

# Syntax natürlicher Sprachen

## 13: Komplexität formaler und natürlicher Sprachen

A. Wisiorek

Centrum für Informations- und Sprachverarbeitung,  
Ludwig-Maximilians-Universität München

30.01.2024

# Themen der heutigen Vorlesung

- 1 Chomsky-Hierarchie
- 2 *Center-embedding*-Konstruktionen (nicht-regulär)
- 3 *Cross-serial dependencies* (Hinweis auf Nicht-Kontextfreiheit)
- 4 *Garden-path*-Sätze (Hinweis auf probabilistische Sprachverarbeitung)

# 1. Chomsky-Hierarchie

- 1 Chomsky-Hierarchie
- 2 *Center-embedding*-Konstruktionen (nicht-regulär)
- 3 *Cross-serial dependencies* (Hinweis auf Nicht-Kontextfreiheit)
- 4 *Garden-path*-Sätze (Hinweis auf probabilistische Sprachverarbeitung)

- Klassifizierung formaler Grammatiken bzgl. **Regeleinschränkung**  
→ *je stärker eine Grammatik eingeschränkt desto geringer ihre Erzeugungsmächtigkeit*  
→ *desto geringer die Komplexität der erzeugten Sprache*
- 4 Typen von **Typ 0** (rekursiv aufzählbar = ohne Einschränkung) bis **Typ 3** (regulär = am stärksten eingeschränkt)
- klassische Phrasenstrukturgrammatiken: **kontext-frei (Typ 2)**
- einige Syntaxformalismen sind **kontextsensitiv (Typ 1)** (TAG, CCG) bzw. **rekursiv aufzählbar (Typ 0)** (HPSG, LFG)

# Die 4 Typen der Chomsky-Hierarchie

- **rekursiv aufzählbar (Typ 0):**  $\alpha \rightarrow \beta$   
→ *ohne Einschränkung bzgl.  $\alpha, \beta$ ;  $\alpha, \beta \in \text{Alphabet} = \{T, NT\}$*
- **kontext-sensitiv (Typ 1):**  $\alpha \rightarrow \beta, \text{length}(\alpha) \leq \text{length}(\beta)$   
→ *bzw. auch:  $l\alpha r \rightarrow l\beta r$  ( $\alpha \in NT$ ;  $l, r, \beta \in \{T, NT\}$ )*
- **kontext-frei (Typ 2):**  $X \rightarrow \beta$  ( $X \in NT$ ;  $\beta \in \{T, NT\}$ )  
→ *LHS: nur 1 Nicht-Terminal*
- **regulär (Typ 3):**  $X \rightarrow a, X \rightarrow aY$  ( $X, Y \in NT, a \in T$ )  
→ *LHS: nur 1 Nicht-Terminal*  
→ *RHS: 0-n Terminale und 0-1 Nicht-Terminale (links oder rechts)*

- **Chomsky:** Kann natürliche Sprache mit regulärer Grammatik (endlichen Automaten) modelliert werden?
- es gibt **nicht-reguläre Phänomene** in natürlicher Sprache
  - z. B. *center-embedding-Rekursion*
  - benötigt *kontextfreie Regel*
- allerdings: die Konstruktionen, die eine natürliche Sprache **nicht-regulär** machen, sind **für den Menschen schwer zu parsen**

- mathematisch-formal: **Großteil der Syntax menschlicher Sprache** mit **regulärer Grammatik** modellierbar
- aber: **kontextfreie Grammatiken** geben **beschreibungsadäquatere Struktur**
  - *linguistisch adäquates Modell*
  - *wichtig für weitere Verarbeitung (semantische Analyse)*
- **einige Sprachen** enthalten **Konstruktionen**, die sie **kontext-sensitiv** machen: *cross-serial dependencies* im Schweizerdeutschen
- **Hinweise**, dass auch **menschliches Parsing Wahrscheinlichkeiten** berücksichtigt: *garden-path-Sätze*

## 2. *Center-embedding*-Konstruktionen (nicht-regulär)

- 1 Chomsky-Hierarchie
- 2 *Center-embedding*-Konstruktionen (nicht-regulär)
- 3 *Cross-serial dependencies* (Hinweis auf Nicht-Kontextfreiheit)
- 4 *Garden-path*-Sätze (Hinweis auf probabilistische Sprachverarbeitung)



- **center-embedding-Rekursion:**  $X \rightarrow \alpha X \beta$   
→ **rekursive Regel:** Nichtterminal erweitert zu selbem Nichtterminal, umgeben von Strings
- **center-embedding-Regel ist nicht-regulär:**  
→ reguläre Grammatik: **nur links- oder rechtslineare Regeln:**  $X \rightarrow Xa$  oder  $X \rightarrow aX$   
→ entsprechende **Einbettung nicht möglich**

# Beispiel einer *center-embedding*-Konstruktion

- **Rekursive Einbettung von Relativsätzen** als nominales Attribut:

*(Das Kind,)*

*das den Hund<sub>N1</sub>, der die Katze<sub>N2</sub>, die den Vogel<sub>N3</sub> jagt<sub>V3</sub>, anbellt<sub>V2</sub>, ausführt<sub>V1</sub>,*

*...*

- Schema:  $N_1(N_2(N_3V_3)V_2)V_1$

- Regeln:

***RELS***  $\rightarrow$  R NP V

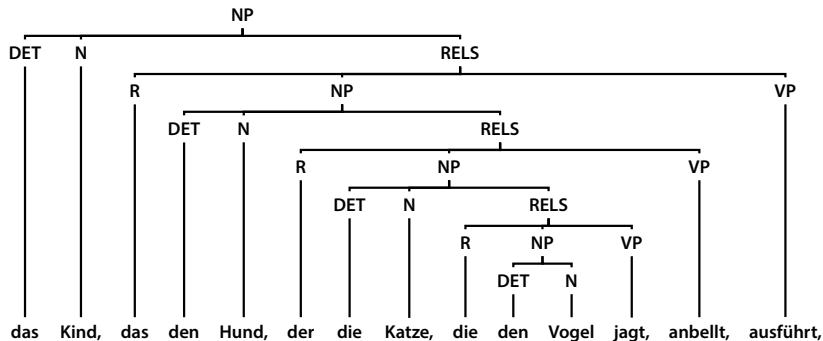
***NP***  $\rightarrow$  DET N ***RELS***

- Ableitungen:

***RELS***  $\rightarrow$  R DET N ***RELS*** V

***RELS***  $\rightarrow$  R DET N R DET N ***RELS*** V V usw.

# Syntaxbaum Center-embedding (Relativsatz)



- psycholinguistische Experimente zeigen **Beschränkung in der Verarbeitung solcher Strukturen durch die menschliche Sprachverarbeitung** aufgrund von *memory limitations*  
→ *mehrfach verschachtelte center-embedding-Strukturen sind nur bis zu einer begrenzten Tiefe verarbeitbar*
- **Korpus-Beispiel für center-embedding der Tiefe 3:**  
[M Er ... war allen Gefahren ...  
    [C-1 welche ein jeder,  
        [C-2 der diese wilde Gegend zu jener Zeit,  
            [C-3 als diese Geschichte dort spielte,]  
            durchstreifte,]  
        gewärtig sein mußte,]  
gewachsen ]  
(vgl. Karlsson 2007, Constraints on multiple center-embedding of clauses, Journal of Linguistics 43/2, 365-392.  
<http://www.ling.helsinki.fi/~fkarlssso/ceb5.pdf>)

### 3. *Cross-serial dependencies* (Hinweis auf Nicht-Kontextfreiheit)

- 1 Chomsky-Hierarchie
- 2 *Center-embedding*-Konstruktionen (nicht-regulär)
- 3 *Cross-serial dependencies* (Hinweis auf Nicht-Kontextfreiheit)
- 4 *Garden-path*-Sätze (Hinweis auf probabilistische Sprachverarbeitung)

- einige Sprachen, z. B. das **Schweizerdeutsche**, besitzen eine **Konstruktion, die nicht mit kontextfreien Grammatikmodellen darstellbar ist**  
→ *cross-serial dependencies, d. h. Dependenzrelationen mit überkreuzenden Kanten:*  
 $N_1 N_2 V_1 V_2$   
 $N_1 N_2 N_3 V_1 V_2 V_3$
- Wörter bzw. Teilkonstituenten sind **seriell überkreuzend angeordnet**

Swiss-German:

...mer    em Hans    es huss    hälfed    aastriiche



English:

...we    helped    Hans    paint    the house



**Abbildung:** Cross-serial dependencies (by Christian Nassif-Haynes - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=28304322>)

# Argument für Nicht-Kontextfreiheit des Schweizerdeutschen

- **Anzahl von Verben mit Dativ-Komplement** muss übereinstimmen mit **Anzahl von Dativ-Komplementen**
- ebenso für Akkusativ-Komplemente
- theoretisch **unbegrenzte Anzahl** solcher *cross-serial dependencies* pro Satz
- **solche Sprachen** enthalten  $L' = a^m b^n c^m d^n$
- die **Sprache  $L'$  ist aber nicht-kontextfrei**  
→ *Nachweis über Pumping Lemma für kontextfreie Sprachen*



## 4. *Garden-path*-Sätze (Hinweis auf probabilistische Sprachverarbeitung)

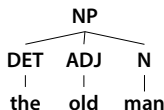
- 1 Chomsky-Hierarchie
- 2 *Center-embedding*-Konstruktionen (nicht-regulär)
- 3 *Cross-serial dependencies* (Hinweis auf Nicht-Kontextfreiheit)
- 4 *Garden-path*-Sätze (Hinweis auf probabilistische Sprachverarbeitung)

- Psycholinguistik: **Parser als Modell menschlicher Sprachverarbeitung**
- **Vergleich mit statistischen Sprachmodellen** gibt Hinweis, dass auch **menschliches Parsing Wahrscheinlichkeiten berücksichtigt**  
→ *Disambiguierung über statistische Informationen*

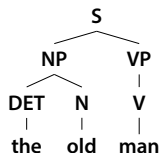
- Beispiel: ***garden-path*-Sätze** = Sätze mit **temporärer Ambiguität**
  - **Gesamter Satz: unambig**, nur eine Ableitung
  - **Teil des Satzes: ambig**, eine strukturelle **Lesart** wird (offensichtlich) **von der menschlichen Sprachverarbeitung bevorzugt**
  - **aber: nicht-präferierte Lesart** für den Teil ist **die für die Ableitung des Satzes korrekte**
- Beobachtung: **wahrscheinlichste Ableitung wird verfolgt**, bis sie fehlschlägt und **Backtracking (Reanalyse)** notwendig ist

# Beispiel: *garden-path*-Satz

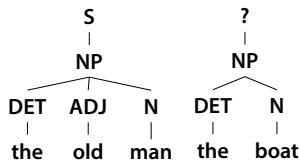
*The old man the boat.*



VS.



- $P(\text{man}|\text{N}) > P(\text{man}|\text{V}), P(\text{old}|\text{ADJ}) > P(\text{old}|\text{N})$



VS.

