Universiteit Utrecht

Extending Alpino with Flexible MWEs
An unexpected possible use of PaQu

Jan Odijk

CLARIN Conference 2015, Bazaar

Wrocław, 2015-10-16

Overview

- Alpino Dutch parser
- PaQU
- Minitreebank
- Recognizing flexible MWEs
- Examples:
  - *de plaat poetsen*, *(een) bok schieten*, (*een) flater slaan,*
  - *(een) bok geschoten hebben, iets (niet) kunnen velen*
- Conclusions

Alpino

- Parser for Dutch
- Does not deal with MWEs except for some fixed MWEs
- http://www.let.rug.nl/vannoord/alp/Alpino/

PaQu

- Search interface to (Dutch) treebanks
- Enables you to
  - upload your own corpus
  - have it parsed by Alpino
  - and then to search in it
- http://portal.clarin.nl/node/4182

Flexible MWEs

- MWE:
  - word combination with idiosyncratic linguistic properties
- Flexible:
  - No fixed word order, not necessarily contiguous, inflectional variation (in principle) possible on multiple words
- Examples
  - *Hij heeft gisteren **de plaat** ge**poets**t*
  - *Hij **poet**ste gisteren **de plaat***
  - *…dat hij **de plaat** wilde **poets**en*

Recognizing MWEs

- MWE same synt str. as literal
- MWE in a sentence
  - replace it in semantics or deep syntax
    - by a single node (opaque idioms)
    - By multiple nodes ((semi-)transparent idioms)
  - Replacement: not here
- MWE identification: XPATH Query

Minitreebank

- 35 example sentences relevant for MWEs (mostly V+NP idioms)
- Go to PaQu
  - Login with your e-mail address
  - Click on the link in the mail sent to you
  - Select Corpus: PARSEME MWE Test
- List all sentences: //node[@cat="top"]
- Download fully parsed corpus

Minitreebank

- Morpho-syntactic phenomena:
  - Verb: Vfinal, V2, V1 (YNQ), VR
  - NP: Topicalisation, Wh-questioning, scrambling, independent occurrence
  - N: (internal) modification, relativisation, diminutives, plural, other dets
  - V+NP: passive

De plaat

‘the plate’

//node[@lemma="plaat"]

De plaat poetsen

‘to polish the plate’ (‘to bolt’)

```
//node[node[@rel="su"] and
      node[@pos="verb" and
@lemma="poetsen" and @root="poets"] and
node[node[@rel="det" and @lemma="de"] and
node[@rel="hd" and @lemma="plaat" and
@getal="ev" and @gen="de"  and
@graad="basis" and @naamval="stan" and
@ntype="soort" and @pos="noun"  ]]]
```

bok

‘male goat’

//node[@lemma="bok"]

(een) bok schieten

To shoot a male goat (‘to make a blunder’)

```
//node[node[@rel="su"] and
      node[@pos="verb" and
@lemma="schieten" and @root="schiet"] and
node[node[@rel="hd" and @lemma="bok" and
@gen="de"  and @naamval="stan" and
@ntype="soort" and @pos="noun"  ]]]
```

flater

blunder

//node[@lemma="flater"]

(een) flater slaan

hit a blunder (make a blunder)

```
//node[node[@rel="su"] and
      node[@pos="verb" and @lemma="slaan"
and @root="sla"] and
node[node[@rel="hd" and @lemma="flater" and
@gen="de"  and @naamval="stan" and
@ntype="soort" and @pos="noun"  ]]]
```

een bok geschoten hebben

To have shot a male goat (‘to have made a blunder)

```
//node[
   node[@rel="su"] and
   node[@lemma="hebben" and @rel="hd" and @pos="verb" ] and
   node[@cat="ppart" and @rel="vc" and
         node[@rel="su" and
              @index=../../node[@rel="su"]/@index]  and
         node[@rel="obj1" and @cat="np" and
                  node[@rel="hd" and @lemma="bok"]
            ] and
         node[ @buiging="zonder" and
               @lemma="schieten" and
               @pos="verb"       and
               @wvorm="vd"
            ]
      ]
]
```

lets (niet) kunnen velen

- (*not) be able to endure something*
- *Velen* as a verb is a *cranberry word*
- Alpino parses it as a pronoun
  - ‘many (persons)’
- Alpino lexicon must be extended

Conclusions

- Extensions/adaptations of Alpino needed:
- Overacceptance (extension needed)
  - Modification of *plaat*
  - Scrambling of *de plaat,* Topicalisation of *de plaat*
  - Passivisation (adaptation+MWE property)
- Underacceptance
  - Relativisation, wh-q of *bok*, *flater* (adaptation)
  - Passivisation (adaptation)
  - MWEs with cranberry words (extension)

Caveats

- Internal Alpino structure / analysis may be different
- Rules in the Alpino system may differ from XPATH queries (esp. wrt matching properties)
- Maybe sufficient for analysis (parsing), but not for production.