Enriching a grammatical database with intelligent links to linguistic resources

Ton van der Wouden, Gosse Bouma, Matje van de Kamp, Marjo van Koppen, Frank Landsbergen, and Jan Odijk
The Taalportaal project (www.taalportaal.org) is a collaboration of the Meertens Institute, the Fryske Akademy, the Institute of Dutch Lexicology and Leiden University, funded, to a large extent, by the Netherlands Organisation for Scientific Research (NWO). The project is aimed at the development of a comprehensive and authoritative scientific grammar for Dutch and Frisian in the form of a virtual language institute.
2.1. Prepositional complements

The examples in (2) show that complements of adjectives are normally PPs, which are often optional.

Example 2

a. Jan is niet boos (over die opmerking).
   Jan is not angry about that remark

b. Jan is niet tevreden (over zijn beloning).
   Jan is not satisfied with his reward

The examples in (3) further show that complements of adjectives can sometimes be clauses, which are introduced by the (often optional) anticipatory pronominal PP er + P'P + it'; given that, due to the phonological weakness of the pronominal element er, this PP is normally split by means of R-extraction, we will use italics to make it easier for the reader to detect the constituting parts of this PP.

Example 3

a. Jan is (er) boos (over) dat Peter niet gekomen is.
   Jan is angry (about) that Peter not come is
   'Jan is angry (about it) that Peter didn’t come.'

b. Jan is (er) tevreden (over) dat hij ontvangen is.
   Jan is satisfied with that he received is
   'Jan is satisfied (about it) that he has been received.'

The option of having an anticipatory pronominal PPs in (3) indicates that these examples are related to the examples in (2). The following two subsections will discuss examples of this sort more extensively.
In his largely positive review of the SoD volumes on NP syntax in *LINGUA* (2013), Jack Hoeksema points out that “There is a growing body of work in empirical studies of judgment variation [...] that future extensions of this grammar could benefit from, especially when coupled to studies of actual usage patterns in corpus material”. By enriching the on-line version of SoD with queries over syntactically annotated corpora, the current project tries to accommodate the needs of researchers like Hoeksema.
Some relevant resources

- The (syntactically annotated part of the) Corpus of Spoken Dutch (manually verified syntactic annotation for 1M words of speech)
- The Lassy Small treebank (manually verified syntactic annotation for 1M words of text from various genres)
- The Lassy Large treebank (automatically created syntactic annotation using the Alpino parser for 700M words of text)
Accessing the corpora

• GrETEL  
http://gretel.ccl.kuleuven.be/gretel-2.0/

• PAQU http://zardoz.service.rug.nl:8067/
Example: the possibility of preceding PP-complements:

*Jan was boos over deze opmerking*
*Jan was angry over this remark*
*Jan was over deze opmerking boos*

```
//node[@cat="ap"]/node[@rel="hd"
    and @pt="adj"
    and ../node[@rel="pc"
        and cat="pp"]/number(@end)=number(@begin)]
```

This query selects the adjectival head of an AP, while the PP-sister of the adjective has to *end* in a position that is equal to the *begin* position of the adjective.
This query has 85 matches in Lassy Small, for 30 different adjectives. However, the adjective *boos* is not among them.

After restricting the query to the lemma *boos*, while removing the linear order constraint, we find that there are 76 hits for *boos*+PP in Lassy Large, but only one for PP+*boos*:

*Leopold II was over die aantasting ... bijzonder boos*

*Leopold II was over that violation ... extremely angry*

This result shows that there is a sharp contrast in frequency between both word orders.
Open issues

• Morphology and phonology: 
  https://portal.clarin.inl.nl/webcelex/
Phonology

• Phonotaxis
  – onsets: hiatus and complex onsets of more than two consonants
  – rhyme: A and B class vowels and diphtongs in open syllables and one consonant coda’s
  – Sonority Sequencing Generalization

• phonological processes
  – final divoicing
  – regressive and progressive assimilation
  – degemination
Morphology

- **Inlection:**
  - Comparative and superlative

- **Derivation:**
  - Suffixes: category-changing (-baar, -heid en -tje) and category-neutral (-schap)
  - Prefixes: be-, ver- (category-changing) and on- and oer- (category-neutral)
  - Non-native prefixes (pro- and sub-) and suffixes (-iteit) and (-esk)

- **Compounding:**
  - NN (stoelpoot), VA (kotsmisselijk) and ANN (blotevrouwenblad)
  - Separable complex verbs (SCVs): schoonmaken and aanvallen
Concluding remarks

• Proof of concept: the Taalportaal can be enriched with intelligent links to linguistic resources.
• The Taalportaal thus becomes an even more valuable tool.
• Linguists are pointed at the existence and usefulness of search interfaces developed in the CLARIN infrastructure such as PaQu, GrETEL, OpenSONAR and WEBCELEX.
• By redirecting the user to these front-ends, the further use of these applications in the CLARIN infrastructure for modifying queries or submitting new queries is stimulated.