

Introduction to Corpus Resources, Annotation and Access: *Tree Tagger*

Sabine Schulte im Walde
Universität des Saarlandes

Heike Zinsmeister
Universität Tübingen

Foundational Course
18th European Summer School in Logic, Language and Information
Málaga, Spain
July 31 - August 4, 2006

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

Schulte im Walde & Zinsmeister

2

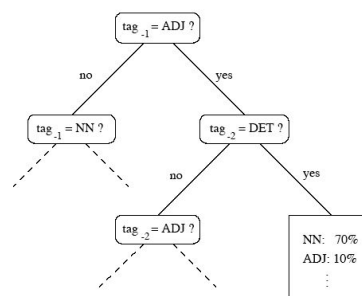
Tree Tagger: Method

- 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 $tag_{-1} \neq DET$.
- 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

Schulte im Walde & Zinsmeister

3

Tree Tagger: Example Decisions



Schulte im Walde & Zinsmeister

4

Tree Tagger: References

- Helmut Schmid (1994): "Probabilistic part-of-speech tagging using Decision Trees". In *Proceedings of the International Conference on New Methods in Language Processing*.
- Helmut Schmid (1995): "Improvements in part-of-speech tagging with an application to German". In *Proceedings of the ACL SIGDAT Workshop*.
- **Tree-Tagger** download: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

Schulte im Walde & Zinsmeister

5