

Compact Course on Tübingen Treebanks

Kathrin Beck and Erhard Hinrichs

kbeck@sfs.uni-tuebingen.de

eh@sfs.uni-tuebingen.de

Dept. of Linguistics
Eberhard Karls University of Tübingen

13/14 December 2010

Course Objectives

- ▶ Provide an introduction of the suite of TüBa treebanks
- ▶ Provide practical advice for constructing a treebank:
 - ▶ useful tools
 - ▶ the do's and don'ts
- ▶ Provide an in-depth look at a treebank with multiple layers of annotation
- ▶ The challenges of treebanking spoken language data

Course Overview

- ▶ Overview of Tübingen Treebanks for Spoken and Written Language
- ▶ From the Treebanker's Cook Book: preparing the data
- ▶ Annotation Layers of the TüBa-D/Z
- ▶ Constructing a valence lexicon from a treebank

Source data of the TüBa-D/Z

- ▶ German daily newspaper *die tageszeitung* (taz)
 - ▶ Comparable corpus
 - ▶ Cooperative editor
 - ▶ Cheap licensing costs for users
 - ▶ Source format: DVD-ROM with HTML files (“scientific edition”)
- ▶ Blocks of newspaper editions between 1992 and 1999
 - ▶ Period before German spelling reform
 - ▶ Same orthography

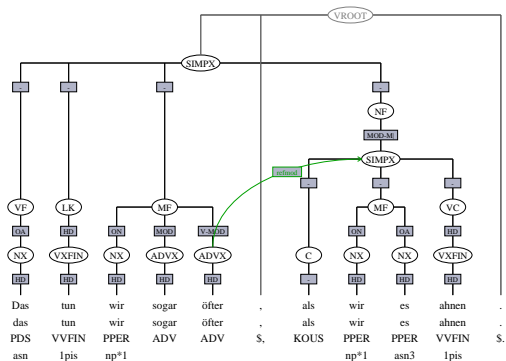
TüBa-D/Z

- ▶ 2,778 newspaper articles
- ▶ 55,814 sentences
- ▶ 976,262 tokens
- ▶ semi-automatic annotation and manual correction of all layers
- ▶ ongoing annotation since 2001

Annotation levels of the TüBa-D/Z Treebank

- ▶ POS tags
- ▶ Morphology
- ▶ Lemmas
- ▶ Syntax
- ▶ Grammatical functions
- ▶ Named Entities
- ▶ Coreference and anaphora

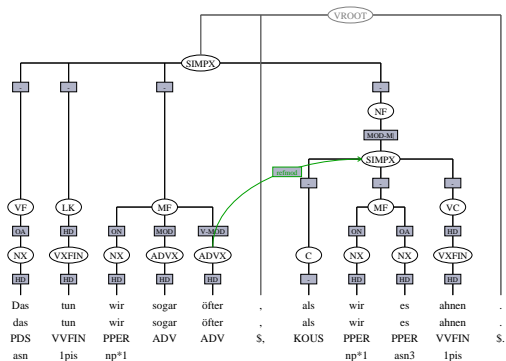
Annotation scheme



That do we even more_often , than we it_expl suspect .

We do that even more often than we suspect.

Annotation scheme



That do we even more_often , than we it_expl suspect .

We do that even more often than we suspect.

The Stuttgart-Tübingen Tagset STTS

- ▶ The STTS is a set of 54 tags for annotating German text corpora with part-of-speech labels.
- ▶ The STTS guidelines (available on the website) explain the use of each tag by illustrative examples to aid human annotators in consistent corpus annotation by STTS tags.
- ▶ It was jointly developed by the Institut für maschinelle Sprachverarbeitung of the University of Stuttgart and the Seminar für Sprachwissenschaft of the University of Tübingen.

The Stuttgart-Tübingen Tagset STTS (2)

- | | |
|--------------------------|--------------------------|
| 1. Nomina (N) | 7. Adverbien (ADV) |
| 2. Verben (V) | 8. Konjunktionen (KO) |
| 3. Artikel (ART) | 9. Adpositionen (AP) |
| 4. Adjektive (ADJ) | 10. Interjektionen (ITJ) |
| 5. Pronomina (P) | 11. Partikeln (PTK) |
| 6. Kardinalzahlen (CARD) | |

Tabelle: Tags for major word classes

STTS Tags

| POS = | Beschreibung | Beispiele |
|---|---|---|
| ADJA ADJD | attributives Adjektiv adverbiales oder prädikatives Adjektiv | <i>[das] große [Haus]</i> <i>[er fährt] schnell</i> <i>[er ist] schnell</i> |
| ADV | Adverb | <i>schon, bald, doch</i> |
| APPR APPRART APPO APZR | Präposition; Zirkumposition links Präposition mit Artikel Postposition Zirkumposition rechts | <i>in [der Stadt], ohne [mich]</i> <i>im [Haus], zur [Sache]</i> <i>[ihm] zufolge, [der Sache] wegen</i> <i>[von jetzt] an</i> |
| ART | bestimmter oder unbestimmter Artikel | <i>der, die, das,</i> <i>ein, eine</i> |

STTS Tags (2)

| POS = | Beschreibung | Beispiele |
|--------------|---|--|
| CARD | Kardinalzahl | <i>zwei [Männer], [im Jahre] 1994</i> |
| FM | Fremdsprachliches Material | <i>[Er hat das mit "] A big fish [" übersetzt]</i> |
| ITJ | Interjektion | <i>mhm, ach, tja</i> |
| KOUI | unterordnende Konjunktion mit "zu" und Infinitiv | <i>um [zu leben], anstatt [zu fragen]</i> |
| KOUS | unterordnende Konjunktion mit Satz | <i>weil, daß, damit, wenn, ob</i> |
| KON | nebenordnende Konjunktion | <i>und, oder, aber</i> |
| KOKOM | Vergleichspartikel, ohne Satz | <i>als, wie</i> |

STTS Tags (3)

| POS = | Beschreibung | Beispiele |
|-------------|--|--|
| NN | normales Nomen | <i>Tisch, Herr, [das] Reisen</i> |
| NE | Eigennamen | <i>Hans, Hamburg, HSV</i> |
| PDS | substituierendes Demonstrativ- pronomen | <i>dieser, jener</i> |
| PDAT | attribuierendes Demonstrativ- pronomen | <i>jener [Mensch]</i> |
| PIS | substituierendes Indefinit- pronomen | <i>keiner, viele, man, niemand</i> |
| PIAT | attribuierendes Indefinit- pronomen ohne Determiner | <i>kein [Mensch], irgendein [Glas]</i> |

STTS Tags (4)

| POS = | Beschreibung | Beispiele |
|-------------------------------|--|---|
| PIDAT | attribuierendes Indefinitpronomen mit Determiner | <i>[ein] wenig [Wasser], [die] beiden [Brüder]</i> |
| PPER | irreflexives Personalpronomen | <i>ich, er, ihm, mich, dir</i> |
| PPOSS | substituierendes Possessivpronomen | <i>meins, deiner</i> |
| PPOSAT | attribuierendes Possessivpronomen | <i>mein [Opa], deine [Oma]</i> |
| PRELS PRELAT | Relativpronomen substituierend attribuierend | <i>[der Hund,] der [der Mann ,] dessen [Hund]</i> |

STTS Tags (5)

| POS = | Beschreibung | Beispiele |
|---------------|---|--|
| PRF | reflexives Personalpronomen | <i>sich, einander, dich, mir</i> |
| PWS | substituierendes Interrogativpronomen | <i>wer, was</i> |
| PWAT | attribuierendes Interrogativpronomen | <i>welche [Farbe], wessen [Hut]</i> |
| PWAV | adverbiales Interrogativ- oder Relativpronomen | <i>warum, wo, wann, worüber, wobei</i> |
| PAV | Pronominaladverb | <i>dafür, dabei, deswegen</i> |
| PTKZU | “zu” vor Infinitiv | <i>zu [gehen]</i> |
| PTKNEG | Negationspartikel | <i>nicht</i> |

STTS Tags (6)

| POS = | Beschreibung | Beispiele |
|-----------------|--------------------------------------|---|
| PTKVZ | abgetrennter Verbzusatz | <i>[er kommt] an, [er fährt] rad</i> |
| PTKANT | Antwortpartikel | <i>ja, nein, danke, bitte</i> |
| PTKA | Partikel bei Adjektiv oder Adverb | <i>am [schönsten], zu [schnell]</i> |
| TRUNC | Kompositions-Erstglied | <i>An- [und Abreise]</i> |
| VVFIN | finites Verb, voll | <i>[du] gehst, [wir] kommen [an]</i> |
| VVIMP | Imperativ, voll | <i>komm [!]</i> |
| VVINFINF | Infinitiv, voll | <i>gehen, ankommen</i> |
| VVIZU | Infinitiv mit "zu", voll | <i>anzukommen, loszulassen</i> |
| VVPP | Partizip Perfekt, voll | <i>gegangen, angekommen</i> |

STTS Tags (7)

| POS = | Beschreibung | Beispiele |
|--------------|--|--------------------------------|
| VAFIN | finites Verb, aux | <i>[du] bist, [wir] werden</i> |
| VAIMP | Imperativ, aux | <i>sei [ruhig !]</i> |
| VAINF | Infinitiv, aux | <i>werden, sein</i> |
| VAPP | Partizip Perfekt, aux | <i>gewesen</i> |
| VMFIN | finites Verb, modal | <i>dürfen</i> |
| VMINF | Infinitiv, modal | <i>wollen</i> |
| VMPP | Partizip Perfekt, modal | <i>[er hat] gekonnt</i> |
| XY | Nichtwort, Sonderzeichen enthaltend | <i>D2XW3</i> |

STTS Tags (8)

| POS = | Beschreibung | Beispiele |
|------------|----------------------------------|-----------|
| \$, | Komma | , |
| \$. | Satzbeendende Interpunktion | . ? ! ; : |
| \$(| sonstige Satzzeichen; satzintern | - [] () |

Morphology in TüBa-D/Z (1)

- ▶ person/number/mood/tense for verbs
- ▶ case/number/gender[/person]
for nouns, adjectives, determiners, pronouns
- ▶ case/number/gender for **names**
 - ▶ different from TiGer, Ancora
 - ▶ person: m/f
 - ▶ rivers: (usually) f
 - ▶ city:n
 - ▶ can be tested with adjective modification
ein schöneres Bremen (a more beautiful Bremen)
eine sauberere Donau (a cleaner Danube)

Morphology in TüBa-D/Z (2)

Inflectional morphology: 54 different tags, 412 combination of morphological properties

Values of morphological features:

| Feature | Values |
|---------|--|
| case | n (nominative), g (genitive), d (dative), a (accusative), * (underspecified) |
| gender | m (masculine), f (feminine), n (neuter), * (underspecified) |
| number | s (singular), p (plural), * (underspecified) |
| mood | i (indicative), k (subjunctive; German 'Konjunktiv') |
| person | 1 (first), 2 (second), 3 (third), * (underspecified) |
| tense | s (present), t (past) |

Levels of Syntactic Annotation

| Level | Inventory |
|----------------|---|
| sentence level | root node labels for different types of sentences |
| field level | node labels for topological fields |
| phrase level | node labels for syntactic categories and edge labels for grammatical functions |
| lexical level | lexical entries tagged with the part-of-speech (POS) tags taken from the STTS tagset |

Major Clause Types in German

- (1) a. Peter wird das Buch gelesen haben.
Peter will the book read have
'Peter will have read the book.'
- b. Wird Peter das Buch gelesen haben?
Will Peter the book have read
'Will Peter have read the book?'
- c. dass Peter das Buch gelesen haben wird.
that Peter the book read have will
'... that Peter will have read the book.'

Flexible Phrase Ordering

- (2) a. Der Mann hat gestern den Roman gelesen.
The man has yesterday the novel read
'The man read the novel yesterday.'
- b. Gestern hat der Mann den Roman gelesen
- c. Den Roman hat der Mann gestern gelesen

Discontinuous Constituents

- (3) Der Mann hat gestern den Roman gelesen, den ihm Peter
The man has yesterday the novel read which him Peter
empfahl.
recommended
'Yesterday the man read the novel which Peter recommended to him.'
- (4) Peter soll dem Mann empfohlen haben, den Roman zu lesen.
Peter is to the man recommended have the novel to read
'Peter is said to have recommended to the man to read the novel.'

Topological Fields

- (5) a. [_{VF} [_{NP} Peter]] [_{LK} wird] [_{MF} [_{NP} das Buch]]
[_{RK} [_{VC} gelesen haben.]]
- b. [_{LK} Wird] [_{MF} [_{NP} Peter] [_{NP} das Buch]]
[_{RK} [_{VC} gelesen haben?]]
- c. [_{LK} [_{CF} dass]] [_{MF} [_{NP} Peter] [_{NP} das Buch]]
[_{RK} [_{VC} gelesen haben wird.]]

Node Labels

| Node Labels | Description |
|---------------------------|--|
| Phrase Node Labels | |
| NX | noun phrase |
| PX | prepositional phrase |
| ADVX | adverbial phrase |
| ADJX | adjectival phrase |
| VXFIN | finite verb phrase |
| VXINF | infinite verb phrase |
| DP | determiner phrase (e.g. <i>gar keine</i>) |
| Root Node Labels | |
| SIMPX | simplex clause |
| R-SIMPX | relative clause |
| P-SIMPX | paratactic construction of simplex clauses |
| DM | discourse marker |

Node Labels (2)

| Node Labels | Description |
|--|--|
| Topological Field Node Labels | |
| LV | resumptive construction (Linksversetzung) |
| VF | initial field (Vorfeld) |
| LK | left sentence bracket (Linke (Satz-)Klammer) |
| MF | middle field (Mittelfeld) |
| VC | verb complex (Verbkomplex) |
| NF | final field (Nachfeld) |
| C | complementizer field (C-Feld) |
| KOORD | field for coordinative particles |
| PARORD | field for coordinative particles |
| FKOORD | coordination consisting of conjuncts of fields |
| Field Conjunct Node Labels | |
| LKM, LKMVC, LKMVCN, LKMN, LKVCN, LKN, MVC, MVCN, MN, VCN, CM, CMVC | combinations of fields - node labels are derived by concatenation of conjunct field labels (V = VF, M = MF, N = NF) e.g. LKM = LK + MF |

Edge Labels

| Edge Labels | Description |
|----------------------------------|-------------------------------|
| Edge Labels denoting Head | |
| HD | head |
| - | non-head |
| Complement Edge Labels | |
| ON | nominative object |
| OD | dative object |
| OA | accusative object |
| OS | sentential object |
| OPP | prepositional object |
| OADVP | adverbial object |
| OADJP | adjectival object |
| PRED | predicate |
| OV | verbal object |
| FOPP | optional prepositional object |
| VPT | separable verb prefix |
| APP | apposition |

Edge Labels (2)

| Edge Labels | Description |
|---|--|
| Edge Labels denoting Head | |
| Modifier Edge Labels | |
| MOD | ambiguous modifier |
| ON-MOD, OA-MOD, OD-MOD, MOD-MOD, V-MOD, OPP-MOD, PRED-MOD, FOPP-MOD | modifiers modifying complements or modifiers e.g. V-MOD = modifier of the verb |
| Edge Labels in Split-up Coordinations | |
| ONK, ODK, OAK, OPPK, FOPPK, OADJPK, PREDK, MODK, OA-MODK, V-MODK, OPP-MODK, PREDMODK, MOD-MODK | second conjunct in split-up coordinations e.g. ONK = second conjunct of a nominative object (subject) |

Edge Labels (3)

| Secondary Edge Labels | |
|-----------------------|---|
| ref1 | first verbal object in VC selected by a verbal object |
| EN | phrase internal relation between two parts of a proper noun |
| refcontr | dependency relation between a control verb and its complement |
| refint | dependency relation between a phrase internal part and its modifier |
| refmod | dependency relation in case of ambiguous modification |
| refvc | dependency relation between two verbal objects in the verb complex |

Constructing the TüBa-D/Z Treebank

- ▶ Choice of source data
- ▶ Preprocessing
- ▶ Annotation
- ▶ Correction
- ▶ Postprocessing

1st processing step – Formatting and Segmentation

- ▶ Script
 - ▶ Extraction of text from HTML files
 - ▶ Sentence splitter (Finite State)
 - ▶ Tokenizer (Finite State)
 - ▶ POS tagger (multi-classifier combination)
 - ▶ Conversion of data into Negra Export format
- ▶ Manual
 - ▶ Correction of sentence boundaries
 - ▶ Annotation of article headlines

Source format

Example:

```
<HTML>
[... ]
<!--TI-->
<H2>Tödliche Nachtschicht</H2>
<!--END-->
<!--TX-->
<P>
In einer Berliner Papierfabrik kam der 26jährige Papiermacher
und Student der Elektrotechnik Thomas H. ums Leben. Ein Arbeitsunfall.
Offizielle Unfallerklärung: Eigenverschulden. Bericht auf der
Hintergrund-Seite 8</P>
<!--END-->
<HR>
[... ]
</HTML>
```

2nd processing step – Syntactic annotation

Annotation Tool: @nnotate

- ▶ Semi-automatic annotation of corpus data
- ▶ Context-free structures
- ▶ Additionally allows crossing edges
- ▶ Labels for terminal nodes, non-terminal nodes, and edges
- ▶ User-defined label inventory

@nnotate

- ▶ Communication with external taggers and parsers
- ▶ POS tagger *TnT*
- ▶ Statistical Parser based on cascaded Markov-Models
- ▶ NP chunker *Chunkie*
- ▶ Relational database for annotated corpora: MySQL

- ▶ Developed at University of Saarbrücken, Germany
- ▶ Developed ca. 1998
- ▶ Operating system: Solaris/Linux
- ▶ Software no longer maintained

Syntactic annotation

- ▶ Manual correction of POS tags
- ▶ Manual correction of spelling errors into a distinct “comment” layer
- ▶ Interactive semi-automatic annotation of
 - ▶ Phrases
 - ▶ Topological fields
 - ▶ Grammatical functions
 - ▶ Secondary Edges
 - ▶ (Complex) Named Entities
 - ▶ Sentences

Named Entities

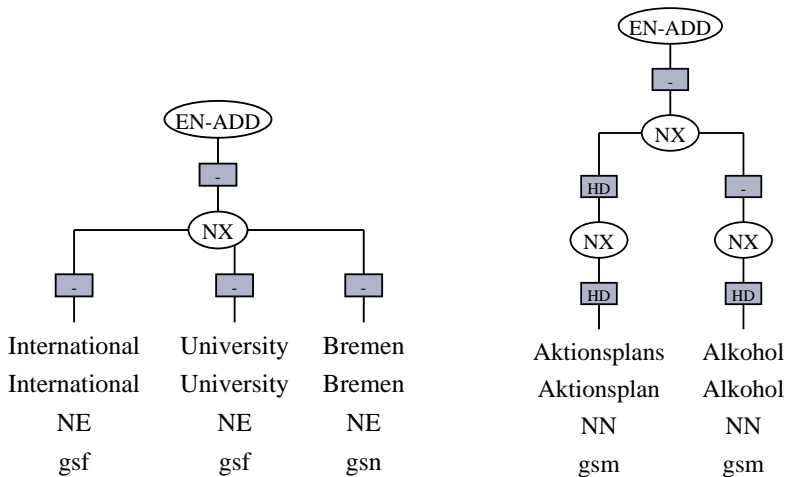
Classes of Named Entities:

- ▶ Names consisting of one lexical element (POS-tagged as NE or POS-tagged according to their distribution + EN-ADD)
- ▶ Complex names consisting of more than one lexical element, each of them POS-tagged as NE
- ▶ Complex names which are POS-tagged according their distribution (EN-ADD or EN secondary edge)

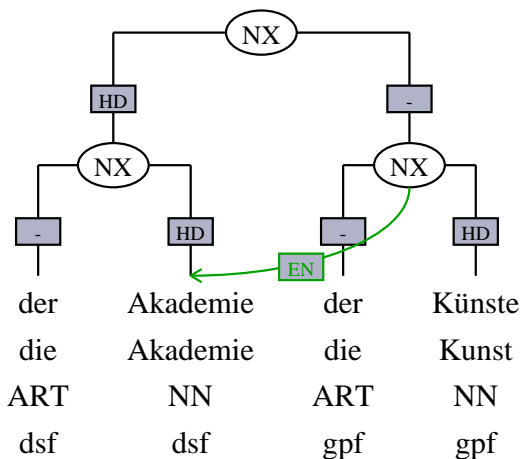
Annotation of Named Entities:

- ▶ On the morpho-syntactic level via STTS tags (NE)
- ▶ On the syntactic level via labels

Named Entities: node label

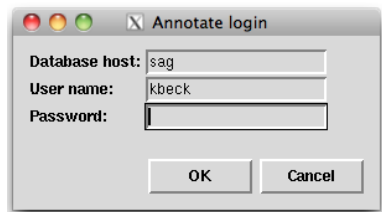


Named Entities: secondary edge



@nnotate

- ▶ Authorization and user management



A screenshot of a Mac OS-style dialog box titled "Annotate login". The dialog has three text input fields: "Database host:" containing "sag", "User name:" containing "kbeck", and "Password:" which is empty. At the bottom, there are two buttons: "OK" and "Cancel".

@nnotate – Annotation Process



Why but shall she herself now still make-up?

But why should she still put on make-up now?

Kathrin Beck and Erhard Hinrichsbeck@sfs.uni-tuebingen.deeh@sfs.uni-tuebingen.de
Dept. of Linguistics Eberhard Karls University of Tübingen

@nnotate – Annotation Process

General:

Corpus: TUEBADZaktuell

Editor: Kathrin

Save Reload Exit Options

Sentence:

No.: 59986 (45059.63151) Last edited: <Automatisch>, 16/12/08, 12:02:10

Comment: ...

Origin: T970808.169

Warum₀ aber₁ soll₂ sie₃ sich₄ jetzt₅ noch₆ schminken₇ ?₈

PWAV ADV VMFIN PPER PRF ADV ADV VVIZU \$.

Move: << >> Go to: Matches: Search for:

Dependency: Selection: Command: Execute

Parentlabel: Node no.: Parentlabel: << >> End

T:1 D:3/4 M

@nnotate – Annotation Process

General:
 Corpus: TUEBADZaktuell
 Editor: Kathrin
 Save Reload Exit Options

Sentence:
 No.: 59986 (45059..63151) Last edited: <Automatisch>, 16/12/08, 12:02:10
 Comment:
 Origin: T970808.169

Warum₀ PWAV aber₁ ADV soll₂ VMFIN sie₃ PPER sich₄ PRF jetzt₅ ADV noch₆ ADV schminken₇ VVIZU ?₈ \$

Move: << >> Go to: Matches: Search for:

Dependency: Selection: Command: Execute

Parentlabel: Node no.: Parentlabel: << >> End

T:1 D:34 M

@nnotate – Annotation Process

The screenshot shows the @nnotate v3.6(sfs-p1) window. The main area displays the sentence: "Warum₀ aber₁ soll₂ sie₃ sich₄ jetzt₅ noch₆ schminken₇ ?₈". Below the words are their grammatical functions: PWAV, ADV, VMFIN, PPER, PRF, ADV, ADV, VVIZU, and \$. Above the words are dependency arcs with labels like ADVP, VXFIN, NCV, and NCV, and node numbers like 505, 500, 501, 503, 502, and 504. A tree structure is visible above the arcs, showing a root node 'HC' (506) connected to 'OJ' (504), which is connected to 'VXINF' (504) and 'ADV' (502). 'VXINF' (504) is connected to 'ADV' (504) and 'ADV' (502). 'ADV' (504) is connected to 'ADV' (504) and 'ADV' (502). 'ADV' (502) is connected to 'ADV' (502) and 'ADV' (502).

General:
 Corpus: TUEBADZaktuell
 Editor: Kathrin
 Buttons: Save, Reload, Exit, Options

Sentence:
 No.: 59986 (45059..63151) Last edited: <Automatisch>, 16/12/08, 12:02:10
 Comment:
 Origin: T970808.169

Move:
 Buttons: <<, >>, Mask..., Search for:
 Go to:
 Matches:

Dependency:
 Selection:
 Command: Parse
 Execute

Edgelabel:
 Node no.: 504 VXINF
 Edgelabel: 77.14% OV (verbales Objekt, 4)
 Buttons: <<, >>, End, Cancel, Parentlabel

Bottom right: T:1 D:¼ M

@nnotate – Annotation Process

The screenshot shows the @nnotate v3.6(sfs-p1) window. The title bar reads "Annotate v3.6(sfs-p1)".

General:

- Corpus: TUEBADZaktuell
- Editor: Kathrin
- Buttons: Save, Reload, Exit, Options

Sentence:

- No.: 59986 (45059..63151) Last edited: <Automatisch>, 16/12/08, 12:02:10
- Comment: [text input]
- Origin: T970808.169

Main Area:

Warum₀ PWAV
 aber₁ ADV
 soll₂ VMFIN
 sie₃ PPER
 sich₄ PRF
 jetzt₅ ADV
 noch₆ ADV
 schminken₇ VVIZU
 ?₈ \$

Tree structures above the words:

- ADUX⁵⁰⁵ | HD
- VXFN⁵⁰⁰ | HD
- NCX⁵⁰¹ | HD
- NCX⁵⁰³ | HD
- ADUX⁵⁰² | HD
- VXNF⁵⁰⁴ | HD
- NC⁵⁰⁶ | HD

Move:

- Buttons: <<, >>, Mask...
- Go to: [text input]
- Matches: [text input]
- Search for: [text input]

Dependency:

- Selection: [text input]
- Command: [text input]
- Execute button

Edglabel:

- Node no.: [text input]
- Edglabel: [text input]
- Buttons: <<, >>, End, Cancel, Parentlabel

Bottom right: T:1 D:3/4 M

@nnotate – Annotation Process

The screenshot shows the @nnotate v3.6(sfs-p1) application window. The interface is divided into several sections:

- General:** Corpus: TUEBADZaktuell, Editor: Kathrin. Buttons: Save, Reload, Exit, Options.
- Sentence:** No.: 59986 (45059.63151), Last edited: <Automatisch>, 16/12/08, 12:02:10. Comment: (empty). Origin: T970808.169.
- Main Area:** A dependency tree for the sentence "Warum_0 aber_1 soll_2 sie_3 sich_4 jetzt_5 noch_6 schminken_7 ?_8". The tree structure shows nodes like LI, MF, NC, AD, ADVS, and VXP, with edges labeled with grammatical relations like HD, NCX, ADVS, and VXP.
- Bottom Panel:**
 - Move:** Navigation buttons (<<, >>) and a "Go to:" field.
 - Dependency:** Selection and Command fields (set to "Parse"), and an "Execute" button.
 - Edglabel:** Node no.: 501, NCX. Edglabel: 64.84% ON (Nominativ-Objekt). Buttons: <<, >>, End, Cancel, Parentlabel.

@nnotate – Annotation Process

The screenshot shows the @nnotate v3.6(sfs-p1) application window. The interface is divided into several sections:

- General:** Corpus: TUEBADZaktuell, Editor: Kathrin. Buttons: Save, Reload, Exit, Options.
- Sentence:** No.: 59986 (45059.63151), Last edited: <Automatisch>, 16/12/08, 12:02:10. Comment: [empty]. Origin: T970806.169.
- Main Area:** A dependency tree diagram for the sentence "Warum₀ aber₁ soll₂ sie₃ sich₄ jetzt₅ noch₆ schminken₇ ?₈". The tree shows nodes for phrases like "ADUX", "UxPN", "NCX", "D2", "NOC", "D19", and "UxNFP" with their respective dependency labels (e.g., HE, HE, HE, HE, HE, HE, HE, HE, HE). Below the tree, the words and their part-of-speech tags are listed: PWAV, ADV, VMFIN, PPER, PRF, ADV, ADV, VVIZU, \$.
- Move:** Navigation buttons (<<, >>), Go to: [input], Mask... button, Search for: [input].
- Dependency:** Selection: [input], Command: Parse [dropdown], Execute button.
- Edgelabel:** Node no.: 503 NCX, Edgelabel: 50.96% OA (Akkusativ-Objekt, 1) [dropdown], End, Cancel, Parentlabel buttons.

@nnotate – Annotation Process

The screenshot shows the @nnotate v3.6(sfs-p1) application window. The interface is divided into several sections:

- General:**
 - Corpus: TUEBADZaktuell
 - Editor: Kathrin
 - Buttons: Save, Reload, Exit, Options
- Sentence:**
 - No.: 59986 (45059..63151) Last edited: <Automatisch>, 16/12/08, 12:02:10
 - Comment: [Empty field]
 - Origin: T970808.169
- Tree Diagram:**
 - Root node: SIMPX (512)
 - Children of SIMPX: UP (514), LX (500), NP (509), NC (506)
 - UP (514) branches into PX (515) and ADUX (505)
 - PX (515) branches into PX (510) and HD (505)
 - LX (500) branches into UXPIN (500) and HD (500)
 - NP (509) branches into ON (501), GA (503), ADUX (507), and ADUX (502)
 - NC (506) branches into UXPIN (504) and HD (504)
 - Leaf nodes (from left to right): Warum₀ (PWAV), aber₁ (ADV), soll₂ (VMFIN), sie₃ (PPER), sich₄ (PRF), jetzt₅ (ADV), noch₆ (ADV), schminken₇ (VVIZU), ?₈ (\$).
- Move:**
 - Navigation buttons: <<, >>, Mask...
 - Go to: [Input field]
 - Matches: [Input field]
 - Search for: [Input field]
- Dependency:**
 - Selection: [Input field]
 - Command: [Dropdown menu]
 - Execute button
- Edge/Node Labels:**
 - Edge label: [Input field]
 - Node no.: [Input field]
 - Edge label: [Input field]
 - Navigation buttons: <<, >>, End, Cancel, Parentlabel

@nnotate – Annotation Process

The screenshot shows the @nnotate v3.6(sfs-p1) interface. The main window displays a syntactic tree for the sentence: "Warum₀ PWAV, aber₁ ADV, soll₂ VMFIN, sie₃ PPER, sich₄ PRF, jetzt₅ ADV, noch₆ ADV, schminken₇ VVIZU, ?₈ \$." The tree is rooted at S₁₃ (SIMPX) and branches into VP₁₁₂ (VP), LK₁₀₉ (LK), MF₁₁₀ (MF), and VC₁₁₁ (VC). The MF node further branches into CN₁₀₅ (CN), OA₁₀₄ (OA), MOG₁₀₅ (MOG), MOG₁₀₆ (MOG), and UNF₁₀₇ (UNF). The VC node branches into OI₁₀₇ (OI) and UNF₁₀₇ (UNF). The UNF nodes further branch into PX₁₀₀ (PX), ADUX₁₀₁ (ADUX), UNF₁₀₂ (UNF), NCX₁₀₃ (NCX), NCX₁₀₄ (NCX), ADUX₁₀₅ (ADUX), ADUX₁₀₆ (ADUX), and UNF₁₀₇ (UNF). The UNF nodes further branch into HD₁₀₀ (HD), HD₁₀₁ (HD), HD₁₀₂ (HD), HD₁₀₃ (HD), HD₁₀₄ (HD), HD₁₀₅ (HD), HD₁₀₆ (HD), and HD₁₀₇ (HD).

The interface includes a "General" panel with fields for "Corpus" (TUEBADZaktuell), "Editor" (Kathrin), and buttons for "Save", "Reload", "Exit", and "Options". The "Sentence" panel shows "No.: 59986 (45059.63151)", "Last edited: Kathrin, 06/12/10, 15:15:00", "Comment:" (with a search icon), and "Origin: T970808.169".

At the bottom, there are three panels: "Move:" with navigation buttons and "Go to:" and "Matches:" fields; "Dependency:" with "Selection:" and "Command:" fields and an "Execute" button; and "Edglabel:" with "Node no.:", "Edglabel:" fields, and navigation buttons. The bottom right corner shows "T:1 D:1".

3rd step – Morphological annotation (1)

Input for morphological pre-annotation scripts:

- ▶ Morphological analyser SMOR (Helmut Schmid, Stuttgart)
- ▶ POS annotation \Rightarrow list of possible morphology
- ▶ Gazetteer of names (person names, cities)
- ▶ Gender/number of unknown names and common nouns (learned)

```
UNK-NN Peptimist      sm
NON-AMB Wolfgang Schuchard NE NE sm sm
NON-AMB Washington Post NE NE  sn sf
```

3rd step – Morphological annotation (2)

Input for morphological pre-annotation scripts:

- ▶ Syntactic annotation
 - ▶ Grammatical functions \Rightarrow case of NPs
 - ▶ PP annotation \Rightarrow case of head NPs
 - ▶ NP annotation \Rightarrow agreement of modifiers
 - ▶ S annotation \Rightarrow subject-verb-agreement
- ▶ Accusative/dative disambiguation of prepositions
 - ▶ Accusative default for verb adjuncts
 - ▶ Dative default for noun adjuncts
 - ▶ List of exceptions (learned)
- ▶ Sentences are separated by empty text lines

Morphology

Example (1):

```
>>> Am (APPRART --)
dsn an
>>> Ende (NN --)
dsn Ende
>>> kehrte (VVFIN --)
3sit kehren
>>> man (PIS --)
ns* --
>>> zum (APPRART --)
dsm zu
>>> Anfang (NN --)
dsm Anfang
>>> zurück (PTKVZ --)
>>> . ($. --)
```

At+the end turned one to+the beginning back .

At the end, one returned to the beginning.

Morphology

Example (2):

```
>>> Katrin (NE --)
asf Katrin
nsf Katrin
dsf Katrin
>>> Bettina (NE --)
asf Bettina
nsf Bettina
dsf Bettina
>>> Müller (NE --)
asf Müller
nsf Müller
dsf Müller
```

- ▶ Manual correction of the morphology text file
- ▶ Manual post-correction of the morphological annotation

Lemmatization in TüBa-D/Z (1)

- ▶ No standards for lemmatization
- ▶ Make use of manual annotation and make a rich annotation

General rules for the TüBa-D/Z annotation

- ▶ Nouns \Rightarrow nominative singular
- ▶ Adjectives \Rightarrow predicative form
- ▶ Verbs \Rightarrow infinitive; suffix for passive and auxiliaries
- ▶ Reflexives \Rightarrow %refl
- ▶ Adverbs, prepositions, cardinal numbers,... \Rightarrow as-is

Lemmatization in TüBa-D/Z (1)

Open-class words

- ▶ Consistency with GermaNet
- ▶ deadjectival nouns → strong form
ein Arbeitsloser [masc, st] vs. *der Arbeitslose* [fem, wk]
eine Arbeitslose [fem, st] vs. *die Arbeitslose* [fem, wk]
- ▶ Attach separable prefixes
*Peter **schließt** sich in der Küche **ein*** → ein#schließen
- ▶ Distinguish (non-)separable verb prefixes
umfahren (drive around) vs. *um#fahren* (drive over)
- ▶ Complete truncated items
Bau- und Verkehrsplanung → Bauplanung%N

Lemmatization in TüBa-D/Z (2)

Closed-class words

- ▶ Distinguish auxiliary/passive from full verb uses
- ▶ articles/possessives/definite and indefinite pronouns:
normalize to nominative singular, but keep gender and root
- ▶ Possibility for underspecification (*der|die|das*)

Pretagging – Closed-class list

| | |
|--------------------------|----------------------------|
| ART d.* .*m der | PDS d.. .** der die das |
| ART d.* .*n das | PDS denen .*m der |
| ART d.* .*f die | PDS denen .*n das |
| ART d.* .** der die das | PDS denen .*f die |
| ART ein.* .*m ein | PDS denen .** der die das |
| ART ein.* .*n ein | |
| ART ein.* .*f eine | PPER .s.1 ich |
| ART ein.* .** ein eine | PPER .s.2 du |
| ART 'n.* .*m ein | PPER .sm3 er |
| ART 'm.* .*m ein | PPER .sn3 es |
| ART 'n.* .*n ein | PPER .sf3 sie |
| ART 'n.* .*f eine | PPER .p.1 wir |
| ART 's .*n das | PPER .p.2 ihr |
| | PPER ihnen .p.3 sie |
| PDS d.. .*m der | PPER Ihnen .p.3 Sie |
| PDS d.. .*n das | PPER sie .p.3 sie |
| PDS d.. .*f die | PPER Sie .p.3 Sie |

Pretagging – Frequency-based Heuristics

- ▶ **Truncated items:** look for coordinate sister and choose completion by frequency
- ▶ **Verb ambiguities:**
 - ▶ Non separable *ge*-prefix
past participle *geraten* → *raten* (guess/advise), *geraten* (turn out, become)
 - ▶ Separable vs. inseparable prefixes
Infinitive *umfahren* → *zu umfahren*, *umzufahren*
- ▶ Use frequency ratio to **keep ambiguities**

4th step – Lemmatization

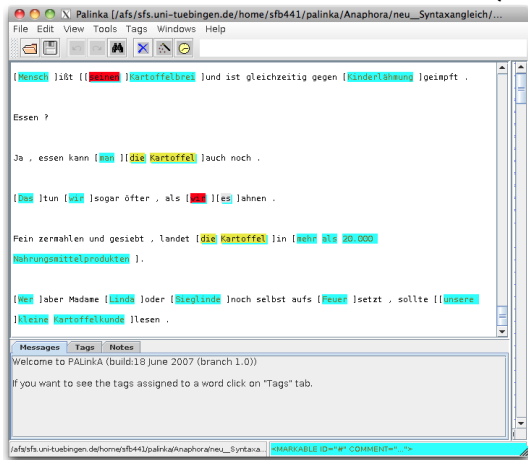
- ▶ Input: Text file with source tokens, POS tags, morphology and suggested lemma
- ▶ Manual correction of the suggested lemma

Lemmatization

```
<s id=13054>
Doch KON doch --
inzwischen ADV inzwischen --
gesteht VFIN zu#gestehen 3sis
die ART die nsf
Türkei NE Türkei nsf
mit APPR mit d
einer ART eine dsf
" $( " --
rosa ADJA rosa dsf
Karte NN Karte dsf
" $( " --
den ART der|die|das dp*
ehemaligen ADJA ehemalg dp*
Staatsangehörigen NN Staatsangehöriger dp*
viele PIDAT viele apn
Bürgerrechte NN Bürgerrecht apn
zu PTKVZ -- --
. $. . --
</s>
```

5th step – Annotation of referential relations

Several terms refer to the same (discourse) entity



Yes, eat can one the potato also even.
Yes, you can even eat the potato.

That do we even more_often , than we
it_expl suspect .

We do that even more often than we
suspect.

Finely ground and sieved, lands the po-
tato in more than 20,000 food_products.
Finely ground and sieved, the potato en-
ters over 20,000 food products.

Coreference, Anaphora, Expletives

Yes, [you] can even eat [₁ **the potato**]. [₂ We] do [that] even more often than [₂ we] suspect [_{expl} it_{expl}]. Finely ground and sieved, [₁ **the potato**] enters [over 20,000 food products].

Several mentions refer to the same (discourse) referent:

- ▶ **the potato** \equiv **the potato** (coreferent)
- ▶ We \equiv we
- ▶ it_{expl}

Coreference, Anaphora, Expletives

Yes, [you] can even eat [₁ the potato]. [₂ **We**] do [that] even more often than [₂ **we**] suspect [_{expl} it_expl]. Finely ground and sieved, [₁ the potato] enters [over 20,000 food products].

Several mentions refer to the same (discourse) referent:

- ▶ the potato \equiv the potato
- ▶ **We** \equiv **we** (anaphoric)
- ▶ it_expl

Coreference, Anaphora, Expletives

Yes, [you] can even eat [₁ the potato]. [₂ We] do [that] even more often than [₂ we] suspect [_{expl} **it_expl**]. Finely ground and sieved, [₁ the potato] enters [over 20,000 food products].

Several mentions refer to the same (discourse) referent:

- ▶ the potato \equiv the potato
- ▶ We \equiv we
- ▶ **it_expl** (expletive it)

Referential relations (1)

| relation | description |
|---------------|--|
| coreferential | links subsequent-mention definite noun phrases to the closest preceding mention |
| anaphoric | links a pronominal anaphor (personal, demonstrative, reflexive, and relative pronouns, including possessive pronouns) to its closest (preceding-in-linear-order) antecedent. |
| cataphoric | links a pronoun to the mention it is resolved to if that mention comes later in the linear order of the text |

Referential relations (2)

| relation | description |
|------------------|--|
| bound | is used for anaphora bound by the same quantifier as their antecedent |
| split_antecedent | links a plural pronoun to the descriptions that together refer to the summum to which the pronoun refers |
| instance | links a first-mention noun phrase to the set-denoting noun phrase of which it denotes a member |
| expletive | correlates of clausal arguments and semantically empty subjects of weather verbs |

Annotation process

- ▶ Annotation tool: PAlinkA
 - ▶ Constantin Orasan, 2005
 - ▶ Java
 - ▶ No longer updated, but developer answers emails
- ▶ Alternative annotation tool: MMAX2
- ▶ Data format: XML
- ▶ Independent double annotation
- ▶ Merge and correction of annotation variance
- ▶ Correction of inherent reflexives and expletive it
- ▶ Adaptation to syntax and tokenization changes

Data format

```

<P ID="P58">
  <S ID="S58">
    <MARKABLE COMMENT="" ID="m320">
      <W ID="w927">Das</W>
    </MARKABLE>
    <W ID="w928">tun</W>
    <MARKABLE COMMENT="" ID="m321">
      <W ID="w929">wir</W>
    </MARKABLE>
    <W ID="w930">sogar</W>
    <W ID="w931">öfter</W>
    <W ID="w932">,</W>
    <W ID="w933">als</W>
    <MARKABLE COMMENT="" ID="m322">
      <anaphoric COMMENT="" ID="43" SRC="m321"/>
      <W ID="w934">wir</W>
    </MARKABLE>
    <MARKABLE COMMENT="" ID="m323">
      <expletive COMMENT="" ID="348"/>
      <W ID="w935">es</W>
    </MARKABLE>
    <W ID="w936">ahnen</W>
    <W ID="w937">.</W>
  </S>
</P>

```

6th step – Integration of annotation layers

- ▶ Integration of morphology and lemma layer into the syntax
Negra Export format
- ▶ Running of syntax/morphology queries for spotting typical
annotation errors
- ▶ Manual correction of syntax, morphology and lemmas of every
sentence
- ▶ Integration of coreferential relations

Discourse annotation (experimental)

- ▶ Coherence relations, taxonomic scheme
- ▶ Annotation of discourse connectives
 - ▶ focus on contribution of (ambiguous) connectives (nachdem *after/since*, während *while*, als *when*, aber *but*)
 - ▶ per-connective
- ▶ Full document annotation
 - ▶ discourse structure of multi-paragraph units (topic segments)
 - ▶ includes 'implicit' relations (without connectives)

Discourse relations (extract)

Temporal

- ▶ 27:24 hatten die Badener das Hinspiel gewonnen , am Samstag siegten sie zu Hause 24:21. “Ich habe gewußt, daß es klappt”, stellte TVW-Trainer Hrovje Horvat auffallend gelassen fest, noch **während** seine Spieler Ringelreien tanzten inmitten der Fans.

*The Baden team won the first leg with 27:24, on Saturday they won the home game 24:21. “I knew that it would turn out well”, TVW’s trainer Hrovje Horvat declares ostentatively relaxed, **while** his players still dance among the fans.*

Discourse relations II (extract)

Cause

- ▶ In Großbritannien geht bei Minderheiten die Angst um, **nachdem** Neonazis am Freitag Abend beim dritten Bombenanschlag innerhalb von zwei Wochen drei Menschen getötet und fast 60 verletzt hatten.

Fear runs among the minorities in Great Britain, after Neonazis killed three people and injured almost 60 on Friday evening on the third bombing within two weeks.

Contrast

- ▶ Doch **während** sich die Gäste amüsieren, hockt Jusef im Nebenzimmer und starrt finster an die Wand.

But while the guests enjoy themselves, Jusef sits in the neighbouring room and glumly stares at the wall.

Connective annotation

von der Halle ist nur ein grüner Hügel zu sehen , aus dem eine Stahlgaskuppel ragt .

1979 Diese Kuppel war im Juni 1996 - zwei Monate vor der Einweihung - zum zweiten Mal eingestürzt , **nachdem** sie im Februar 1997 erstmals eingebrochen war .

| Temporal |
|-----------------|
| Result |
| enable |
| cause |
| epistemic_cause |
| speech_act |
| Comparison |
| parallel |
| contrast |

Düsseldorf (dpa) - Der Aufsichtsrat der Mannesmann AG ist künftig nicht mehr nach den Vorschriften des Montanmitbestimmungsgesetzes zu besetzen .

1979 **Nachdem** das Bundesverfassungsgericht mit einer Grundsatzentscheidung den Weg hierzu geebnet hatte , befreite der 19. Zivilsenat des Düsseldorfer Oberlandesgerichts Ende April den Mannesmann-Konzern von dieser Regelung , da die Firma in sechs aufeinanderfolgenden Jahren die Wertschöpfungsquote von mehr als 20 Prozent nicht mehr erreicht habe .

| Temporal |
|-----------------|
| Result |
| enable |
| cause |
| epistemic_cause |
| speech_act |
| Comparison |
| parallel |
| contrast |

Implicit relations

Look at discourse relations independent of connectives

- ▶ Thematically coherent segments (*topics*)
- ▶ Elementary discourse units (EDUs)
- ▶ Discourse relations, coordination/subordination

| |
|--|
| T0 Die Fusion von Repsol und YPF |
| 3.0 Madrid (taz) - |
| 3.1 Repsol will hoch hinaus . Elaboration(3.1,4) |
| 4.0 Der spanische Energiekonzern hat am Freitag ein Kaufangebot an den argentinischen Öl- und Gasförderkonzern YPF gerichtet . Elaboration(4,2) Result-Enable(4,6) |
| 5.0 Für 85,1 Prozent der Aktien bietet Repsol 2,08 Billionen Peseten (24,5 Milliarden Mark) , Background(5.1,1) |
| 5.1 25,4 Prozent mehr als der aktuelle Börsenwert von YPF . |
| 6.0 Die Händler an der Börse in Buenos Aires waren dann am Freitag auch begeistert . Result-Cause(6,7) |
| 7.0 Der YPF-Titel kletterte um 16 Prozent Result-Enable(7,1) |

7th step – Data formats – Negra Export

```

#BOS 17726 2 1113315998 857
Das      das      PDS asn      HD 500
tun      tun      VVFIN 1pis    HD 501
wir      wir      PPER      np*1      HD 502
sogar    sogar    ADV --      HD 503
öfter    öfter    ADV --      HD 504
,        ,        $, -- -- 0
als      als      KOUS -- -- 505
wir      wir      PPER      np*1      HD 506
es       es       PPER      asn3      HD 507
ahnen    ahnen    VVFIN 1pis    HD 508
.        .        $. -- -- 0
#500    --      NX -- -- 0A 509
#501    --      VVFIN -- -- HD 510
#502    --      NX -- -- ON 511
#503    --      ADVX -- -- MOD 511
#504    --      ADVX -- -- V-MOD 511 refmod 514
#505    --      C -- -- 514
#506    --      NX -- -- ON 512 %% R=anaphoric.17726:502
#507    --      NX -- -- OA 512 %% R=expletive
#508    --      VVFIN -- -- HD 513
#509    --      VF -- -- 516
#510    --      LK -- -- 516
#511    --      MF -- -- 516
#512    --      MF -- -- 514
#513    --      VC -- -- 514
#514    --      SIMPX -- -- MOD-MOD 515
#515    --      NF -- -- 516
#516    --      SIMPX -- -- 0
#EOS 17726

```

Data formats – Penn

```

%% sent. no. 17726
(
  (SIMPX
    (VF
      (NX:OA
        (PDS:HD Das)
      )
    )
  )
  (LK
    (VXFIN:HD
      (VVFIN:HD tun)
    )
  )
  (MF
    (NX:ON
      (PPER:HD wir)
    )
    (ADVX:MOD
      (ADV:HD sogar)
    )
    (ADVX:V-MOD
      (ADV:HD öfter)
    )
  )
  ($, .)
  (NF
    (SIMPX:MOD-MOD
      (CC
        (KOUS als)
      )
      (MF
        (NX:ON
          (PPER:HD wir)
        )
        (NX:OA
          (PPER:HD es)
        )
      )
      (VC
        (VXFIN:HD
          (VVFIN:HD ahnen)
        )
      )
    )
  )
  ($, .)
)

```

Data formats – Negra Export XML

```

<sentence origin="T990430_42" date="2005041215:26:38" editor="hschulz">
<node cat="SIMPX" parent="0" comment="" func="-" id="s_17726_n_516">
<node cat="VF" comment="" func="-" id="s_17726_n_589">
<node cat="NK" comment="" func="0A" id="s_17726_n_588">
<word comment="" form="das" func="HD" pos="POS" morph="asn" id="s_17726_n_8"/>
</node>
</node>
<node cat="LK" comment="" func="-" id="s_17726_n_518">
<node cat="VFIN" comment="" func="HD" id="s_17726_n_581">
<word comment="" form="tun" func="HD" pos="VFIN" morph="ipis" id="s_17726_n_1"/>
</node>
</node>
<node cat="NF" comment="" func="-" id="s_17726_n_511">
<node cat="NK" comment="" func="0N" id="s_17726_n_582">
<word comment="" form="wir" func="HD" pos="PPER" morph="np1" id="s_17726_n_2"/>
</node>
</node>
<node cat="ADVV" comment="" func="MD" id="s_17726_n_583">
<word comment="" form="sogar" func="HD" pos="ADV" morph="-" id="s_17726_n_3"/>
</node>
</node>
<node cat="ADVV" comment="" func="V-MD" id="s_17726_n_584">
<ceedge parent="514" cat="refrod"/>
<word comment="" form="öfter" func="HD" pos="ADV" morph="-" id="s_17726_n_4"/>
</node>
</node>
<word parent="0" comment="" form="," func="-" pos="S," morph="-" id="s_17726_n_5"/>
<node cat="NF" comment="" func="-" id="s_17726_n_515">
<node cat="SIMPX" comment="" func="MOD-MD" num="514" id="s_17726_n_514">
<node cat="IC" comment="" func="-" id="s_17726_n_585">
<word comment="" form="als" func="-" pos="MOUS" morph="-" id="s_17726_n_6"/>
</node>
</node>
<node cat="NF" comment="" func="-" id="s_17726_n_512">
<node cat="NK" comment="" func="0N" id="s_17726_n_586">
<anaphora>
<relation type="anaphoric" antecedent="s_17726_n_582"/>
</anaphora>
<word comment="" form="wir" func="HD" pos="PPER" morph="np1" id="s_17726_n_7"/>
</node>
<node cat="NK" comment="" func="0A" id="s_17726_n_587">
<anaphora>
<relation type="expletive" antecedent=""/>
</anaphora>
<word comment="" form="es" func="HD" pos="PPER" morph="asn3" id="s_17726_n_8"/>
</node>
</node>
<node cat="VC" comment="" func="-" id="s_17726_n_513">
<node cat="VFIN" comment="" func="HD" id="s_17726_n_588">
<word comment="" form="ahnen" func="HD" pos="VFIN" morph="ipis" id="s_17726_n_9"/>
</node>
</node>
</node>
</node>
</node>
<word parent="0" comment="" form="," func="-" pos="S," morph="-" id="s_17726_n_18"/>
</sentence>

```

Data formats – TigerXML / SynAF

```

<: id="s17726">
<graph root="s17726_s16">
<terminals>
<: id="s17726_1" word="Das" lemma="das" pos="PDS" morph="asn">
</>
<: id="s17726_2" word="tun" lemma="tun" pos="VFIN" morph="Ipis">
</>
<: id="s17726_3" word="wir" lemma="wir" pos="PPER" morph="np1">
</>
<: id="s17726_4" word="sogar" lemma="sogar" pos="ADV" morph="...">
</>
<: id="s17726_5" word="&#x00f6;" fter" lemma="&#x00f6;" fter" pos="ADV" morph="...">
</>
<: id="s17726_6" word="," lemma="," pos="S," morph="...">
</>
<: id="s17726_7" word="als" lemma="als" pos="KOUS" morph="...">
</>
<: id="s17726_8" word="wir" lemma="wir" pos="PPER" morph="np1">
</>
<: id="s17726_9" word="es" lemma="es" pos="PPER" morph="asn3">
</>
<: id="s17726_10" word="ahnen" lemma="ahnen" pos="VFIN" morph="Ipis">
</>
<: id="s17726_11" word="," lemma="," pos="S," morph="...">
</>
</terminals>
<nonterminals>
<: id="s17726_500" cat="NX">
<: edge label="HD" idref="s17726_1" />
</>
<: id="s17726_501" cat="VFIN">
<: edge label="HD" idref="s17726_2" />
</>
<: id="s17726_502" cat="NX">
<: edge label="HD" idref="s17726_3" />
</>
<: id="s17726_503" cat="ADV">
<: edge label="HD" idref="s17726_4" />
</>
<: id="s17726_504" cat="ADV">
<: edge label="HD" idref="s17726_5" />
<: edge label="refmod" idref="s17726_514" />
</>
<: id="s17726_505" cat="C">
<: edge label="-" idref="s17726_7" />
</>
<: id="s17726_506" cat="NX">
<: edge label="HD" idref="s17726_8" />
</>
<: id="s17726_507" cat="NX">
<: edge label="HD" idref="s17726_9" />
</>
<: id="s17726_508" cat="VFIN">
<: edge label="HD" idref="s17726_10" />
</>
<: id="s17726_509" cat="VF">
<: edge label="OA" idref="s17726_500" />

```

Tübingen Treebank of German/Spontaneous Speech (TüBa-D/S)

- ▶ transliterated dialogues of recorded, spontaneous speech
- ▶ subject domain: appointment scheduling
- ▶ consists of 38.000 dialogue turns

Task-oriented Verbmobil Dialogues

- ▶ Task
 - ▶ schedule a date for a one-and-a-half-day business trip
 - ▶ settle on mode of transportation and hotel
 - ▶ schedule meetings
 - ▶ arrange evening entertainment
- ▶ Recording
 - ▶ a close microphone
 - ▶ a room microphone
 - ▶ telephone

Data Transcription

The BAS Partitur Format:

- ▶ SAM compatible structure and entries.
- ▶ easy to extend by simple UNIX cat.
- ▶ open format, that is extensions to the format can be implemented without necessary alterations to the software reading the older format.

Data Transcription

- ▶ Time-aligned independent description of as many different levels of the speech signal as necessary. For instance: orthography, canonical transcript, phonology, phonetics, prosody, dialog acts, syntax tagging, semantics, ...
- ▶ Symbolic links between the independent levels allow logical assignments aside to the physical time scale. These links are based on the word units of the utterance.
- ▶ For more information see: www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html

A Sample Dialogue

N: Guten Tag, Frau Heinicke. Wie wir bereits ausgemacht haben, wollten wir auf eine eineinhalbtägige Geschäftsreise nach Hamburg fahren.

Hello, Mrs. Heinicke, as we have already arranged, we wanted to go on a business trip to Hamburg for one and a half days.

H: Ja , Grü Gott, Herr Nishimoto. Wir wollten jetzt, glaube ich , noch mal die Termine besprechen.

N: Well , hello , Mr. Nishimoto . now we want, I think , to discuss the times once again.

N: Ja , genau.

H: Yes , exactly.

Segmentation of Dialogue Data

Primary segmentation unit: dialogue turn

- ▶ a single, typically uninterrupted contribution to the dialog by one of the dialog participants
- ▶ may consist of one or more sentences in the grammatical sense and/or phrases
- ▶ preprocessed into syntactic units delimited by full stops and question marks

Characteristics of Spontaneous Speech

- ▶ Fragmentary Utterances
- ▶ Repetitions
- ▶ False starts
- ▶ Speech errors (with correction)
- ▶ Interruptions
- ▶ Parentheticals
- ▶ Discourse markers
- ▶ Hesitation noises

Fragmentary Utterances

| | | | | | |
|----------------|------------|-------|-----|----------|---------|
| meinetwegen | von | zehn | bis | dreizehn | Uhr |
| for me | from | ten | to | one | a'clock |
| Vorbereitungen | eigentlich | nicht | | | |
| preparations | really | not | | | |

Repetitions

Theater **wäre** **wäre** mal nicht schlecht
theater would be would be surely not bad
ja, das ist **das** **das** ist in Ordnung, genau
yes that is that that is all right exactly

False Starts

ja, also, das, wenn wir allerdings,
yes well that if we though

wenn wir mit dem Flugzeug fliegen
if we fly by plane

wie kommen wir dann nach Hannover rein ?
how do we then get into Hannover?

False Starts (2)

das ist schade, ich hätte diese erste Juniwoche
that is a pity I would have this first week of June

habe ich mehrere Besprechungen,
I have several meetings

die ich nicht verschieben möchte
which I do not want to move

Speech Errors

| | | | | |
|---------|-------|------------|----|----------|
| trotz | Nebel | Nebels | im | November |
| despite | fog | of the fog | in | November |

| | | | | | |
|------|--------|--------|-----|-----|---------|
| dann | machen | nehmen | wir | den | Flieger |
| then | make | take | we | the | plane |

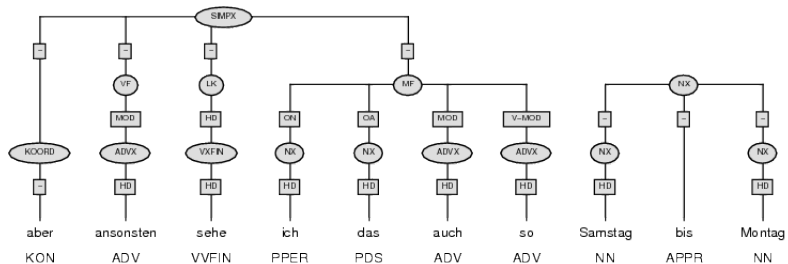
Interruptions

| | | | | | | |
|--------|---------|-----|----|------|------|------------|
| sieben | Uhr | fnf | am | | | |
| 7 | (hours) | 05 | at | das | heit | Moment |
| | | | | that | is | one moment |

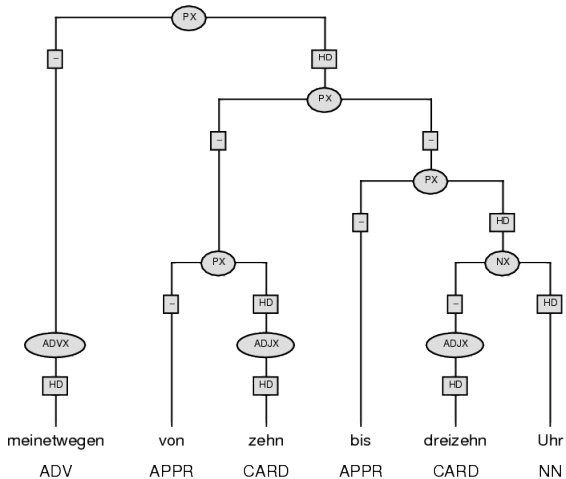
Annotation Principles

- ▶ Longest Match Principle
 - ▶ as many daughter nodes as possible are combined into a single mother node, provided that the resulting construction is syntactically as well as semantically well-formed.
 - ▶ Speech errors, repetitions, corrections, and hesitations are structured as much as possible, but are not typically connected to surrounding constituents as a whole.
- ▶ Flat Clustering Principle
 - ▶ Keeps the number of hierarchy levels in a syntactic structure as small as possible.
 - ▶ Any branching factor is allowed.

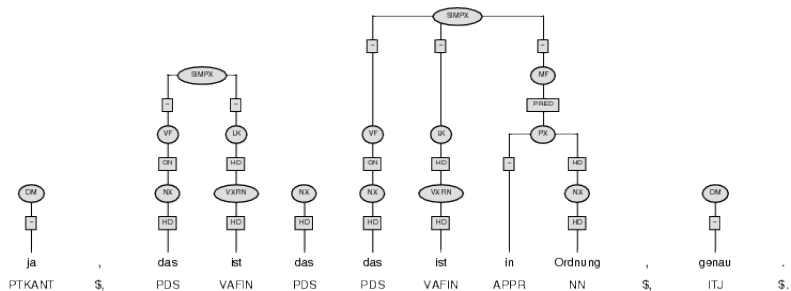
A Sample Sentence



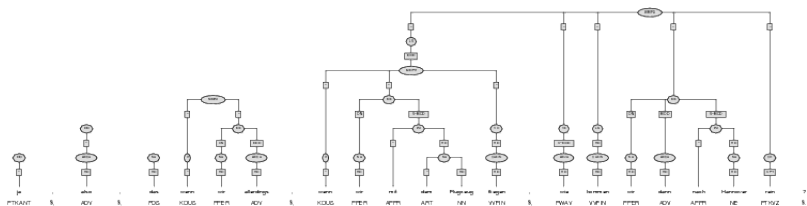
Fragmentary Utterances



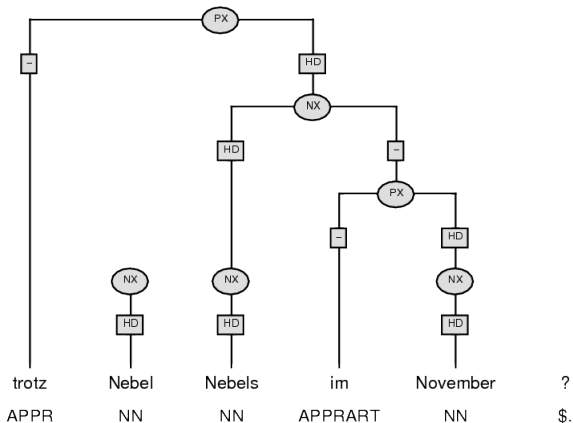
Repetitions



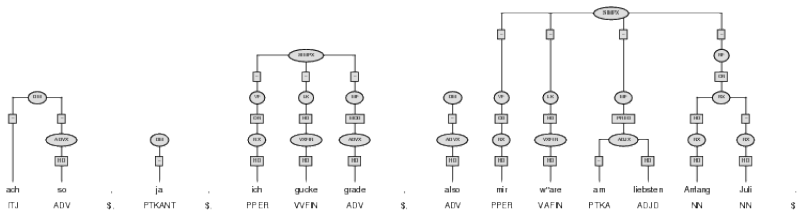
False Starts



Speech Errors



Discourse Markers



Some Concluding Remarks

- ▶ Treebanking is extremely labor-intensive (i.e. costly).
 - ▶ Good planning is therefore necessary.
 - ▶ Good tools are crucial.
 - ▶ For annotation, I recommend the tool Annotate.
- A detailed stylebook is essential.
- ▶ Every time you hire a well-trained linguist, your treebank will get better.

References

- ▶ @nnotate. <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html>
- ▶ Hinrichs, E., S. Kübler. (2005). Treebank Profiling of Spoken and Written German. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*. Barcelona, Spain.
- ▶ Hinrichs, E., S. Kübler, K. Naumann. (2005). A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations. *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, MI, June 2004.
- ▶ Hinrichs, E. and H. Wunsch. (2008). Selectional Preferences for Anaphora Resolution. Erhard Hinrichs and John Nerbonne (Eds.): *Theory and Evidence in Semantics*. CSLI Publications, Stanford University.

Readings (2)

- ▶ MMAX2. <http://mmax2.sourceforge.net/>
- ▶ PALinkA. Constantin Orasan (2005).
<http://clg.wlv.ac.uk/projects/PALinkA/>.
- ▶ Schiller, A., S. Teufel, C. Stöckert, and C. Thielen. (1999).
Guidelines für das Tagging deutscher Textcorpora mit STTS.
<http://www.ifi.uzh.ch/~siclemat/man/SchillerTeufel99STTS.pdf>
- ▶ Schmid, H., A. Fitschen, and U. Heid (2004). SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. *Proceedings of LREC 2004*.
<http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/LREC04/smor.pdf>.

Readings (3)

- ▶ Telljohann, H., E. Hinrichs, S. Kübler, H. Zinsmeister, and K. Beck (1999). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). <http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-0911.pdf>
- ▶ Versley, Y. (2008). Vagueness and Referential Ambiguity in a Large-scale Annotated Corpus. Massimo Poesio and Ron Artstein (Eds.): *Ambiguity in Anaphora*. Special Issue of the Journal on Research in Language and Computation.
- ▶ Versley, Y., K. Beck, E. Hinrichs, and H. Telljohann. (2010). A Syntax-first Approach to High-quality Morphological Analysis and Lemma Disambiguation for the TüBa-D/Z Treebank. *Proceedings of TLT9, Tartu*.

Readings (4)

- ▶ Zinsmeister, H., E. Hinrichs, S. Kübler, A. Witt. (2008). Linguistically annotated corpora: Quality assurance, reusability and sustainability. A. Lüdeling, M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

TIGERSearch Queries for TüBa-D/Z

Series of search queries on different types of accusative objects

1) premodifiers

```
#field >OA #nx & #nx > [pos="NN"]
```

2) several modifiers

```
#field >OA #nx & #nx > [pos="NN"] &  
#nx > #a1:[cat="ADJX"] & #nx > #a2:[cat="ADJX"]  
& #a1 .* #a2
```

3) complex modifiers

```
#field >OA #nx & #nx > [pos="NN"] & #nx > #a:[cat="ADJX"]  
& #a > [cat="NX"]
```

4) postmodifiers

```
#field >OA #nx & #nx > [cat="PX"]
```


5) discontinuous constituents

```
#field >0A\ -MOD #nx
```

6) discontinuous constituents with secondary edge

```
#field >0A\ -MOD #nx & #x >~#nx
```

Two more searches on interesting phenomena:

7) really big sentences

```
#vroot:[cat="VROOT"] >@l #l:[T] & #vroot >@r #r:[T]
& #l .80,100 #r
```

8) exampleon errors in the source data

```
[word="hierhin"] . [word="bitte"]
```

If you would like to get support on TIGERSearch queries, Kathrin Beck will be happy to assist you.