Tree Tagger: Overview

- Tool for annotating text with part-of-speech and lemma information
- Easily adaptable to languages if a lexicon and a manually tagged training corpus are available
- Languages so far: German, English, (old) French, Italian, Spanish, Bulgarian, Russian, Greek, Portuguese
- Executable code for Sparc workstations, Linux and Windows PCs, and Macs
- Project Textual Corpora and Tools for their Exploitation, Institut für Maschinelle Sprachverarbeitung, Stuttgart

Tree Tagger: Method

- The decision tree automatically determines the appropriate size of the context to estimate the part-of-speech transition probabilities.
- Possible contexts include bigrams, trigrams, etc., as well as negations of them, such as \( \text{tag}_{-1} \neq \text{DET} \).
- The probability of an \( n \)-gram is determined by following the corresponding path through the tree until a leaf is reached.
- Improves on sparse data, compared to Markov Models; avoids zero frequencies

Tree Tagger: Example Decisions

```
+----------------+-------------------+
|               |       |Device         |
|               |  yes  |               |
|               |  no   |               |
+----------------+-------------------+
```

Tree Tagger: References

- Tree-Tagger download: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html