The IMS Corpus WorkBench

Marco Baroni

University of Bologna

Granada “Morphology and Corpora” Seminar

The IMS Corpus WorkBench

- 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/`

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)
- Supported on most Unix platforms: Linux, Mac OS X, Solaris
- Programmatic interface to develop, e.g., Web-based front-end

Advantages over alternatives

- Alternatives: WordSketch 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

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

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
```

Possible input 7...

...

The IMS corpus creation pipe

- Save corpus document(s) as plain text
- Tag and lemmatize with TreeTagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`)
- Index with CWB
- Enjoy!
- Often, literally a matter of minutes