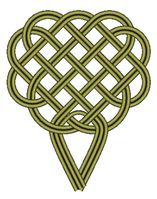
Syntactic Annotation of Large Corpora in STEVIN - LASSY

Gertjan van Noord, Erik Tjong Kim Sang, Gosse Bouma (University of Groningen)
Ineke Schuurman, Vincent Vandeghinste, Frank van Eynde (KU Leuven)

Background

- STEVIN: `taalunieversum.org/taal/technologie/stevin/`
- Goal: 500-million-word reference corpus of written Dutch
- Lassy: 1 million words syntactically annotated (manual)
- Lassy: 500 million words syntactically annotated (automatic)
- Syntactic annotations in Lassy are input for semantic annotations in SoNaR

Alpino parser



- wide-coverage HPSG for Dutch
- large lexicon; rules for named entities
- heuristics for unknown words
- left-corner parser constructs parse forest; best-first search selects best parse using statistical disambiguation model (trained on treebank)

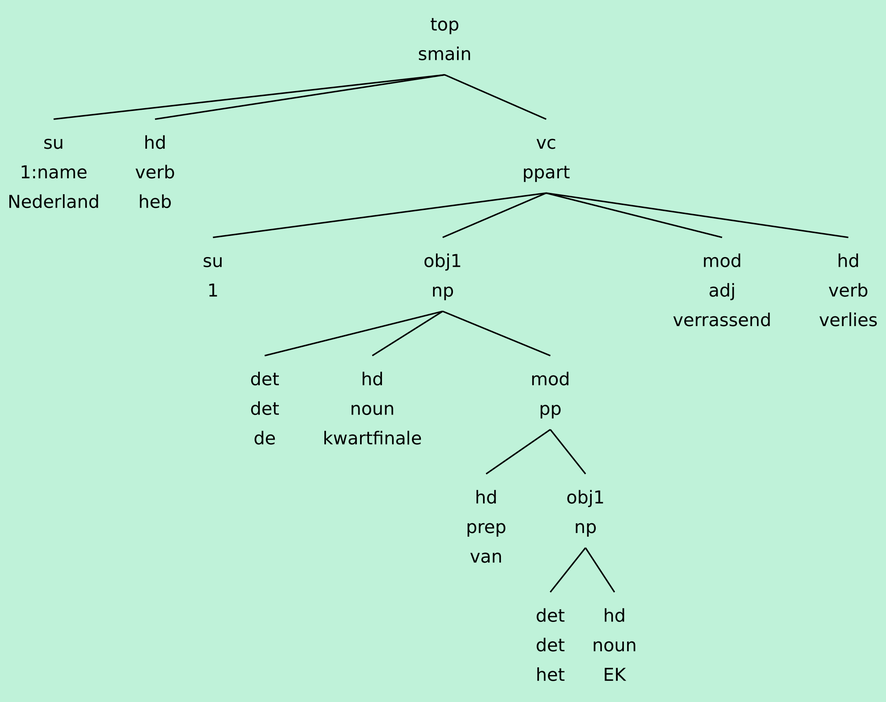
corpus	sents	len	Acc%	F-sc%
Alpino treebank (Eindhoven cdbl)	7136	20	89.88	90.24
D-Coi	12390	16	85.98	86.78

Accuracy of Alpino on Alpino treebank, and on all manually annotated D-Coi treebanks. The table lists the number of sentences, mean sentence length (in tokens), labeled dependency accuracy and F-score.

Annotation Guidelines

- inherited from CGN (Corpus of Spoken Dutch)
- Additions for written language inherited from D-Coi
- Some minor differences remain (annotation of subject in auxiliary, modal constructions; object in passive constructions)

(1) Nederland heeft de kwartfinale van het EK verrassend verloren
Netherlands has the quarter-final of the EC surprisingly lost
The Netherlands surprisingly lost the quarter final of the EC



XML Format

- straightforward representation in XML where recursive nature of dependency structure is represented using embedding in XML
- allows full XPATH/Xquery search with standard tools
- CGN treebank available in this format as well
- For automatically annotated treebanks, we use compressed archives of XML files, with pseudo random access (`dictd`)
- Fully parsed TwNC-04, Wikipedia, Europarl, Mediargus

Annotation Tools

Off-line annotation: Use Alpino off-line to construct dependency structures for potentially large set of corpus sentences. Manually verify and edit any mistakes using TrEd editor.

Interactive annotation with Alpino: Use Alpino interactively to parse corpus sentences one at the time. Guide Alpino to the correct parse, using various tools:

- Interactive assignment and selection of lexical categories
- Assignment of *some* syntactic brackets in the input

```
I saw [ @np the man ] with the telescope
I saw [ @np the man with the telescope ]
```

- Find all or best N parses, select best parse using parse selection tool

Additional checks: Sets of manually verified dependency structures undergo further checks:

- apply XML tools for consistency checking, spot frequent mistakes
- XML tools for browsing and searching in large archives of dependency structures, using XPATH and XQUERY
- Second manual verification check

Annotation Efforts

layer	annotated	target
lemmatization	560	500
POS-tagging	560	500
Syntactic	614	800

Progress of Annotation Efforts (April 2008). Numbers in Kilo-words.

Applications of Huge Treebanks

- Question Answering
  - Joost: Dutch QA system based on Alpino
  - best result for Dutch in 2005, 2006 and 2007
- corpus linguistics
- machine translation
- disambiguation for parsing: find selectional restrictions

high scoring verb-obj1 pairs bijltje neergooien, duimschroeven aandraaien, peentjes zweten, traantjes wegpinken, boontjes doppen, centjes bijverdienen, champagneflessen ontkurken, dorst lessen ...

high scoring verb-mod pairs overlangs doorsnijden, welig tieren, dunnetjes overdoen, stiefmoederlijk bedelen, onzedelijk betasten, stierlijk vervelen, cum-laude afstuderen, hermetisch afgrendelen, ingespannen turen, instemmend knikken, kostelijk amuseren ...

high scoring noun-mod pairs in-vitro fertilisatie, Hubble ruimtetelescoop, zelfrijzend bakmeel, bezttelijk voornaamwoord, ingegroeide teennagels, knapperend haarvuur, levendbarende hagedis, onbevlekte ontvangenis, ongebluste kalk ...

- information extraction & ontology building
  - distributional similarity, vector-based methods
  - use dependency relations as contexts
  - display 20 nearest neighbors:

Beatles Rolling Stones, Stones, John Lennon, Jimi Hendrix, Tina Turner, Bob Dylan, Elvis Presley, Michael Jackson, The Beatles, David Bowie, Prince, Genesis, Mick Jagger, The Who, Elton John, Barbra Streisand, Led Zeppelin, Eric Clapton, Diana Ross, Janis Joplin

Sony Matsushita, Toshiba, Time Warner, JVC, Hitachi, Nokia, Samsung, Motorola, Philips, Siemens, Apple, Canon, IBM, PolyGram, Thomson, Mitsubishi, Kodak, Pioneer, AT&T, Sharp

Hinault Kübler, Vermandel, Bruyère, Depredomme, Mottiat, Merckx, Depoorter, De Bruyne, Argentin, Schepers, Criquielion, Dierickx, Van Steenbergen, Kint, Bartali, Ockers, Coppi, Fignon, Kelly, De Vlaeminck

More Info

- `www.let.rug.nl/~vannoord/Lassy/`