OPUS – Parallel Corpora for Everyone

Jörg TIEDEMANN

Department of Modern Languages, University of Helsinki, 00014 Helsinki, Finland

jorg.tiedemann@helsinki.fi

Abstract. Abstract. OPUS is a large collection of freely available parallel corpora that we provide in various formats and packages. All data sets are completely aligned at the sentence level for all possible language pairs. OPUS covers over 200 languages and language variants with a total of about 3.2 billion sentences and sentence fragments containing over 28 billion tokens. The collection contains data from various sources and domains and each sub-corpus is provided in common data formats to make it easy to integrate them in research and development. OPUS also provides tools and on-line interfaces for exploring parts of the collection and is continuously growing in terms of size and coverage.

Description

Parallel data sets in OPUS[1] are freely available and cover various domains. The largest collections (in terms of volume) come from political and administrative sources such as the European Commission and user-provided movie subtitles in various languages. Other sources include software localisation, multilingual news providers, translated descriptions of medical products, religious texts and multilingual wikis and other websites. OPUS is organised by source and one of the main principles of the collection is to preserve the original data structures (file structure, formatting, meta-data) as much as possible. The goal of our project is to make the collection applicable as widely as possible. Currently, OPUS comprises 3.2 billion sentences with over 28 billion tokens in total. An important principle is complete sentence alignment for all language pairs, thus, supporting even low-density languages and unusual combinations. There are bitexts such as Arabic–Korean or Indonesian–Latvian with over one million translation units among the 12,572 language pairs with their 10.8 billion translation units in total. We provide the data in standalone XML and stand-off alignment (as its native format) but also commonly used formats such as TMX and aligned plain Unicode text. The latest edition of OPUS contains parallel sentences extracted from Wikipedia and a significantly extended collection of movie subtitles, now also including intra-lingual alignments of alternative translations. Furthermore, we also provide word alignments and phrase-tables for statistical machine translation (SMT) ready to be used in common SMT toolboxes. A subset of our data is also available via on-line search interfaces. Feedback and contributions are welcome.

[1] http://opus.lingfil.uu.se