

CS 498

Expressive Grammars for Natural Language Processing: Theory and applications

Lecture 9

Julia Hockenmaier
juliahmr@cs.uiuc.edu
3324 Siebel Center

What we've done so far

We can build a parser for the Wall Street Journal

- The Penn Treebank
- (Probabilistic) context-free grammars
- Statistical Treebank parsers
- Trace recovery systems

➔ Do these techniques transfer to other languages?

Today's lecture: Parsing German

- **German syntax:**
 - Verb-final & verb-second word order
 - Topological fields
 - Scrambling
- **German treebanks:**
 - Negra
 - Tiger
 - TüBa/DZ
- **German parsers:**
 - Dubey/Keller '03
 - Kübler et al '06

German syntax (+ morphology)

German morphology

- **Inflectional morphology** (changes in word form):
 - **Verbs are inflected for person/number:** *ich sehe, du siehst,...*
Some have separable prefixes: *aufgeben*, but *ich gebe auf* and an infix 'zu' in the infinitive (*aufzugeben*, vs. *zu sehen*)
 - **Nouns/Adjectives have 4 cases:** nominative [subject], genitive [possessive], dative [ind. object], accusative [direct object]
- **Derivational morphology** (word₁(+word₂) -> new word):
 - **Compound nouns:** *Donaudampfschiffahrtskapitän* (Danube steam boat captain)
- **Some prepositions + determiner merge:**
 - *in + dem* (*in + the*) = *im*

German word order

- **Scrambling:** Verb is fixed, arguments/modifiers can move
- **Main clauses are verb-second:**
(*Peter gives Mary the book*)
... VERB ...
Peter gibt Maria das Buch
Maria gibt Peter das Buch
Das Buch gibt Peter Maria
- **Subordinate clauses are verb-final:**
(*[I know] that Peter gives Mary the book*)
COMP ... VERB
dass Peter Maria das Buch gibt
dass Peter das Buch Maria gibt
dass das Buch Peter Maria gibt

Topological fields

Vorfeld	Left Bracket	Mittelfeld	Right Bracket	Nachfeld
Maria	liest	das Buch morgen		das Peter ihr gegeben hat
Das Buch		Maria morgen		
Morgen		Maria das Buch		
Maria	wird	das Buch morgen	gelesen haben	
Das Buch		Maria morgen		
Morgen		Maria das Buch		
	dass	Maria das Buch morgen	gelesen haben wird	

Translation:

1. *Maria* reads the book that Peter has given her tomorrow.
2. *Maria* will have read the book that Peter has given her tomorrow.
3. ...that *Maria* will have read the book that Peter has given her tomorrow

German treebanks I: Negra and Tiger

Stuttgart-Tübingen Tag Set (STTS) (Schiller *et al.* 1995)

- **Similar level of granularity as Penn Treebank:**
 - 2 kinds of nouns: common nouns vs. proper nouns
 - 2 kinds of adjectives: attributive vs predicative
 - 5 finite verb forms, 4 infinitival, 3 perfect participle
 - 12 kinds of pronouns
 - 3 kinds of conjunctions
 - 5 kinds of particles
- **54 POS tags**

Negra

- **355,096 tokens / 20,602 sentences of newspaper text** (Frankfurter Rundschau)
- **25 node labels = syntactic categories** (similar to Penn Treebank, but with different labels for coordination)
- **45 edge labels = grammatical functions** (every edge is labeled, more detailed than Penn TB)
- **Crossing branches** (can be converted to Penn TB style traces)

Negra edge labels

- head, predicate
- subject, subject or predicate, accusative object, second accusative object, dative object, genitive object, clausal object
- modifier, apposition, postnominal modifier,
- complementizer
- conjunction, conjunct, junctor
- discourse marker, discourse-level head
- reported speech
- ...

Tiger corpus

- **900,00 tokens (50,000 sentences) of newspaper text** (Frankfurter Rundschau)
- **Annotation: further development of Negra** (parsed with shallow parser or XLE grammar (??), then manually disambiguated)
- **Full morphological analysis + lemma** (& POS tags)
- **27.5% of graphs are discontinuous**
- **“Secondary edges” for coordination**

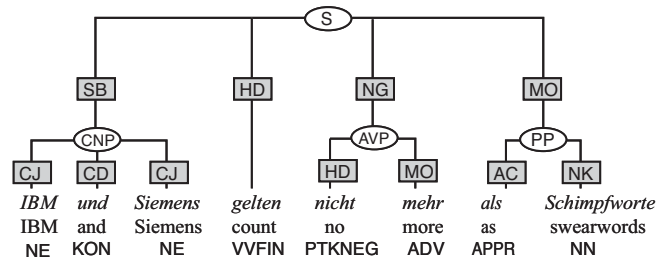
Negra vs. Tiger

(Brants & Hansen, LREC '02)

	Negra	Tiger
Verb subcategorization of PPs	All PPs are modifiers	PPs are: - objects - modifiers - part of collocation
Shared arguments in coordination (right-node raising, arg. cluster coord.)	only one link, one dependency is missing	'secondary edges' to indicate all dependencies
NP1, NP2 appositives (coref) vs. parentheticals (no coref)	same edge label	two different edge labels

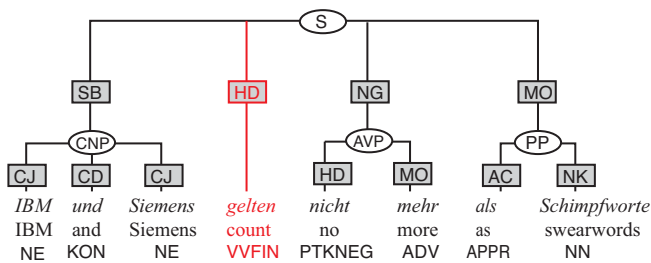
13

A simple example



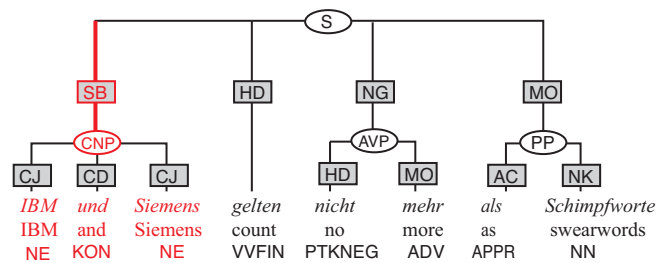
14

A simple example



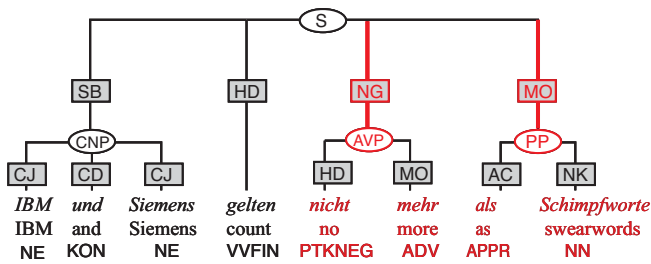
15

A simple example



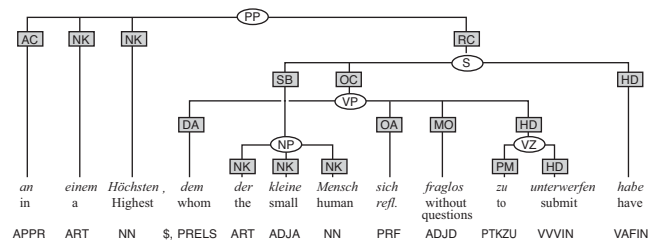
16

A simple example



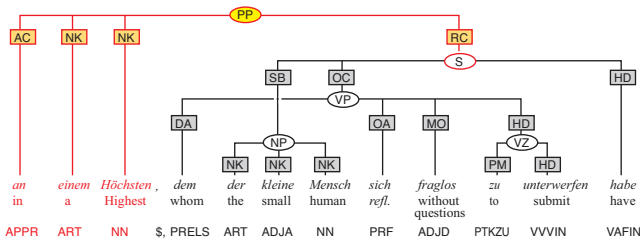
17

Wh-extraction in Tiger



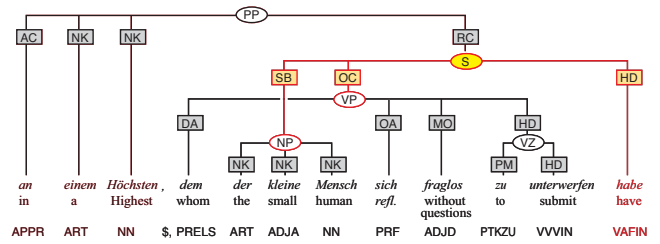
18

Wh-extraction in Tiger



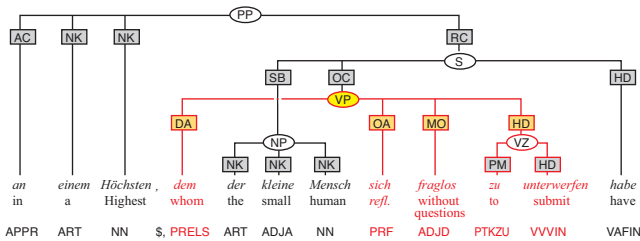
19

Wh-extraction in Tiger



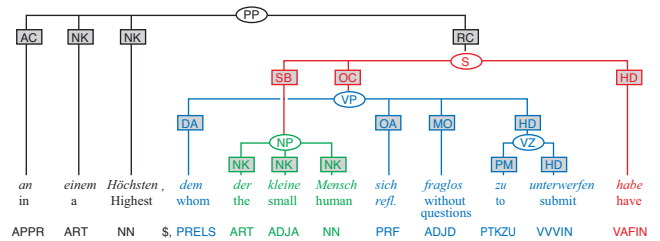
20

Wh-extraction in Tiger



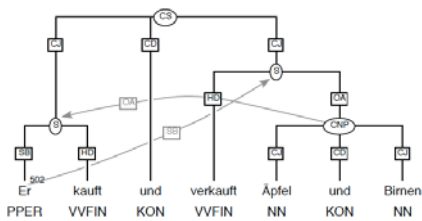
21

Crossing branches



22

Coordination in Tiger



- Explicit annotation (CS/CNP nodes, CJ/CD edges)
- Secondary edges indicate non-local dependencies
- Negra: same, but without secondary edges

23

Parsing Negra/Tiger

- In order to use PCFGs etc., the original graphs are translated into trees without crossing branches
- This is done by lifting crossing branches to the highest node.
- This creates very flat trees, and changes the dependency structure !!!

24

Negra, "Penn Treebank"-style

```
((CS (S-CJ (NP-SB (PPER-PH Es)
                (*Tl*-RE -))
            (VVFIN-HD spielt)
            (ADV-MO eben)
            (NP-OA (PIAT-NK keine)
                  (NN-NK Rolle))
            ($, .)
            (S-*Tl* (KOUS-CP ob)
                   (NP-SB (ART-NK die)
                           (NN-NK Musik))
                   (ADJD-PD gefällig)
                   (VAFIN-HD ist)))
            ($*LRB* -)
            (S-CJ (ADV-MO nur)
                  (NP-*T2* (PIAT-NK etwas)
                          ($*LRB* ")
                          (NN-NK Neues))
                  ($*LRB* ")
                  (VMFIN-HD muß)
                  (PPER-SB sie)
                  (CVP-OC (VP-CJ (*T2*-PD -)
                                (VAINF-HD sein))
                          (KON-CD und)
                          (VP-CJ (NP-OA (ART-NK eine)
                                      (ADJA-NK eigene)
                                      (NN-NK Handschrift))
                                (VVINF-HD aufweisen)))))))
```

Negra, "Penn Treebank"-style

Sentential coordination

```
((CS (S-CJ (NP-SB (PPER-PH Es)
                (*Tl*-RE -))
            (VVFIN-HD spielt)
            (ADV-MO eben)
            (NP-OA (PIAT-NK keine)
                  (NN-NK Rolle))
            ($, .)
            (S-*Tl* (KOUS-CP ob)
                   (NP-SB (ART-NK die)
                           (NN-NK Musik))
                   (ADJD-PD gefällig)
                   (VAFIN-HD ist)))
            ($*LRB* -)
            (S-CJ (ADV-MO nur)
                  (NP-*T2* (PIAT-NK etwas)
                          ($*LRB* ")
                          (NN-NK Neues))
                  ($*LRB* ")
                  (VMFIN-HD muß)
                  (PPER-SB sie)
                  (CVP-OC (VP-CJ (*T2*-PD -)
                                (VAINF-HD sein))
                          (KON-CD und)
                          (VP-CJ (NP-OA (ART-NK eine)
                                      (ADJA-NK eigene)
                                      (NN-NK Handschrift))
                                (VVINF-HD aufweisen)))))))
```

Negra, "Penn Treebank"-style

Sentential coordination

VP coordination

```
((CS (S-CJ (NP-SB (PPER-PH Es)
                (*Tl*-RE -))
            (VVFIN-HD spielt)
            (ADV-MO eben)
            (NP-OA (PIAT-NK keine)
                  (NN-NK Rolle))
            ($, .)
            (S-*Tl* (KOUS-CP ob)
                   (NP-SB (ART-NK die)
                           (NN-NK Musik))
                   (ADJD-PD gefällig)
                   (VAFIN-HD ist)))
            ($*LRB* -)
            (S-CJ (ADV-MO nur)
                  (NP-*T2* (PIAT-NK etwas)
                          ($*LRB* ")
                          (NN-NK Neues))
                  ($*LRB* ")
                  (VMFIN-HD muß)
                  (PPER-SB sie)
                  (CVP-OC (VP-CJ (*T2*-PD -)
                                (VAINF-HD sein))
                          (KON-CD und)
                          (VP-CJ (NP-OA (ART-NK eine)
                                      (ADJA-NK eigene)
                                      (NN-NK Handschrift))
                                (VVINF-HD aufweisen)))))))
```

Negra, "Penn Treebank"-style

Sentential coordination

VP coordination

Expletive "trace"

```
((CS (S-CJ (NP-SB (PPER-PH Es)
                (*Tl*-RE -))
            (VVFIN-HD spielt)
            (ADV-MO eben)
            (NP-OA (PIAT-NK keine)
                  (NN-NK Rolle))
            ($, .)
            (S-*Tl* (KOUS-CP ob)
                   (NP-SB (ART-NK die)
                           (NN-NK Musik))
                   (ADJD-PD gefällig)
                   (VAFIN-HD ist)))
            ($*LRB* -)
            (S-CJ (ADV-MO nur)
                  (NP-*T2* (PIAT-NK etwas)
                          ($*LRB* ")
                          (NN-NK Neues))
                  ($*LRB* ")
                  (VMFIN-HD muß)
                  (PPER-SB sie)
                  (CVP-OC (VP-CJ (*T2*-PD -)
                                (VAINF-HD sein))
                          (KON-CD und)
                          (VP-CJ (NP-OA (ART-NK eine)
                                      (ADJA-NK eigene)
                                      (NN-NK Handschrift))
                                (VVINF-HD aufweisen)))))))
```

Negra, "Penn Treebank"-style

Sentential coordination

VP coordination

Expletive "trace"

Topicalization trace

```
((CS (S-CJ (NP-SB (PPER-PH Es)
                (*Tl*-RE -))
            (VVFIN-HD spielt)
            (ADV-MO eben)
            (NP-OA (PIAT-NK keine)
                  (NN-NK Rolle))
            ($, .)
            (S-*Tl* (KOUS-CP ob)
                   (NP-SB (ART-NK die)
                           (NN-NK Musik))
                   (ADJD-PD gefällig)
                   (VAFIN-HD ist)))
            ($*LRB* -)
            (S-CJ (ADV-MO nur)
                  (NP-*T2* (PIAT-NK etwas)
                          ($*LRB* ")
                          (NN-NK Neues))
                  ($*LRB* ")
                  (VMFIN-HD muß)
                  (PPER-SB sie)
                  (CVP-OC (VP-CJ (*T2*-PD -)
                                (VAINF-HD sein))
                          (KON-CD und)
                          (VP-CJ (NP-OA (ART-NK eine)
                                      (ADJA-NK eigene)
                                      (NN-NK Handschrift))
                                (VVINF-HD aufweisen)))))))
```

Dubey & Keller '03: Parsing Negra

Dubey & Keller '03

Do standard parsing techniques carry over to German?

Models:

- Unlexicalized PCFG (+ grammatical functions)
- Lexicalized PCFG (+ grammatical functions, + pooling)
- Collins Model 1 with standard head-head dependencies,
- Collins Model 1 with sister-head dependencies (like baseNP)

Corpus:

- PTB-style Negra (gold and TnT POS tags)
- PTB-style Negra with NPs inserted into PPs

27

Two lexicalized models

Lexicalized PCFG (Carroll/Root):

- PCFG with headed rules: $S \rightarrow NP VP'$
- $P(X \rightarrow \alpha | X) P(w_{sister} | Sister, Parent, w_{Head})$

Markov PCFG (Collins Model 1):

- $P(Head | Parent)$
- $\times \prod \{ P(Sister | Parent, Head, dir)$
- $\times P(w_{sister}, t_{sister} | Sister, Parent, w_P)$

28

Dubey/Keller'03: Experiment 1

	F-score TnT tags	Coverage TnT tags	F-score Gold tags	Coverage Gold tags	
PCFG	68.6	94.4%	71.5	95.3%	Best
PCFG + GF	70.5	79.2%	79.7	65.4%	
Lexicalized PCFG	63.8	94.4%	66.9	95.3%	
Lexicalized PCFG + GF	63.8	79.2%	78.9	65.4%	
Collins Model 1	67.0	95.2%	67.8	96.2%	

Note: In GF-setting, gold tags include grammatical functions

- Unlexicalized PCFG performs best.
- Grammatical functions decrease coverage.

29

Learning curves

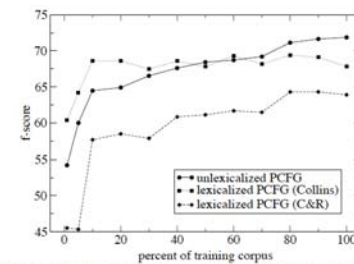


Figure 1: Learning curves for all three models

- Collins initially outperforms PCFG, but tapers off
- Lex.PCFG consistently much lower than PCFG
- steep initial rise "evidence against data sparsity" (????)

30

Experiment II

	F-score TnT tags	Coverage TnT tags	F-score Gold tags	Coverage Gold tags
Collins Model 1	67.0	95.2%	67.8	96.2%
PP+NP train+eval	73.8	95.1%	75.6	93.8%
PP+NP only train	66.3	95.1%	67.8	93.8%
Sister-head NP	66.9	95.1%	70.9	94.6%
Sister-head PP	69.3	94.8%	72.8	94.5%
Sister-head all	71.1	95.9%	74.1	95.2%

Insert NPs into PPs: Change flat PP into PP \rightarrow P NP structure.
Is low performance due to annotation style?

Sister-head: Dependencies between adjacent sisters, not head and other children
(this also helps with NPs in the Penn Treebank)

31

Dubey'05: Parsing German with suffix analysis and smoothing

- Same data set as Dubey & Keller '03.
- Changes to the grammatical function labels, a simple model of morphology, and smoothing of the unlexicalized part of the model
- Overall F-score: 76.3
(vs. 74.1 in Dubey/Keller'03)
Better, but still much lower than English!

32

German treebanks II: TüBa/DZ

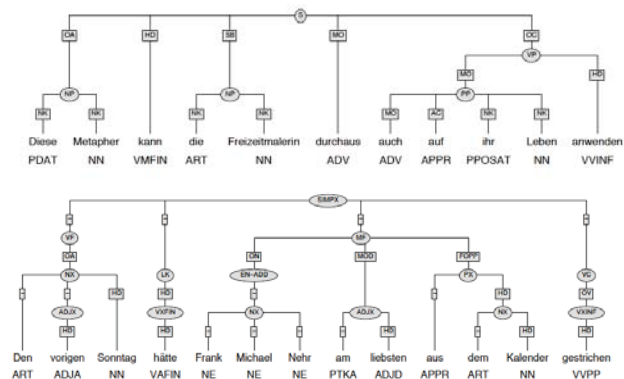
TüBa/DZ

- **15,000 sentences newspaper (taz)**
 - now 27,000 sentences, 470,000 words;
 - 23,500 sentences with co-reference and anaphora
- **Like Negra/Tiger, mixed phrase-structure/dependency annotation...**
 - ...but with **topological field** information
 - ... and **without discontinuous constituents** (crossing branches)
- **36 grammatical functions**

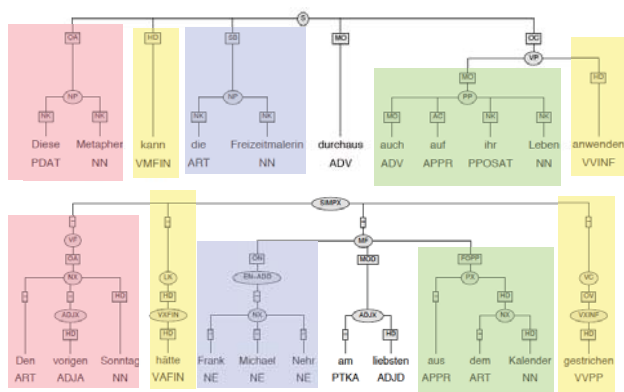
Negra vs TüBa-D/Z

	Negra	TüBa-D/Z
Unary branching	No: One-word constituents have no non-terminal label	Yes: topological fields and one-word constituents
Phrase-internal structure	flat (e.g PPs, S(BAR))	more internal structure
Discontinuous constituents	crossing branches	function labels (no coindexation?)

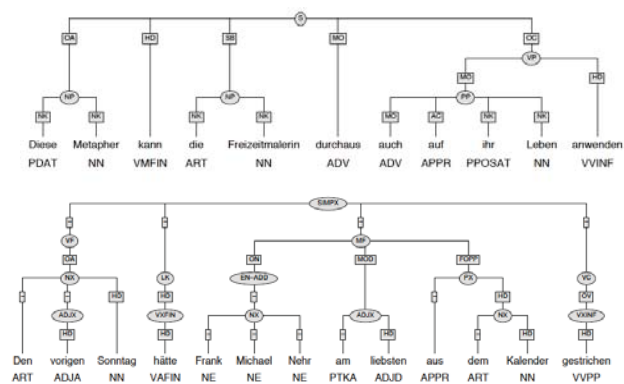
Negra vs. TüBa/DZ



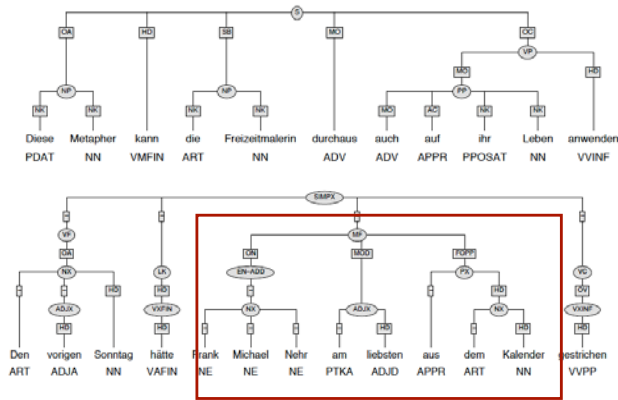
Negra vs. TüBa/DZ



Negra vs. TüBa/DZ

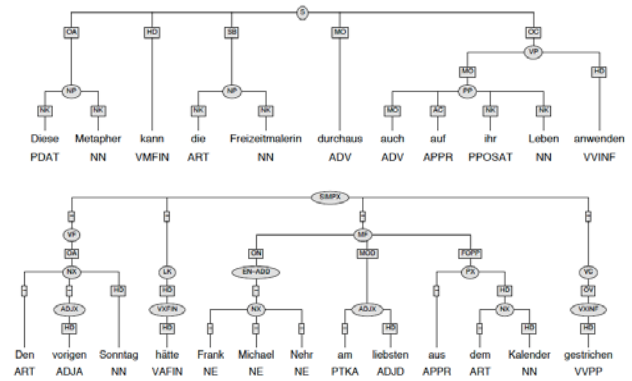


Negra vs. TüBa/DZ



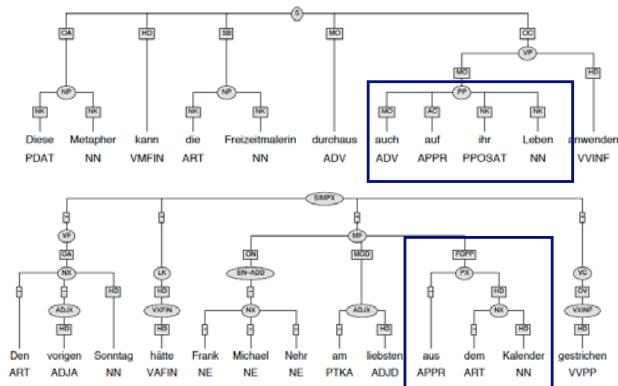
36

Negra vs. TüBa/DZ



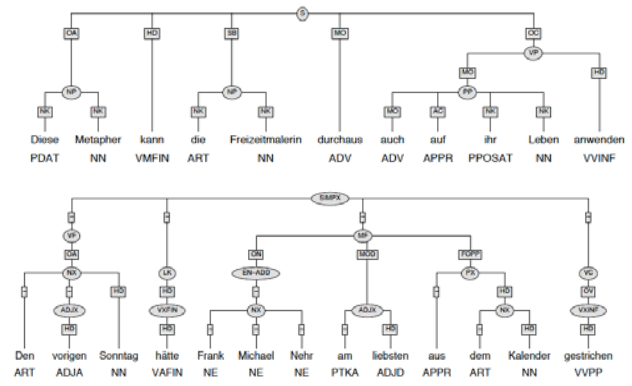
36

Negra vs. TüBa/DZ



36

Negra vs. TüBa/DZ



36

Kübler et al. '06: Parsing Negra and TüBa/DZ

37

Kübler et al., 2006

Is it really that difficult to parse German?

- **Models:**
 - Unlexicalized PCFG (Stanford parser and Lopar)
 - Lexicalized PCFG (Stanford parser: factored model)
- **Corpus:**
 - PTB-style Negra (gold and TnT POS tags)
 - TüBa/DZ
- **Evaluation:**
 - evalb
 - evalb(?) plus grammatical functions (just labels, not dependencies)

38

Factored lexicalized PCFGs

(Klein/Manning '02)

Probability of dependency tree D and CFG tree T is assumed to be independent:

$$P(D, T) = P(D)P(T)$$

$$P(D): \prod P(w_{\text{dep}}, t_{\text{dep}} \mid w_{\text{head}}, t_{\text{head}}, \text{DIR})$$

39

Results (constituents)

	Negra			TüBa/DZ			
LoPar	70.8	72.5	71.7	92.6	88.6	90.6	U
	65.9	67.4	66.6	87.4	83.6	85.4	L
Stanford PCFG	71.2	72.7	72.0	93.1	89.4	91.2	U
	66.3	67.6	66.9	88.3	84.8	86.5	L
Stanford PCFG + Markov	74.1	74.1	74.1	92.3	90.9	91.6	U
	70.0	70.0	70.0	89.9	88.5	86.5	L
Stanford lex. PCFG	71.3	73.1	72.2	91.6	91.2	91.4	U
	66.3	68.0	67.1	89.1	88.6	88.9	L
	P	R	F	P	R	F	

40

Results (grammatical functions)

	Negra			TüBa/DZ		
	LP	LR	F	LP	LR	F
no GFs	70.0	70.0	70.0	89.9	88.5	89.1
all GFs	47.2	56.4	51.4	75.7	74.9	75.3
subjects	52.5	58.0	55.1	66.8	75.9	71.1
acc. objects	35.1	36.3	35.7	43.8	47.3	45.5
dat. objects	8.4	3.6	5.0	24.5	10.0	14.1

Evaluation: constituent label (+ grammatical function label)
Note: these are not dependencies!

41

What can we take away from this?

- With TüBa/DZ, scores are in the same range as Penn Treebank results for English.
 - Is Negra the wrong representation?
 - Are the numbers skewed? (because there are fewer constituents overall or more easy constituents)
 - What about dependency recovery?

42