OPUS parallel corpora (version 2 with m:n alignment)

The parallel corpora available here have been collected, prepared and aligned by Joerg Tiedermann in the OPUS project (see http://opus.lingfil.uu.se/, where the source data is also available). We are most grateful to Joerg Tiedermann for his great work and co-operation. The data was prepared for the Sketch Engine using a range of lemmatisers, part-of-speech taggers and Sketch Grammars. Unlike the first version, the alignment is now m:n, which allows for just one corpus per language.

List of Sub-corpora:

- ECB - European Central Bank corpus
- EMEA - European Medicines Agency documents
- EUconst - The European constitution
- Europarl3 - European Parliament Proceedings (v3)
- PHP - PHP manual corpus
- SETIMES2 - A parallel corpus of the Balkan languages
- SPC - Stockholm Parallel Corpora
- RF - Regeringsförklaringen - Declarations of Government Policy by the Swedish Government
- MBS - Belgisch Staatsblad corpus
- OfisPublik
- TedTalks
- hrenWaC
- TEP - The Tehran English-Persian subtitle corpus
- KDE4 - KDE4 localization files
- KDEdoc - KDE manual corpus
- OpenOffice
- OpenOffice3
- OpenSubtitles2011 - Open Subtitles corpus (2011 version)
- Tatoeba
- UN - Translated UN documents
- MultiUN - Translated UN documents