OPUS - an open source parallel corpus

http://logos.uio.no/opus/

Lars Nygaard
Tekstlaboratoriet HF
University of Oslo
Postboks 1102 Blindern
0317 Oslo
lars.nygaard@ilf.uio.no

Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527
SE-751 20 Uppsala, Sweden
joerg@stp.ling.uu.se

1 Introduction

Parallel corpora are very valuable for a wide variety of research, particularly in machine translation and lexicography. However, parallel corpora have been few, unrepresentative, and not generally available. The aim of the OPUS project is to provide a public collection of parallel corpora which can be freely used and distributed. This makes it possible for everyone to run experiments on bitexts and their results can be easily compared.

We base our corpus collection on open source documentation and their translations. Many open source projects include a large amount of textual data and invite people around the world to localise products and their documentation. Similarly to the software itself, the entire documentation is freely available and may be used in any way by anybody.

The idea of using translated textual data which can be collected from the web is not new (e.g. [Res98]). In OPUS, we collect translated texts from the web, convert and align the entire collection, add linguistic data, and provide

the community with a publicly available parallel corpus. OPUS is based on open source products and will also be delivered as an open source package.

2 OPUS v0.1

In the current stage, we concentrate on a specific domain, namely software documentation, which can be found in open source projects. OPUS consists so far of the documentation of the office package OpenOffice [1] with its original collection of 2014 file in English and 5 collections of translated texts, French, Spanish, Swedish, German, and Japanese. The English part comprises about 500,000 words. Not all files have been translated yet. Each translation contain between 400,000 and 500,000 words. The entire corpus includes about 2.6 million words in its current version.

2.1 Sentence alignment

All documents have been tokenised and aligned on the sentence level for all possible language pairs. For alignment, we used a length-based approach based on the algorithm and software by Gale&Church [GC93]. Alignments are stored in XML according to the XCES recommendations. We decided to keep all the original mark-up in the text documents with additional mark-up for linguistic information. Existing mark-up improves, e.g. the quality of tokenisation and may be useful for later investigations. Furthermore, we kept the structure of the text collection with all its sub-directories and separate files. This makes it easy to identify the origin of each text segment and to use parts of varying size. It also helped the sentence alignment a lot because each pair of files has been aligned separately: In this way, the amount of follow-up errors could be kept at a low level.

2.2 Linguistic markup

We also applied available tools for language specific mark-up, namely part-of-speech taggers (TnT, TreeTagger, Grok, ChaSen) and a shallow parser (Grok). Grok [Bal01] and ChaSen [MKY+00] are freely available from the web. Grok is an implementation of the OpenNLP interfaces and comes with modules for tagging and chunking English texts. Both modules are trained on the Penn Tree Bank using the Penn tagset [MSM93]. ChaSen is a tokeniser and morphological analyser for Japanese. It provides several kinds

[1] http://www.openoffice.org/

of linguistic information such as readings, parts of speech, and base forms. TnT [Bra00] and the TreeTagger [Sch94] are freely available for research purposes and their usage on OPUS files has been granted by the authors. They come with ready-to-use modules for tagging English and German (TnT & TreeTagger) and for French and Italian (TreeTagger only). Furthermore, the TreeTagger also comes with a lemmatiser for all supported languages. Both taggers can be trained on other material. A module for tagging Swedish with TnT trained on the SUC corpus has been provided by Beáta Megyesi [Meg01].

Using external tools requires several conversions in order to fulfil format and encoding requirements. Several tailored scripts and tools have been used to achieve this, including recode (converts between many character encoding standards) [Pin00], tidy (validates and pretty-prints HTML and XML files) [Rag03], and the Uplug toolbox (basic mark-up, sentence splitting, tokenisation, XML processing) [Tie02].

2.3 Availability

The corpus is available from the OPUS home page [2]. We also converted the corpus (except for the Japanese part) to be indexed by the corpus work bench [Chr94] including the alignment information. The corpus is accessible via web-interfaces [3] and can be searched for multiple languages in parallel [4].

OPUS is meant to be open for extensions. We are currently working on adding further material. The content of the collection will be updated gradually which will be announced via the project webpage.

3 References

References

[Bal01] Jason Baldrigde. Grok - an open source natural language processing library, 2001.

[Bra00] Thorsten Brants. Tnt - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA, 2000. http://www.coli.uni-sb.de/ thorsten/.

[2] http://logos.uio.no/opus/
[3] http://logos.uio.no/opus/search.html
[4] http://logos.uio.no/opus/oo.html

[Chr94] Oliver Christ. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94*, Budapest, 1994.

[GC93] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19:75–102, 1993.

[Meg01] Beáta Megyesi. Comparing data-driven learning algorithms for POS tagging of swedish. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 151–158, Carnegie Mellon University, Pittsburgh, PA, USA, 2001.

[MKY+00] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. Morphological analysis system chasen version 2.2.1 manual. http://chasen.aist-nara.ac.jp/chasen/bib.html.en, December 2000.

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19, 1993.

[Pin00] François Pinard. The recode reference manual. http://www.iro.umontreal.ca/contrib/recode/HTML/recode.html, 2000.

[Rag03] Dave Raggett. Clean up your web pages with html tidy. http://www.w3.org/People/Raggett/tidy/, 2003. http://tidy.sourceforge.net/.

[Res98] Philip Resnik. *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA*, chapter Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. Number 1529 in Lecture Notes in Artificial Intelligence. Springer, October 1998.

[Sch94] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September 1994. http://www.ims.uni-stuttgart.de/ schmid/.

[Tie02] Jörg Tiedemann. *Parallel Corpora, Parallel Worlds*, chapter Uplug - a modular corpus tool for parallel corpora. Rodopi, 2002. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden.