Treebanks –
Formats, Tools and Usages

Martin Volk
volk@ling.su.se
Stockholm University

First Treebank Course

in March 2004 sponsored by NorFa's Nordic Treebank Network

- Participants from
  - Danmark
  - Estonia
  - Finland
  - Iceland
  - Norway
  - Sweden

Course Overview

- The program / how to get credit points
- Goal: By the end of the course you should know
  - some important treebanks (and treebank projects)
  - some treebank representation formats
  - the reasons for building treebanks
  - some tools for building and searching treebanks
  - how to train and evaluate a parser on a treebank

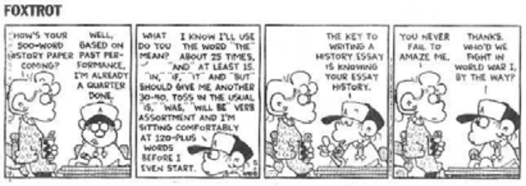
What is a Treebank?

- A Treebank is a corpus with linguistic annotation beyond the word level. The annotation is typically
  - a syntax tree and
  - manually checked and corrected.
- Not a treebank:
  - A corpus with manually checked PoS labels only.
  - An automatically parsed corpus.

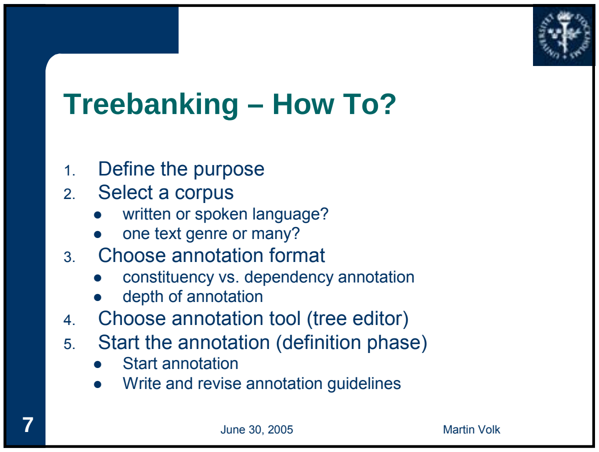
Why Treebanking?

- Providing training material for Machine Learning → NLP systems
- Building Gold Standards for the evaluation of NLP systems.
- Advocating linguistic empiricism against 'other' linguistic theories.
- Providing material for human grammar exploration and learning.

Linguistic empiricism

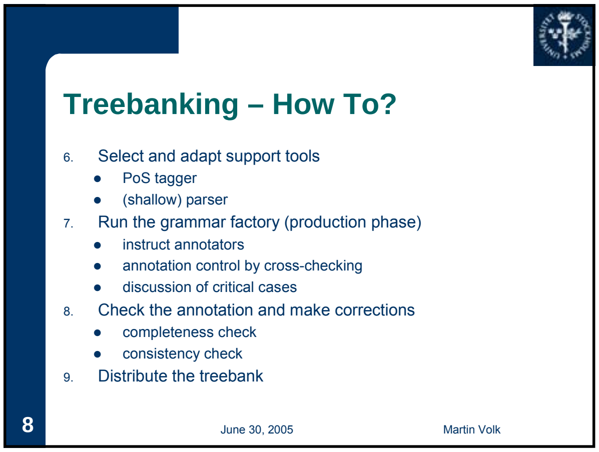

FOXTROT
HOW'S YOUR 500-WORD HISTORY PAPER COMING?
WELL, BASED ON PAST PERFORMANCE, I'M ALREADY A QUARTER DONE.
WHAT DO YOU MEAN?
I KNOW I'LL USE THE WORD "THE" ABOUT 25 TIMES. "AND" AT LEAST 15. "IN," "IF," "IT" AND "BUT" SHOULD GIVE ME ANOTHER 30-40. TOSS IN THE USUAL "IS," "WAS," "WILL BE" VERB ASSORTMENT AND I'M SITTING COMFORTABLY AT 120-PLUS WORDS BEFORE I EVEN START.
THE KEY TO WRITING A HISTORY ESSAY IS KNOWING YOUR ESSAY HISTORY.
YOU NEVER FAIL TO AMAZE ME.
THANKS. WHO'D WE FIGHT IN WORLD WAR I, BY THE WAY?

[Cartoon found by Gerold Schneider, Zurich]



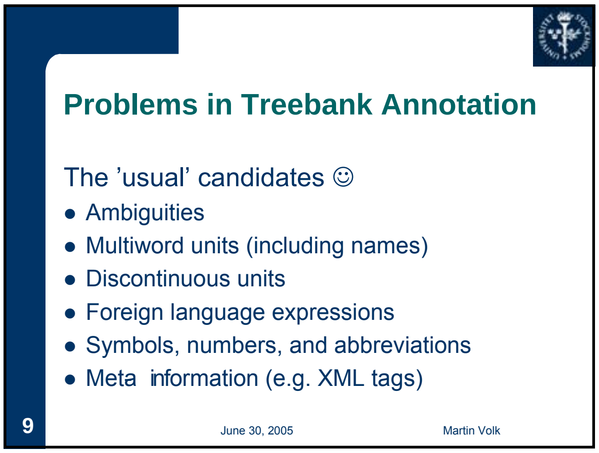
Treebanking – How To?

1. Define the purpose
2. Select a corpus
   - written or spoken language?
   - one text genre or many?
3. Choose annotation format
   - constituency vs. dependency annotation
   - depth of annotation
4. Choose annotation tool (tree editor)
5. Start the annotation (definition phase)
   - Start annotation
   - Write and revise annotation guidelines



Treebanking – How To?

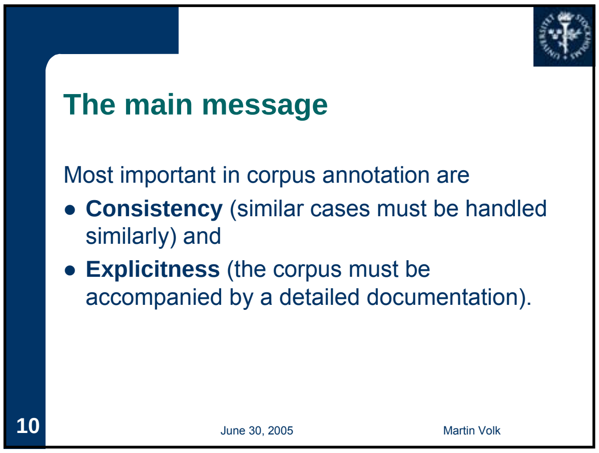
6. Select and adapt support tools
   - PoS tagger
   - (shallow) parser
7. Run the grammar factory (production phase)
   - instruct annotators
   - annotation control by cross-checking
   - discussion of critical cases
8. Check the annotation and make corrections
   - completeness check
   - consistency check
9. Distribute the treebank



Problems in Treebank Annotation

The ’usual’ candidates ☺

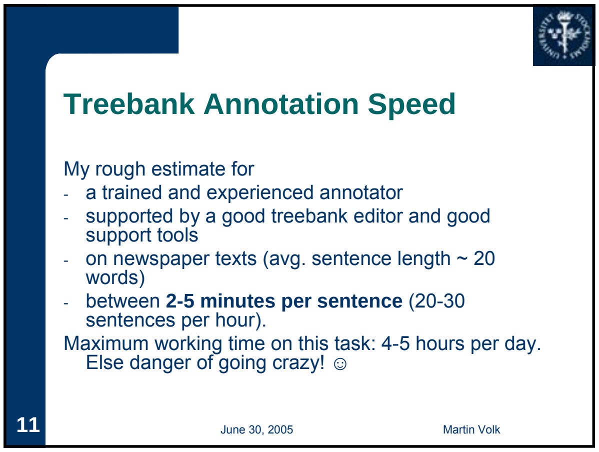
- Ambiguities
- Multiword units (including names)
- Discontinuous units
- Foreign language expressions
- Symbols, numbers, and abbreviations
- Meta information (e.g. XML tags)



The main message

Most important in corpus annotation are

- **Consistency** (similar cases must be handled similarly) and
- **Explicitness** (the corpus must be accompanied by a detailed documentation).

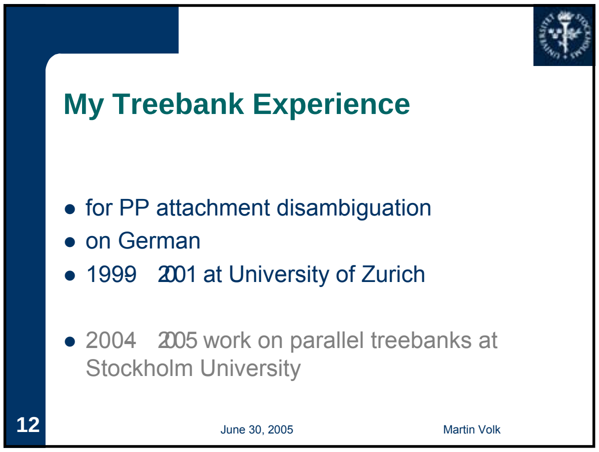


Treebank Annotation Speed

My rough estimate for

- a trained and experienced annotator
- supported by a good treebank editor and good support tools
- on newspaper texts (avg. sentence length ~ 20 words)
- between **2-5 minutes per sentence** (20-30 sentences per hour).

Maximum working time on this task: 4-5 hours per day. Else danger of going crazy! ☺

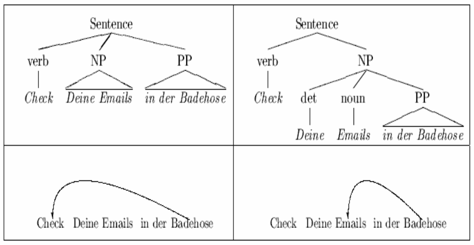


My Treebank Experience

- for PP attachment disambiguation
- on German
- 1999 2001 at University of Zurich

- 2004 2005 work on parallel treebanks at Stockholm University

PP Attachment Disambiguation



Example of Cooccurrence Measure

*For: **Check deine Emails in der Badehose***

freq(*Emails, in*) = 50

freq(*Emails*) = 10'000

→cooc(*Emails, in*) = 0.005

freq(*check, in*) = 15

freq(*check*) = 1'000

→cooc(*check, in*) = 0.015

Training Corpus

Annotate a 6 million words computer journal corpus (raw text) through

1. Proper name recognition
2. PoS-Tagging
3. Lemmatisation
4. NP/PP chunking
5. Clause boundary detection

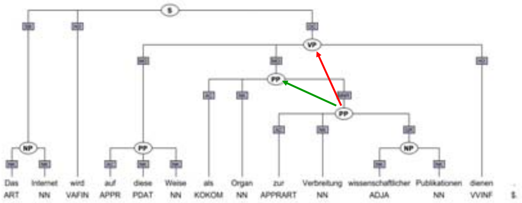
→ Learn cooc(noun,prep) and
cooc(verb,prep)

Evaluation Corpus

The CZ Treebank

- 3000 **manually** annotated German sentences with PPs in ambiguous positions
- from the 1996 ComputerZeitung (CZ)
- annotated at the University of Zurich in 1999
- following the NEGRA guidelines

Sentence with an 'ambiguous' PP

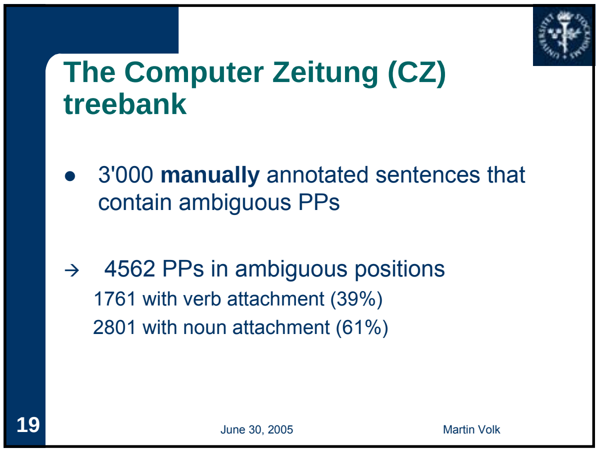


Extraction of 5-tuples from treebank sentences

Sentence:

… als (***Organ (zur Verbreitung (****wiss. Publikationen****))) dienen***

1.	Verb:	*dienen*
2.	Reference noun N1:	*Organ*
3.	Preposition:	*zur*
4.	PP noun N2:	*Verbreitung*
5.	Function:	noun attachment



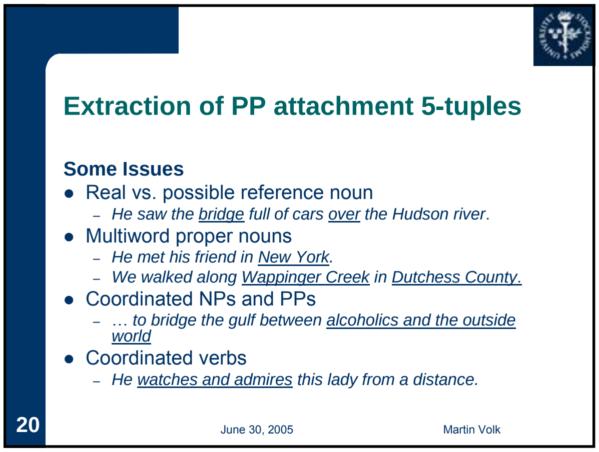
The Computer Zeitung (CZ) treebank

- 3'000 **manually** annotated sentences that contain ambiguous PPs

→ 4562 PPs in ambiguous positions

1761 with verb attachment (39%)

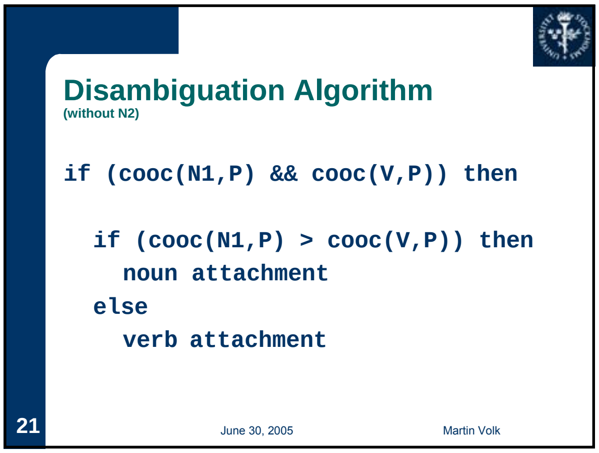
2801 with noun attachment (61%)



Extraction of PP attachment 5-tuples

Some Issues

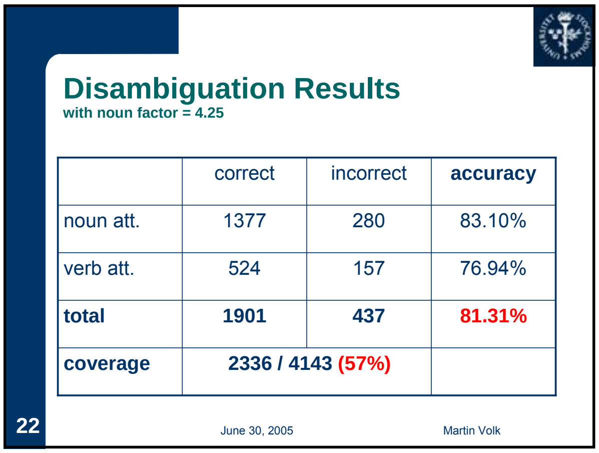
- Real vs. possible reference noun
  - *He saw the <u>bridge</u> full of cars <u>over</u> the Hudson river.*
- Multiword proper nouns
  - *He met his friend in <u>New York</u>.*
  - *We walked along <u>Wappinger Creek</u> in <u>Dutchess County</u>.*
- Coordinated NPs and PPs
  - *… to bridge the gulf between <u>alcoholics and the outside world</u>*
- Coordinated verbs
  - *He <u>watches and admires</u> this lady from a distance.*



Disambiguation Algorithm

(without N2)

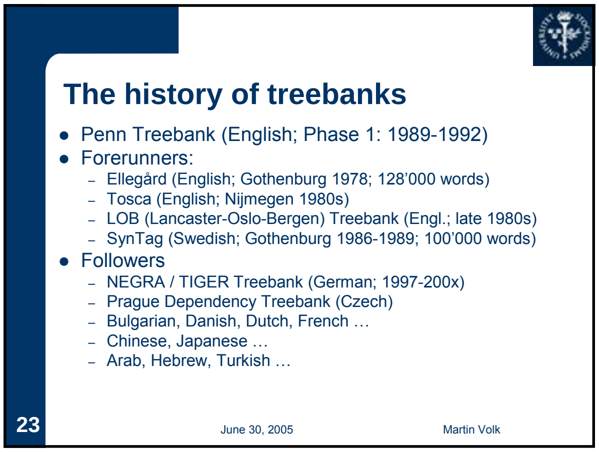
```
if (cooc(N1,P) && cooc(V,P)) then

  if (cooc(N1,P) > cooc(V,P)) then
    noun attachment
  else
    verb attachment
```



Disambiguation Results

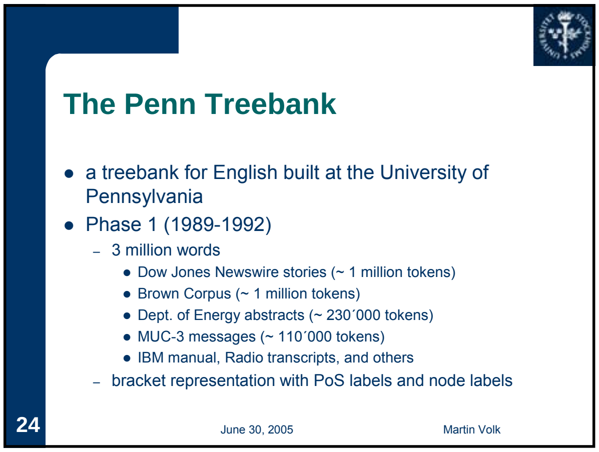
with noun factor = 4.25

	correct	incorrect	**accuracy**
noun att.	1377	280	83.10%
verb att.	524	157	76.94%
total	**1901**	**437**	**81.31%**
coverage	**2336 / 4143 (57%)**		



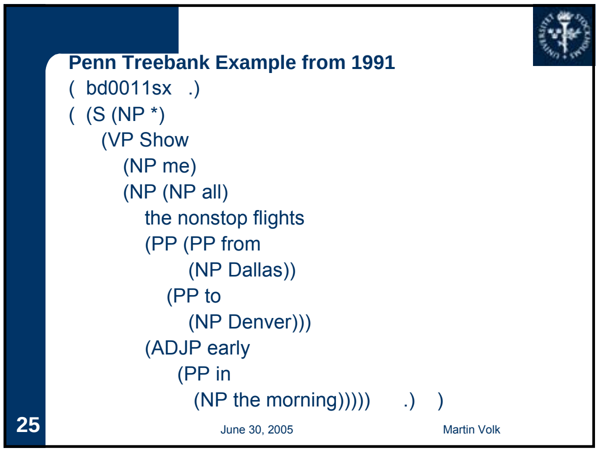
The history of treebanks

- Penn Treebank (English; Phase 1: 1989-1992)
- Forerunners:
  - Ellegård (English; Gothenburg 1978; 128'000 words)
  - Tosca (English; Nijmegen 1980s)
  - LOB (Lancaster-Oslo-Bergen) Treebank (Engl.; late 1980s)
  - SynTag (Swedish; Gothenburg 1986-1989; 100'000 words)
- Followers
  - NEGRA / TIGER Treebank (German; 1997-200x)
  - Prague Dependency Treebank (Czech)
  - Bulgarian, Danish, Dutch, French …
  - Chinese, Japanese …
  - Arab, Hebrew, Turkish …



The Penn Treebank

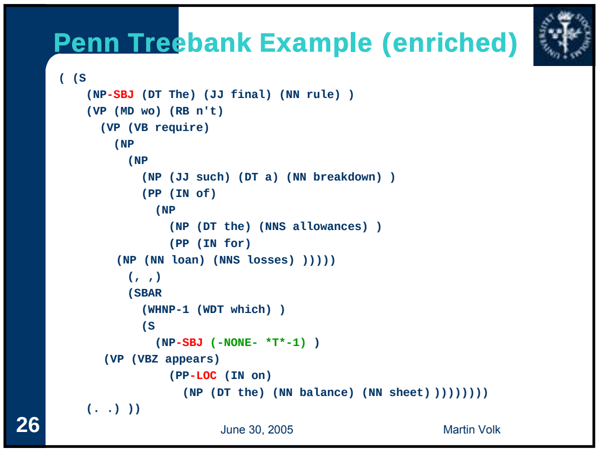
- a treebank for English built at the University of Pennsylvania
- Phase 1 (1989-1992)
  - 3 million words
    - Dow Jones Newswire stories (~ 1 million tokens)
    - Brown Corpus (~ 1 million tokens)
    - Dept. of Energy abstracts (~ 230´000 tokens)
    - MUC-3 messages (~ 110´000 tokens)
    - IBM manual, Radio transcripts, and others
  - bracket representation with PoS labels and node labels



Penn Treebank Example from 1991

```
( bd0011sx   .)
( (S (NP *)
     (VP Show
        (NP me)
        (NP (NP all)
           the nonstop flights
           (PP (PP from
                 (NP Dallas))
              (PP to
                 (NP Denver)))
           (ADJP early
              (PP in
                 (NP the morning)))))     .)   )
```

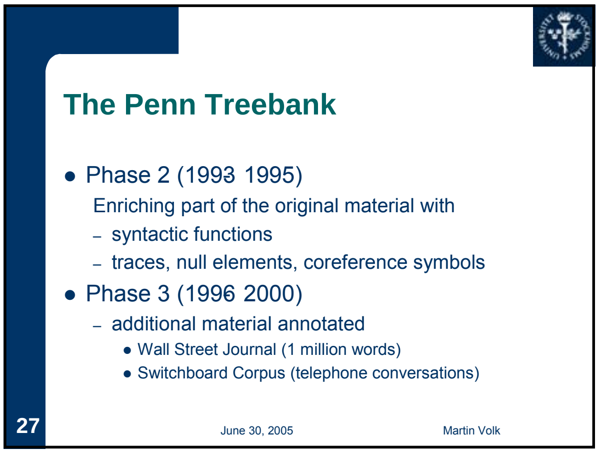
25 June 30, 2005 Martin Volk



Penn Treebank Example (enriched)

```
( (S
    (NP-SBJ (DT The) (JJ final) (NN rule) )
    (VP (MD wo) (RB n't)
      (VP (VB require)
        (NP
          (NP
            (NP (JJ such) (DT a) (NN breakdown) )
            (PP (IN of)
              (NP
                (NP (DT the) (NNS allowances) )
                (PP (IN for)
       (NP (NN loan) (NNS losses) )))))
          (, ,)
          (SBAR
            (WHNP-1 (WDT which) )
            (S
              (NP-SBJ (-NONE- *T*-1) )
     (VP (VBZ appears)
                (PP-LOC (IN on)
                  (NP (DT the) (NN balance) (NN sheet) ))))))))
    (. .) ))
```

26 June 30, 2005 Martin Volk

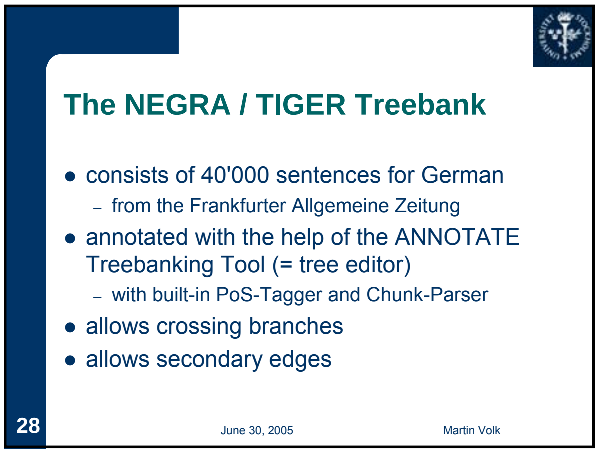


The Penn Treebank

- Phase 2 (1993 1995)

  Enriching part of the original material with
  - syntactic functions
  - traces, null elements, coreference symbols
- Phase 3 (1996 2000)
  - additional material annotated
    - Wall Street Journal (1 million words)
    - Switchboard Corpus (telephone conversations)

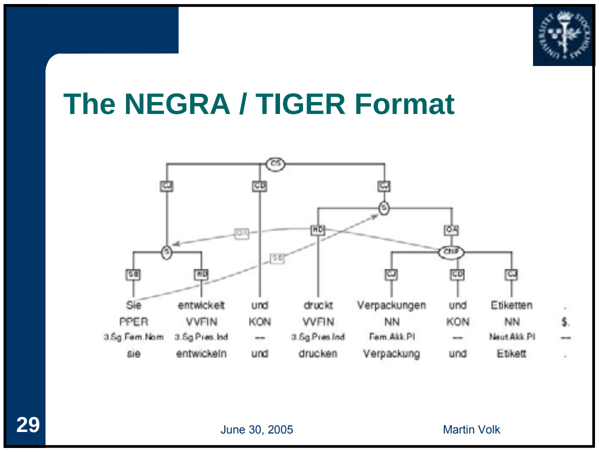
27 June 30, 2005 Martin Volk



The NEGRA / TIGER Treebank

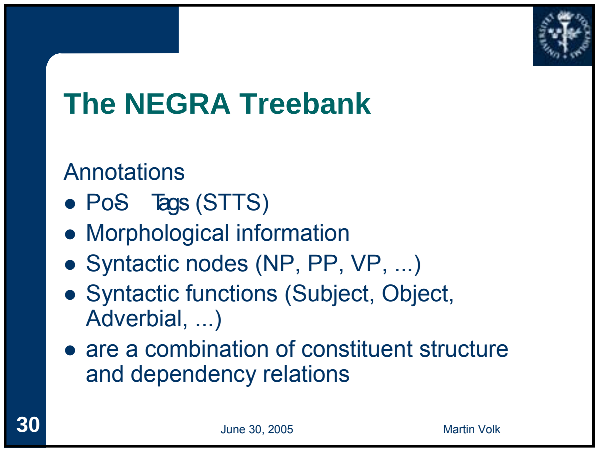
- consists of 40'000 sentences for German
  - from the Frankfurter Allgemeine Zeitung
- annotated with the help of the ANNOTATE Treebanking Tool (= tree editor)
  - with built-in PoS-Tagger and Chunk-Parser
- allows crossing branches
- allows secondary edges

28 June 30, 2005 Martin Volk



The NEGRA / TIGER Format

29 June 30, 2005 Martin Volk



The NEGRA Treebank

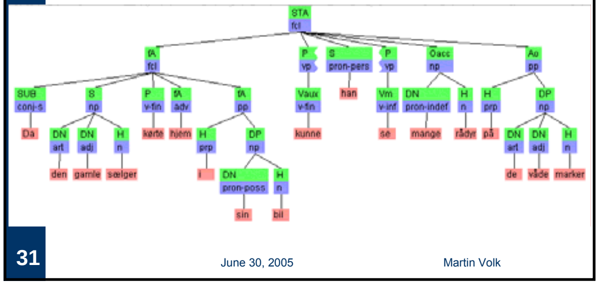
Annotations

- PoS-Tags (STTS)
- Morphological information
- Syntactic nodes (NP, PP, VP, ...)
- Syntactic functions (Subject, Object, Adverbial, ...)
- are a combination of constituent structure and dependency relations

30 June 30, 2005 Martin Volk

Constituent trees for Danish

from Eckhard Bick



Why Treebanking?

Treebanks are at the heart of the Machine Learning paradigm.

My believe: NLP will only make progress

1. if we can combine rule-based systems with machine learning, and
2. if we have standards for evaluation.

Summary

- Central to treebank building
  - Clear annotation guidelines
  - Good treebank editor and support tools
- The Penn Treebank has been the most influential in our field.
- Treebanks have been built for many languages in various formats.