# Towards an Interactive Platform for Linguistic Corpus Exploration

Jonas Kuhn, Kathrin Spreyer und Bettina Schrader
Institut für Linguistik, Universität Potsdam, Germany
kuhn@ling.uni-potsdam.de, {spreyer,bschrade}@uni-potsdam.de

We propose a methodology for explorative linguistic search on large unannotated corpora, not presupposing any higher-level NLP analysis tools for the language under consideration.[1] The approach relies on existing annotated corpora and analysis tools for other languages, the use of parallel corpora to bridge across languages, and machine learning techniques for bootstrapping linguistic search tools in an interactive framework involving the linguistic researcher. This presentation will include a programmatic discussion of the planned framework, which is under development in a recently launched longer-term project (as part of SFB 632 on Information Structure); but we also present experimental results from a pilot study on one of the core modules of the proposed architecture, which has been fully implemented.

Our point of departure is that many phenomena of interest to syntactic, semantic and pragmatic research are too infrequent to expect a sufficient number of occurrences in subcorpora that can be realistically hand-annotated using the standard methodology, which we may call "*a priori* annotation", i.e., designing annotation schemata and performing annotation of a corpus sample well in advance and independently of the use of this annotation during corpus search. Moreover, carefully annotated corpora are only available for a small number of languages; and even when annotated corpora for different languages exist, contrastive studies are often complicated by differences in annotation schemata and/or in the types of genres sampled in the corpora. As a consequence, linguistic research often has to fall back to larger unannotated corpora. In practice, search on unannotated corpora is often based on *ad hoc* decisions, such as formulating search expressions with particular lexemes that are deemed to represent a whole class of items. This leads to an often tedious turnaround cycle of manually assessing search results and refining search expressions. Our hypothesis is that with a systematic interactive platform exploiting a combination of technologies, human effort can be channeled much more effectively.

[1] This research has been supported by SFB 632 (project D4) of the Deutsche Forschungsgemeinschaft (DFG).

Ultimately, the platform we are developing will provide management facilities for a collection of modular search components focusing on specific linguistic aspects (e.g., identification of grammatical relations among two linguistic entities or animacy classification of a given nominal). Modules can be combined to perform higher-level search tasks (in particular those related to linguistic research on information structure), but quality assessment is based on individual modules, as they are reused and constantly improved in a large, multilingual research network with partially overlapping interests. It should be noted that the platform is not intended to ever provide fully automatic annotation, but to streamline the expert assessment of linguistic data. Our pilot study focuses on one component of the envisaged platform that is representative for the types of challenges we expect overall: identification of comparable argument-head relations in a multi-lingual parallel corpus, including languages lacking a broad-coverage parser. The module is straightforwardly usable as a stepping stone in a contrastive study on factors triggering specific aspects of argument realization (such as constituent order).

The goal is to use machine learning techniques to bootstrap an argument-head classifier for some language C lacking a parser, given a parallel corpus of languages A, B, C, where parallel parsers exist for A and B. For our experiments we pretend that Dutch is a language of type C and we use the broad-coverage LFG parsers from the ParGram project (Butt et al., 2002) for English and German as A and B. The parallel corpus we use is Europarl (Koehn, 2005), with GIZA++ word alignments for all language pairs (www.fjoch.com/GIZA++.html). Our approach is inspired by the ideas of annotation projection (Yarowsky et al., 2001) and triangulation (Kay, 1997). Heads and (the lexical heads of) their arguments are projected to C from both the parse in language A and the one in B, using the word alignments on the parallel corpus. Where the A- and the B-based projections to C coincide in the same word in language C, we have a candidate for a head-argument pair in that language. This is illustrated in (1), where the alignments of the arguments in the German and English sentences coincide on the Dutch *wij ('we')* and *wetten ('laws')*, respectively. Various further filtering conditions for candidate selection are conceivable, including interactive assessment of the candidate pairs by the linguistic researcher.

(1) DE: … *wenn* [*wir*]$_{\text{subj}}$ [*Gesetze*]$_{\text{obj}}$ ***erlassen****, die wir anschließend nicht anwenden.* (grammar A)
EN: ... *if* [*we*]$_{\text{subj}}$ ***make*** [*laws*]$_{\text{obj}}$ *which we then do not enforce.* (grammar B)
NL:... als [*wij*]$_{\text{subj}}$ [*wetten*]$_{\text{obj}}$ ***maken*** *die wij vervolgens niet handhaven.* (projected)

In the pilot study described below (which was focused on setting up the general machinery) we relied on the projected candidates directly to train a head-argument classifier for C, i.e. we applied a single cycle of fully unsupervised bootstrapping, the simplest possible scenario. The bootstrapping procedure can rely on a rich set of features (in addition to surface-level features within language C). Under projection across word alignments, the LFG analyses of languages A and B provide tentative morpho-syntactic and lexical information for C such as projected tense or verb type of the head, and person or number of the head and argument. (We are currently not attempting to determine the exact grammatical relation between a head and its dependent, but we plan to address these tasks in the future. Here, more aspects of the LFG analyses in A and B will become relevant.) Given that we cannot expect the word alignments to be free of noise, we explicitly include features that refer to the alignment topology. The modular architecture makes it very easy to add further features from any other available analysis tools or linguistic resources.

For our first investigations of the classification task, we considered 707 sentence triples where the German part contains one of 10 German verbs of medium frequency. For the 707 triples, we projected the German and English argument structures onto the Dutch sentences as described above and discarded those sentences for which the German and English grammars disagree. The projected classifications in the remaining 154 triples achieve 83.0% precision and 59.2% recall when evaluated against gold standard annotations for the subset of 83 sentences with correctly aligned heads and comparable argument structures. We then trained a loglinear model directly from the projected annotations using the MegaM software package (www.cs.utah.edu/~hal/megam/). This model exhibits 81.1% precision and 57.5% recall in 5-fold cross-validation. For comparison, a model based on gold standard annotations for the 83 comparable argument structures indicates an upper bound for performance (given the data set) at 86.2% precision and 60.8% recall.

Although the performance based on this small training set and a first set of simple local features is still moderate, we consider the results of the pilot study a good starting point to build up the interactive bootstrapping framework, and we will explore directions for improvements on a larger scale.

## References

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Kay, Martin (1997): The Proper Place of Men and Machines in Language Translation. *Machine Translation*. 12 (1-2), 3-23.

Koehn, P. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit, Phuket, Thailand, pp. 79-86.

Yarowsky, D., G. Ngai and R. Wicentowski (2001): Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In: Proceedings of the first International Conference on Human Language Technology Research (HLT), San Diego, USA, pp. 200-207.