Large-scale Annotation and Alignment of Parallel Corpora for the Investigation of Linguistic Variation

Marianne Hundt and Martin Volk, University of Zurich
mhundt@es.uzh.ch, volk@cl.uzh.ch

May 13, 2013

Summary of the Project Proposal

Translated documents in multiple languages (here: parallel documents) are highly regarded as valuable resources for various tasks in natural language processing and linguistic research. Parallel corpora are useful for tasks as diverse as word sense disambiguation, terminology extraction and contrastive corpus linguistics. The usefulness of these resources for contrastive linguistics, in particular, has increased tremendously with the possibility to automatically align the texts not only on the sentence level but also on the word level.

We propose to align (sentence alignment and word alignment) and annotate (PoS-Tagging and Parsing) a large parallel corpus for the language pairs English-German, English-French, and English-Spanish. For this purpose we will use two large parallel corpora: the Europarl corpus and the UN-Corpus. In addition, we will align the corpus for the language pairs English-Russian and English-Finnish in order to include comparison with languages that do not have articles. Moreover, the latter language pair enables us to include comparisons with a non-Indoeuropean language in the linguistic research.

Linguistic variation at times involves the choice between the use of an element and its omission (articles, relativizers, pronouns etc.). Modelling such zero or null contexts in a corpus-driven approach is particularly challenging because variable zero elements are difficult to retrieve even from annotated monolingual corpora: While it is possible to extract noun phrases without an article from a parsed corpus, such algorithms have poor precision because a vast number of instances will not allow for a definite or indefinite article. Therefore we propose to use parallel annotated and word-aligned corpora. That will enable us to precisely target constructions with variable optional elements in one of the languages. Our goal is to prove the usefulness of such a large aligned and annotated corpus (LAAC) for the investigation of linguistic variation. As a case in point we will investigate variable article use in these languages, and zero articles in English, in particular. We believe that such a LAAC provides a number of advantages for new insights in these areas.

Studying articles in English is of interest and importance because of the growing influence of non-native English speakers whose first languages do not have articles or that use articles in ways clearly different from English. With the help of our LAAC we will retrieve instances of zero articles of the language pairs in the corpus where both languages have definite and indefinite articles (English, French, German and Spanish). The language pairs English-Russian and English-Finnish will serve to model variable article use in English against the background of typological differences.

In addition, we will build a rich database that models factors influencing variable article choice in the language pairs English and German. These factors include lexical variation in the head noun, the internal structure of the noun phrase (pre- and postmodification), syntactic function of the noun phrase but also discourse-pragmatic functions (given vs. new). The aim for the linguistics part of the project is to arrive

at a detailed description of variable article use in English and German using a multi-variate analysis. This will prove useful for purposes of language teaching and machine translation. Because of the element of lexico-grammatical variation, large parallel corpora are a prerequisite for this kind of research.

The challenge for computational linguistics lies in the high-quality alignment and annotation of large corpora and the construction of an efficient and powerful corpus query tool that is able to handle these corpora. Efficient query tools for large monolingual corpora exist but the development of such a tool for parallel corpora is highly innovative.[1]

[1]The authors would like to thank Jeanette Isele for researching numerous details and helping with the first draft of this proposal.