Natural Language Processing

Timo Baumann

OSTBAYERISCHE
TECHNISCHE HOCHSCHULE
REGENSBURG

Heute

- consolidation: FSAs and FSTs
- syntax and constituent grammars
- parsing (=syntactic analysis) with CFGs

Principle of Compositionality

- Speech sounds (or letters) form words
- Words follow each other, forming sequences
- sequences of words have an inner structure (syntax)
- meaning (semantics) can be derived from the structure

automaton-based methods

- finite state automata correspond to linear grammar (and regular expressions)
- transducers provide the means to convert strings into other strings
- applications:
  - (de)composition of words (morphological analysis)
  - pronunciation modeling
  - sentence chunking (in a few slides)

Syntax and Parsing: Sentence structure and its analysis

Sentence structure / Syntax

- Syntax is the subfield of linguistics that is concerned with the structure of sentences
  - not: structure of words (that's morphology)
    (not parts of speech – Wortarten – either)
  - not the meaning of sentences either
  - understanding syntax is an important step in understanding the meaning of sentences
  - syntax is strongly lexicalized, i.e., what construction is allowed can depend on individual words (“Döner mit alles” is perfectly normal); therefore, syntax often relates to meaning

Syntax and NLP

Intersection of:

- linguistic perspective:
  what kinds of structures must be represented?
  - introspection
  - corpus studies
  - psycholinguistics / neurolinguistics
- formal languages and grammars (theory of computer science):
  - what formal models are available to represent the required kinds of structure?
- language resources (empirical / data science):
  - collections of (syntax-annotated) sentences: *treebanks*
  - syntactic lexica, which describe syntactic requirements of words (such as valence for verbs)
  -

Syntax

- describes why some word sequences are structurally sound (“grammatical”) while others are not

  The man reads the book.

  The book reads the man

  *The book the reads man.

  *The man reads the.

  The man reads.

  *The man the book.

Constituents

- some parts of a sentence belong together more closely than others. They form a constituent. Tests for constituency:
  - distribution (exchangability): Peter isst Eis. / Anne isst Eis. → “isst Eis” forms a constituent; Peter and Anne are of the same kind (grammatically)
  - coordination: Peter *und* Anne essen Eis.
  - exchangeable with a pronoun: Sie essen Eis. Sie essen etwas.
  - omission test (ellipsis): Peter isst Eis aus dem Eiscafé und Anne vom Kiosk.
  - question tests: Wer isst Eis? Was isst Anne? Was tut Peter?
  - repositioning tests: Eis essen Anne und Peter.
- constituents build hierarchies, i.e., constituents themselves consits of words or (lower-level) constituents

some constituents and their mapping to grammar rules

- **nominal phrases** such as:

  names, determiners+nouns, det.+adjektives+nouns, ...

  - NP → NN | Det N | Det A N | ...
- determiner beyond the definite/indefinite articles (the/a):

  Peter's Eis, Peter's Freundin's Eis, Peter's Freundin's Schwester's Eis, ...

  - Det → “der” | “die” | “das” | “ihr” ... | NP “'s”

automaton-based syntax analysis

[I begin] [with an intuition]: [when I read] [a sentence], [I read it] [a chunk] [at a time]
(Example from S. Abney, Parsing by Chunks)

- Chunks correspond to prosodic phrases when reading / speaking aloud
  - stuff that belongs together is spoken together (=in-between short breaths)
- are *flat*, i.e., they don't build hierarchies
  - can be modelled by FSA/regular expressions via types of speech (=Wortarten)
  - e.g. NP = Det N | NN | Det Adj N
- chunking is a good basis for probabilistic approaches and knowledge extraction
- correspond to parts of constituents but do not *relate* constituents to each other

  → no full support for hierarchical modelling
- are insufficient to model all of syntax → we'll ignore FSA-based syntax chunking

kontextfreie Grammatiken

$G = (\Phi, \Sigma, R, S)$

Φ Nonterminalsymbole

Σ Terminalsymbole

R Regeln $A \rightarrow \alpha$, wobei $A \in \Phi$ und $\alpha \in (\Phi \cup \Sigma)^*$

S Startsymbol aus Φ

S $\rightarrow$ NP VP

NP $\rightarrow$ Pr | Det N | NP PP

Nominalphrase!

VP $\rightarrow$ V NP | V | VP PP

Verbalphrase!

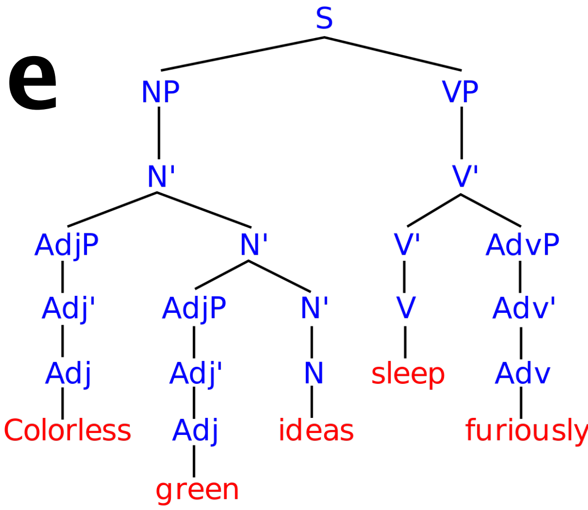
PP $\rightarrow$ P NP

Präpositionalphrase!
(P = mit/bei/für/wegen/an/auf/...)

theoretical issues of automaton-based syntax analysis

- FSAs can do some hierarchies, such as:
  - einfache Rekursion wie bei Anne's Freund's Eis
- however, you remember that FSA can't do $a^n b^n$ – do you?
  - die Frau
  - die Frau die den Hund führte
  - die Frau die den Hund der den Mann biss führte
  - die Frau die den Hund der den Mann der auf die Uhr sah biss führte
  - d Frau d d Hund d d Mann d a d Uhr die die Zeit anzeigte sah biss führte
  - “center embedding” is unlimited in theory
- in practice embedding depth is not unlimited, at least not in spontaneously observed language

constituents and context-free grammars



- exchangeability is fundamental idea of constituents

  → that's what “context-free” is about

- CFGs therefore are a reasonable model for constituent structures

Colorless green ideas sleep furiously
– Noam Chomsky

context-free grammars and productions

- S
- NP VP
- Pr VP PP
- Pr V NP P NP
- Pr V Det N P Det N

S → NP VP

NP → Pr | Det N | NP PP

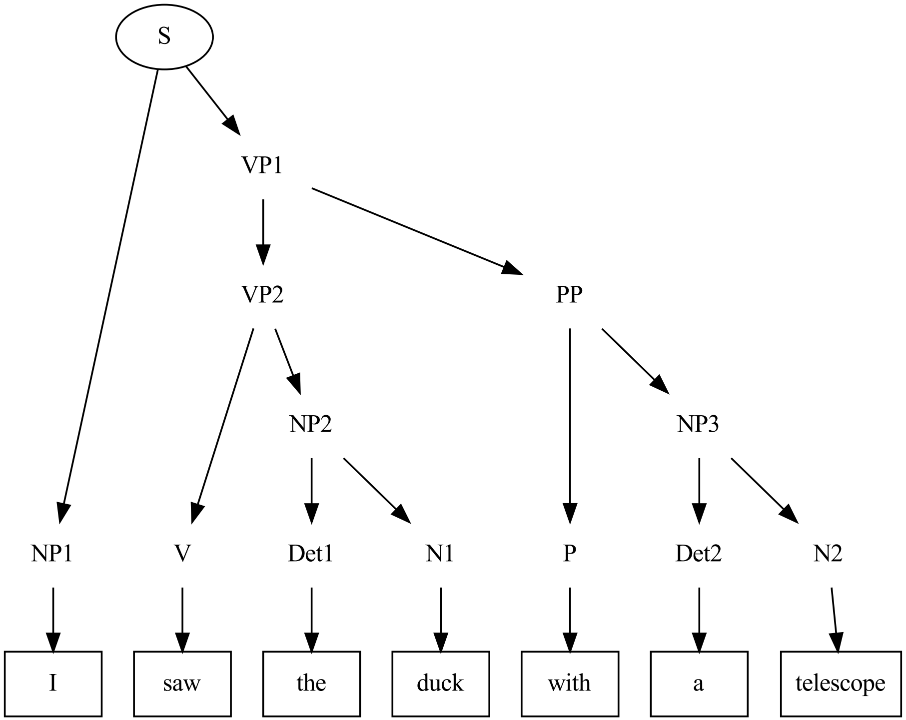
VP → V NP | V | VP PP

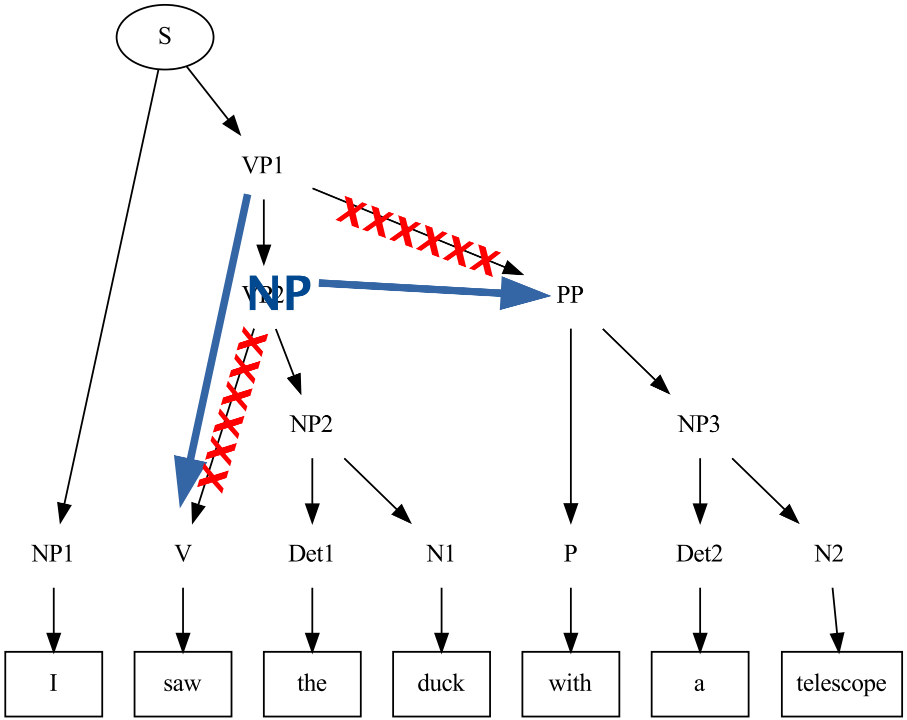
PP → P NP

I saw her duck with the telescope

could we have achieved the same results with different derivations?

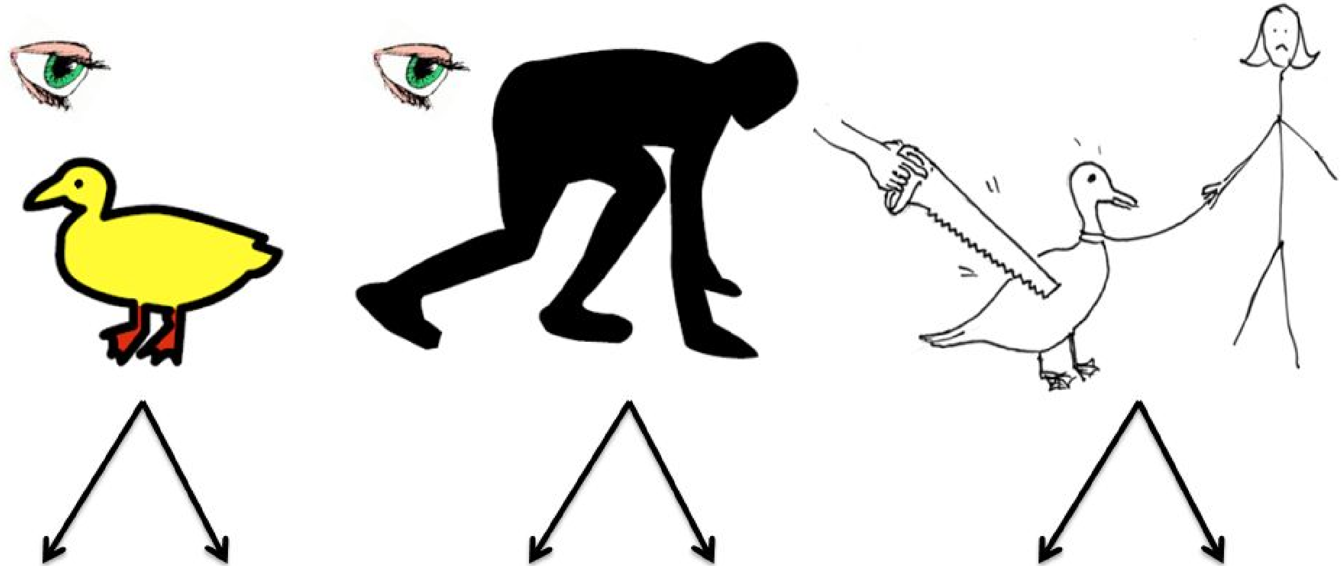
visualization of derivation trees



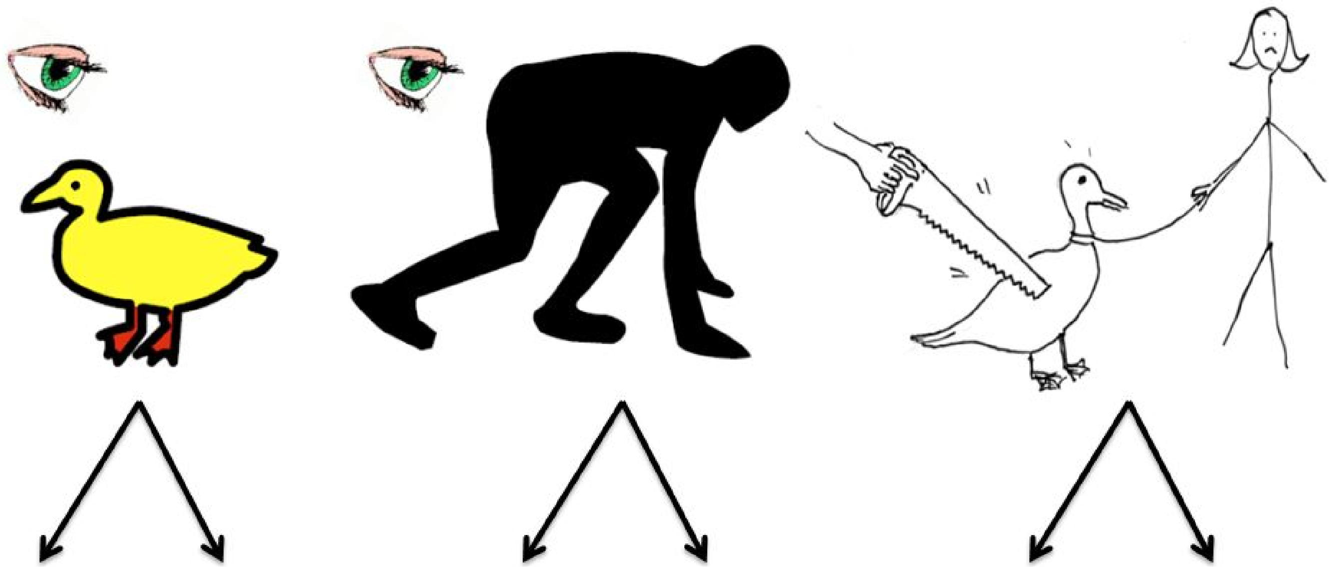


(syntactic) ambiguity in language

"I saw her duck with a telescope…"



“I saw her duck with a telescope…”



$S \rightarrow NP\ VP$

$NP \rightarrow Pr \mid Det\ N \mid NP\ PP$

$VP \rightarrow V\ NP \mid V \mid VP\ PP$

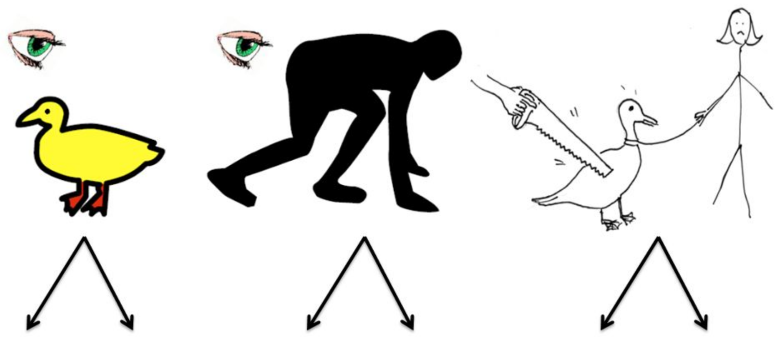
$PP \rightarrow P\ NP$

Slide credit: Dhruv Batra, figure credit: Liang Huang

(syntactic) ambiguity in language

- natürlich sind nicht alle Interpretationen sinnvoll
  - aber Syntax schert sich nicht um Sinn
  - erstmal alle Möglichkeiten zulassen, später sieben
- im Ergebnis meist sehr viele mögliche Ergebnisse
  - auf Effizienz achten!

"I saw her duck with a telescope..."



Slide credit: Dhruv Batra, figure credit: Liang Huang

parsing

- “parsing a sentence”
  = infer the syntactic structure of a sentence from the sequence of words (=surface structure)
- “the parse” / parse tree = generated structure
- parser = computer programm for parsing
- default: syntax (however: “semantic parsing” is also something which you may hear)

challenge: efficient management of ambiguity

- search complexity: 2^n (i.e., prohibitively high)
- dynamic programming (=storage and re-use of partial results)
- store partial results in a chart
  - complexity → $O(n^3)$ to find out if sentence is grammatical (still exponential if we want to get all possible parse trees)

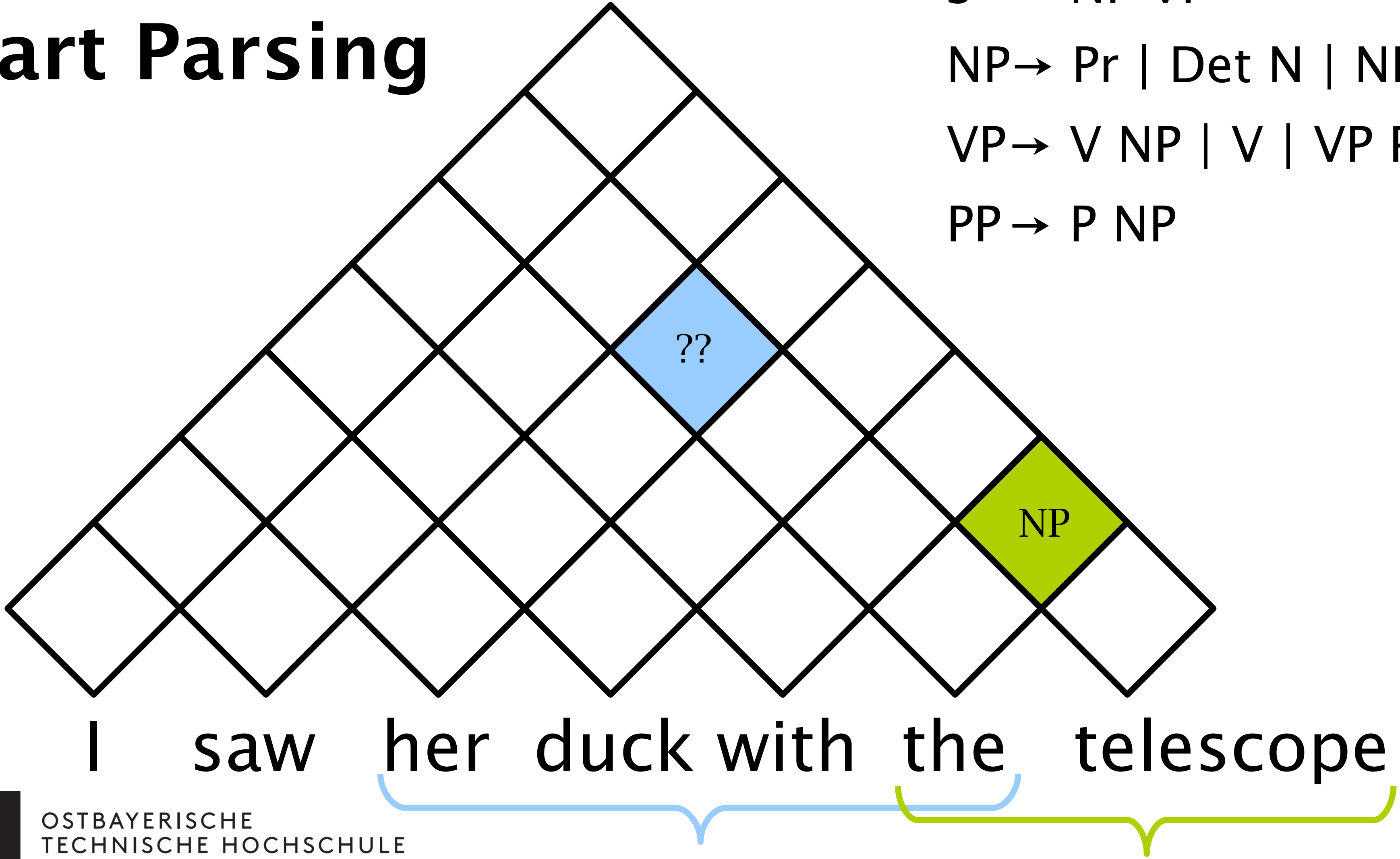
Chart Parsing

S → NP VP

NP→ Pr | Det N | NP PP

VP→ V NP | V | VP PP

PP→ P NP



Parsing strategies

- (purely) top-down:
  - expand S until we find the full sentence (hopefully)
- bottom-up (Cocke-Kasami-Younger 1967)
  - check what can be combined (e.g. NP → ... → “the telescope”), hoping that we'll eventually get an S at the top
- mixed: left-corner parsing (Earley 1970)
  - expand S (top-down), such that the first (then second, third, ...) words fit the derived structure

S → NP VP

NP → Pr | Det N | NP PP

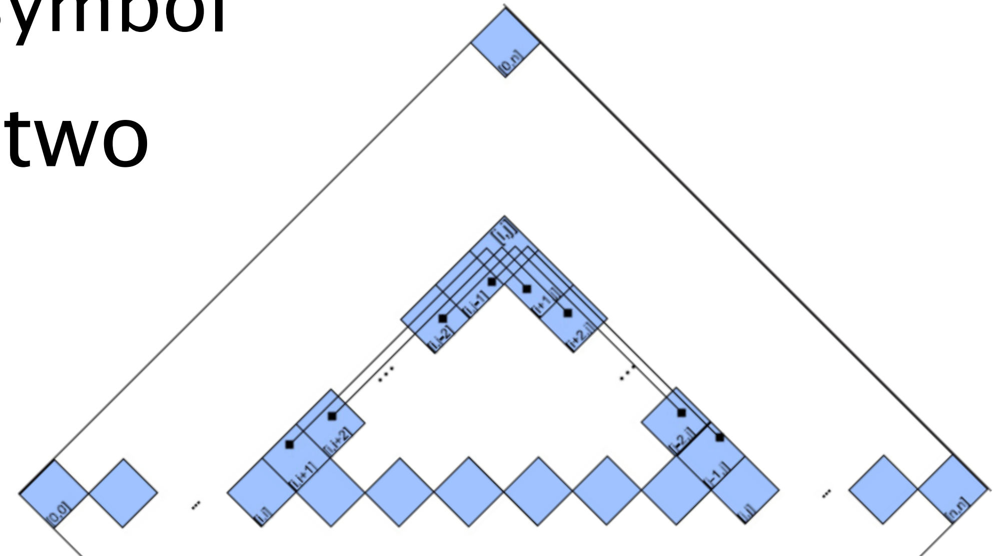
VP → V NP | V | VP PP

PP → P NP

I saw her duck with the telescope

Cocke-Kasami-Younger Algorithm

- erquires grammar to be in Chomsky-Normal-Form (CNF)
  - each rules produces (exactly) two non-terminals OR exactly one terminal symbol
- idea: always combine in two directions of the chart



Cocke-Kasami-Younger Algorithmus

function CKY-PARSE(*words, grammar*) **returns** *table*

for $j \leftarrow$ **from** 1 **to** LENGTH(*words*) **do**
    $table[j-1, j] \leftarrow \{A \mid A \rightarrow words[j] \in grammar\}$
    for $i \leftarrow$ **from** $j-2$ **downto** 0 **do**
        for $k \leftarrow i+1$ **to** $j-1$ **do**
            $table[i,j] \leftarrow table[i,j] \cup$
                $\{A \mid A \rightarrow BC \in grammar,$
                    $B \in table[i,k],$
                    $C \in table[k,j]\}$

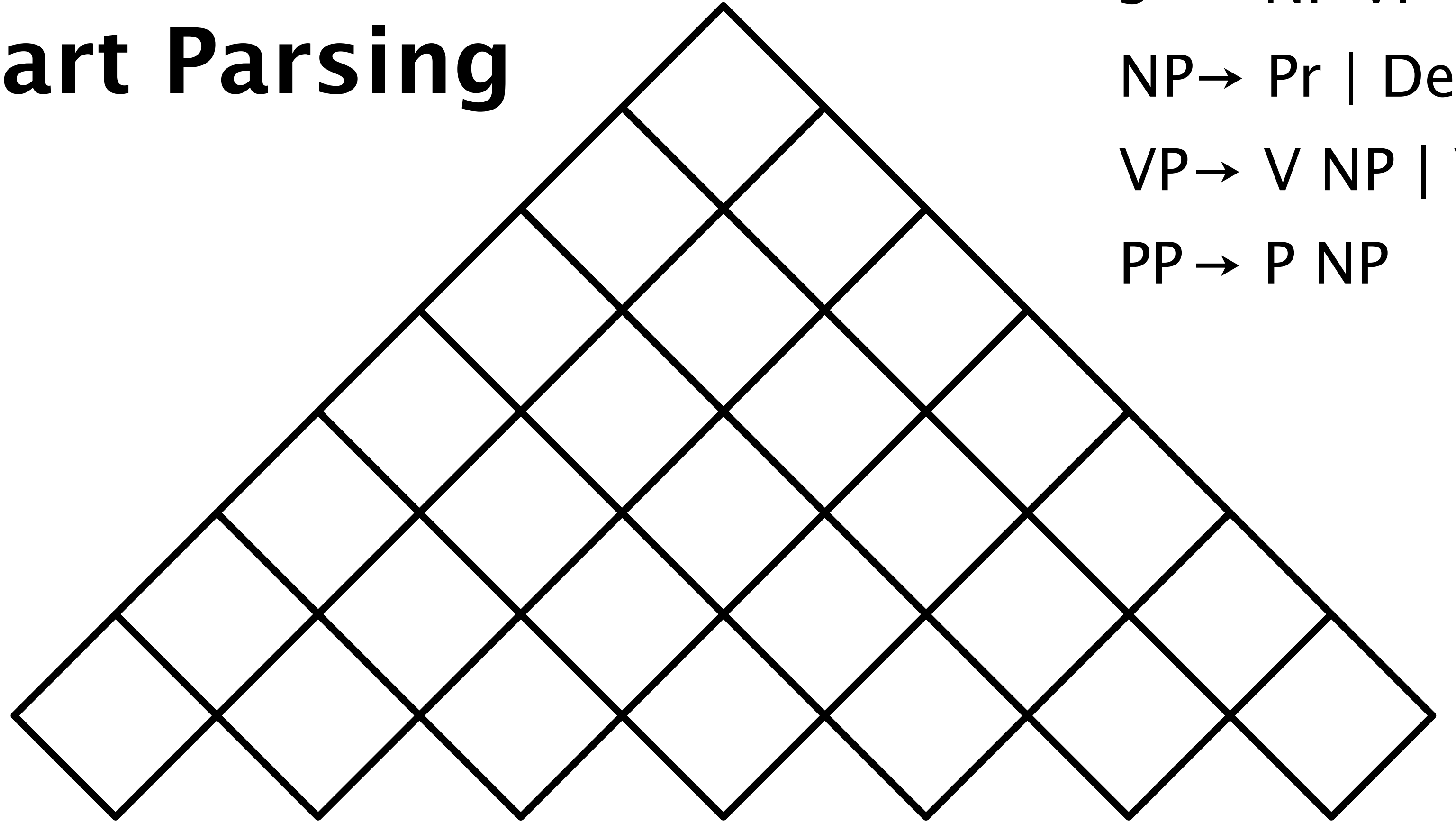
Chart Parsing

S → NP VP

NP→ Pr | Det N | NP PP

VP→ V NP | VP PP

PP→ P NP



I saw her duck with the telescope

CKY

- starting bottom up: search for ever longer constituents and add these higher up in the chart
  - use the fact that longer constituents are composed of shorter ones
- if in the end, the start symbol convers the full sentence, then the grammar supports the sentence
- need additional book-keeping to be able to retrieve the parse trees

limits of CFGs for syntax in NLP

- language contains structures which go beyond CFGs (if you're looking for examples of highly complex language, Swiss German is the language of choice)
  - see separate slide set
- agreement is difficult to model in CFGs
  - e.g.: subject-verb have to agree in person and number
    → solution: unification based grammars
- subcategorization, e.g. arity (number of open semantic slots) of verbs
  - *John found. *John disappeared the ring.

summary

- parse trees capture syntactic structure and help in deducing semantics / meaning
- CFGs are useful to describe the grammar of natural language (by and large)
- efficient computation of parse trees via dynamic programming

applications of CFGs

- most NLP applications profit from syntactic structure (it's, however, not always clear whether automatically derived structure yields a benefit)
- speech recognition grammars that can readily be semantically interpreted (e.g. in command&control applications)
- beyond NLP (e.g. parsing programming languages)

ambiguity

"One morning I shot an elephant in my pajamas."

–"How he got in my pajamas, I don't know."

- ambiguous: yes. equally likely? no.
  - probabilistic models in 2. third of the term

Thank you. Your questions?

timo.baumann@oth-regensburg.de

weiterführende Literatur

- Kozen (1997): Automata and Computability
- Jurafsky and Martin (2009): Kapitel 12,13
- Carstensen et al. (2004): Seiten 79ff, 264–329
- Grewendorf, Hamm, Sternefeld (1989) "Sprachliches Wissen": Seiten 150ff

Lehrziele

- die Studierenden kennen kontextfreie Grammatiken als strukturbildendes Mittel in der Sprachverarbeitung
- die Studierenden erfassen den Vorteil der Strukturbildung durch Syntaxparsing gegenüber flachen Verfahren
- die Studierenden kennen Verfahren zur manuellen Grammatikinferenz
- die Studierenden kennen einen Algorithmus zum Parsen syntaktischer Konstituentengrammatiken und können diesen anwenden