Computerlinguistik II

Vorlesung im SoSe 2018
(M-GSW-10)

Prof. Dr. Udo Hahn

Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

http://www.julielab.de

Allgemeine Hinweise

- Vorlesung: Mi, 10-12h (Humboldt 8, SR 1)
- Übung zV: Mo 10-12h (Johannisfr.hof.3, SR 2)
  - beginnt am 16.4.2018
- Vorlesungsmaterialien im Netz
  - `http://www.julielab.de/` ⇨ „Students“
- M-GSW-10 besteht aus VL+ÜB und Seminar!
- Sprechstunde: Mi, 12-13h (bA) (FG 30, R 004)
- Email: `udo.hahn@uni-jena.de`
- URL: `http://www.julielab.de`
- Fachliteratur ist überwiegend in Englisch

Veranstaltungen im SS 2018

- Seminar „Linked Open Data“
  - Do, 16-18 Uhr
- Software-Praktikum: „Softwaretechnologien für Natürlichsprachliche Systeme“
  - Di, 14-16 Uhr
- Theoreticum: Methoden der Computer-linguistik,
- Technicum: Praxis sprachtechnologischer Systeme
  - Alterierend: Fr, 10-12 Uhr

Two Paradigms for NLP

- Symbolic Specification Paradigm
  - Manual acquisition procedures
  - Lab-internal activities
  - Intuition and (few!) subjectively generated examples drive progress based on individual (competence) judgments
    - “I have a system that parses all of my nine-teen sentences!”

Symbolic Specification Paradigm

- Manual rule specification
  - Source: linguist´s intuition
- Manual lexicon specification
  - Source: linguist´s intuition
- Each lab has its own (home-grown) set of NLP software
  - Hampers reusability
  - Limits scientific progress
  - Waste of human and monetary resources (we “burnt” thousands of Ph.D. student all over the world ☹)

Shortcomings of the "Classical" Linguistic Approach

- Huge amounts of background knowledge req.
  - Lexicons (approx. 100,000 – 150,000 entries)
  - Grammars (>> 15,000 – 20,000 rules)
  - Semantics (>> 15,000 – 20,000 rules)
- As the linguistic and conceptual coverage of classical linguistic systems increases (slowly), it still remains insufficient; systems also reveal 'spurious' ambiguity, and, hence, tend to become overly "brittle" and unmaintainable
- More fail-soft behavior is required at the expense of … ? (e.g., full-depth understanding)

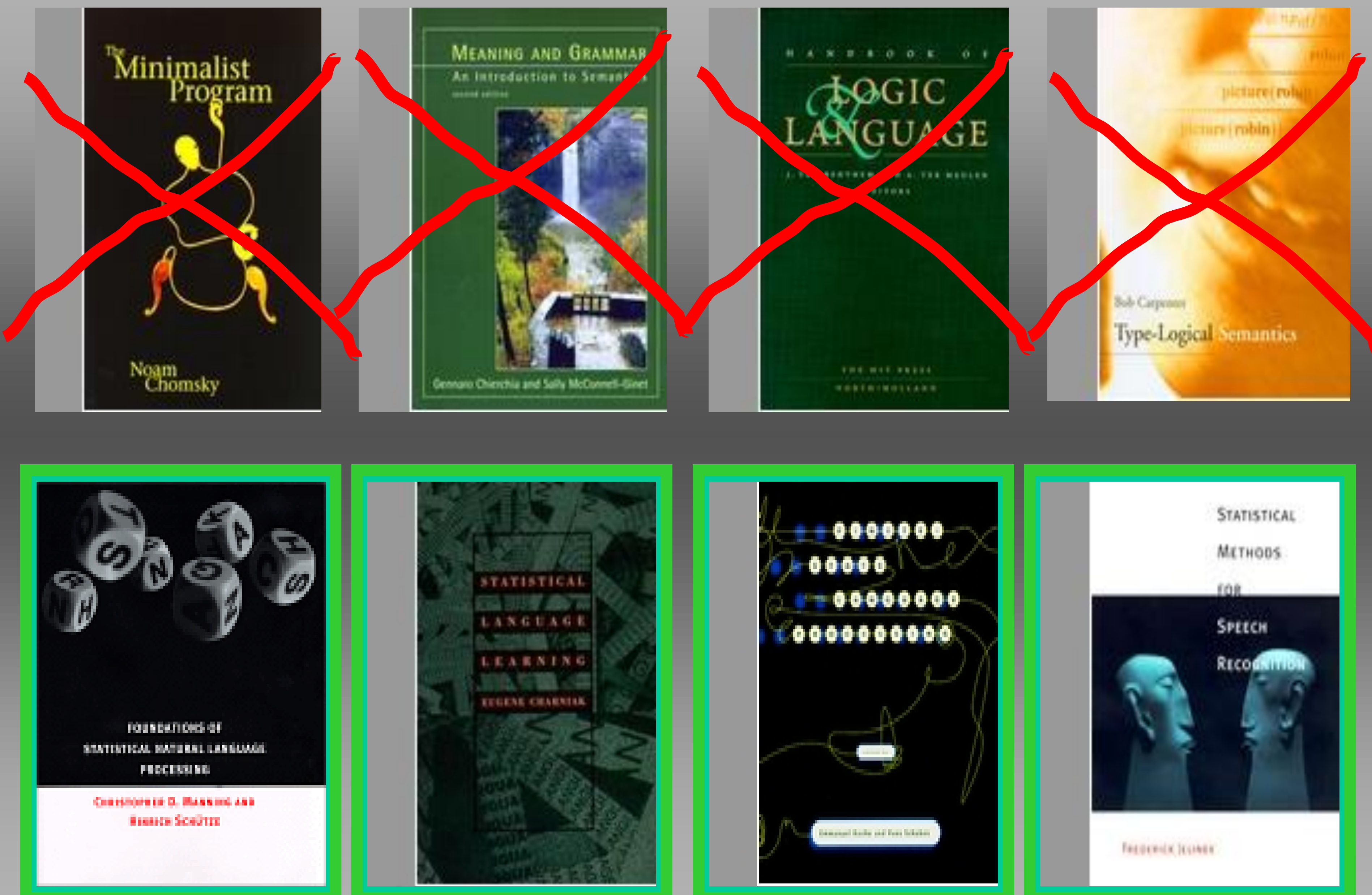
Two Paradigms for NLP

- Symbolic Specification Paradigm
  - Manual acquisition procedures
  - Lab-internal activities
  - Intuition and (few!) subjectively generated examples drive progress based on individual (competence) judgments
    - “I have a system that parses all of my nine-teen sentences!”
- Empirical (Learning) Paradigm
  - Automatic acquisition procedures
  - Community-wide sharing of common knowledge and resources
  - Large and ‚representative‘ data sets drive progress according to experimental standards
    - “The system was tested on 1,7 million words taken from the WSJ segment of the MUC-7 data set and produced 4.9% parsing errors, thus yielding a statistically significant 1.6% improvement over the best result by parser X on the same data set & a 40.3% improvement over the baseline system!”

Empirical Paradigm

- Large repositories of language data
  - Corpora (plain or annotated, i.e., enriched by meta-data)
- Large, community-wide shared repositories of language processing modules
  - Tokenizers, POS taggers, chunkers, NE recognizers, ...
- Shared repositories of machine learning algos
- Automatic acquisition of linguistic knowledge
  - Applying ML algos to train linguistic processors by using large corpora with valid linguistic metadata (linguist as educated data supplier, „language expert“) rather than manual intuition (linguist as creative rule inventor)
- Shallow analysis rather than deep understanding
- Large, community-wide self-managed, task-oriented competitions, comparative evaluation rounds
- Change of mathematics:
  - Statistics rather than algebra and logics

Paradigm Shift – We Exchanged our Textbooks…



POS Tagging

A	severe	infection	ended	the	pregnancy	.
↓	↓	↓	↓	↓	↓	↓
DET	ADJ	NOUN	VERB	DET	NOUN	ST

Penn Treebank Tag Set

Tag	Description	Examples
.	sentence terminator	. ! ?
DT	determiner	all an many such that the them these this
JJ	adjective, numeral	first oiled separable battery-powered
NN	common noun	cabbage thermostat investment
PRP	personal pronoun	herself him it me one oneself theirs they
IN	preposition	among out within behind into next
VB	verb (base form)	ask assess assign begin break bring
VBD	verb (past tense)	asked assessed assigned began broke
WP	WH-pronoun	that what which who whom

In total,
45 tags

Transformation Rules for Tagging [Brill, 1995]

- Initial State: Based on a number of features, guess the most likely POS tag for a given word:
  - die/DET Frau/NOUN ,/COMMA die/DET singt/VFIN
- Learn transformation rules to reduce errors:
  - *Change* DET *to* PREL *whenever the preceding word is tagged as* COMMA
- Apply learned transformation rules:
  - die/DET Frau/NOUN,/COMMA die/PREL singt/VFIN

First 20 Transformation Rules

	Change Tag		
#	From	To	Condition
1	NN	VB	Previous tag is *TO*
2	VBP	VB	One of the previous three tags is *MD*
3	NN	VB	One of the previous two tags is *MD*
4	VB	NN	One of the previous two tags is *DT*
5	VBD	VBN	One of the previous three tags is *VBZ*
6	VBN	VBD	Previous tag is *PRP*
7	VBN	VBD	Previous tag is *NNP*
8	VBD	VBN	Previous tag is *VBD*
9	VBP	VB	Previous tag is *TO*
10	POS	VBZ	Previous tag is *PRP*
11	VB	VBP	Previous tag is *NNS*
12	VBD	VBN	One of previous three tags is *VBP*
13	IN	WDT	One of next two tags is *VB*
14	VBD	VBN	One of previous two tags is *VB*
15	VB	VBP	Previous tag is *PRP*
16	IN	WDT	Next tag is *VBZ*
17	IN	DT	Next tag is *NN*
18	JJ	NNP	Next tag is *NNP*
19	IN	WDT	Next tag is *VBD*
20	JJR	RBR	Next tag is *JJ*

Taken from: Brill (1995), Transformation-Based Error-Driven Learning

Towards Statistical Models of Natural Language Processing …

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
-

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- W

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- Wh

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- Wha

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- What

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- What d

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- What do

Letter-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
-

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
- We

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
- We are

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
- We are now

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
- We are now entering

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
- We are now entering statistical

Word-based Language Models

- Shannon's Game
- Guess the next letter:
- What do you think the next letter is?
- Guess the next word:
- We are now entering statistical territory

Approximating Natural Language Words

- zero-order approximation: letter sequences are independent of each other and all equally probable:
  - xfoml rxkhrjffjuj zlpwcwkcy ffjeyvkcqsghyd

Approximating Natural Language Words

- first-order approximation: letters are independent, but occur with the frequencies of English text:
  - ocro hli rgwr nmielwis eu ll nbnesebya th eei alhenhtppa oobttva nah

Approximating Natural Language Words

- second-order approximation: the probability that a letter appears depends on the previous letter
  - on ie antsoutinys are t inctore st bes deamy achin d ilonasive tucoowe at teasonare fuzo tizin andy tobe seace ctisbe

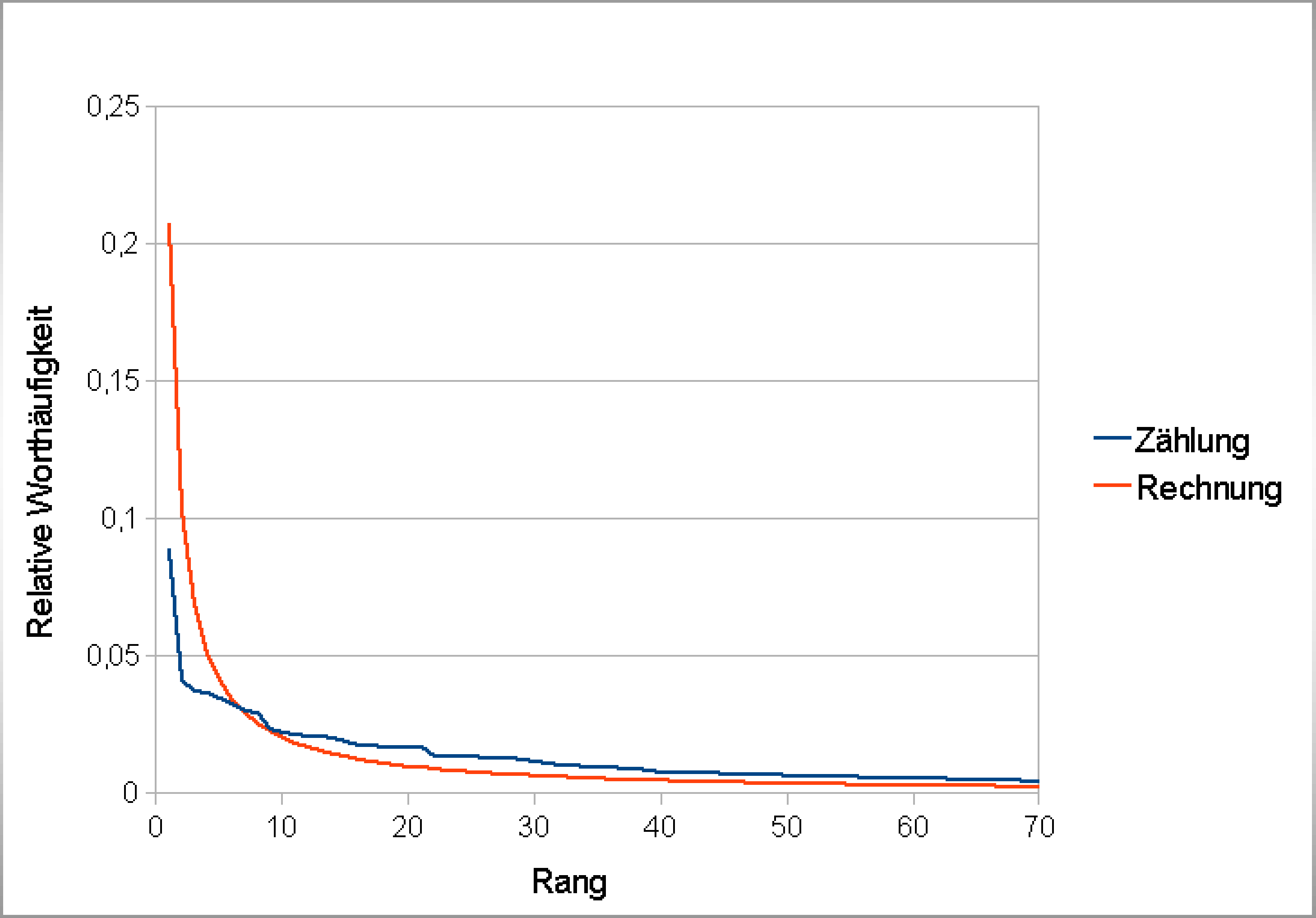
Approximating Natural Language Words

- third-order approximation: the probability that a certain letter appears depends on the two previous letters
  - in no ist lat whey cratict froure birs grocid pondenome of demonstures of the reptagin is regoactiona of cre

Approximating Natural Language Words

- Higher frequency trigrams for different languages:
  - English: THE, ING, ENT, ION
  - German: EIN, ICH, DEN, DER
  - French: ENT, QUE, LES, ION
  - Italian: CHE, ERE, ZIO, DEL
  - Spanish: QUE, EST, ARA, ADO

Zipfsches Gesetz



Wortverteilung im Vergleich zu einer einfachen Zipf-Verteilung (~1/n. Wortanzahl: 70; Texte aus: `http://www.gutenberg.org/dirs/etext04/8effi10.txt`)

Terminology

- **Sentence**: unit of written language
- **Utterance**: unit of spoken language
- **Word Form**: the inflected form that appears literally in the corpus
- **Lemma**: lexical forms having the same stem, part of speech, and word sense
- **Types (V)**: number of distinct words that might appear in a corpus (vocabulary size)
- **Tokens (N_T)**: total number of words in a corpus (note: $V << N_T$)
- **Types seen so far (T)**: number of distinct words seen so far in corpus (note: $T \leq V << N_T$)

Word-based Language Models

- A model that enables one to compute the probability, or likelihood, of a sentence S, P(S).
- Simple: Every word follows every other word with equal probability (0-gram)
  - Assume |V| is the size of the vocabulary V
  - Likelihood of sentence S of length n is $1/|V| \times 1/|V| \ldots \times 1/|V|$
  - If English has 100,000 words, the probability of each next word is 1/100000 = .00001

Relative Frequency vs. Conditional Probability

- Smarter: *Relative* Frequency

  Probability of each next word is related to word frequency within a corpus (unigram)

  - Likelihood of sentence S = $P(w_1) \times P(w_2) \times \ldots \times P(w_n)$
  - Assumes probability of each word is independent of probabilities of other words

Relative Frequency vs. Conditional Probability

- Smarter: *Relative* Frequency

  Probability of each next word is related to word frequency within a corpus (unigram)

  - Likelihood of sentence S = $P(w_1) \times P(w_2) \times \ldots \times P(w_n)$
  - Assumes probability of each word is independent of probabilities of other words

- Even smarter: *Conditional* Probability

  Look at probability given previous words (n-gram)

  - Likelihood of sentence S = $P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_{n-1})$
  - Assumes probability of each word is dependent on probabilities of previous words

Generalization of Conditional Probability via Chain Rule

- Conditional Probability for Two Events, A_1 and A_2
  - $P(A_1,A_2) = P(A_1) \cdot P(A_2|A_1)$
- Chain Rule generalizes to multiple (*n*) events
  - $P(A_1, \ldots, A_n) =$

    $P(A_1) \times P(A_2|A_1) \times P(A_3|A_1,A_2) \times \ldots \times P(A_n|A_1 \ldots A_{n-1})$

    - Examples:
      - P(the dog) = P(the) × P(dog | the)
      - P(the dog bites) = P(the) × P(dog | the) × P(bites| the dog)

Relative Frequencies and Conditional Probabilities

- Relative word frequencies are better than equal probabilities for all words
  - In a corpus with 10K word types, each word would have P(w) = 1/10K
  - Does not match our intuitions that different words are more likely to occur
    - (e.g. “the” vs. “shop” vs. “aardvark”)
- Conditional probability is more useful than individual relative word frequencies
    - dog may be relatively rare in a corpus
    - but if we see barking, P(dog|barking) may be large

Probability for a Word String

- **In general, the probability of a complete string of words $w_1^n = w_1 \ldots w_n$ is**

$$P(w_1^n)$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1\, w_2)\ldots P(w_n|w_1\ldots w_{n-1})$$

$$= \prod_{k=1}^{n} P(w_k \mid w_1^{k-1})$$

- **But this approach to determining the probability of a word sequence gets to be computationally very expensive <u>and</u> suffers from sparse data**

Markov Assumption (basic idea)

- How do we (efficiently) compute $P(w_n|w_1^{n-1})$?
- <u>Trick</u> (!): Instead of P(rabbit|I saw <u>a</u>), we use P(rabbit|<u>a</u>).
  - This lets us collect statistics in practice via a bigram model: P(the barking dog) = P(the|<start>) × P(barking|the) × P(dog|barking)

Markov Assumption (the very idea)

- Markov models are the class of probabilistic language models that assume that we can predict the probability of some future unit *without looking too far* into the past
  - Specifically, for N=2 (bigram):
  - $P(w_1^n) \approx \Pi_{k=1}^{n} P(w_k|w_{k-1})$; w_0 := <start>
- Order of a Markov model: length of prior context
  - bigram is first order, trigram is second order, …

Statistical HMM-based Tagging

[Brants, 2000]

- *State transition probability*: Likelihood of a tag immediately following n other tags
  - $P_1(Tag_i \mid Tag_{i-1} \ldots Tag_{i-n})$
- *State emission probability*: Likelihood of a word given a tag
  - $P_2(Word_i \mid Tag_i)$

    - die/DET Frau/NOUN ,/COMMA die/DET or PREL singt/VFIN

Trigrams for Tagging

- *State transition probabilities (trigrams)*:
  - P_1(DET | COMMA NOUN) = 0.0007
  - P_1(PREL | COMMA NOUN) = 0.01
- *State emission probabilities*:
  - P_2(die | DET) = 0.7
  - P_2(die | PREL) = 0.2
- Compute probabilistic evidence for the tag being
  - DET: $P_1 \cdot P_2 = 0.0007 \cdot 0.7 = 0.00049$
  - PREL: $P_1 \cdot P_2 = 0.01 \cdot 0.2 = 0.002$

  - die/DET Frau/NOUN ,/COMMA die/PREL singt/VFIN

Inside (most) POS Taggers

- Lexicon look-up routines
- Morphological processing (not only deflection!)
- Unknown word handler, if lexicon look-up fails (based on statistical information)
- Ambiguity ranking (priority selection)

Chunking

Arginine methylation of STAT1 modulates IFN induced transcription

Chunking

[Arginine methylation] of [STAT1] modulates [IFN induced transcription]

Shallow Parsing

[Arginine methylation of STAT1]$_{NP}$ [modulates]$_{VP}$ [IFN induced transcription]$_{NP}$

Shallow Parsing

[[Arginine methylation]$_{NP}$ [of STAT1]$_{PP}$]$_{NP}$

[Arginine methylation of STAT1]$_{NP}$ [modulates]$_{VP}$ [IFN induced transcription]$_{NP}$

Shallow Parsing

[[IFN induced]$_{AP}$ [transcription]$_{N}$]$_{NP}$

[[Arginine methylation]$_{NP}$ [of STAT1]$_{PP}$]$_{NP}$

[Arginine methylation of STAT1]$_{NP}$ [modulates]$_{VP}$ [IFN induced transcription]$_{NP}$

Deep Parsing

$[\ [\text{IFN induced}]_{AP}\ [\text{transcription}]_{N}\ ]_{NP}$

$[\ [[\text{Arginine}]_{N}\ [\text{methylation}]_{N}]_{NP}\ [[\text{of}]_{P}\ [\text{STAT1}]_{N}]_{PP}\ ]_{NP}$

$[\ [\text{Arginine methylation}]_{NP}\ [\text{of STAT1}]_{PP}\ ]_{NP}$

$[\text{Arginine methylation of STAT1}]_{NP}\ [\ [\text{modulates}]_{V}\ [\text{IFN induced transcription}]_{NP}\ ]_{VP}$

Deep Parsing

[[[IFN]$_{N}$ [induced]$_{A}$]$_{AP}$ [transcription]$_{N}$]$_{NP}$

[[IFN induced]$_{AP}$ [transcription]$_{N}$]$_{NP}$

[[[Arginine]$_{N}$ [methylation]$_{N}$]$_{NP}$ [[of]$_{P}$ [STAT1]$_{N}$]$_{PP}$]$_{NP}$

[[Arginine methylation]$_{NP}$ [of STAT1]$_{PP}$]$_{NP}$

[Arginine methylation of STAT1]$_{NP}$ [[modulates]$_{V}$ [IFN induced transcription]$_{NP}$]$_{VP}$

Chunking Principles

- Goal: divide a sentence into a sequence of chunks (ako phrases)
- Chunks are non-overlapping regions of a text
  - *[I] saw [a tall man] in [the park]*
- Chunks are non-exhaustive
  - not all words of a sentence are included in chunks
- Chunks are non-recursive
  - a chunk does not contain other chunks
- Chunks are mostly base NP chunks

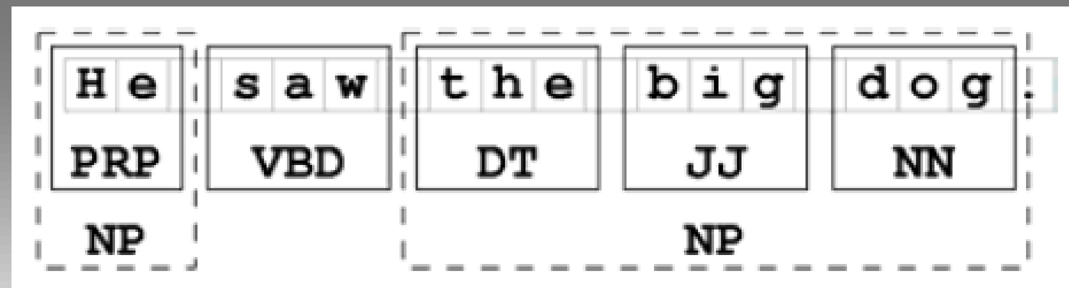
[[*the synthesis*]$_{\text{NP-base}}$ *of* [*long enhancer transcripts*]$_{\text{NP-base}}$]$_{\text{NP-complex}}$

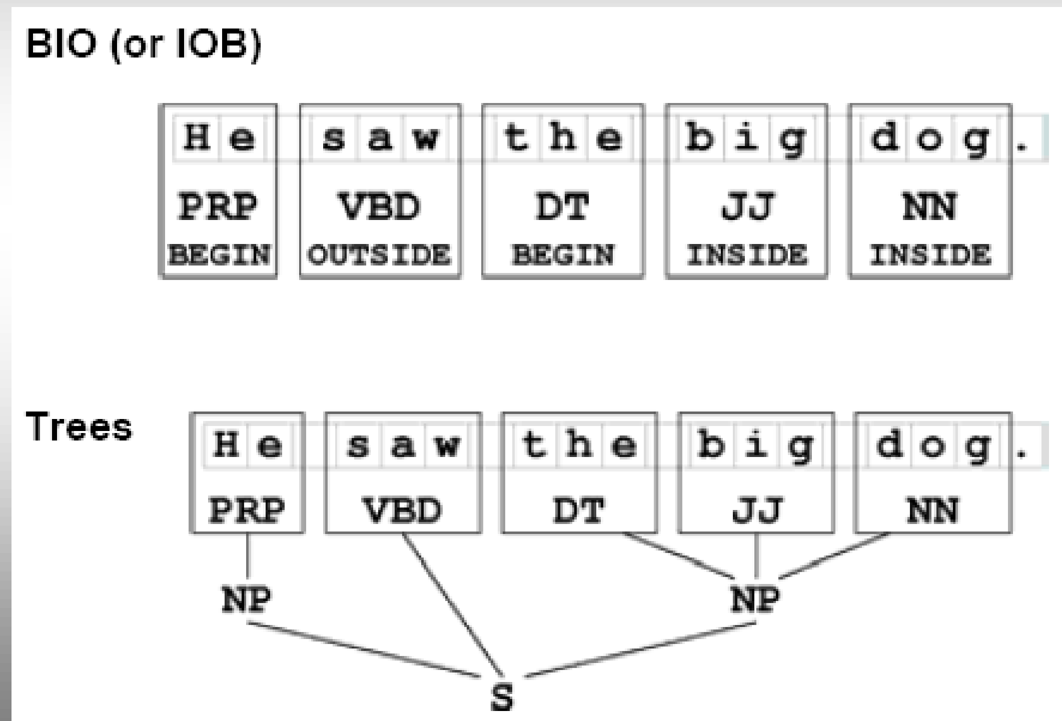
The Shallow Syntax Pipeline

Tagging

Chunking

Parsing





BIO Format for Base NPs

a	DT	I
mechanism	NN	I
that	WDT	B
increases	VBZ	O
NF-kappa	NN	I
B/I	NN	I
kappa	NN	I
B	NN	I
dissociation	NN	I
without	IN	O
affecting	VBG	O
the	DT	I
NF-kappa	NN	I
B	NN	I
translocation	NN	I
step	NN	I

A Simple Chunking Technique

- Simple chunkers usually ignore lexical content
  - Only need to look at part-of-speech tags
- Basic steps in chunking
  - Chunking / Unchunking
  - Chinking
  - Merging / Splitting

Regular Expression Basics

- “|” OR operator (explicit OR-ing)
  - “[a|e|i|o|u]” matches any occurrence of vowels
- “[abc]” matches any occurrence of either “a”, “b” or “c” (implicit OR-ing)
  - “gr[ae]y” matches “grey” or “gray” (but not “graey”)
- “.” matches arbitrary char
  - “d.g” matches “dag”, “dig”, “dog”, “dkg” …
- “?” preceding expression/char may or may not occur
  - “colou?r” matches “colour” and “color”
- “+” preceding expression occurs at least one time
  - “(ab)+” matches “ab”, “abab”, “ababab”, …
- “*” preceding expression occurs null time or arbitrary often
  - “(ab)*” matches “_”, “ab”, “abab”, “ababab”, …

Chunking

- Define a regular expression that matches the sequences of tags in a chunk
  - <DT>? <JJ>* <NN.?>
- Chunk all matching subsequences
  - *A*/DT *red*/JJ *car*/NN *ran*/VBD *on*/IN *the*/DT *street*/NN
  - [*A*/DT *red*/JJ *car*/NN] *ran*/VBD *on*/IN [*the*/DT *street*/NN]
- If matching subsequences overlap, the first one gets priority
- Unchunking is the opposite of chunking

Chinking

- A chink is a subsequence of the text that is not a chunk
- Define a regular expression that matches the sequences of tags in a chink
  - (<VB.?> | <IN>)+
- Chunk anything that is *not* a matching sub-sequence
  - *A*/DT *red*/JJ *car*/NN *ran*/VBD *on*/IN *the*/DT *street*/NN
  - [*A*/DT *red*/JJ *car*/NN] *ran*/VBD *on*/IN [*the*/DT *street*/NN]

chink

Merging

- Combine adjacent chunks into a single chunk
- Define a regular expression that matches the sequences of tags on both sides of the point to be merged
  - Merge a chunk ending in "JJ" with a chunk starting with "NN", i.e. left: <JJ>, right: <NN.>
- Chunk all matching subsequences
  - [*A*/DT *red*/JJ] [*car*/NN] *ran*/VBD
    on/IN *the*/DT *street*/NN
  - [*A*/DT *red*/JJ *car*/NN] *ran*/VBD
    on/IN *the*/DT *street*/NN
- Splitting is the opposite of merging

What are Named Entities?

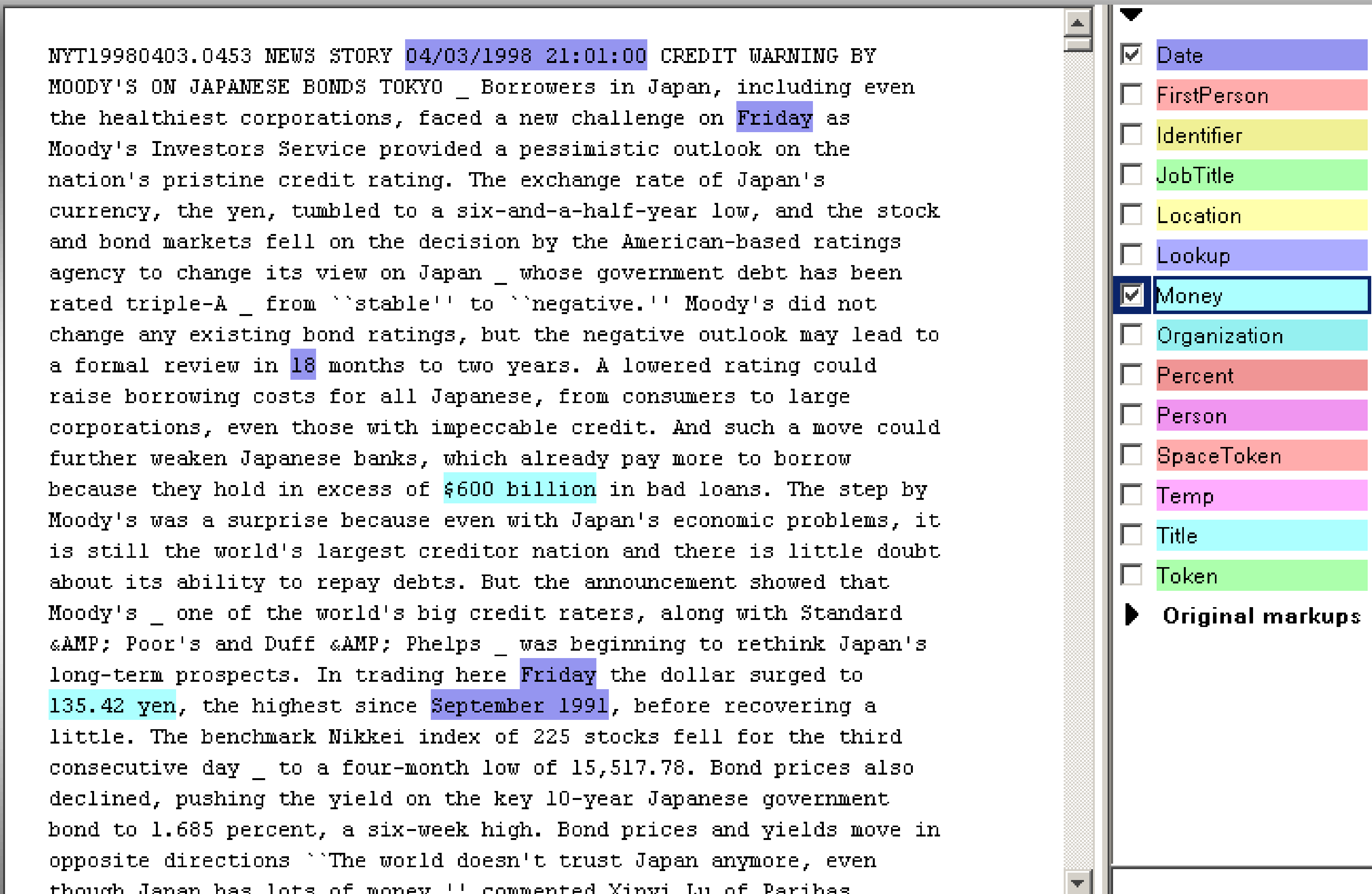
- Names of persons
  - *Dr. Jonathan Peeko, Professor Johnson*
- Names of companies or organizations
  - *Sony, United Nations, Texas Instruments, General Motors*
- Names of locations
  - *Paris, San Francisco, Rocky Mountains, Yellowstone Park*
- Date and time expressions
  - *Feb 17, 1973; 4.40p.m.; 16.40 Uhr; autumn 2000; last year*
- Addresses
  - *7 Ugly Way, Wolverhampton WH0 1Q5*
  - udo.hahn@uni-jena.de
- Names of proteins or genes or diseases,
  - *chloramphenicol acetyltransferase, NF-kappa B, SARS*
- Measure expressions
  - 420 kp, 21 l/m^2, 37%, 900€

What are Named Entities?

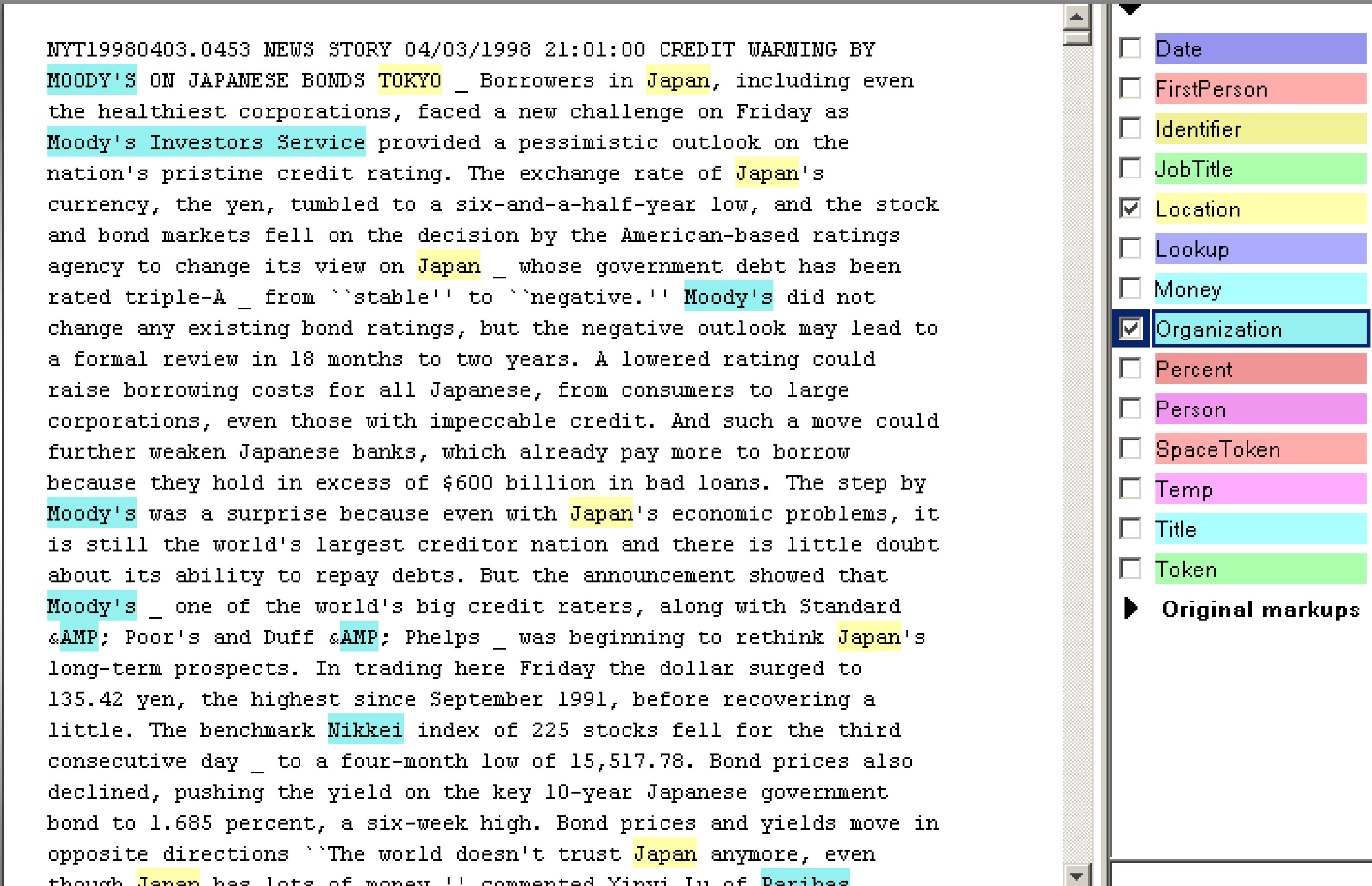
- Names of persons
  - *Dr. Jonathan Peeko, Prof… …ssor Johns…*
- Names of c… …ns
  - *Sony, Un… …al Motors*
- Na…
  - *P… …e Park*
- Date…
  - *…ear*
- Addre…
  - *…pton …40 1Q5*
  - udo.hah… …ni-jena.de
- Names of pr oteins or genes or diseases,
  - *chloramphenicol acetyltransferase, NF-kappa B, SARS*
- Measure expressions
  - 420 kp, 21 l/m^2, 37%, 900€

named entities are intentionally <u>excluded from</u> the lexicon

GATE: NER – Examples (1/3)

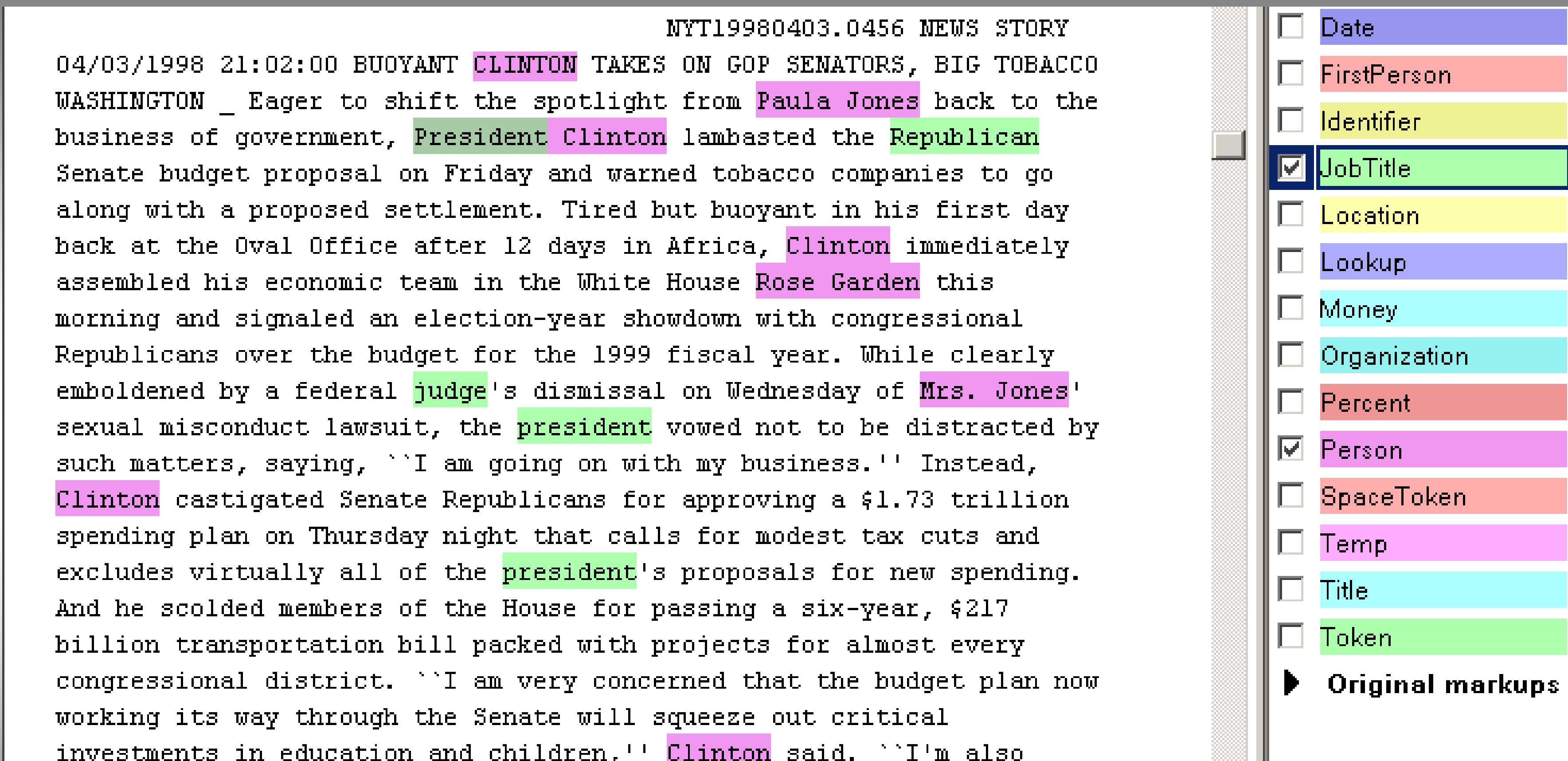

NYT19980403.0453 NEWS STORY 04/03/1998 21:01:00 CREDIT WARNING BY MOODY'S ON JAPANESE BONDS TOKYO _ Borrowers in Japan, including even the healthiest corporations, faced a new challenge on Friday as Moody's Investors Service provided a pessimistic outlook on the nation's pristine credit rating. The exchange rate of Japan's currency, the yen, tumbled to a six-and-a-half-year low, and the stock and bond markets fell on the decision by the American-based ratings agency to change its view on Japan _ whose government debt has been rated triple-A _ from ``stable'' to ``negative.'' Moody's did not change any existing bond ratings, but the negative outlook may lead to a formal review in 18 months to two years. A lowered rating could raise borrowing costs for all Japanese, from consumers to large corporations, even those with impeccable credit. And such a move could further weaken Japanese banks, which already pay more to borrow because they hold in excess of $600 billion in bad loans. The step by Moody's was a surprise because even with Japan's economic problems, it is still the world's largest creditor nation and there is little doubt about its ability to repay debts. But the announcement showed that Moody's _ one of the world's big credit raters, along with Standard & Poor's and Duff & Phelps _ was beginning to rethink Japan's long-term prospects. In trading here Friday the dollar surged to 135.42 yen, the highest since September 1991, before recovering a little. The benchmark Nikkei index of 225 stocks fell for the third consecutive day _ to a four-month low of 15,517.78. Bond prices also declined, pushing the yield on the key 10-year Japanese government bond to 1.685 percent, a six-week high. Bond prices and yields move in opposite directions ``The world doesn't trust Japan anymore, even though Japan has lots of money,'' commented Xinyi Lu of Paribas

GATE: NER – Examples (2/3)



NYT19980403.0453 NEWS STORY 04/03/1998 21:01:00 CREDIT WARNING BY MOODY'S ON JAPANESE BONDS TOKYO _ Borrowers in Japan, including even the healthiest corporations, faced a new challenge on Friday as Moody's Investors Service provided a pessimistic outlook on the nation's pristine credit rating. The exchange rate of Japan's currency, the yen, tumbled to a six-and-a-half-year low, and the stock and bond markets fell on the decision by the American-based ratings agency to change its view on Japan _ whose government debt has been rated triple-A _ from ``stable'' to ``negative.'' Moody's did not change any existing bond ratings, but the negative outlook may lead to a formal review in 18 months to two years. A lowered rating could raise borrowing costs for all Japanese, from consumers to large corporations, even those with impeccable credit. And such a move could further weaken Japanese banks, which already pay more to borrow because they hold in excess of $600 billion in bad loans. The step by Moody's was a surprise because even with Japan's economic problems, it is still the world's largest creditor nation and there is little doubt about its ability to repay debts. But the announcement showed that Moody's _ one of the world's big credit raters, along with Standard & Poor's and Duff & Phelps _ was beginning to rethink Japan's long-term prospects. In trading here Friday the dollar surged to 135.42 yen, the highest since September 1991, before recovering a little. The benchmark Nikkei index of 225 stocks fell for the third consecutive day _ to a four-month low of 15,517.78. Bond prices also declined, pushing the yield on the key 10-year Japanese government bond to 1.685 percent, a six-week high. Bond prices and yields move in opposite directions ``The world doesn't trust Japan anymore, even though Japan has lots of money.'' commented Xinyi Lu of Paribas

GATE: NER – Examples (3/3)



Two Types of NER Methods

Human Knowledge Engineering (symbolic p.)

- rule based
- developed by experienced language engineers
- based on human intuition
- requires only small amount of plain training data
- development can be very time consuming
- some changes may be hard to accommodate

(Supervised) Machine Learning Systems (empir.p.)

- use statistics or other machine learning technique
- developers do (almost) not need linguistic expertise
- fully automatic
- requires large amounts of annotated training data
- annotators are cheap (but you get what you pay for!)
- some changes may require re-annotation of the entire training corpus

Naïve NER Method: List Look-up

- Recognize entities stored in given lists
  - *gazetteers*, e.g., online phone directories, yellow pages)
- Advantages:
  - simple, fast, language independent, easy to retarget (just create lists)
- Disadvantages:
  - impossible to enumerate all names and name variants, collection and maintenance of lists

NER by Pattern Recognition

- Names often have internal structure - these components can be either stored or guessed, e.g., for "Location" we have RegEx-style constraints such as:

Capitalized Word + {City, Forest, Center, River}

which yields: *Sherwood Forest, Manchester City, Rhine River*

Capitalized Word + {Street, Boulevard, Avenue, Road}

which yields: *Portobello Street, Washington Avenue*

NER by Expressive Rules

- Context-sensitive rules of the kind:

$$A \rightarrow B \setminus C \; / \; D$$

  - A is a set of attribute-value expressions and optional score, the attributes refer to elements of the input token feature vector
  - B, C, D are sequences of attribute-value pairs and regular expressions; variables are also supported
  - B and D are left and right context, respectively, and can be empty (hint: read backwards!)

Example: [syn=NP, sem=ORG] (0.9) →
\ [norm="university"], [token="of"],
[sem=REGION|COUNTRY|CITY] / ;

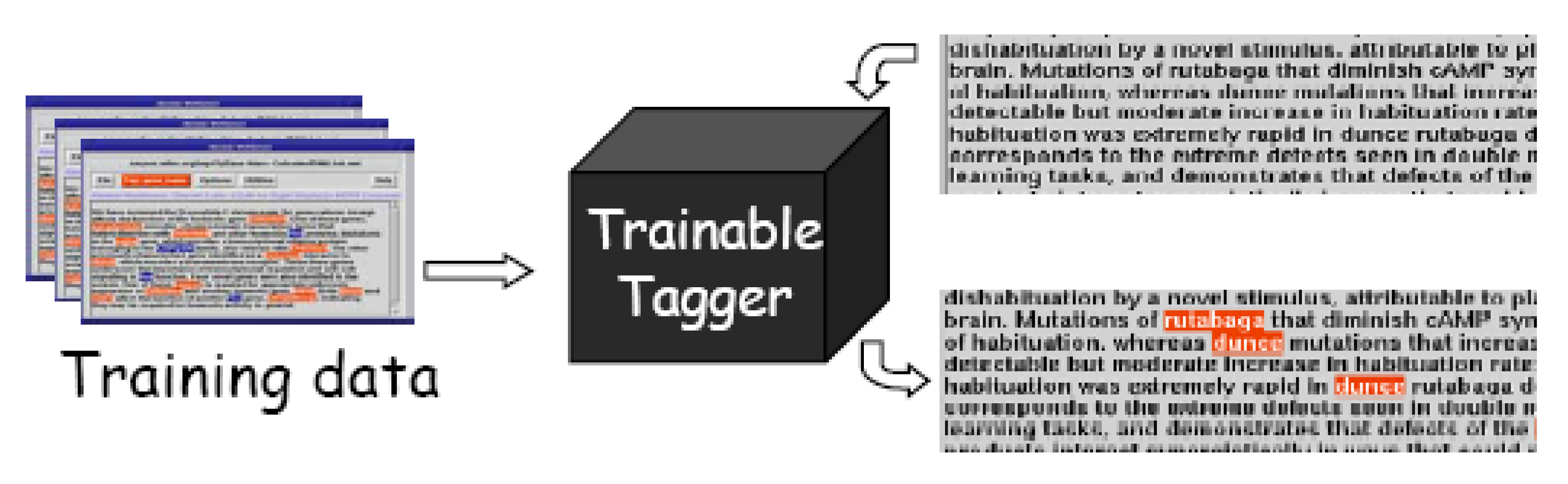
NER by Machine Learning

- NE task is frequently broken down in two parts:
  - Recognizing the entity boundaries
  - Classifying the entities in the NE categories
- Features are at least as important as the choice of the ML method
  - Simple pattern matching of orthographic features: capitalization, punctuation marks, numerical symbols
  - Windows for lexical features (e.g., “Mr.” for persons)
  - Affix features (“-ase” for proteins, “”-ectomy” for medical procedures, etc.”)
  - POS info (and chunks)
- Major Approaches (ML is a study topic on its own!)
  - Maximum Entropy [Chieu & Ng, 2002]
  - Hidden Markov Models [Bikel et al., 1999]
  - Support Vector Machines [Takeuchi & Collier, 2002]

Systemarchitektur für (überwachtes) Maschinelles Lernen

Merkmale
= beobachtbare Indikatoren
(in den Trainingsdaten)

Algorithmen für Maschinelles Lernen
= Rechenverfahren zur Bestimmung von
(statistischen) Modellen über die Verteilung von
Merkmalen (in den Trainingsdaten)



Resources for NLP

- Empirical (Learning) Paradigm for NLP
- Types of Resources
  - Language data (plain, annotated)
  - Systems for acquiring and maintaining language data
  - Computational lexicons and ontologies
  - NLP Core Engines
  - NLP Application Systems
  - Machine Learning Resources
- Methodological Issues of NLP Resources

Ressourcen für die Sprachverarbeitung

- Referenzkorpora (Nationalkorpora)
  - Standardsprache (Zeitungen, Belletristik)
- Non-Standard-Korpora
  - Informelle Sprache (Chats, Blogs, E-Mails)
  - Fachsprachen (z.B.: klinische Berichte)
- Rohdaten vs. Annotation
  - Linguistische Metadaten
    - Morphologie, Syntax, Semantik, Pragmatik

Language Data

- Plain language data
  - Just text or speech
    - ASCII/UTF-8-compatible, pdf, HTML/SGML
- Annotated language data
  - Enriched by linguistic meta-data
    - Linguistic annotation languages (XML)

Plain Language Data

- Mixed/Balanced text collections
  - British National Corpus (BNC)
  - American National Corpus (ANC)
- Newspaper collections
  - Wall Street Journal
  - IdS-Korpora (DeReKo*)
- The Web

British National Corpus (BNC)

- 100M word collection (some 4,050 texts) of 20th century British English
- Written part (90%)
  - Regional and national newspapers
  - Specialist periodicals and journals (various genres)
  - Academic books and popular fiction
  - Letters, memoranda, school and university essays
- Spoken part (10%)
  - Informal conversations (different ages, regions, social classes)
  - Formal business and government meetings
  - Radio shows and phone-ins

- http://www.natcorp.ox.ac.uk/

British National Corpus (BNC)

- Encoding based on ‚Guidelines of the Text Encoding Initiative‘ (TEI),
  - using ISO standard 8879 (SGML: Standard Generalized Markup Language)
- Whole collection is POS-tagged
  - using the CLAWS tagger for the C5 tag set (C7 is much more elaborate)
  - Error rate: 1.7%
  - Tagging ambiguity for 4.7% of all tags

American National Corpus (ANC)

- 15M word collection (texts) of 20th century American English
- Annotated for structural markup (sections, chapters, etc.) down to the level of paragraph, sentence boundaries, words (tokens) with part of speech annotations and lemma using the Penn tagset, noun and verb chunks, named entities (Person, Location, Organization, Date)
- Written part (80%)
  - Regional and national newspapers
  - Specialist periodicals and journals (various genres)
  - Academic books and popular fiction
  - Governmental docs
- Spoken part (20%)
  - Informal face-to-face conversations (different ages, regions, social classes)
  - Telephone conversations
- http://www.anc.org/

Große deutsche Textkorpora

(verschriftlichte Sprache)

- Deutsches Referenzkorpus – DeReKo
  - Institut für deutsche Sprache (IdS) Mannheim
  - Zeitungen, Belletristik, Handbücher, Parlamentsprotokolle (seit 1956)
  - Umfang: ca. 42 Mrd. Tokens
  - http://www1.ids-mannheim.de/kl/projekte/korpora/
- Digitales Wörterbuch der deutschen Sprache – DWDS
  - Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
  - Zeitungen, Belletristik, Gebrauchsliteratur, Wissenschaft (20./21. Jahrhundert)
  - Umfang: 12,8 Mio Dokumente, ca. 5,5 Mrd. Tokens
  - https://www.dwds.de/
- Deutsches Textarchiv – DTA
  - Historisches Referenzkorpus: 1600-1900
  - 3352 Werke (639K Seiten), 157 Mio. Tokens
  - Annotiert mit Tokens, Lemmata, POS
  - http://www.deutschestextarchiv.de/

Language Data Repositories

- Linguistic Data Consortium
  - „Catalog“ option
  - „LDC Online“ provides you a guest account

`http://www.ldc.upenn.edu/`

Language Data Repositories

- Linguist List
  - Open Language Archives Community
  - „Text & Computer Tools“ button
    - Texts and Corpora
  - „Language Resources“ button
    - Texts and Corpora

http://linguistlist.org/olac

Language Data Repositories

- European Language Resources Association (ELRA)
  - „R&D Catalog“ option
  - Spoken LRs
    - Telephone recordings
    - Desktop/mircophone recordings
    - Broadcast resources
    - Speech related resources
  - Written LRs
    - Corpora
    - Mono- and multilingual lexicons
  - (Domain-specific) Terminological resources
  - Multimodal/multimedia LRs

http://www.elra.info/

Language Data Repositories

- Natural Language Software Registry
  - Annotation tools
  - Evaluation tools
  - Language Resources
  - Multimedia
  - Multimodality
  - NLP Development Aid
  - Spoken Language
  - Written Language

http://registry.dfki.de

Annotated Language Data

- Levels of annotation
  - Formal text structure processing
    - Paragraphs, sentences, tokens
  - Syntactic mark-up
    - Parts of speech
    - Shallow syntactic structures: chunks
    - Deep syntactic structures: parses
  - Semantic mark-up
    - Named entities
    - Propositions, predicate-argument structures
  - Discourse mark-up
    - Referential relations
    - Rhetorical relations

Annotation Styles

- In-line annotation
  - Mark-ups appear as integral part of the original text
    - This is an <XMLTag> in-line <\XMLTag> annotation
- Stand-off annotation
  - Mark-ups appear distinct from the original text (e.g., in a different window)
    - This is a stand-off annotation
      - <XMLTag StartChar: 11, XMLTag EndChar: 19, XMLTag Type STAND-OFF>

General Language Corpora for Syntactic Annotation

- Penn Treebank (U Penn)
  - language: English (general language)
  - text genre: mostly newspaper articles (*Wall Street Journal*)
  - size: 1,200,000 (annotated) tokens
  - Syntactic tagging based on set of 45 tags
  - Syntactic phrase structures (parse trees) based on Government-Binding grammar
  - <u>No</u> named entity annotation
  - But propositional annotation: PropBank

http://www.cis.upenn.edu/~treebank/

General Language Corpora for Proposition Annotation

- PropBank (U Penn)
  - language: English (general language)
  - text genre: financial newspaper articles (*Wall Street Journal*)
  - size: 300,000 (annotated) tokens
  - proposition format:
    - [subject - predicate - object]
  - "semantic" counterpart of Penn Treebank

http://www.cis.upenn.edu/~ace/

General Language Corpora for Discourse Annotation

- Penn Discourse TreeBank (PDTB; U Penn)
  – language: English (general language)
  – text genre: financial newspaper articles (*Wall Street Journal*)
  – size: 1 M tokens (WSJ) and 40k relations
  – Annotated with information related to discourse structure and discourse semantics, i.e., temporal, contingency, comparison, and expansion discourse relations (**after, when, but, although, if**)
  – "discourse" counterpart of Penn Treebank

http://www.cis.upenn.edu/~pdtb/

General Language Corpora for Discourse Analysis

- RST Corpus (ISI/USC, USA)
  - language: English
  - size: 385 documents, i.e., 176,000 tokens; 21,789 elementary discourse units (EDUs)
  - text genre: newspaper articles (*Wall Street Journal*)
  - Rhetorical Structure Theory (RST)
    - 90 coherence relations

Penn TreeBank: Sizes and Genres

Table 4:
Penn Treebank
(as of 11/92)

Description	**Tagged for Part-of-Speech** (Tokens)	**Shallow Parsing** (Tokens)
Dept. of Energy abstracts	231,404	231,404
Dow Jones Newswire stories	3,065,776	1,061,166
Dept. of Agriculture bulletins	78,555	78,555
Library of America texts	105,652	105,652
MUC-3 messages	111,828	111,828
IBM Manual sentences	89,121	89,121
WBUR radio transcripts	11,589	11,589
ATIS sentences	19,832	19,832
Brown Corpus, retagged	1,172,041	1,172,041
Total:	4,885,798	2,881,188

Penn TreeBank POS Tag Set

Table 2:
The Penn Treebank POS tagset

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

PTB POS Annotation Process

- Four annotators: Grad students of linguistics
- Comparison of two annotation styles on a 16,000 word sample:
  - „Tagging“:
    - completely manual annotation
  - „Correcting“:
    - automatical POS tagging and subsequent manual correction
- Inter-annotator disagreement:
  - „Tagging“: 7,2%
  - „Correcting“: 4,1%
- Comparison of accuracy with benchmark version (disagreement):
  - „Tagging“: 5,4%
  - „Correcting“: 4,0%

Illustration of the „Correcting“ Mode

Battle-tested/NNP*/JJ Japanese/NNP*/JJ industrial/JJ managers/NNS here/RB always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.

- Training of annotators took 15h
- Annotation speed (after one month of training): > 3000 words/h
- Double as fast as „Tagging“ !

Syntactic Annotation of PTB

- Correction of false automatic parser output as provided by the FIDDITCH parser (Hindle 1989):
  - Outputs only one analysis per sentence
  - No attachments when parser is unsure about attachment decision
  - Alternative solution: decomposition of sentence structure into sets of partial trees

    ➔ partial sentence structure description
  - Good lexicon and grammar coverage
- Task of annotators is mainly to „glue“ (i.e., to attach) partial phrase structure trees
  - Less time-consuming than re-bracketing the entire parser output

Penn Treebank Phrasal Tag Set

1.	ADJP	Adjective phrase
2.	ADVP	Adverb phrase
3.	NP	Noun phrase
4.	PP	Prepositional phrase
5.	S	Simple declarative clause
6.	SBAR	Clause introduced by subordinating conjunction or *0* (see below)
7.	SBARQ	Direct question introduced by *wh*-word or *wh*-phrase
8.	SINV	Declarative sentence with subject-aux inversion
9.	SQ	Subconstituent of SBARQ excluding *wh*-word or *wh*-phrase
10.	VP	Verb phrase
11.	WHADVP	*Wh*-adverb phrase
12.	WHNP	*Wh*-noun phrase
13.	WHPP	*Wh*-prepositional phrase
14.	X	Constituent of unknown or uncertain category

Null elements

1.	*	"Understood" subject of infinitive or imperative
2.	0	Zero variant of *that* in subordinate clauses
3.	T	Trace—marks position where moved *wh*-constituent is interpreted
4.	NIL	Marks position where preposition is interpreted in pied-piping contexts

Partially bracketed output from FIDDITCH

```
( (S
    (NP (NBAR (ADJP (ADJ "Battle-tested/JJ")
                    (ADJ "industrial/JJ"))
              (NPL "managers/NNS")))
    (? (ADV "here/RB"))
    (? (ADV "always/RB"))
    (AUX (TNS *))
    (VP (VPRES "buck/VBP")))
    (? (PP (PREP "up/RP")
           (NP (NBAR (ADJ "nervous/JJ")
                     (NPL "newcomers/NNS")))))
    (? (PP (PREP "with/IN")
           (NP (DART "the/DT")
               (NBAR (N "tale/NN"))
                     (PP of/PREP
                         (NP (DART "the/DT")
                             (NBAR (ADJP
                                    (ADJ "first/JJ"))))))))
    (? (PP of/PREP
           (NP (PROS "their/PP$")
               (NBAR (NPL "countrymen/NNS")))))
    (? (S (NP (PRO *))
              (AUX to/TNS)
              (VP (V "visit/VB")
                  (NP (PNP "Mexico/NNP")))))
    (? (MID ",/,"))
    (? (NP (IART "a/DT")
           (NBAR (N "boatload/NN"))
                 (PP of/PREP
                     (NP (NBAR
                         (NPL "warriors/NNS"))))
                 (VP (VPPRT "blown/VBN")
                     (? (ADV "ashore/RB"))
                     (NP (NBAR (CARD "375/CD")
                               (NPL "years/NNS"))))))
    (? (ADV "ago/RB"))
    (? (FIN "./.")))
```

Automatic simplification of the output from Fidditch

```
( (S
    (NP (ADJP Battle-tested industrial)
         managers)
    (? here)
    (? always)
    (VP buck))
    (? (PP up
           (NP nervous newcomers)))
    (? (PP with
           (NP the tale
               (PP of
                   (NP the
                       (ADJP first))))))
    (? (PP of
           (NP their countrymen)))
    (? (S (NP *)
          to
          (VP visit
              (NP Mexico))))
    (? ,)
    (? (NP a boatload
           (PP of
               (NP warriors))
           (VP blown
               (? ashore)
               (NP 375 years))))
    (? ago)
    (? .))
```

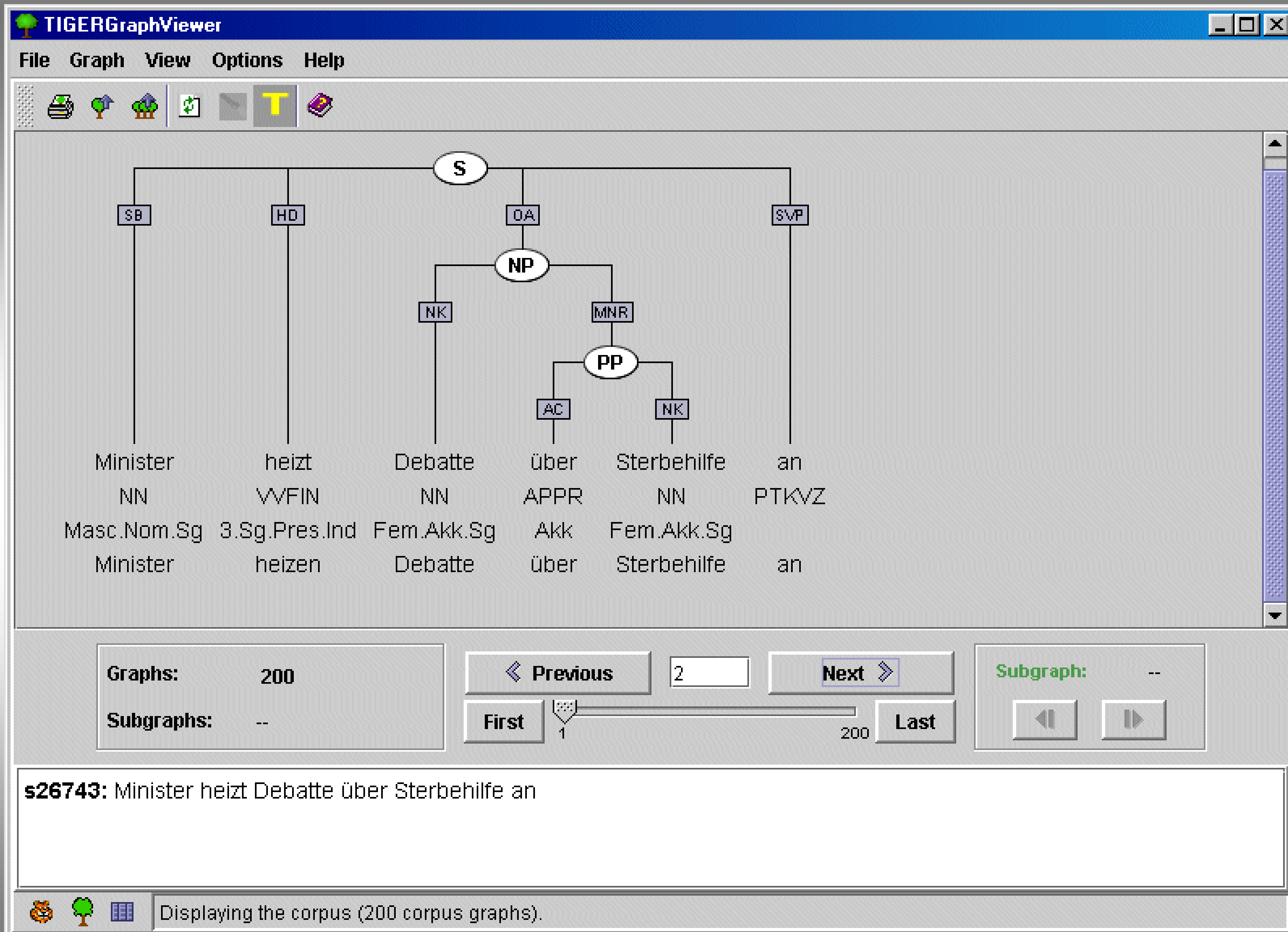
After „Correcting“ by the annotators

```
( (S
    (NP Battle-tested industrial managers
        here)
     always
    (VP buck
        up
        (NP nervous newcomers)
        (PP with
            (NP the tale
                (PP of
                    (NP (NP the
                            (ADJP first
                                  (PP of
                                      (NP their countrymen)))
                            (S (NP *)
                               to
                               (VP visit
                                   (NP Mexico))))
                        ,
                        (NP (NP a boatload
                                (PP of
                                    (NP (NP warriors)
                                        (VP-1 blown
                                            ashore
                                            (ADVP (NP 375 years)
                                                  ago)))))
                            (VP-1 *pseudo-attach*))))))))
  .)
```

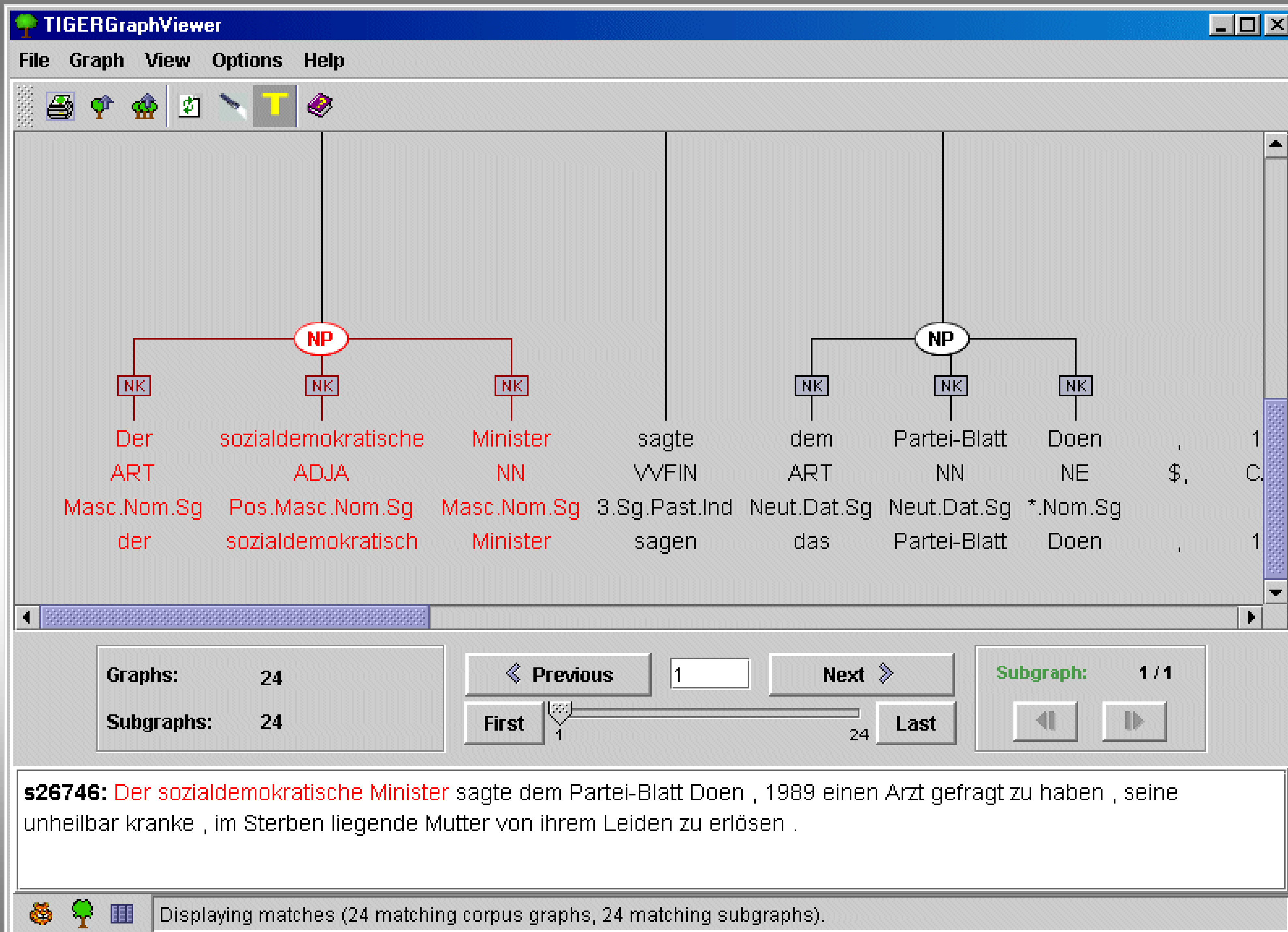
TiGer Corpus

- 0,9M word collection (50K sentences) of German language newspaper articles (FR)
- http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
- morphological, POS, parse tree tagging
- Treebank query tool TiGer Search

TiGer Corpus



TiGer Search (NP)



STTS Tag Set for German (1/2)

- ADJA attributives Adjektiv [das] große [Haus]
- ADJD adverbiales oder [er fährt] schnell prädikatives Adjektiv [er ist] schnell
- ADV Adverb schon, bald, doch
- APPR Präposition; Zirkumposition links in [der Stadt], ohne [mich]
- APPRART Präposition mit Artikel im [Haus], zur [Sache]
- APPO Postposition [ihm] zufolge, [der Sache] wegen
- APZR Zirkumposition rechts [von jetzt] an
- ART bestimmter oder der, die, das, unbestimmter Artikel ein, eine, ...
- CARD Kardinalzahl zwei [Männer], [im Jahre] 1994 (Ordinalzahlen sind als ADJA getaggt)
- FM Fremdsprachliches Material [Er hat das mit ``] A big fish ['' übersetzt]
- ITJ Interjektion mhm, ach, tja
- KOUI unterordnende Konjunktion um [zu leben], mit ``zu'' und Infinitiv anstatt [zu fragen]
- KOUS unterordnende Konjunktion weil, daß, damit, mit Satz wenn, ob
- KON nebenordnende Konjunktion und, oder, aber
- KOKOM Vergleichskonjunktion als, wie
- NN normales Nomen Tisch, Herr, [das] Reisen
- NE Eigennamen Hans, Hamburg, HSV
- PDS substituierendes Demonstrativ- dieser, jener pronomen
- PDAT attribuierendes Demonstrativ- jener [Mensch] pronomen
- PIS substituierendes Indefinit- keiner, viele, man, niemand pronomen
- PIAT attribuierendes Indefinit- kein [Mensch], pronomen ohne Determiner irgendein [Glas]
- PIDAT attribuierendes Indefinit- [ein] wenig [Wasser], pronomen mit Determiner [die] beiden [Brüder]
- PPER irreflexives Personalpronomen ich, er, ihm, mich, dir
- PPOSS substituierendes Possessiv- meins, deiner pronomen
- PPOSAT attribuierendes Possessivpronomen mein [Buch], deine [Mutter]

STTS Tag Set for German (2/2)

- PRELS substituierendes Relativpronomen [der Hund ,] der
- PRELAT attribuierendes Relativpronomen [der Mann ,] dessen [Hund]
- PRF reflexives Personalpronomen sich, einander, dich, mir
- PWS substituierendes wer, was Interrogativpronomen
- PWAT attribuierendes welche [Farbe], Interrogativpronomen wessen [Hut]
- PWAV adverbiales Interrogativ- warum, wo, wann, oder Relativpronomen worüber, wobei
- PAV Pronominaladverb dafür, dabei, deswegen, trotzdem
- PTKZU ``zu'' vor Infinitiv zu [gehen]
- PTKNEG Negationspartikel nicht
- PTKVZ abgetrennter Verbzusatz [er kommt] an, [er fährt] rad
- PTKANT Antwortpartikel ja, nein, danke, bitte
- PTKA Partikel bei Adjektiv am [schönsten], oder Adverb zu [schnell]
- TRUNC Kompositions-Erstglied An- [und Abreise]
- VVFIN finites Verb, voll [du] gehst, [wir] kommen [an]
- VVIMP Imperativ, voll komm [!]
- VVINF Infinitiv, voll gehen, ankommen
- VVIZU Infinitiv mit ``zu'', voll anzukommen, loszulassen
- VVPP Partizip Perfekt, voll gegangen, angekommen
- VAFIN finites Verb, aux [du] bist, [wir] werden
- VAIMP Imperativ, aux sei [ruhig !]
- VAINF Infinitiv, aux werden, sein
- VAPP Partizip Perfekt, aux gewesen
- VMFIN finites Verb, modal dürfen
- VMINF Infinitiv, modal wollen
- VMPP Partizip Perfekt, modal gekonnt, [er hat gehen] können
- XY Nichtwort, Sonderzeichen 3:7, H2O, enthaltend D2XW3
- $, Komma , $. Satzbeendende Interpunktion . ? ! ; : $(sonstige Satzzeichen; satzintern - [ ]()

Penn Proposition (Prop) Bank (2000 –)

- Predicate/Argument structure (PAS) along syntactic subcategorization frames
  - P:`Drink (`A`: Agent: x)`
  - P:`Drink (`A`: Patient: y)`
- Focus on verbs (*events*) and their syntactic arguments (*participants*)
  - later phases: nominalizations, adjectives and prepositions
- Linguistic heritage:
  - Verb classes for the English language (Levin 1993)
  - with focus on semantic considerations (semantic or theta roles)
- Large coverage is a major goal

Example for Propositions (PPB)

Bush met Blair

Bush and Blair met

Bush met with Blair

Bush and Blair had a meeting

. . .

battle

wrestle

join

debate

consult

Proposition: meet(Bush, Blair)

meet(Somebody1, Somebody2)

When Bush met Blair on Thursday

they discussed the stabilization of the Iraq.

meet(Bush, Blair) **discuss([Bush, Blair], stabilize(X, Iraq))**