6.10 Alpino: computational syntactic analysis of Dutch

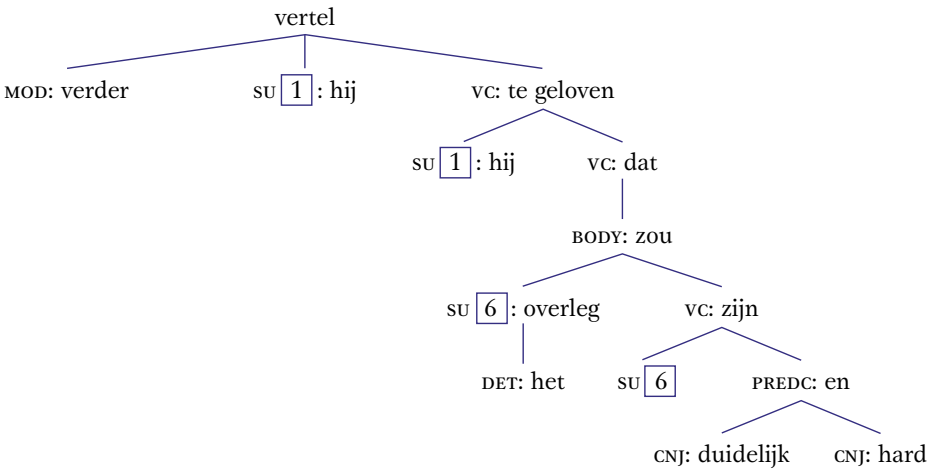
van der Beek L.J., Bouma G., van Noord G.J.M.

Alpino is a wide-coverage computational syntactic analyzer of Dutch which aims at accurate, full, parsing of unrestricted text. Syntactic analysis is a necessary prerequisite for understanding the meaning of a natural language utterance. Therefore, syntactic analysis is used in software applications such as information retrieval, information extraction, question answering, automated translation, spoken dialogue systems, etc. Consider the following sentence:

Verder vertelde hij te geloven dat het overleg duidelijk en hard zou zijn.
(lit. Furthermore he said to believe that the discussion would be clear and tough.)

Fort this sentence, Alpino produces the abstract syntactic structure given in the illustration. This structure makes explicit, for instance, that the thing that was "told" by "him" is his belief that the discussion would be clear and tough. In addition, "clear and tough" is a description of the discussion, and not of e.g. "him".

Figure 16 ▸



In order to create such abstract syntactic structures adequately, we have implemented a large computational grammar and a large computational dictionary for Dutch (containing more than 100.000 words). The architecture of the grammar and dictionary are based on sophisticated insights from linguistics. Techniques from computer science and statistics, on the other hand, are applied to optimize the application of these linguistic knowledge sources to analyze a given text.

Although the system performs reasonably well on many sentences a number of unsolved problems remain. In order to measure progress, it is important to be able to evaluate the coverage and accuracy of the system. Therefore we have constructed a treebank: a set of sentences for which the correct syntactic structure has been assigned manually. The treebank currently contains about 110.000 words. We can now compare syntactic structures assigned by Alpino to the correct structures in the treebank. This allows us to identify problematic constructions, and to evaluate various proposed solutions for these problems.