LASSY for Beginners

Erik Tjong Kim Sang, Gosse Bouma and Gertjan van Noord

5 February 2010

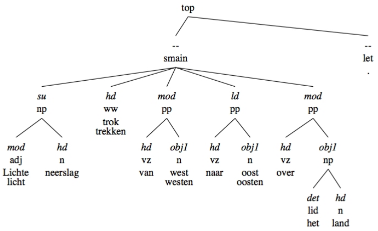
LASSY

Currently a large corpus of parsed Dutch text is being developed by the universities of Groningen and Leuven in the LASSY project (STEVIN)

The corpus will consist of a manually checked part of 1 million words and an automatically annotated part of 500 million words

The annotations are specified in XML (see next example)

Example tree from the LASSY corpus



Corpus example in XML

```
...
  <node cat="np" rel="su">
     <node postag="adj" rel="mod" lemma="licht" word="Lichte"/>
     <node postag="n" rel="hd" lemma="neerslag" word="neerslag"/>
  </node>
  <node postag="ww" rel="hd" lemma="trekken" word="trok"/>
  <node cat="pp" rel="mod">
    <node postag="vz" rel="hd" lemma="van" word="van"/>
    <node postag="n" rel="obj1" lemma="westen" word="west"/>
  </node>
...
<sentence>Lichte neerslag trok van west naar oost...</sentence>
```

Corpus tasks

Preliminary versions of the LASSY corpus have already been used in a variety of computational linguistics studies:

- automatic extraction of hypernym-hyponym pairs (Tjong Kim Sang & Hofmann, 2009)
- distribution of strong and weak reflexive pronouns (Bouma & Spenader, 2009)
- improvement of parse selection process (Van Noord, 2010)

Each of these studies required extracting different relevant material from the corpus

Corpus queries

We used XQuery for extracting relevant material from the XML files in the LASSY corpus

XQuery is a programming language which can be used for processing XML data.

The language contains instructions for traversing hierarchical structures encoded in XML and selecting XML nodes and their attributes

XQuery example

```
(: basic XQuery commands for extracting pairs of nouns :)
for $node1 in //node, $node2 in //node
where
  $node1/@postag = "noun" and    (: $node1 should contain noun :)
  $node2/@postag = "noun" and    (: $node1 should contain noun :)
  (: require words of $node1 to precede those of $node2 :)
  $node1/@end cast as xs:integer <=
    $node2/@begin cast as xs:integer
return
  <node>
    <pair word1="$node1/@word" word2="$node2/@word />
  </node>
```

Alternatives for XQuery

XQuery is a strong programming environment which can be used for extracting arbitrary interesting information from XML data

However, it is too complex for novice users

Furthermore, when applied to a large corpus, it requires processing times of minutes or even hours

Can we make the contents of the LASSY available in an easy and fast way?

What kind of basic queries can we expect?

- searching for a word
- searching for a word with a specific part-of-speech tag
- searching for words in a particular relation
- searching for different forms of a certain lemma
- searching for a phrase

These are the types of queries which should be able to be applied to the corpus in an easy way

LASSY Search Demo

The LASSY Search Demo makes all these queries possible without having to download the LASSY corpus

It is available at

`http://www.let.rug.nl/erikt/lassy`

Concluding remarks

- We have presented a search facility for the LASSY corpus
- It offers access to individual dependency relations
- Search queries are shaped like five-tuples or SQL database queries
- Complex queries will still require more complex search environments like dtsearch and XQuery

THE END