The IMS Corpus WorkBench

1 CWB basics

1.1 History

- Institut für Maschinelle Sprachverarbeitung of the University of Stuttgart
- Early to mid 90s: Oliver Christ
- Late 90s to 2005: Stefan Evert
- From 2006: open source project led by Stefan Evert, hosted on SourceForge
- `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`
  `http://cwb.sourceforge.net/`

1.2 The CWB toolkit

- Toolkit of command-line programs
- Tools to encode/index corpus
- Tools to explore corpus; in particular:
  - CQP, the *corpus query processor* for interactive exploration of corpus
  - `cwb-scan-corpus` for non-interactive extraction of frequency data
- Supported on most Unix platforms: Linux, Mac OS X, Solaris
- Programmatic interface to develop, e.g., Web-based front-end

1.3 Advantages over alternatives

- Alternatives: Word Sketch Engine, Xaira, WordSmith...
- Only CWB satisfies all of following requirements:
  - Scaling up to very large corpora
  - Flexible, annotation-aware queries
  - Flexible input format

  - Central storage of corpora
  - Command-line interface for easy interaction with other tools
  - Free, open source, active support and documentation community

1.4 Problems

- At the moment, corpora larger than about 400M tokens will have to be split into sub-corpora
- No standard Web interface supporting full (or even sizable subset of) CQP options
- (Virtually) no query optimization, i.e., `[pos="V.*"][lemma="dog"]` will be much slower than `[lemma="dog"][pos="V.*"]`
- Ongoing work on first two issues

2 Corpus representation

- Positional attributes: properties of words, e.g., pos and lemma
- Structural attributes: meta-data and constituency information
- Possible input 1:

```
The
dog
barks
```

- Possible input 2

```
The    ART  the
dog    NN   dog
barks  VV   bark
```

- Possible input 3

```
<s>
The    ART  the
dog    NN   dog
barks  VV   bark
</s>
```

- Possible input 4

  ```
  <text title="poem" author_sex="m">
  <s>
  The    ART   the
  dog    NN    dog
  barks  VV    bark
  </s>
  </text>
  ```

- Possible input 5

  ```
  <text title="poem" author_sex="m">
  <s>
  <np>
  The    ART   the
  dog    NN    dog
  </np>
  <vp>
  barks  VV    bark
  </vp>
  </s>
  </text>
  ```

- Possible input 6

  ```
  The     n
  dog     y
  barks   n
  ```

- and so on!

3 The IMS corpus creation pipe

- Save corpus document(s) as plain text
- Tag and lemmatize with TreeTagger[1]
- Index with CWB
- Enjoy!
- Often, literally a matter of minutes

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html