2,176,537	47,236,849
Greek	1,517,141	-
English	2,218,201	53,974,751
Spanish	2,123,835	54,806,927
Estonian	692,210	11,358,009
Finnish	2,119,515	33,708,706
French	2,190,579	54,202,850
Hungarian	658,824	12,606,986
Italian	2,081,669	50,259,169
Lithuanian	678,665	11,512,131
Latvian	666,026	12,085,228
Dutch	2,333,816	53,487,257
Polish	387,490	7,087,016
Portuguese	2,121,889	52,300,149
Romanian	402,904	9,663,544
Slovak	674,359	13,116,301
Slovene	634,488	12,665,974
Swedish	2,241,386	45,665,947

Sizes for parallel corpora after sentence aligning and removing XML.

Parallel Corpus (L1-L2)	Sentences	L1 Words	English Words
Bulgarian-English	406,934	-	9,886,291
Czech-English	646,605	12,999,455	15,625,264
Danish-English	1,968,800	44,654,417	48,574,988
German-English	1,920,209	44,548,491	47,818,827
Greek-English	1,235,976	-	31,929,703
Spanish-English	1,965,734	51,575,748	49,093,806
Estonian-English	651,746	11,214,221	15,685,733
Finnish-English	1,924,942	32,266,343	47,460,063
French-English	2,007,723	51,388,643	50,196,035
Hungarian-English	624,934	12,420,276	15,096,358
Italian-English	1,909,115	47,402,927	49,666,692
Lithuanian-English	635,146	11,294,690	15,341,983
Latvian-English	637,599	11,928,716	15,411,980
Dutch-English	1,997,775	50,602,994	49,469,373
Polish-English	632,565	12,815,544	15,268,824
Portuguese-English	1,960,407	49,147,826	49,216,896
Romanian-English	399,375	9,628,010	9,710,331
Slovak-English	640,715	12,942,434	15,442,233
Slovene-English	623,490	12,525,644	15,021,497
Swedish-English	1,862,234	41,508,712	45,703,795

Test Sets

Several test sets have been released for the Europarl corpus. In general, the Q4/2000 portion of the data (2000-10 to 2000-12) should be reserved for testing. All released test sets have been selected from this quarter. The shared tasks for the 2006 and 2007 ACL Workshops on Statistical Machine Translation provide test sets from the Europarl corpus.

The original common test set from the Koehn/Och/Marcu ACL 2003 Paper is available in the archives.

Extended versions of these test sets are available in the Evaluation Matrix of the EuroMatrix project.

Known Bugs

- Some special HTML entities and noisy characters are not removed from the data.
- Some recent Greek data has only parts of transcripts in the files.

Terms of Use

We are not aware of any copyright restrictions of the material. If you use this data in your research, please contact phi@jhu.edu. Please let us know if you find problems with the data or if you want the data for other language pairs. We recommend using the last quarter of 2000 for testing (2000-10 until