Lumberjacking with GrETEL

Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde
CCL, KU Leuven
{liesbeth,vincent,ineke,frank}@ccl.kuleuven.be

GrETEL (Greedy Extraction of Trees for Empirical Linguistics [Augustinus et al., 2012] is a linguistic search engine enabling linguists to consult a syntactically annotated corpus (or treebank) in a user-friendly way, as it accepts a natural language example instead of a complex search instruction. Therefore, limited or no knowledge about tree representations and formal query languages is needed.

We will investigate the case of collective noun constructions, e.g. een aantal bomen ‘a number of trees’. Such constructions are possibly discontinuous, e.g. een groot aantal oude bomen ‘a large amount of old trees’. Making use of a treebank instead of a ‘flat’ corpus facilitates retrieving those interrupted examples as well.

If one would query the treebank for collective noun constructions using a formal query language, the query would be something like:

```xml
//node[@cat="np" and node[@rel="det" and @cat="np" and node[@rel="det" and @pos="det" and @root="een"] and node[@rel="hd" and @pos="noun"]]
and node[@rel="hd" and @pos="noun"]]
```

But for GrETEL, a natural language example such as een aantal bomen is sufficient. In our presentation we will show how collective noun constructions similar to the examples above can easily be extracted from the Dutch LASSY [van Noord et al., 2013] and CGN [Hoekstra et al., 2003] treebanks using GrETEL. Moreover, we will indicate how some simple fine-tuning of the input construction can give you control over the search results.

References


http://nederbooms.ccl.kuleuven.be/eng/gretel