Syntactic Annotation of Large Corpora in STEVIN

Gertjan van Noord Ineke Schuurman Vincent Vandeghinste
RU Groningen KU Leuven KU Leuven

Background

The Dutch Language Corpus Initiative (D-Coi) is one of the projects funded within the current STEVIN programme. The construction of a 500-million-word reference corpus of written Dutch has been identified as one of the priorities in the programme. In D-Coi, a 50-million-word pilot corpus is being compiled, parts of which will be enriched with (verified) linguistic annotations. In particular, syntactic annotation of a representative subcorpus of 200,000 words is envisaged. The focus is on written language in order to complement the Spoken Dutch Corpus (CGN).

CGN contains a subcorpus of 1 million words with syntactic annotations. During the construction of this corpus, no syntactically annotated corpus of Dutch was available to train a statistical parser on, nor an adequate parser for Dutch (requirements: wide-coverage, theory-neutral output, access to both functional and categorial information). This situation has changed considerably since then. Over the last few years, Alpino was developed at the University of Groningen. Alpino is a wide-coverage computational analyzer of Dutch which aims at accurate, full parsing of unrestricted text. For D-Coi, we have decided to use Alpino for the semi-automatic syntactic annotation of our subcorpus.

Alpino is based on a head-driven lexicalized grammar and a large lexical component. Ambiguity resolution is provided by a disambiguation component which consists of a beam-search procedure to select the best parse from a parse forest of all possible parses [3]. This beam search procedure is guided by a maximum-entropy model which is trained on the Alpino Treebank. Efficiency is greatly enhanced by an implementation of a HMM trigram POS-tagger, which removes all unlikely lexical categories for a given input sentence Alpino produces dependency structures, similar to those employed in the CGN (Corpus of Spoken Dutch).

1http://taalunieversum.org/taal/technologie/stevin/
2http://www.let.rug.nl/~vannoord/papers/alpino.pdf
3http://www.let.rug.nl/~vannoord/trees/
Annotation Guidelines

The original annotation scheme deployed in Alpino was not exactly the same as the one used in CGN. In order to enhance the possibilities to compare results found in D-Coi on the one hand and CGN on the other, we have adapted the Alpino scheme in such a way that it more closely resembles the CGN annotation scheme. For instance, the treatment of multi-word-units, punctuation tokens, ordinal numbers and te-infinitives has been adapted in Alpino and now conforms to the CGN-standard. A few remaining differences are documented exhaustively for the benefit of the users of both corpora. These differences include, for instance, the annotation of subjects of the embedded verb in auxiliary, passive, modal and control structures. In CGN, these are not expressed. In D-Coi we follow the convention of Alpino to encode these subject relations explicitly.

In D-Coi, we also inherit from Alpino the XML-format in which syntactic annotations are stored. This format directly allows the use of full XPATH and/or Xquery search queries for linguistically interesting queries. Therefore, we can employ standard tools for the exploitation of the syntactic annotations, and there is no need to dedicate resources for the development of specialized query languages. Note that the existing CGN corpus has been translated to the same XML-format, so that the same tools can be used for both corpora.

Annotation Tools

For interactive annotation, Alpino provides a variety of tools. These tools include optional interactive assignment and selection of lexical categories. The annotator can pick, if desired, the correct lexical categories for some or all of the words in the input, or add additional lexical categories on the fly.

Another tool is the optional and interactive assignment of syntactic brackets. The annotator can indicate, for instance, that a particular sequence of words must be analysed as a particular syntactic category, in order to direct the parser to the correct analysis in the case of ambiguities. Both labeled and unlabeled brackets are supported.

Alpino can be used to obtain the best N or all parses. A parse selection tool is available to select the correct parse or the best parse from a potentially large set of parses without the need to consider each of these parses individually (similar to the SRI Treebanker⁴). In this parse selection tool, the annotator makes a number of binary decisions about particular properties of

⁴http://acl.ldc.upenn.edu/W/W97/W97-1502.pdf
the desired parse. Based on each decision, the tool computes the remaining set of candidate parses, and reduces the number of binary decisions.

The annotator has access to the Thistle editor\(^5\) for intuitive editing CGN-type dependency structures.

In addition, a number of XML-based tools is available for automatic consistency checking of the annotations, for browsing the annotations, and for searching the annotations.\(^6\) The tools are all freely available.

### Future Directions

One of the ultimate goals of the STEVIN-programme is the construction of a 500-million-word reference corpus of written Dutch. In a future project we hope to be able to provide verified syntactic annotations for at least 1 million words. In addition, we intend to provide syntactic annotations of this full 500-million-word corpus. Such a large syntactically annotated corpus is useful for a wide variety of applications in information extraction, question answering, corpus linguistics, automated ontology building, lexicography etc. Recently, a number of experiments have been performed with the automatically constructed syntactic annotations of the 75-million word CLEF corpus\(^2\), which indicate the potential of such huge resources.

### References

1. Gosse Bouma and Geert Kloosterman. Querying dependency treebanks in xml. In *Proceedings of the Third international conference on Language Resources and Evaluation (LREC)*, Gran Canaria, 2002.

2. Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann. Question answering for Dutch using dependency relations. In *Proceedings of the CLEF2005 Workshop*, 2005.

3. Robert Malouf and Gertjan van Noord. Wide coverage parsing with stochastic attribute value grammars. In *Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*, Hainan China, 2004. IJCNLP.

\(^5\)http://www.ltg.ed.ac.uk/software/thistle/home/index.html
\(^6\)http://www.let.rug.nl/~vannoord/alp/