- Jurafsky, D. and Martin, J. H. (2009): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Pearson: New Jersey: Chapter 14
- Manning, C. D. and Schütze, H. (1999): Foundations of Statistical Natural Language Processing. MIT Press: Cambridge, Massachusetts. Chapters 11, 12.
- with further examples by Ray Mooney, UT at Austin

PCFGs, probabilistic CYK, dependency parsing

STATISTICAL PARSING

STATISTICAL PARSING

- Statistical parsing uses a probabilistic model of syntax in order to assign probabilities to each parse tree.
- Provides principled approach to resolving syntactic ambiguity.
- Allows supervised learning of parsers from tree-banks of parse trees provided by human linguists.
- Also allows unsupervised learning of parsers from unannotated text, but the accuracy of such parsers has been limited.

PROBABILISTIC CONTEXT FREE GRAMMAR (PCFG)

A **probabilistic context free grammar** $PCFG=(W,N,N_1,R,P)$ consists of

- terminal vocabulary $W=\{w_1,\ldots, w_V\}$
- set of non-terminals $N=\{N_1,., N_n\}$
- start symbol $N_1 \in N$
- set of rules $R:\{N_i \rightarrow D_j\}$, where D_j is a sequence over $W \cup N$
- corresponding set of probabilities on rules P such that the sum of probabilities per LHS is 1

- A PCFG is a probabilistic version of a CFG where each production has a probability.
- Probabilities of all productions rewriting a given non-terminal must add to 1, defining a distribution for each non-terminal.
- String generation is now probabilistic where production probabilities are used to non-deterministically select a production for rewriting a given non-terminal.

SIMPLE PCFG FOR A SUBSET OF ENGLISH

Grammar / Prob. / Lexicon

Grammar	Prob.	
S → NP VP	0.8	
S → Aux NP VP	0.1	+ 1.0
S → VP	0.1	
NP → Pronoun	0.2	
NP → Proper-Noun	0.2	+ 1.0
NP → Det Nominal	0.6	
Nominal → Noun	0.3	
Nominal → Nominal Noun	0.2	+ 1.0
Nominal → Nominal PP	0.5	
VP → Verb	0.2	
VP → Verb NP	0.5	+ 1.0
VP → VP PP	0.3	
PP → Prep NP	1.0	

Lexicon

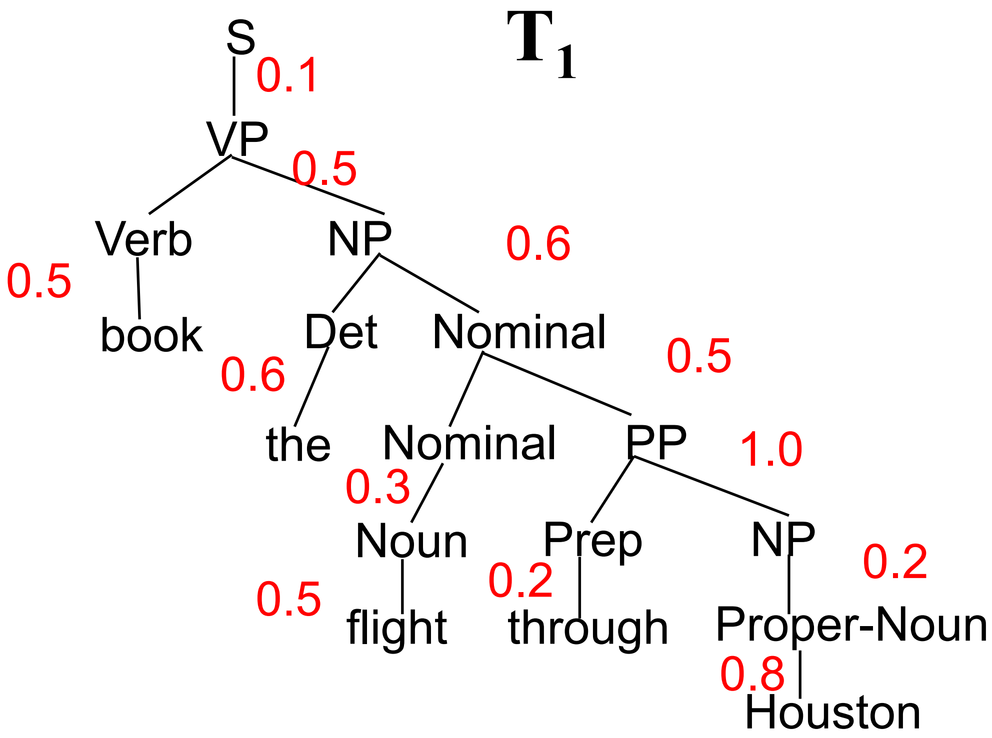
Det →	the	a	that	this	
	0.6	0.2	0.1	0.1	
Noun →	book	flight	meal	money	
	0.1	0.5	0.2	0.2	
Verb →	book	include	prefer		
	0.5	0.2	0.3		
Pronoun →	I	he	she	me	
	0.5	0.1	0.1	0.3	
Proper-Noun →	Houston	NWA			
	0.8	0.2			
Aux →	does				
	1.0				
Prep →	from	to	on	near	through
	0.25	0.25	0.1	0.2	0.2

DERIVATION PROBABILITY

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Assume productions for each node are chosen independently.
- Probability of derivation is the product of the probabilities of its productions.

$$P(T_1) = 0.1 \times 0.5 \times 0.5 \times 0.6 \times 0.6 \times 0.5 \times 0.3 \times 1.0 \times 0.2 \times 0.2 \times 0.5 \times 0.8 = 2.16\ E\text{-}5$$

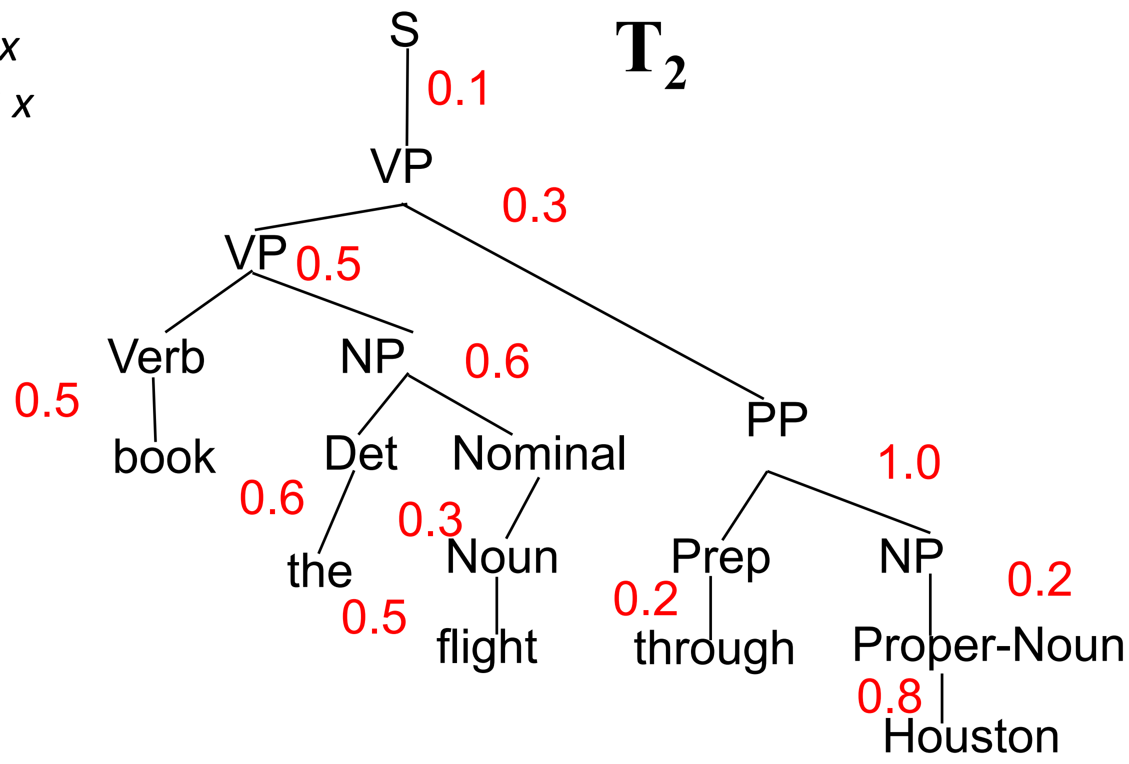


SYNTACTIC DISAMBIGUATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Resolve ambiguity by picking most probable parse tree.

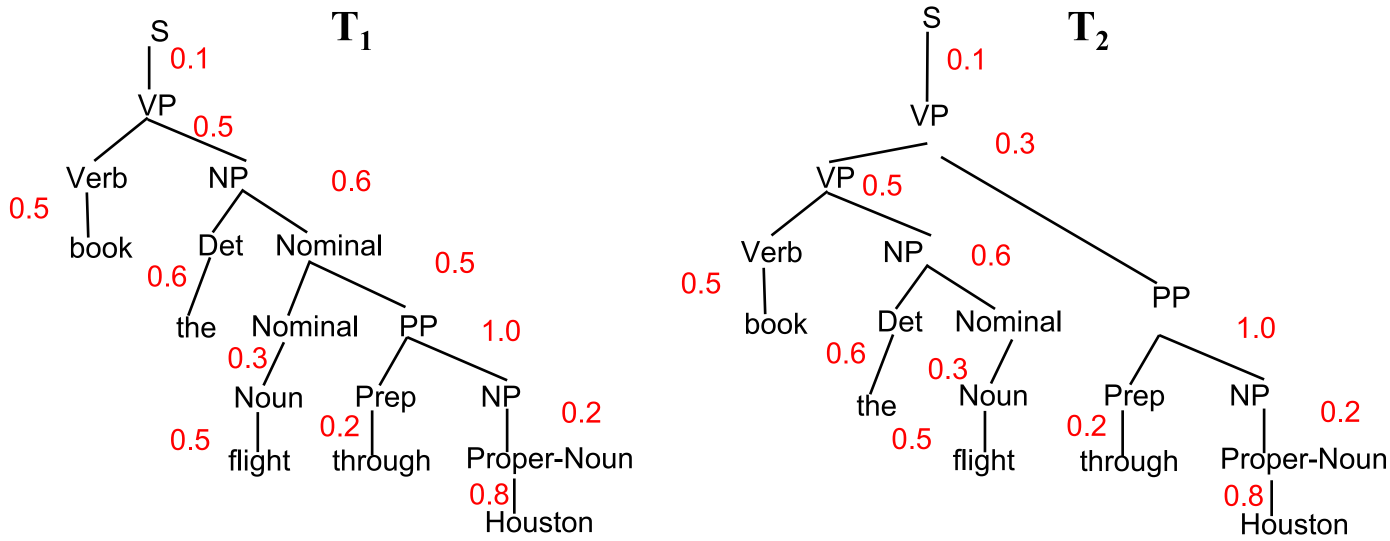
$$P(T_2) = 0.1 \times 0.3 \times 0.5 \times 0.6 \times 0.5 \times 0.6 \times 0.3 \times 1.0 \times 0.5 \times 0.2 \times 0.2 \times 0.8 = 1.296\ E\text{-}5$$



SENTENCE PROBABILITY

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Probability of a sentence is the sum of the probabilities of all of its derivations



*P(“*book the flight through Houston*”) = P(T_1)+P(T_2)=2.16 E-5+1.296 E-5*

= 3.456 E-5

THREE TASKS FOR PCFGS

- observation likelihood: how do we **efficiently compute** the probability of a sentence, given a PCFG?
- most likely derivation: given a PCFG and a sentence, how do we find the **derivation that best explains** the sentence?
- Given a set of sentences and a space of possible PCFGs, how do we find the **PCFG parameters that best explain** the observations? This is is called **training** of the PCFG

Sounds familiar?

PROBABILISTIC CKY

- An analog to the Viterbi algorithm to efficiently determine the most probable derivation (parse tree) for a sentence.
- CKY can be modified for PCFG parsing by including in each cell a probability for each non-terminal.
- Cell[i,j] must retain the most probable derivation of each constituent (non-terminal) covering words i +1 through j together with its associated probability.
- When transforming the grammar to CNF, must set production probabilities to preserve the probability of derivations.

PROBABILISTIC CONVERSION TO CNF

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Original Grammar

S → NP VP	0.8
S → Aux NP VP	0.1
S → VP	0.1
NP → Pronoun	0.2
NP → Proper-Noun	0.2
NP → Det Nominal	0.6
Nominal → Noun	0.3
Nominal → Nominal Noun	0.2
Nominal → Nominal PP	0.5
VP → Verb	0.2
VP → Verb NP	0.5
VP → VP PP	0.3
PP → Prep NP	1.0

Chomsky Normal Form

S → NP VP	0.8
S → X1 VP	0.1
X1 → Aux NP	1.0
S → book | include | prefer	0.01 0.004 0.006
S → Verb NP	0.05
S → VP PP	0.03
NP → I | he | she | me	0.1 0.02 0.02 0.06
NP → Houston | NWA	0.16 .04
NP → Det Nominal	0.6
Nominal → book | flight | meal | money	0.03 0.15 0.06 0.06
Nominal → Nominal Noun	0.2
Nominal → Nominal PP	0.5
VP → book | include | prefer	0.1 0.04 0.06
VP → Verb NP	0.5
VP → VP PP	0.3
PP → Prep NP	1.0

PROBABILISTIC CYK PARSING

Book the flight through Houston

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None			
	Det:.6	NP:.6*.6*.15 =.054		
		Nominal:.15 Noun:.5		

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	VP:.5*.5*.054 =.0135		
	Det:.6	NP:.6*.6*.15 =.054		
		Nominal:.15 Noun:.5		

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

Language Technology

PROBABILISTIC CYK PARSING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135		
	Det:.6	NP:.6*.6*.15 =.054		
		Nominal:.15 Noun:.5		

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

Language Technology

PROBABILISTIC CYK PARSING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Book the flight through Houston

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	
		Nominal:.15 Noun:.5	None	
			Prep:.2	

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

Language Technology

PROBABILISTIC CYK PARSING

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	
		Nominal:.15 Noun:.5	None	
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.05*.5* .000864 =.0000216
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.03*.0135*.032 =.00001296 S:.0000216
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6*.0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

- S → NP VP 0.8
- S → X1 VP 0.1
- X1 → Aux NP 1.0
- S → book | include | prefer
  0.01 0.004 0.006
- S → Verb NP 0.05
- S → VP PP 0.03
- NP → I | he | she | me
  0.1 0.02 0.02 0.06
- NP → Houston | NWA
  0.16 .04
- NP → Det Nominal 0.6
- Nominal → book | flight | meal | money
  0.03 0.15 0.06 0.06
- Nominal → Nominal Noun 0.2
- Nominal → Nominal PP 0.5
- VP → book | include | prefer
  0.1 0.04 0.06
- VP → Verb NP 0.5
- VP → VP PP 0.3
- PP → Prep NP 1.0

- Aux → does
  1.0
- Det → the | a | that | this
  0.6 0.2 0.1 0.1
- Pronoun → I | he | she | me
  0.5 0.1 0.1 0.3
- Verb → book | include | prefer
  0.5 0.2 0.3
- Noun → book | flight | meal | money
  0.1 0.5 0.2 0.2
- Proper-Noun → Houston | NWA
  0.8 0.2
- Prep → from | to | on | near | through
  0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.0000216
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

Pick most probable parse, i.e. take max to combine probabilities of multiple derivations of each constituent in each cell.

S → NP VP 0.8
S → X1 VP 0.1
X1 → Aux NP 1.0
S → book | include | prefer
0.01 0.004 0.006
S → Verb NP 0.05
S → VP PP 0.03
NP → I | he | she | me
0.1 0.02 0.02 0.06
NP → Houston | NWA
0.16 .04
NP → Det Nominal 0.6
Nominal → book | flight | meal | money
0.03 0.15 0.06 0.06
Nominal → Nominal Noun 0.2
Nominal → Nominal PP 0.5
VP → book | include | prefer
0.1 0.04 0.06
VP → Verb NP 0.5
VP → VP PP 0.3
PP → Prep NP 1.0

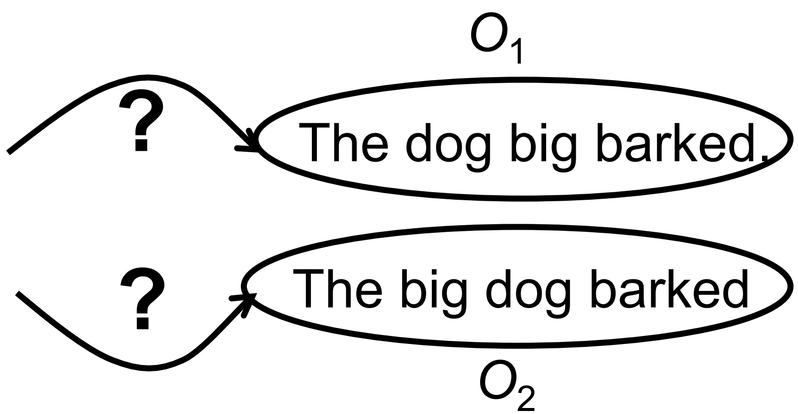
Aux → does
1.0
Det → the | a | that | this
0.6 0.2 0.1 0.1
Pronoun → I | he | she | me
0.5 0.1 0.1 0.3
Verb → book | include | prefer
0.5 0.2 0.3
Noun → book | flight | meal | money
0.1 0.5 0.2 0.2
Proper-Noun → Houston | NWA
0.8 0.2
Prep → from | to | on | near | through
0.25 0.25 0.1 0.2 0.2

PCFG: OBSERVATION LIKELIHOOD

- There is an analog to Forward algorithm for HMMs called the **Inside algorithm** for **efficiently** determining how likely a string is to be produced by a PCFG.
- Can use a PCFG as a syntax-based **language model** to choose between alternative sentences for speech recognition or machine translation.

S → NP VP	0.9
S → VP	0.1
NP → Det A N	0.5
NP → NP PP	0.3
NP → PropN	0.2
A → ε	0.6
A → Adj A	0.4
PP → Prep NP	1.0
VP → V NP	0.7
VP → VP PP	0.3

Grammar



$$P(O_2 \mid \text{Grammar}) > P(O_1 \mid \text{Grammar})\ ?$$

INSIDE ALGORITHM

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Book	the	flight	through	Houston
S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.00001296 +.0000216 =.00003456 S:.00001296 S:.0000216
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

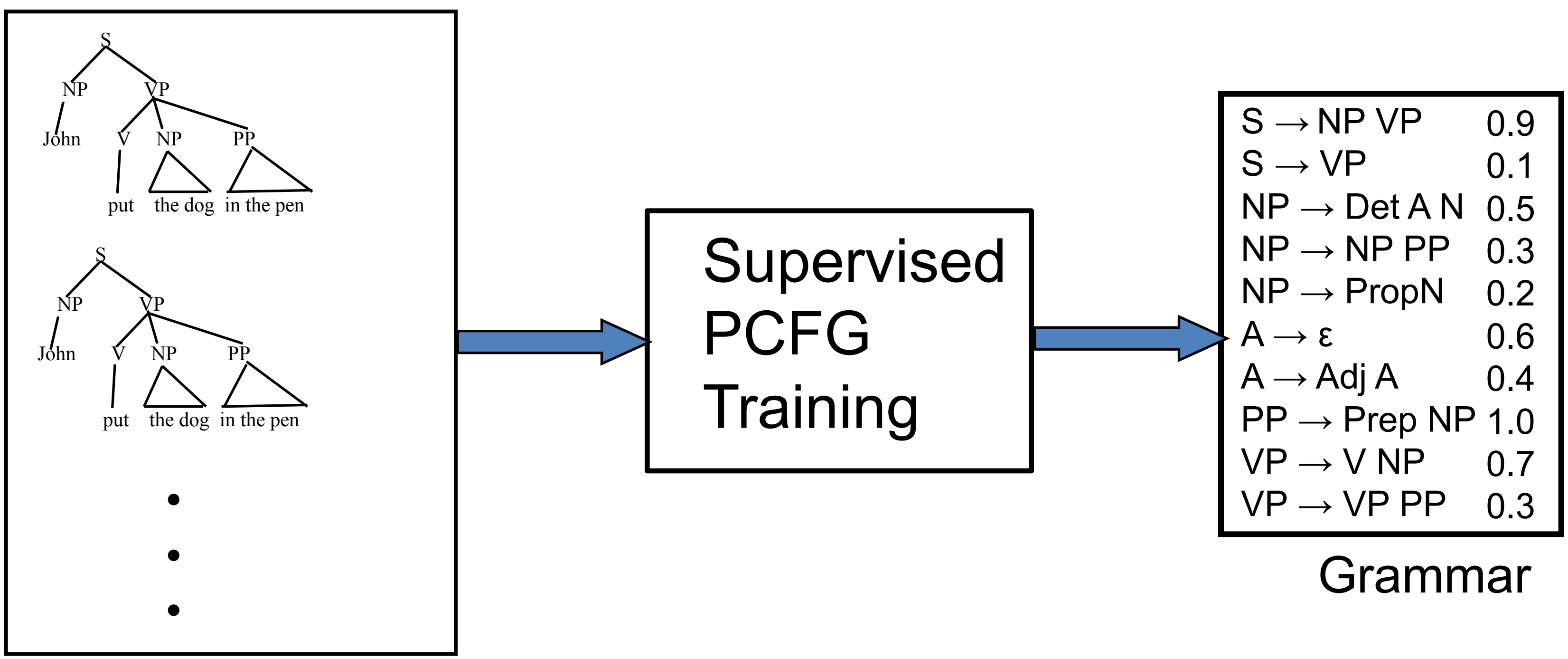
Sum probabilities of multiple derivations of each constituent in each cell.

- Like CYK for PCFGs, but **sum** probabilities of multiple derivations per constituents instead of taking **max**

PCFG: SUPERVISED TRAINING

- If parse trees are provided for training sentences, a grammar and its parameters can all be estimated directly from counts accumulated from the **tree-bank** (with appropriate smoothing).

Tree Bank

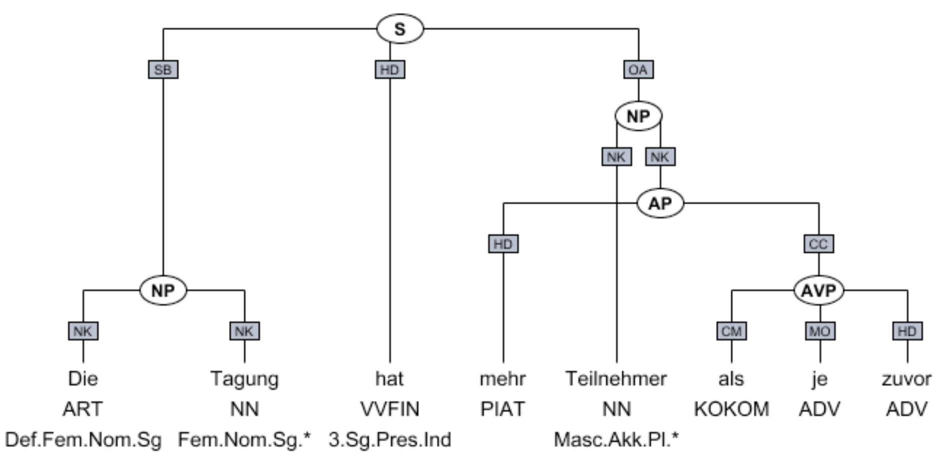


Grammar

TREEBANKS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- analog to annotated POS corpora, but for syntax trees
- **English** Penn Treebank: Standard corpus for testing syntactic parsing consists of 1.2 M words of text from the Wall Street Journal (WSJ).
  - Typical to train on about 40,000 parsed sentences and test on an additional standard disjoint test set of 2,416 sentences.
- **German**
  - TIGER/Negra Treebank: 900K FrankfurterRundschau
  - TüBa-D/Z: 470K words, taz



PENN TREEBANK BRACKETED FORMAT

Every production rule is represented by

- (
- left hand side
- sequence of right hand side symbols
  - non-terminals expanded by production rule
  - terminals
-)

```
( (S
  (NP-SBJ (DT The) (NNP Illinois) (NNP Supreme) (NNP Court) )
  (VP (VBD ordered)
   (NP-1 (DT the) (NN commission) )
   (S
    (NP-SBJ (-NONE- *-1) )
    (VP (TO to)
     (VP
      (VP (VB audit)
       (NP
        (NP (NNP Commonwealth) (NNP Edison) (POS 's) )
        (NN construction) (NNS expenses) ))
      (CC and)
      (VP (VB refund)
       (NP (DT any) (JJ unreasonable) (NNS expenses) ))))))
  (. .) ))
```

Traces: -NONE- and trace-number

ESTIMATING PROBABILITIES OF PRODUCTIONS

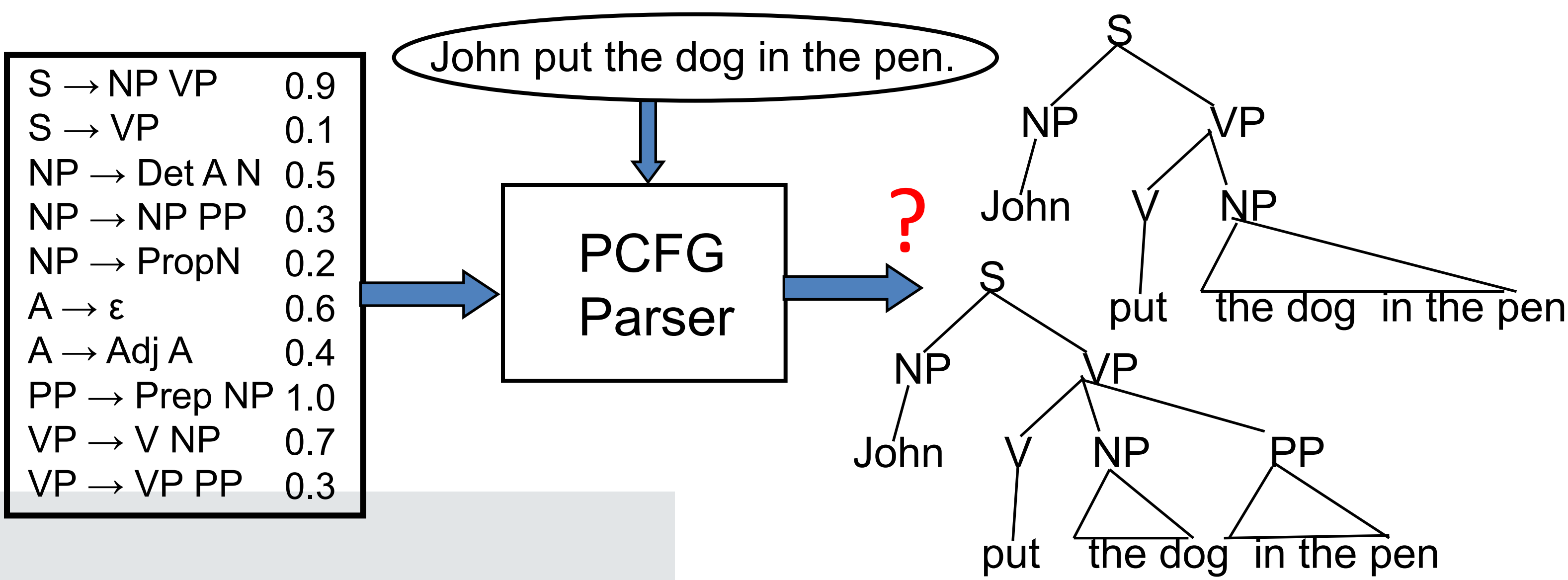
- Set of production rules can be taken directly from the set of rewrites in the treebank.
- Parameters can be directly estimated from frequency counts in the treebank.

$$P(\alpha \rightarrow \beta \mid \alpha) = \frac{\mathrm{C}(\alpha \rightarrow \beta)}{\sum_{\gamma} \mathrm{C}(\alpha \rightarrow \gamma)} = \frac{\mathrm{C}(\alpha \rightarrow \beta)}{\mathrm{C}(\alpha)}$$

- This might result in a grammar that linguists do not like:
  - e.g. Penn Treebank: flat, long RHSs
  - no recursion: will have rules like NP→Det NN, NP→Det JJ NN, NP→Det JJ JJ NN, NP→Det JJ JJ JJ NN, …

VANILLA PCFG LIMITATIONS

- Independence assumptions miss structural dependencies between rules
- Since probabilities of productions do not rely on specific words or concepts, only general structural disambiguation is possible.
- Consequently, vanilla PCFGs cannot resolve syntactic ambiguities that require semantics to resolve, e.g. ate with fork vs. meatballs.
- In order to work well, PCFGs must be lexicalized, i.e. productions must be specialized to specific words by including their head-word in their LHS non-terminals (e.g. VP-ate).
- A general preference for attaching PPs to NPs rather than VPs can be learned by a vanilla PCFG - but the desired preference can depend on specific words.



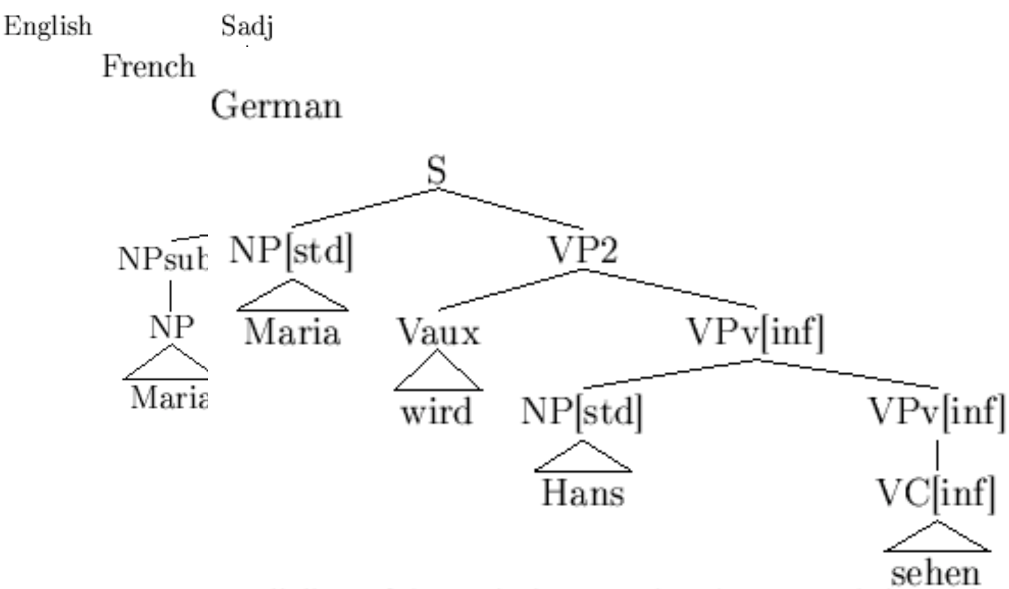
UNIFICATION GRAMMARS

- In order to handle agreement issues more effectively, each constituent has a list of features such as number, person, gender, etc. which may or not be specified for a given constituent.
- In order for two constituents to combine to form a larger constituent, their features must unify, i.e. consistently combine into a merged set of features.
- Expressive grammars and parsers (e.g. HPSG) have been developed using this approach and have been partially integrated with modern statistical models of disambiguation.
- Massive optimization techniques necessary, still only rudimentary support for semantic features

LEXICAL FUNCTIONAL GRAMMAR (LFG)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- generative grammar
- separation between surface structure and deep structure makes it possible to have same representation for several languages
- Constituent Structure (c-structure)
- Functional Structure (f-structure)



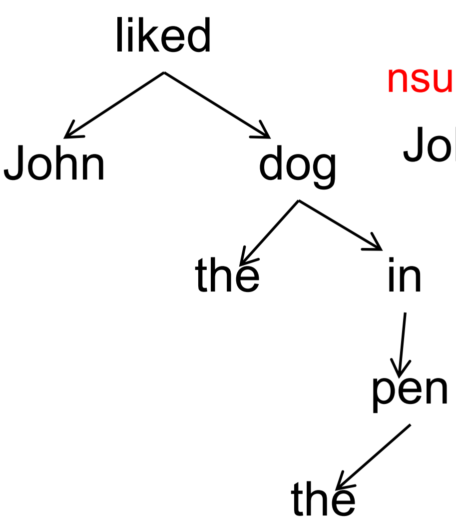
PRED	'see/voir/sehen<(↑ SUBJ),(↑ OBJ>'		
TENSE	FUT		
SUBJ	PRED	'Maria'	
	NTYPE	PROPER	NAME
	PERS	3	
	GEND	FEM	
	NUM	SG	
	CASE	NOM	
OBJ	PRED	'Hans'	
	NTYPE	PROPER	NAME
	PERS	3	
	GEND	MASC	
	NUM	SG	
	CASE	ACC	
PASSIVE	–		
STMT-TYPE	DECLARATIVE		
VTYPE	MAIN		

MILDLY CONTEXT-SENSITIVE GRAMMARS

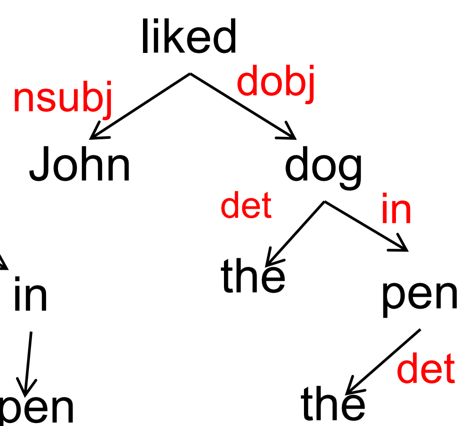
- Some grammatical formalisms provide a degree of **context-sensitivity** that helps capture aspects of NL syntax that are not easily handled by CFGs.
- **Tree Adjoining Grammar** (TAG) is based on combining tree fragments rather than individual phrases.
- **Combinatory Categorial Grammar** (CCG) consists of:
  - **Categorial Lexicon** that associates a syntactic and semantic category with each word.
  - **Combinatory Rules** that define how categories combine to form other categories.

DEPENDENCY PARSING

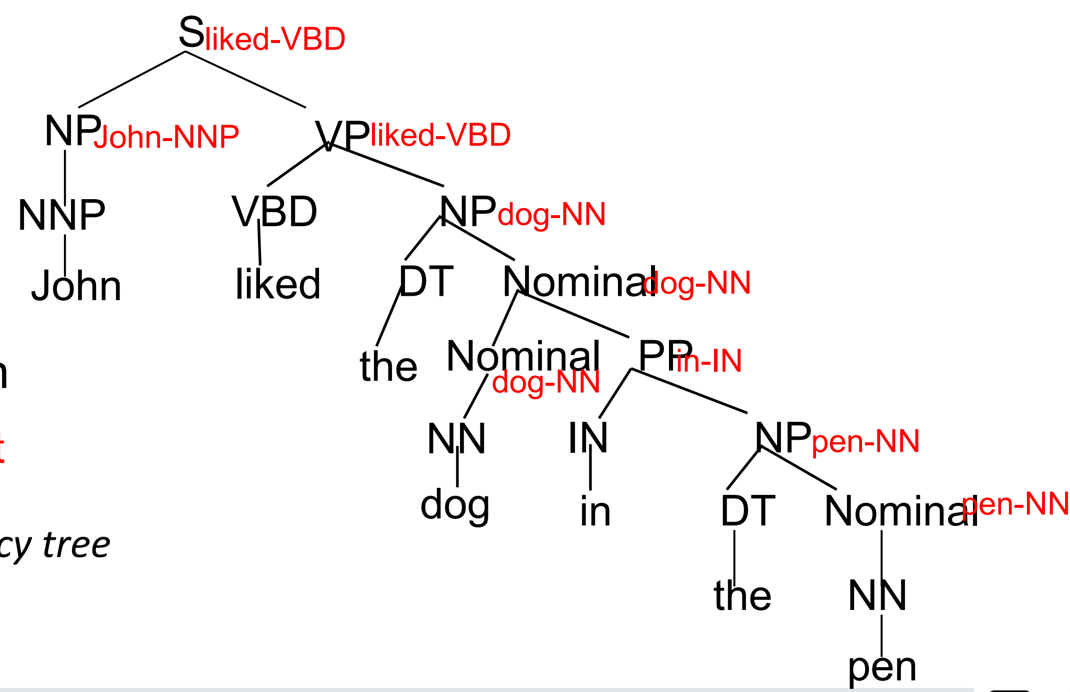
- Alternative to phrase-structure grammar define a parse as directed graph between words of a sentence representing dependencies between words
- No nodes for phrasal structure
- Can convert a phrase structure parse to a dependency tree by making the head of each non-head child of a node depend on the head of the head child



unlabeled dependency tree



labeled dependency tree

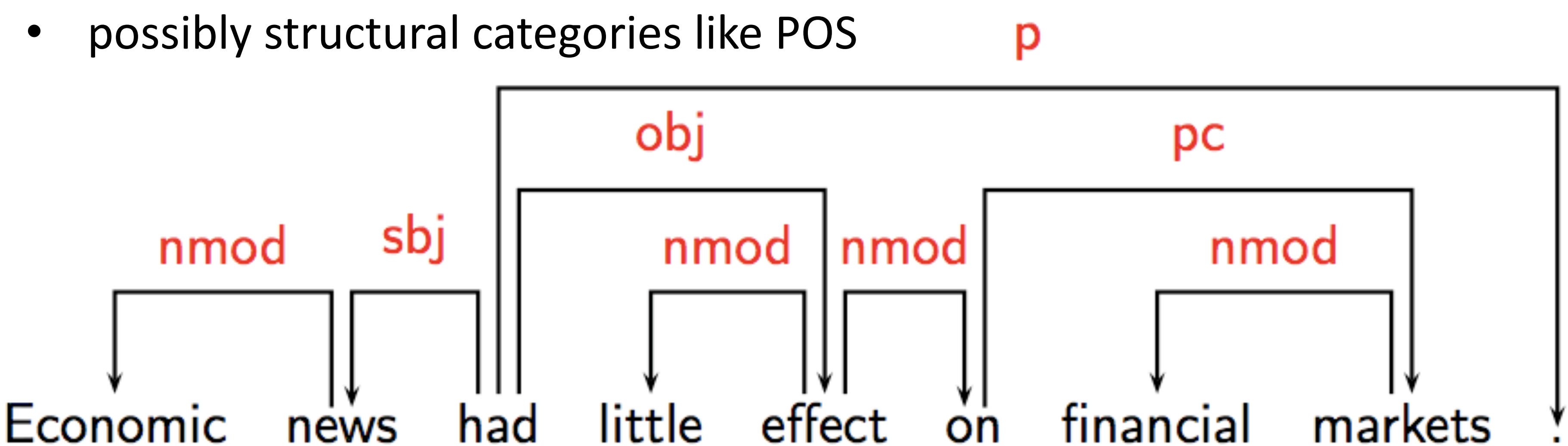


INTUITION BEHIND DEPENDENCY PARSING

- Syntactic structure consists of **lexical items**, linked by binary asymmetric relations called **dependencies**.
- Superior (start of arc) is called **head**, inferior is called **dependent**

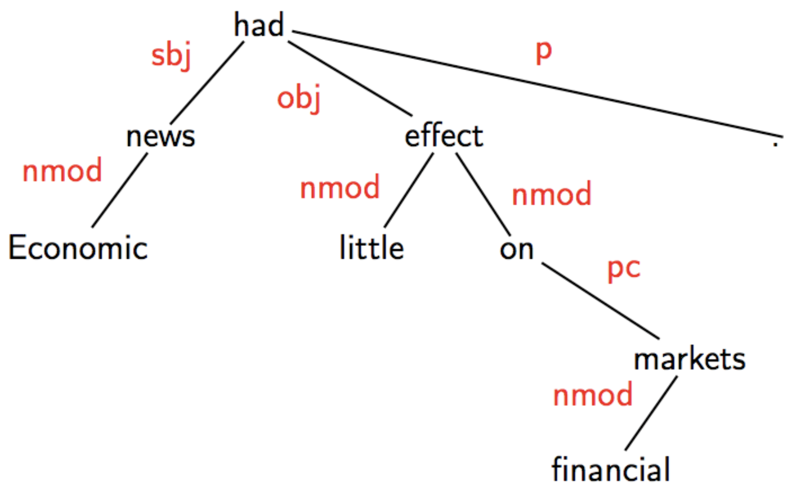
Dependency grammars explicitly represent:

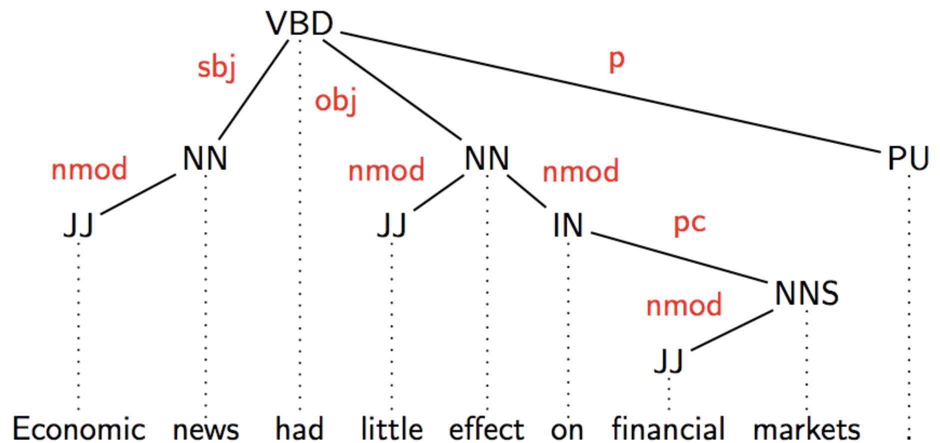
- head-dependency relations (directed arcs)
- functional categories (arc labels)
- possibly structural categories like POS

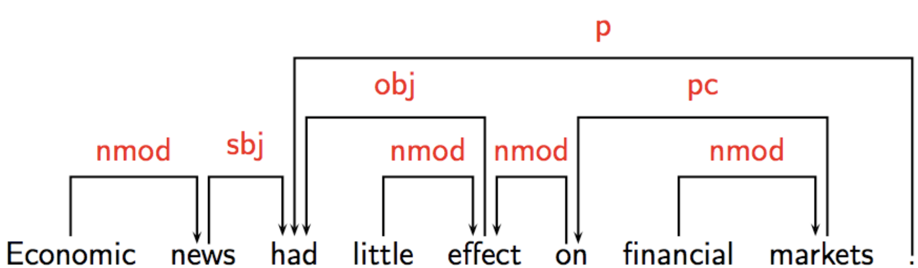


DEPENDENCY PARSING: NOTATIONAL VARIANTS

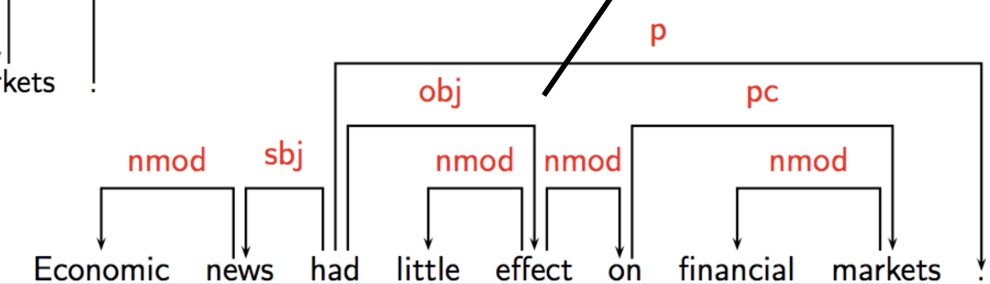
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG







used in this lecture



CRITERIA FOR HEADS AND DEPENDENTS

Criteria for a syntactic relation between a head H and a dependent D in a construction C:

1. H determines the syntactic category of C ; H can replace C .
2. H determines the semantic category of C ; D specifies H .
3. H is obligatory; D may be optional.
4. H selects D and determines whether D is obligatory.
5. The form of D depends on H (agreement or government).
6. The linear position of D is specified with reference to H .

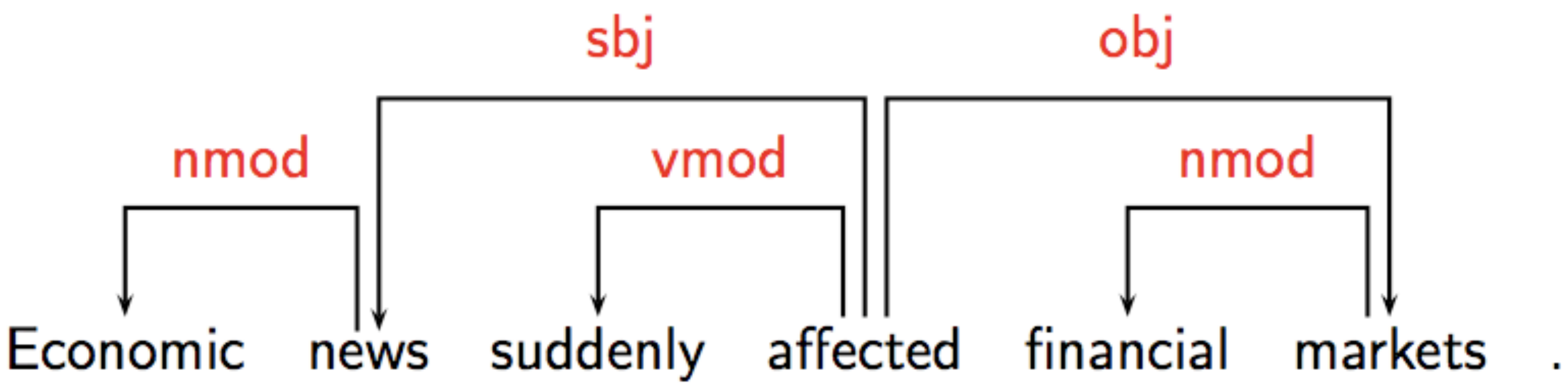
Issues:

- Syntactic (and morphological) versus semantic criteria
- Exocentric versus endocentric constructions

SOME CLEAR CASES

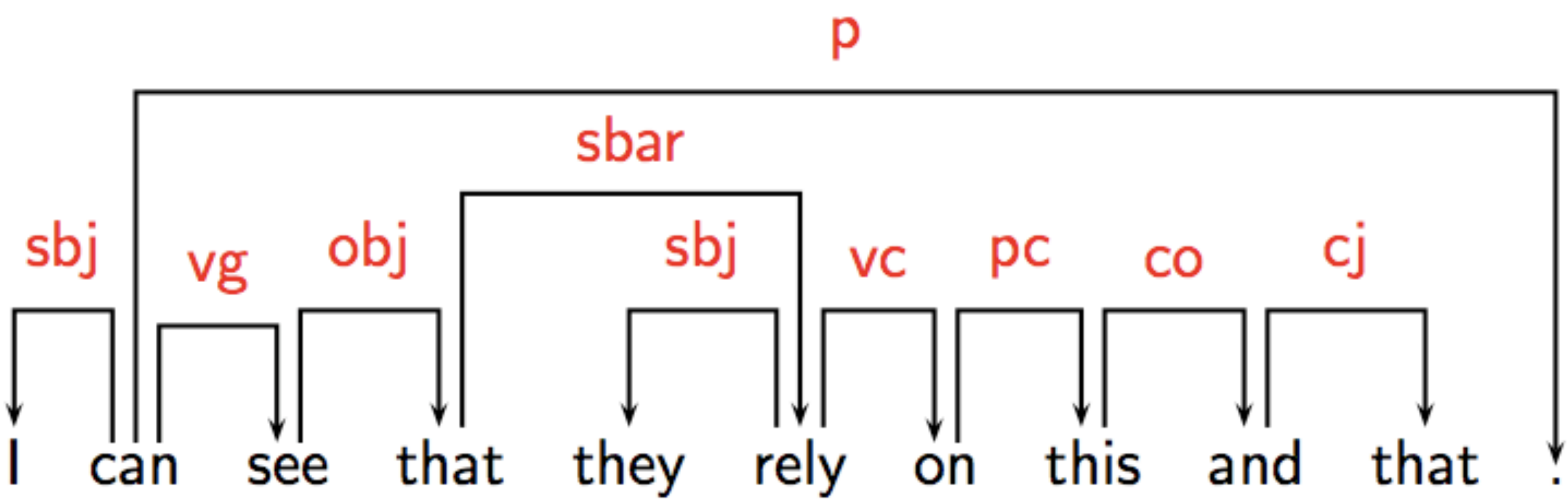
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Construction	Head	Dependent
Exocentric	Verb	Subject (sbj)
	Verb	Object (obj)
Endocentric	Verb	Adverbial (vmod)
	Noun	Attribute (nmod)



SOME TRICKY CASES: CONVENTIONS

- Complex verb groups (auxiliary ↔ main verb)
- Subordinate clauses (complementizer ↔ verb)
- Coordination (coordinator ↔ conjuncts)
- Prepositional phrases (preposition ↔ nominal)
- Punctuation

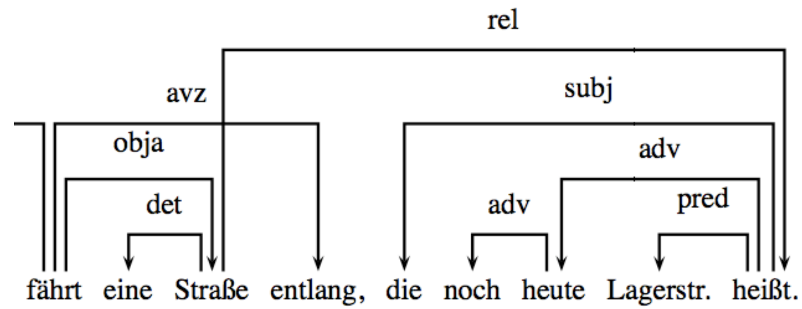


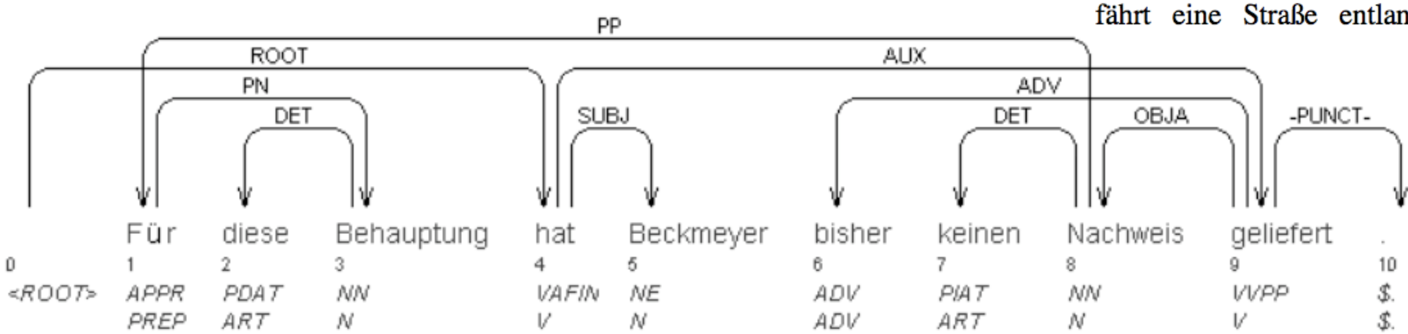
PROPERTIES OF THE DEPENDENCY GRAPH

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Dependency graph *G(V,E)* with *V:* word nodes, *E:* directed edges

- **Connected**: All words in a sentence are connected to the root node (complete structure)
- **Acyclic**: The syntactic structure is hierarchical
- **Single-head**: every word has only one head; a word can be the head of several other words
- **Projective**: no crossing edges. This condition does not hold for all languages, e.g. with free word order





DETERMINISTIC DEPENDENCY PARSING

- Basic idea
  - derive a single syntactic representation through a deterministic sequence of elementary parsing actions
  - possibly combine with a light amount of backtracking for alternatives
- Motivation
  - psycholinguistic plausibility
  - efficiency
  - simplicity
- Incremental algorithm with $O(n^2)$:

```
PARSE(w¹…wⁿ):
  for i=1 to n
     for j=i-1 downto 1
        LINK(wⁱ…wʲ)
```

$$\mathrm{LINK}(w^i, w^j) = \begin{cases} E := E \cup (i,j) & \text{if } w^j \text{ is dependent of } w^i \\ E := E \cup (j,i) & \text{if } w^i \text{ is dependent of } w^j \\ E := E & \text{otherwise} \end{cases}$$

Conditions like projectivity, single head can be incorporated in the LINK operation

TRANSITION-BASED: NIVRE'S ALGORITHM (2003)

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Four parsing actions:

$$\frac{S^{t-1}:[\ldots] \qquad Q^{t-1}:[w^i,\ldots]}{S^t:\ [\ldots,w^i]\ Q^t:\ [\ldots]}$$

shift

$$\frac{S^{t-1}:[\ldots,w^i]\ \ Q^{t-1}:[w^j,\ldots]\ \ \neg\exists w^k:\ w^k\rightarrow w^i}{S^t:\ [\ldots] \qquad Q^t:\ [w^j,\ldots] \qquad w^i\leftarrow w^{j}\ \mathrm{l}}$$

Left-Arc$_r$

$$\frac{S^{t-1}:[\ldots,w^i]\ \ Q^{t-1}:[\ldots]\ \ \exists w^k:\ w^k\rightarrow w^i}{S^t:\ [\ldots] \qquad Q^t:\ [\ldots]}$$

reduce

$$\frac{S^{t-1}:[\ldots,w^i] \qquad Q^{t-1}:[w^j,\ldots]\ \ \neg\exists w^k:\ w^k\rightarrow w^j}{S^t:\ [\ldots,w^i,w^j,]\ \ Q^t:\ [\ldots] \qquad w^i\rightarrow w^{j}r}$$

Right-Arc$_r$

Characteristics

- labeled dependencies as different arc operations
- arc-eager processing of right dependents
- Stack contains tokens that did not get assigned a head yet
- Single pass over the input, time O(n): max operations is 2n

Nivre, J. (2003) An Efficient Algorithm for Projective Dependency Parsing. In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03), Nancy, France, 23-25 April 2003, pp. 149-160.

EXAMPLE: NIVRE'S ALGORITHM

shift
$S^{t-1}:[\ldots] \quad Q^{t-1}:[w^i,\ldots]$
$S^t: [\ldots,w^i] \quad Q^t: [\ldots]$

Left-Arc$_r$
$S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[w^j,\ldots] \quad \neg\exists w^k: w^k \rightarrow w^i$
$S^t: [\ldots] \quad Q^t: [w^j,\ldots] \quad w^i \leftarrow w^j \; r$

reduce
$S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[\ldots] \quad \exists w^k: w^k \rightarrow w^i$
$S^t: [\ldots] \quad Q^t: [\ldots]$

Right-Arc$_r$
$S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[w^j,\ldots] \quad \neg\exists w^k: w^k \rightarrow w^j$
$S^t: [\ldots,w^i,w^j,] \quad Q^t: [\ldots] \quad w^i \rightarrow w^j \; r$

[root]$_S$ [Economic news had little effect on financial markets .]$_Q$

shift: [root Economic]$_S$ [news had little effect on financial markets .]$_Q$

nmod

Left-Arc$_{nmod}$: [root]$_S$ Economic [news had little effect on financial markets .]$_Q$

nmod

shift: [root Economic news]$_S$ [had little effect on financial markets .]$_Q$

nmod sbj

Left-Arc$_{sbj}$: [root]$_S$ Economic news [had little effect on financial markets .]$_Q$

EXAMPLE: NIVRE'S ALGORITHM

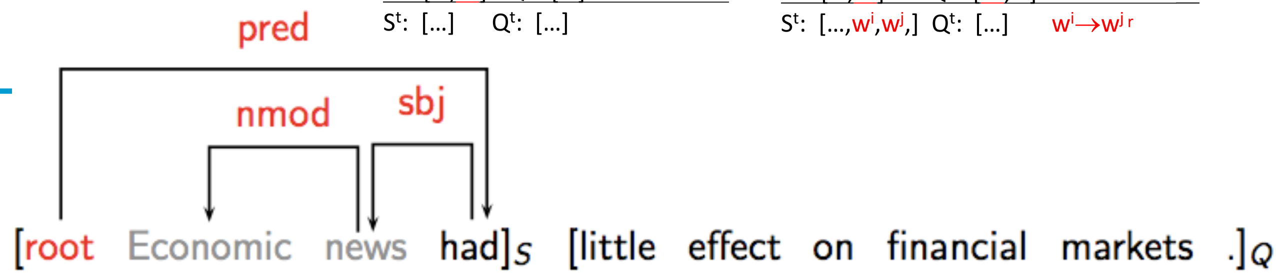
shift
$$\frac{S^{t-1}:[\ldots] \quad Q^{t-1}:[w^i,\ldots]}{S^t:\ [\ldots,w^i] \quad Q^t:\ [\ldots]}$$

Left-Arc$_r$
$$\frac{S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[w^j,\ldots] \quad \neg\exists w^k: w^k\rightarrow w^i}{S^t:\ [\ldots] \quad Q^t:\ [w^j,\ldots] \quad w^i \leftarrow w^j\ r}$$

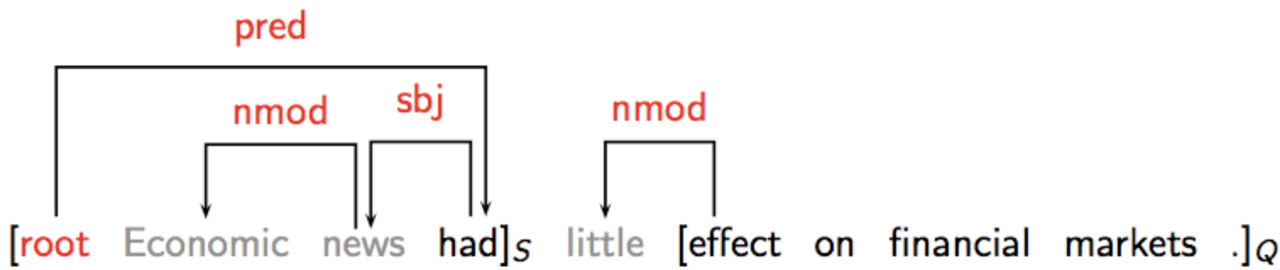
reduce
$$\frac{S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[\ldots] \quad \exists w^k: w^k\rightarrow w^i}{S^t:\ [\ldots] \quad Q^t:\ [\ldots]}$$

Right-Arc$_r$
$$\frac{S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[w^j,\ldots] \quad \neg\exists w^k: w^k\rightarrow w^j}{S^t:\ [\ldots,w^i,w^j] \quad Q^t:\ [\ldots] \quad w^i \rightarrow w^j\ r}$$

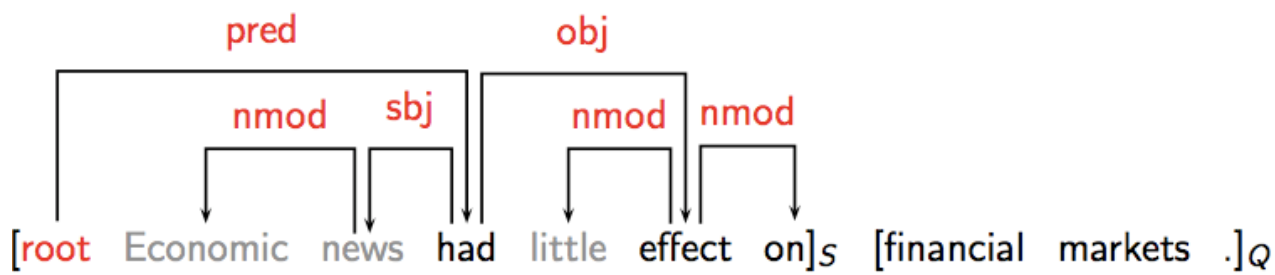
Right-Arc$_{pred}$



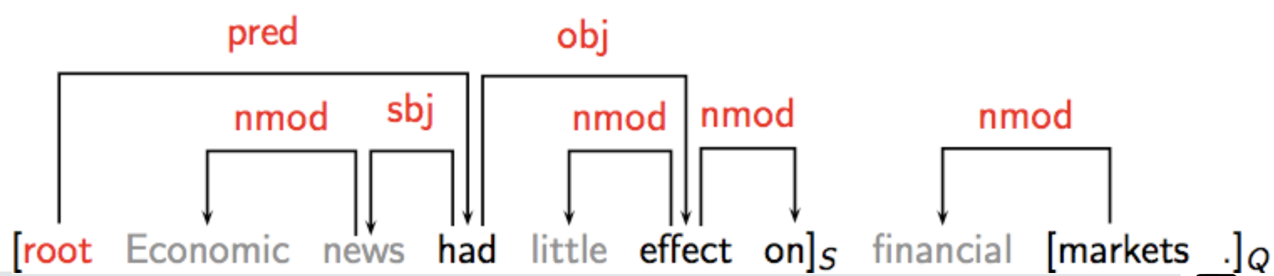
shift Left-Arc$_{nmod}$



Right-Arc$_{obj}$ Right-Arc$_{nmod}$



shift Left-Arc$_{nmod}$



EXAMPLE: NIVRE'S ALGORITHM

shift

$$\frac{S^{t-1}:[\ldots] \quad Q^{t-1}:[w^i,\ldots]}{S^t:[\ldots,w^i] \quad Q^t:[\ldots]}$$

Left-Arc$_r$

$$\frac{S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[w^j,\ldots] \quad \neg\exists w^k: w^k \rightarrow w^i}{S^t:[\ldots] \quad Q^t:[w^j,\ldots] \quad w^i \xleftarrow{r} w^j}$$

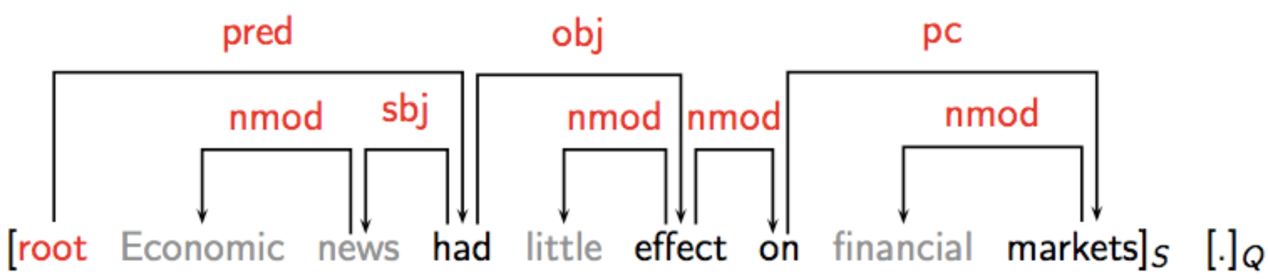
reduce

$$\frac{S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[\ldots] \quad \exists w^k: w^k \rightarrow w^i}{S^t:[\ldots] \quad Q^t:[\ldots]}$$

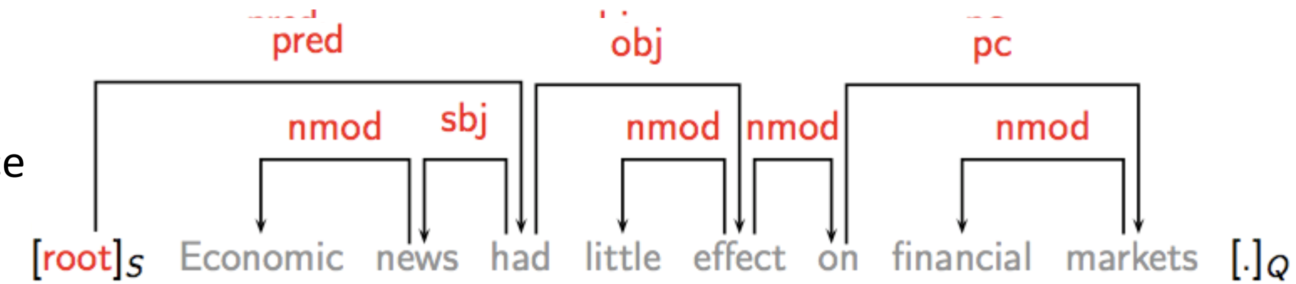
Right-Arc$_r$

$$\frac{S^{t-1}:[\ldots,w^i] \quad Q^{t-1}:[w^j,\ldots] \quad \neg\exists w^k: w^k \rightarrow w^j}{S^t:[\ldots,w^i,w^j,] \quad Q^t:[\ldots] \quad w^i \xrightarrow{r} w^j}$$

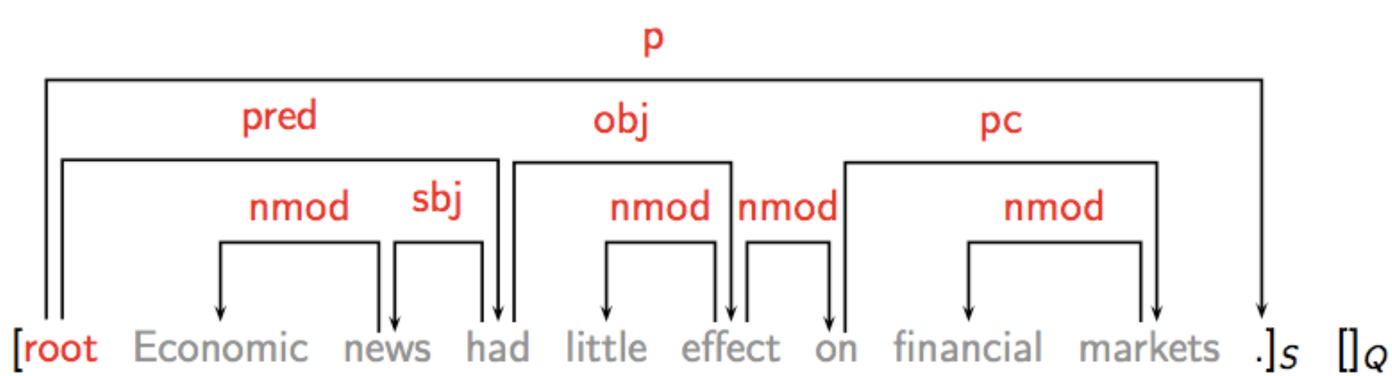
Right-Arc$_{pc}$



reduce reduce reduce reduce



Right-Arc$_p$



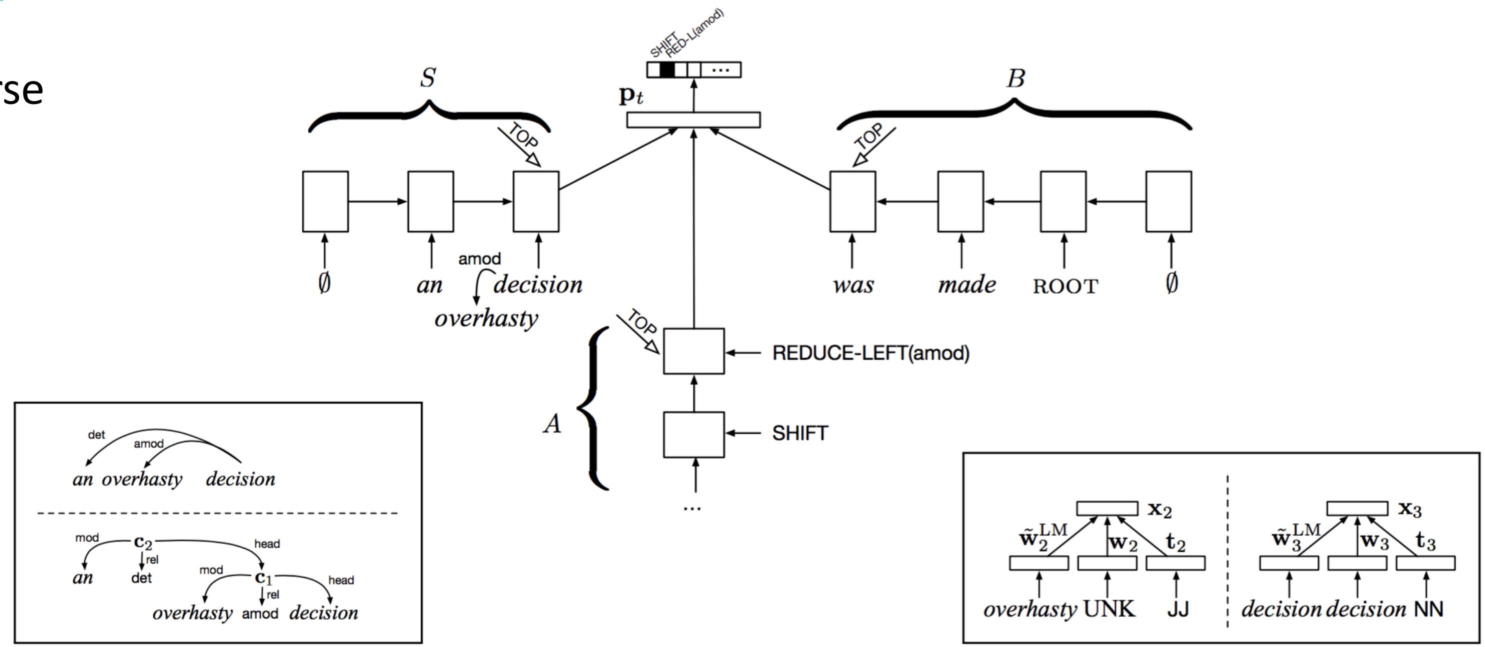
ORACLE APPROXIMATION BY MACHINE LEARNING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Data-driven deterministic parsing
  - deterministic parsing needs an **oracle** that tells us which of the possible steps to take
  - an oracle can be approximated by a **classifier**
  - the classifier can be **trained** from treebank data
- Learning method for dependency parsing: Approximate a function from parser state to parser action. Classifiers used:
  - Support Vector Machines
  - Memory-based learning
  - Maximum Entropy modeling
- Typical features:
  - word and POS of tokens on top of stack and next in queue
  - word and POS of tokens in certain distances and in structural relations
  - dependency types of heads, left/right children, siblings of tokens
- Results come very close to PCFG-based parsing, are obtained much faster

NEURAL DEPENDENCY PARSING

S: stack, collects parse
A: parser actions
B: buffer of words



- Transitions learned by three “Stack” LSTMs:
  - can push/pop elements
  - maintains a continuous representation of the stack state
  - ‘infinite’ history and lookahead

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pages 334–343.

MULTILINGUAL DEPENDENCY PARSING

- 2006 CoNLL-X shared task: 12 languages
- Data sources: dependency treebanks and phrase structure treebanks converted to dependency structures
- Main evaluation metric: labeled accuracy per word
- Top scores range from 91.7 (Japanese) to 65.7 (Turkish)
- Top systems over all languages:
  - approximate second-order non-projective spanning trees with online learning
  - labeled deterministic pseudo-projective parsing with support vector machines

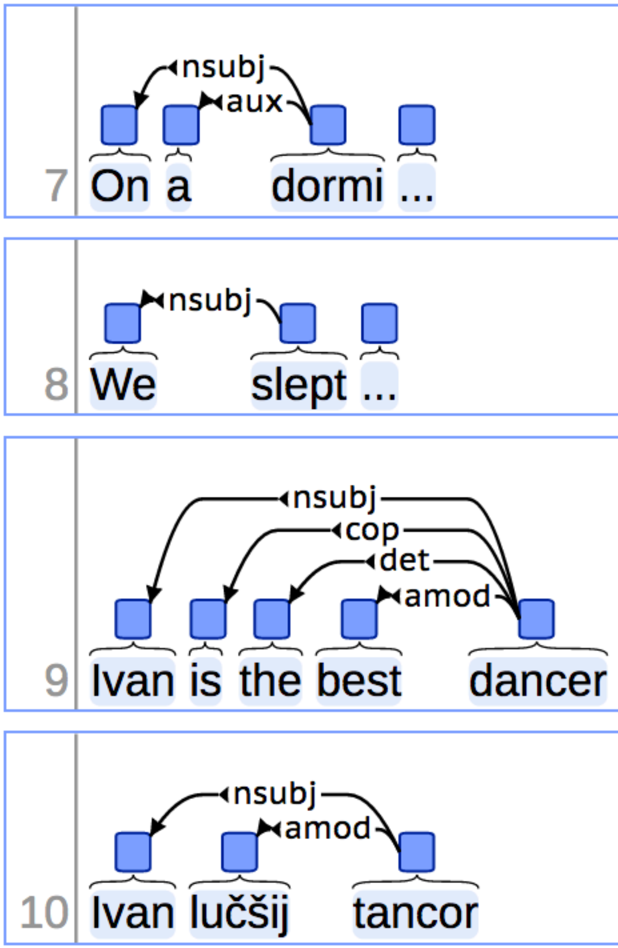
For English, phrase structure grammar parsers score slightly higher than dependency parsers. For some other languages, dependency parsers score higher. How much has parser development been influenced by English-specific phenomena?

UNIVERSAL DEPENDENCIES

HTTP://UNIVERSALDEPENDENCIES.ORG/

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

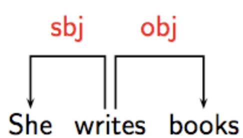
- Attempt to have the same simplified set of 17 POS tags and 37 dependency types for all languages
- Currently available for 50 languages, 15 more announced
- guiding principles: cross-language applicability
- greatly simplifies multilingual applications

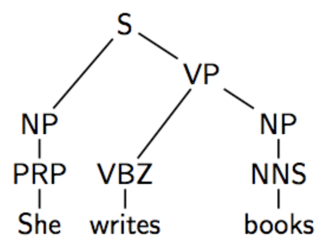


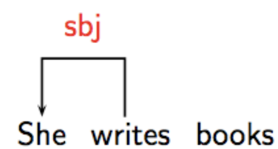
ADVANTAGES OF DEPENDENCY PARSING

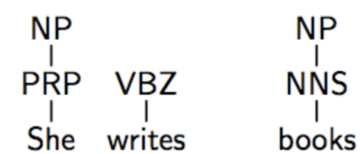
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Complexity: Projective parsing in O(n), non-projective parsing in O(n^2)
- Transparency (for labeled dependency graphs):
  - direct encoding of argument structure
  - interpretability of fragments

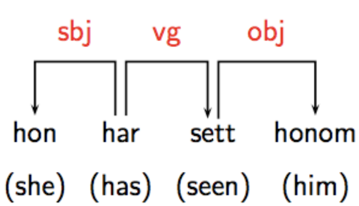


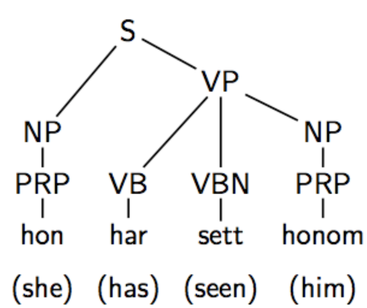


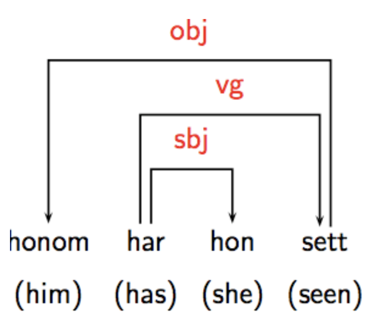


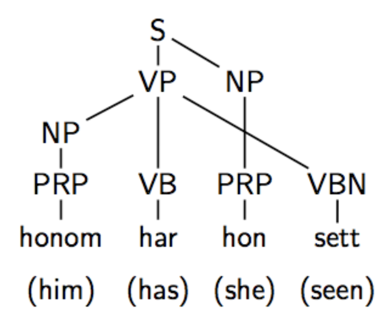


- Suitable for free word order languages (for non-projective approaches)



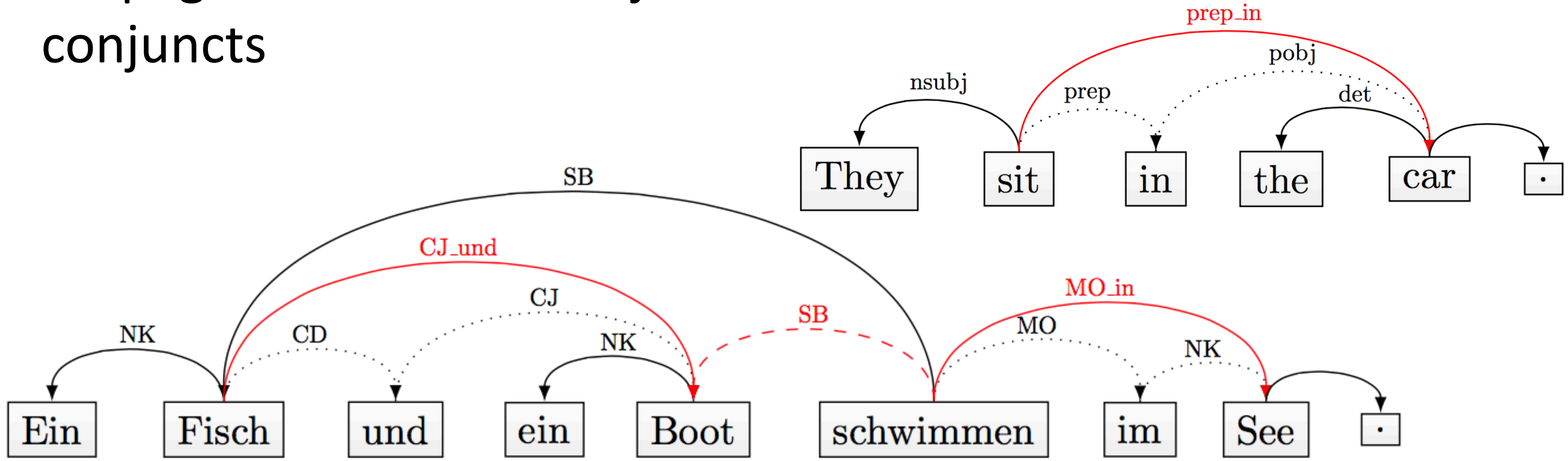






DEPENDENCY PARSING TOWARDS SEMANTICS

- Collapsing rules: Move preposition into relation
- Propagation rules for conjunctions: relation applies to all conjuncts



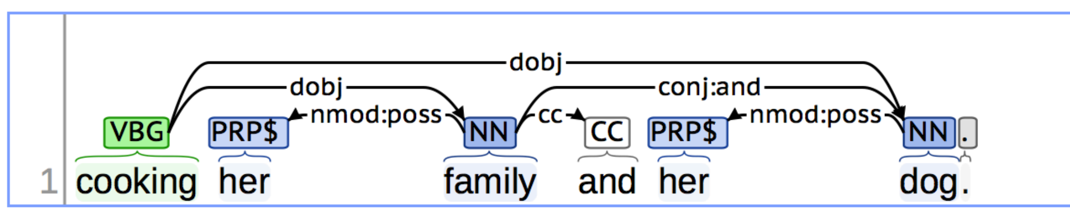
These “collapsed and conjunction-propagated” dependency parses proved advantageous for semantic tasks.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In Proceedings of LREC-2006, pages 449–454, Genova, Italy.

THIS COULD GO WRONG … AND COMMAS SAVE LIVES!

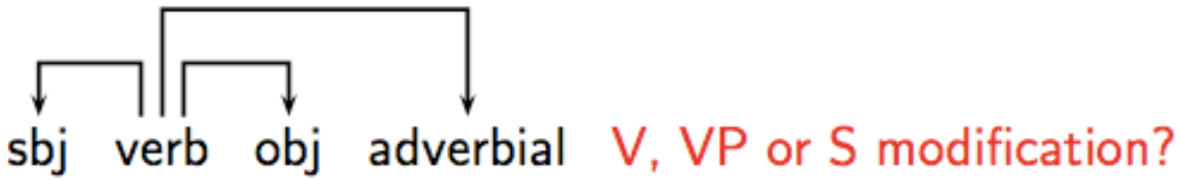
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Enhanced Dependencies: http://nlp.stanford.edu:8080/corenlp/process



DISADVANTAGE OF DEPENDENCY PARSING

- Limited Expressivity
  - Every projective dependency grammar has a strongly equivalent context-free grammar, but not vice versa
  - In unlabeled dependency structure, it is impossible to distinguish between head modification and phrase modification



  - These limits are unclear for labeled, non-projective dependency structures, but this configuration sometimes lacks accuracy and efficiency

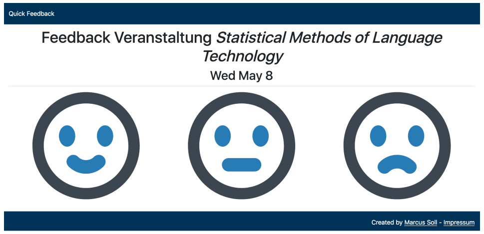
STATISTICAL PARSING CONCLUSIONS

- Statistical models such as PCFGs allow for probabilistic resolution of ambiguities.
- PCFGs can be easily learned from treebanks.
- Current statistical parsers are quite accurate for English but not yet at the level of human-expert agreement.
- For other languages, only dependency parsers are more or less reliable
- dependency parsers are faster, and contain different information than phrase structure grammar parsers, and try to stay as deterministic as possible
- Recent advances in transition-based parsing using neural networks

Main challenge:

- Treebanking is very expensive

IMMEDIATE FEEDBACK



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

coming up next

LSA, LDA, word2vec

DENSE VECTOR REPRESENTATIONS