Universität Hamburg, Department of Informatics

Because Size Does Matter: The Hamburg Dependency Treebank

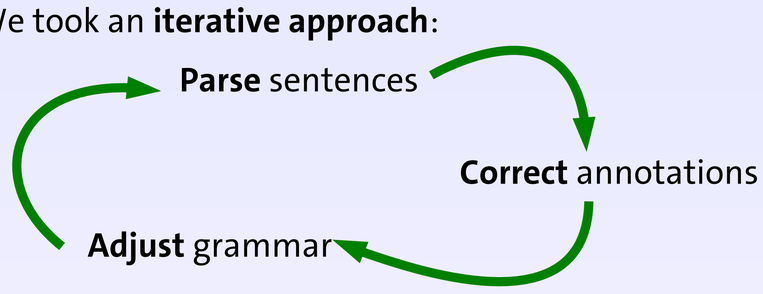
Kilian A. Foth, Arne Köhn, Niels Beuck, Wolfgang Menzel
{foth, koehn, beuck, menzel}@informatik.uni-hamburg.de

HDT in a nutshell

- **Source**: IT-news articles from 1996 to 2001 (heise.de)
- **Largest dependency treebank** available
  - Twice as large as the Prague Dependency Treebank
  - Three times as large as the TIGER treebank and the PTB2
- **Free for scientific/academic use**
- ~ **261,000** German sentences with syntax annotation
- ~ **4 million** hand-annotated tokens
- In development since 2001
- **Genuine dependency annotation**, i.e. not converted from phrase structure
- **Three classes of annotation**:
  - 102k sentences manually annotated and cross-checked (A)
  - 105k sentences manually annotated (B)
  - 55k sentence automatically parsed (C)

The Annotation Process

Main goal: a Weighted Constraint Dependency Grammar for German
We took an **iterative approach**:



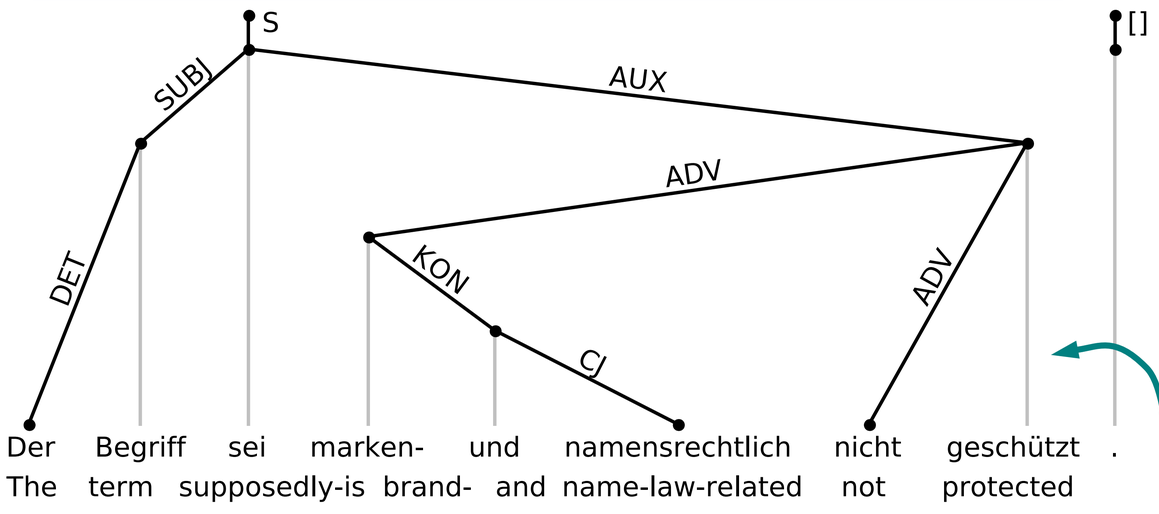
- Parsed sentences are inspected & annotations corrected
- WCDG is adapted to favor the corrected analysis
- Parsing continues with the adapted WCDG
- Regularly re-parse old sentences to make sure that no errors are introduced into the WCDG

Result: A grammar and an annotated corpus (the HDT)

The Annotation Scheme

Target: provide **robust coverage** of phenomena that occur repeatedly in normal written text, **reflect the limit of** the disambiguating decisions **syntax-based dependency parsers** can reasonably make.

- **PoS annotated** using the **S**tuttgart-**T**übingen **T**ag**S**et
- 34 **dependency labels** on the syntax level
- One level for pronouns attached to their antecedent
- **Morphological information**
  - Case
  - Gender
  - Number
  - Etc.



Quality Assurance

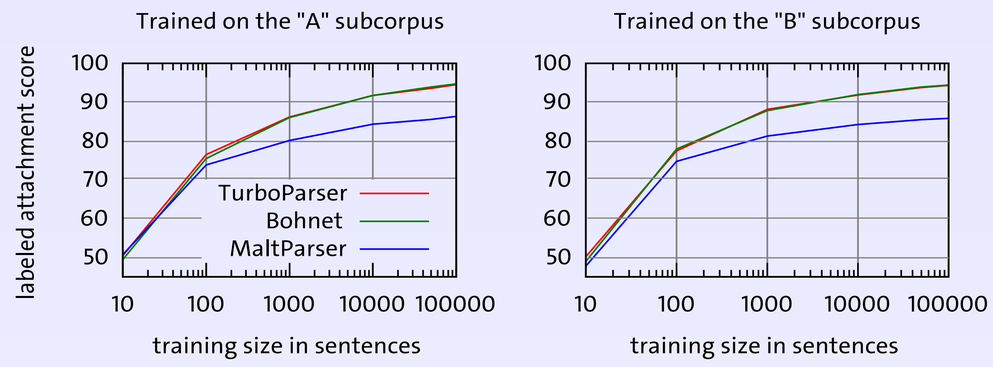
- Part A was cross-checked with the DECCA toolkit (Boyd et al., 2008)
  - Checks for consistency of PoS tags and dependency labels
  - Highlights different annotations in similar context
- 8495 word pairs pointed out
- In 1931 of them at least one occurrence was indeed erroneous
- Resulting precision of the automatic consistency check: 22.7%
- Checking with DECCA led to adjustments of 4% of the sentences

Statistics

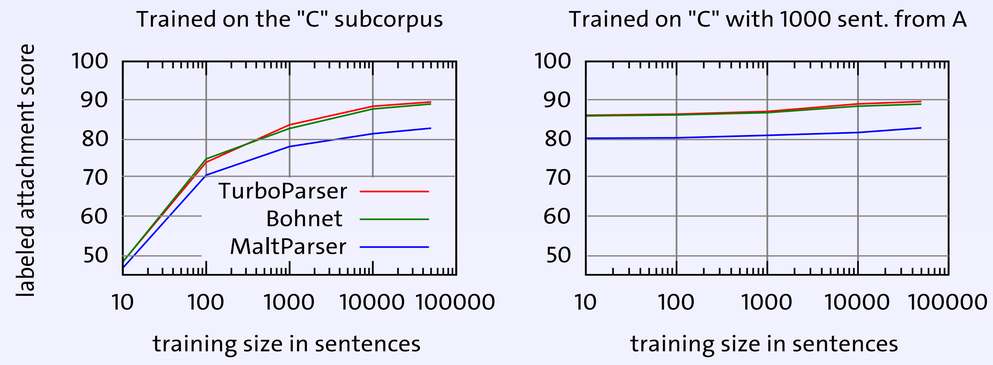
- **Average sentence length**: 18.4 tokens
- 130,933 **different word forms**
- 77,397 of them appear **only once** (e.g. 3,5-ZOLL-Wechselplatte, 3.5 inch removable hard disk drive)
- 12.52% **non-projective**, 10.89% **non-planar**, 0,51% **ill-nested**
- Dependency label highly correlated with PoS of head & dependent
  - can be guessed with an accuracy of 91% from that alone
  - Prediction of head PoS with dependent PoS: 49% accuracy

Parser evaluation

What effect does data **quality** and **quantity** have on parsing performance?



- No big difference between A and B subcorpus
- Parsers **differ** in their **ability to profit from additional data**
  - **More training data** is clearly **beneficial**
- High parsing accuracies suggest **low noise in annotation**



- Parsers were all able to achieve WCDG level accuracy trained on WCDG-parsed sentences
- Small set of high-quality annotations worth more than low quality ones

Tools

- Transformation to CoNNL-X format
- Statistics generation scripts
- Web-based corpus search with WCDG constraints
- SVG generator for “real” trees

Get the HDT at http://nats-www.informatik.uni-hamburg.de/HDT/