Using parallel treebanks for comparative syntax with Poly-GrETEL

Liesbeth Augustinus
University of Leuven

In contrastive linguistics and translation studies it is common to use parallel corpora, see amongst others Johansson (2007). For studies in comparative syntax, however, we need syntactically annotated parallel corpora: “Exploring grammatical phenomena in a multilingual corpus is a difficult and time-consuming task involving manual intervention (...) In general, there is a lack of large multilingual corpora with advanced syntactic annotation. Developing such data is an important task for the future." (Johansson 2007: 37) In the years since Johansson wrote this, some syntactically annotated parallel corpora or "parallel treebanks" have been created, e.g. the SMULTRON parallel treebank (Volk et al. 2010), and the parallel treebanks included in the INESS platform (http://clarino.uib.no/iness). In contrast to (flat) parallel corpora, parallel treebanks are only available for a limited number of languages, and they are often small in size. Moreover, one typically needs sophisticated tools and methods to query such treebanks. Some users are deterred by this, which means that the potential of parallel treebanks will not be realized.

In order to make parallel treebanks accessible to non-technical users, Poly-GrETEL (http://gretel.ccl.kuleuven.be/poly-gretel) has been developed. It is an online tool which enables syntactic querying in parallel treebanks. Similar to the monolingual GrETEL environment (http://gretel.ccl.kuleuven.be/gretel3), Poly-GrETEL allows users to query parallel treebanks by means of natural language examples instead of a formal query. This approach is called "example-based querying", which is a stepwise query procedure to create a search instruction with limited knowledge of the treebank annotations and the exact layout of the syntax trees. As an alternative one can skip this procedure and use the XPath query language to search the treebanks.

Currently Poly-GrETEL provides access to the Europarl parallel treebank for Dutch and English (Koehn 2005), which is automatically parsed and aligned on sentence and node level. Soon the German-Dutch parallel treebank will be added as well. By combining the example-based query functionality with node alignments, Poly-GrETEL limits the need for users to be familiar with the query language and the structure of the trees in the source and target language. In this way, the tool facilitates the use of parallel treebanks for comparative linguistics and translation studies.

This paper illustrates the use of Poly-GrETEL for comparative syntactic research. As a case study the occurrence of substitute infinitives (also known as "Ersatzinfinitiv" or "Infinitivus Pro Participio") in Dutch and German will be considered. The phenomenon refers to the appearance of an infinitive instead of the (expected) past participle, e.g. "John hätte einen Roman schreiben wollen/*gewollt" ("John had wanted to write a novel"). Such constructions occur in German and Dutch, but the languages differ with respect to the type of verbs that appear as substitute infinitives. Moreover, the

languages differ with respect to the set of verbs that obligatorily or optionally occur in such constructions. Therefore, it is an interesting topic for a cross-linguistic syntactic study. In addition to the case study, a number of methodological aspects regarding the use of automatically created parallel treebanks for comparative syntactic research will be discussed.

References:

S. Johansson. (2007) Seeing through multilingual corpora: on the use of corpora in contrastive studies. Amsterdam: John Benjamins.

M. Volk, A. Göhring, T. Marek and Y. Samuelsson. (2010) “SMULTRON (version 3.0) — The Stockholm MULtilingual parallel Treebank.” Institute of Computational Linguistics, University of Zürich.