

Syntax natürlicher Sprachen

11: Grammatikinduktion

A. Wisiorek

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

13.01.2026

Themen der heutigen Vorlesung

1 Induzierte PCFG-Modelle

- *Grammar Induction* aus Treebank
- Normalisierung und *Parent-Annotation*
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based* PCFGs (*Parent Annotation*)

1. Induzierte PCFG-Modelle

1 Induzierte PCFG-Modelle

- *Grammar Induction* aus Treebank
- Normalisierung und *Parent-Annotation*
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based* PCFGs (*Parent Annotation*)

Data-driven vs. Grammar-driven Models

- Grammatikentwicklung (*grammar writing*) ist aufwendig
 - Grammatiken mit von Experten geschriebenen Regeln mit hoher Abdeckung
- Alternative: Induktion von Grammatikregeln aus Korpora
 - *empirisches Syntaxmodell*
 - Berücksichtigung *relativer Häufigkeiten der Regeln* ⇒ PCFG
 - als *statistisches Modell*: direkte Verwendung zur *Disambiguierung*

1.1. Grammar Induction aus Treebank

1 Induzierte PCFG-Modelle

- **Grammar Induction aus Treebank**
- Normalisierung und *Parent-Annotation*
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

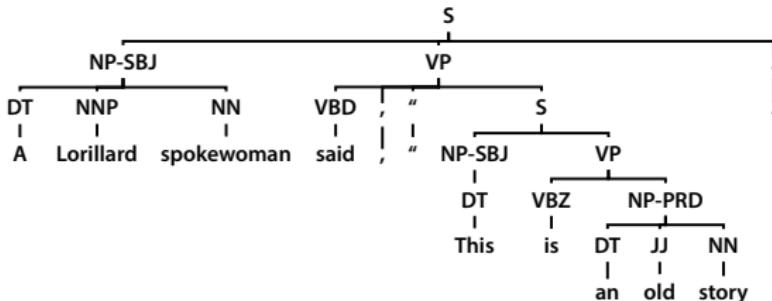
3 history-based PCFGs (*Parent Annotation*)

- Treebank als **implizite Grammatik**
 - jeder **Teilbaum der Tiefe 1 als implizite CFG-Regel**
 - **Expansion eines Nonterminals**
- Extraktion von **CFG-Regeln** aus den Ableitungen der Treebank
- Frequenzbestimmung der Regeln und Berechnung
Regelwahrscheinlichkeiten über **relative Häufigkeiten** (\Rightarrow PCFG)
 - **Gewichtung insbesondere bei induzierter Grammatik notwendig:**
viele Regeln \Rightarrow hohe Ambiguität
- Anwendung von **Smoothing** und **Normalisierung**

- **Form der induzierten Grammatik** hängt stark vom **Annotationsschema** der dem Training des Modells zugrundeliegenden Treebank ab:
 - flache Grammatik = viele Regel-types:
 - Penn-Treebank: 1 Mill. Worttokens, 1 Mill. nicht-lexikalische Regel-tokens, 17.500 Regel-types
 - z. B. jedes PP-Adjunkt mit eigener Regel:
$$VP \rightarrow V PP, VP \rightarrow V PP PP, VP \rightarrow V PP PP PP \text{ usw.}$$
 - tiefere Bäume: mehr Nonterminale, weniger Regel-types:
 - z. B. X-Bar:
$$VP \rightarrow V', V' \rightarrow V' PP, V' \rightarrow V$$

Extrahierte Regeln aus Penn-Treebank-Baum

```
NP-SBJ -> DT NNP NN [0.5]
DT -> 'A' [0.333333]
NNP -> 'Lorillard' [1.0]
NN -> 'spokewoman' [0.5]
VP -> VBD , `` S [0.5]
VBD -> 'said' [1.0]
, -> ',' [1.0]
`` -> '''' [1.0]
S -> NP-SBJ VP [1.0]
NP-SBJ -> DT [0.5]
DT -> 'This' [0.333333]
VP -> VBZ NP-PRD [0.5]
VBZ -> 'is' [1.0]
NP-PRD -> DT JJ NN [1.0]
DT -> 'an' [0.333333]
JJ -> 'old' [1.0]
NN -> 'story' [0.5]
. -> '.' [1.0]
```



- basiert auf **aus Treebanks extrahierten PCFG-Modellen**
→ <https://nlp.stanford.edu/software/lex-parser.shtml>
- Trainingskorpus des englischen Modells (`englishPCFG.ser.gz`):
Penn Treebank
- Trainingskorpus des deutschen Modells (`germanPCFG.ser.gz`):
NEGRA Korpus

1.2. Normalisierung und *Parent-Annotation*

1 Induzierte PCFG-Modelle

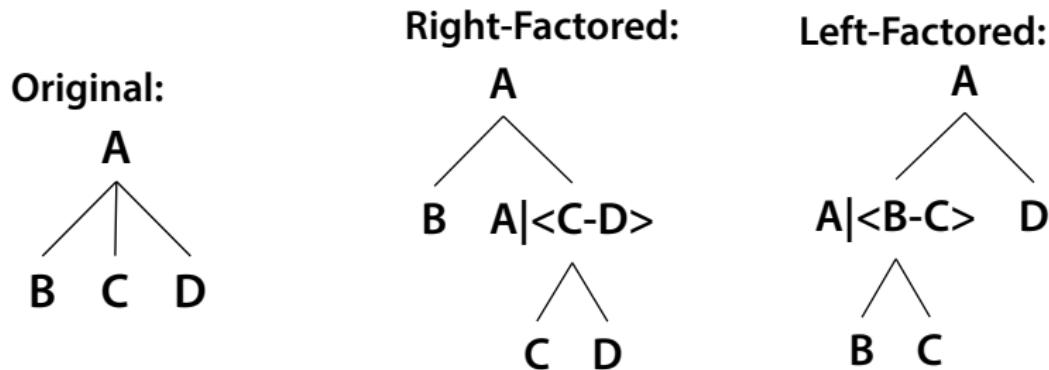
- *Grammar Induction* aus Treebank
- **Normalisierung und *Parent-Annotation***
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based* PCFGs (*Parent Annotation*)

Normalisierung durch Chomsky-Normalform

- Einschränkung der Form von CFG-Regeln:
⇒ RHS: 2 Nichtterminale oder 1 Terminal: $A \rightarrow BC, A \rightarrow a$
- Binärbäume (bis Präterminalknoten, dort: unäre Bäume)
- jede CFG kann in CNF umgewandelt werden:
 $A \rightarrow BCD \Rightarrow A \rightarrow BX, X \rightarrow CD$ (**Right-Factored**)
 $A \rightarrow BCD \Rightarrow A \rightarrow XD, X \rightarrow BC$ (**Left-Factored**)



Anwendungsgebiete Chomsky-Normalform

- notwendig für **CYK-Chart-Parsing**
- zur **Reduktion von extrahierten Grammatikregeln** aus flach annotiertem Korpus:

- $VP \rightarrow V PP$

- $VP \rightarrow V PP PP$

- $VP \rightarrow V PP PP PP$ usw.

- mit **Chomsky-adjunction** ($A \rightarrow A B$):

- $VP \rightarrow V PP$

- $VP \rightarrow VP PP$

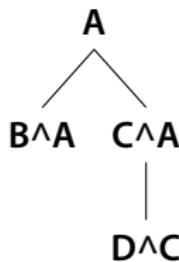
Parent Annotation

- Kategorie des **Mutterknoten** in **Kategoriensymbol** aufnehmen
- Modellierung von **Kontext** → *history-based PCFGs*
- ergibt anderes PCFG-Modell: **mehr Nichtterminale, andere Gewichtung**

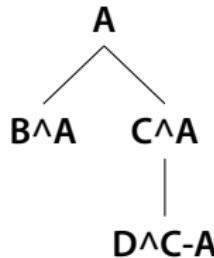
Original:



Parent Annotation:



+Grandparent Annot.:



1.3. Evaluation von PCFG-Modellen

1 Induzierte PCFG-Modelle

- *Grammar Induction* aus Treebank
- Normalisierung und *Parent-Annotation*
- **Evaluation von PCFG-Modellen**
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based* PCFGs (*Parent Annotation*)

- Messen der **Güte von Grammatikmodellen/Parsern** durch Parsen von Sätzen einer **Testmenge**
 - **Teilmenge einer hand-annotierten Treebank** = **gold-standard-Ableitungen**, z. B. von *Penn-Treebank*
- **PARSEVAL-Maße** (Black et al. 1991): Übereinstimmung von Konstituenten in den Ableitungen von **geparsten Daten** (**Ableitungshypothese H**) mit denen der **Test-Daten** (**Referenz-Ableitung R**)
 - **Konstituente ist korrekt** wenn **Übereinstimmung in Nichtterminal-Symbol und Spanne (gleicher Start- und Endpunkt)**

- **Recall** =
$$\frac{\text{(Anzahl von korrekten Konstituenten in Hypothese)}}{\text{(Anzahl von Konstituenten in Referenz-Ableitung)}}$$
- **Precision** =
$$\frac{\text{(Anzahl von korrekten Konstituenten in Hypothese)}}{\text{(Anzahl von allen Konstituenten in Hypothese)}}$$
 - Hypothese: (A) (B C D)
 - Referenz: (A) (B) (C) (D)
 - Recall = 1/4; Precision: 1/2
- **cross-brackets**: Anzahl an Konstituenten mit ((A B) C) in Ableitungshypothese aber (A (B C)) in Referenz-Ableitung
- **moderne Parser**: ca. **90% Precision und Recall**, ca. 1% **cross-brackets**-Konstituenten (trainiert und getestet mit Penn-Treebank)

1.4. Unabhängigkeitsannahmen

1 Induzierte PCFG-Modelle

- *Grammar Induction* aus Treebank
- Normalisierung und *Parent-Annotation*
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based* PCFGs (*Parent Annotation*)

2 Unabhängigkeitsannahmen von PCFGs

- ① **Annahme Unabhängigkeit von lexikalischem Material**
→ Wahrscheinlichkeiten von Teilbäumen sind unabhängig von Terminalen
 - ② **Annahme Unabhängigkeit von Kontext**
→ Wahrscheinlichkeiten von Teilbäumen sind unabhängig von Elternknoten
- Zurücknahme von Unabhängigkeitsannahmen:
 - ⇒ beschreibungsadäquatere Syntaxmodelle
 - ⇒ Berücksichtigung linguistischer Abhängigkeiten

- ① Berücksichtigung lexikalischer Abhängigkeiten:
 - ⇒ *lexikalierte PCFGs*
 - ⇒ Auflösung lexikalischer Ambiguität
- ② Berücksichtigung struktureller Abhängigkeiten zwischen Regeln:
 - ⇒ *history-based PCFGs*
 - ⇒ Auflösung kontextabhängiger struktureller Ambiguität

2. Lexikalisierte PCFGs (Kopfannotation)

1 Induzierte PCFG-Modelle

- *Grammar Induction* aus Treebank
- Normalisierung und *Parent-Annotation*
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based PCFGs (Parent Annotation)*

- PCFGs basierend auf einfachen CFG-Regeln:
⇒ nur **strukturelle Disambiguierung**
- Probleme mit **lexikalisch determinierter Ambiguität**, z. B. bei **Subkategorisierung** oder **PP-Attachment**
- **statisches Modellierung lexikalischer Abhängigkeiten**
- bekannter lexikalisierte Parser: Collins Parser (Collins, 1999)

Vorgehen Lexikalisierung

- **bottom-up-Annotation** nichtterminaler Kategorien **mit lexikalischer Information** (Kopf-Perkolation): VP (kennt)
- auch Annotation mit **Part-of-Speech-Tag** möglich: NP (er, PRON)

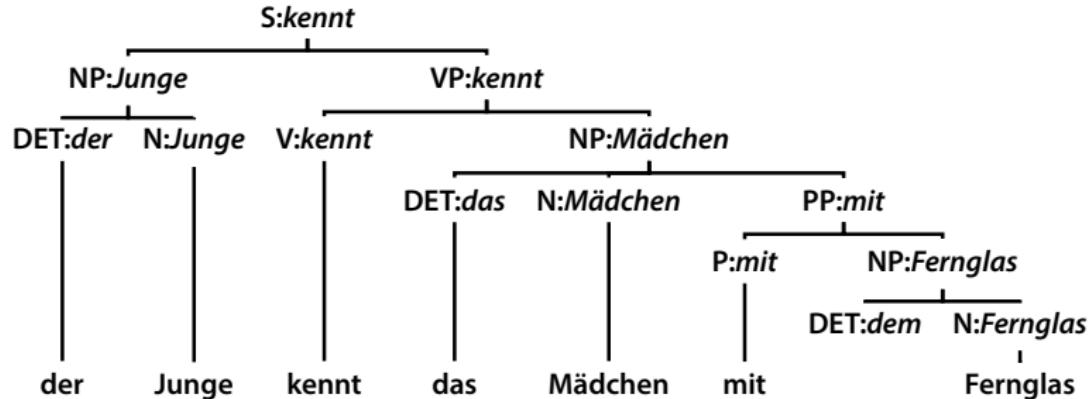
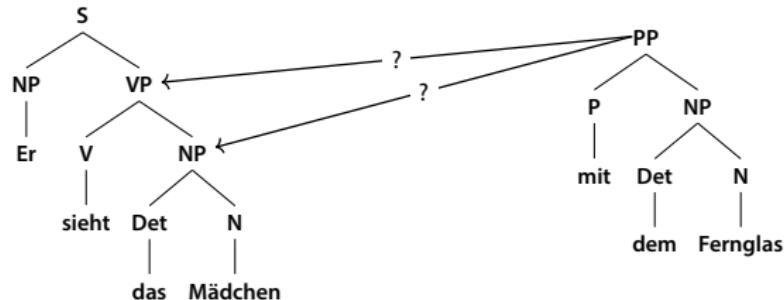


Abbildung: Beispiel für lexikalisierte Phrasenstruktur

PP-Attachment

- strukturelle Ambiguität:
NP- oder VP-Anbindung?
⇒ 2 strukturelle Lesarten:
→ $(VP\ V\ (NP\ N\ PP))$
→ $(VP\ V\ (NP\ N)\ PP)$



- unlexikalisierte PCFG: immer Entscheidung für eine Variante
→ z. B. *englisches Trainingskorpus: NP-Attachment-Frequenz etwas höher*
- häufig: Anbindung lexikalisch konditioniert (**lexikalische Abhängigkeit**):
 - **Bevorzugung von VP-Anbindung:** *Sie stellt die Blumen ins Wasser.*
→ *ins Wasser ist Adverbial*
 - **Bevorzugung von NP-Anbindung:** *Der Junge kennt das Mädchen mit dem Fernglas.*
→ *mit dem Fernglas ist nominales Attribut*

- **statisches Modellierung Subkategorisierung** statt regelbasiert über Subkategorisierungsrahmen
- **transitive Verben:** hohe Wahrscheinlichkeit $P(VP \rightarrow V\ NP)$
 $\rightarrow P(V\ NP / VP, \text{ sehen}) > P(V / VP, \text{ sehen})$
- **intransitive Verben:** hohe Wahrscheinlichkeit $P(VP \rightarrow V)$
 $\rightarrow P(V / VP, \text{ laufen}) > P(V\ NP / VP, \text{ laufen})$

- **Modell wird sehr groß**
 - Grund: *viel mehr Ereignisse durch lexikalierte Regeln*
 - *Regelvervielfachung:*
 $VP(sieht) \rightarrow V(sieht) \; NP(Mädchen)$
 $VP(kennt) \rightarrow V(kennt) \; NP(Mädchen)$
- **umfangreiche Trainingsdaten** notwendig für Parameterabschätzung des Modells
- **neue Abschätzung** für Regelwahrscheinlichkeiten notwendig
 - *MLE-Abschätzung über* $P(\alpha \rightarrow \beta | \alpha) = \frac{\text{count}(\alpha \rightarrow \beta)}{\text{count}(\alpha)}$ ist zu spezifisch
 - *geht meistens gegen 0, da nur sehr wenige Instanzen der lexikalierten Regeln in Trainingskorpus vorhanden*

Backoff wegen *sparse data*

- ***sparse data*-Problem** aufgrund von in Trainingsdaten **ungesehenen Wörtern/Instanzen** (\Rightarrow keine Regel vorhanden)
 \rightarrow Lösung: **Backoff = Verzicht auf Lexikalisierung bei unbekanntem lexikalischen Kopf**
- dazu notwendig: **Smoothing** (Glättung der Regelwahrscheinlichkeiten)
 \rightarrow **Reservierung von Wahrscheinlichkeitsmasse für Regeln bei Backoff bei ungesehenen Köpfen**
 \rightarrow Zuordnung von Wahrscheinlichkeit für Regel mit **ungesehenem Kopf**
 \rightarrow z. B. **Laplace-Smoothing**: zu jeder Häufigkeit im Korpus: **Wert addieren (1 = Add-One-Smoothing)** \Rightarrow Backoff-Regel: $P > 0$
- **Backoff bei Collins Parser**: unbekannte Köpfe aus Testmenge und aus Trainingsmenge mit Frequenz < 6 werden mit UNKNOWN ersetzt

3. *history-based PCFGs (Parent Annotation)*

1 Induzierte PCFG-Modelle

- *Grammar Induction* aus Treebank
- Normalisierung und *Parent-Annotation*
- Evaluation von PCFG-Modellen
- Unabhängigkeitsannahmen

2 Lexikalisierte PCFGs (Kopfannotation)

3 *history-based PCFGs (Parent Annotation)*

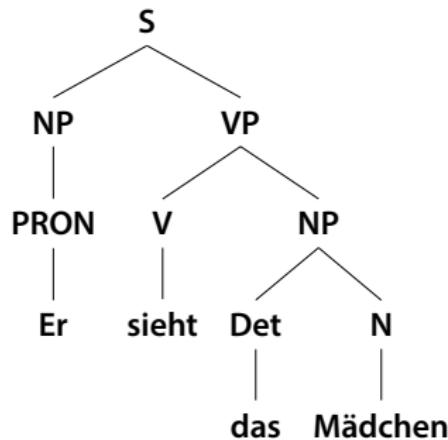
- Berücksichtigung **Abhängigkeit Expansion von Kontext**
 - **Regelauswahl abhängig von vorheriger Regelauswahl**
 - Wahrscheinlichkeit einer Expansion ist abhängig von der **Position im Strukturbaum**
- z. B. unterschiedliche Expansionswahrscheinlichkeiten für NPs in Subjekt- bzw. Objektposition
 - **Subjekt-NP (S-dominiert) erweitert wahrscheinlicher zu Pronomen als Objekt-NP (VP-dominiert)**
 - $P(NP \rightarrow PRON/S) > P(NP \rightarrow PRON/VP)$
 - $P(PRON/NP, S) > P(PRON/NP, VP)$

- Grund = **Informationsstruktur**

→ **Subjekt** typischerweise **Topik** = **bekannte Information**, die durch **Pronomen** ausgedrückt wird

	Pronomen	Nicht-Pronomen
Subjekt	91%	9%
Objekt	34%	66%

Abbildung: Verteilung der Form von Subjekt und Objekt in englischem Korpus (nach Francis et al., 1999, vgl. SLP2, 502)



erwünschte Regelgewichtung Subjekt (S-dominiert):

NP → PRON **0.91**

NP → DET N **0.09**

erwünschte Regelgewichtung Objekt (VP-dominiert):

NP → PRON **0.34**

NP → DET N **0.66**

normale PCFG (keine Differenzierung, Daten aus Korpus):

NP → PRON **0.25**

NP → DET N **0.28**

Lösung: Splitting NP-Kategoriensymbol (*parent annotation*):

NP^{AS} → PRON **0.91**

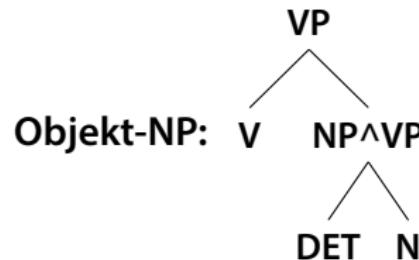
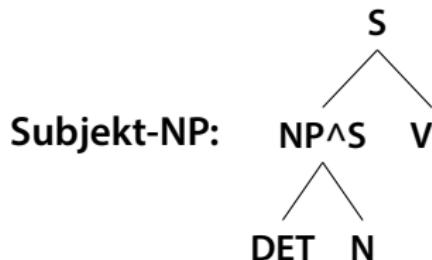
NP^{AS} → DET N **0.09**

NP^{AVP} → PRON **0.34**

NP^{AVP} → DET N **0.66**

Vorgehen

- Annotation nichtterminaler Kategorien mit Kategorie des Mutterknotens (= *history*)
 - **Parent Annotation** (Splitting von Nicht-Terminalen)
 - Subjekt-NP: $NP \wedge S$
 - Objekt-NP: $NP \wedge VP$



Probleme von *history-based* PCFGs

- **ähnlich** wie bei **Lexikalisierung**, aber weniger stark ausgeprägte Regelvervielfachung durch *parent annotation*
→ *sparse data: unbekannte Vorgängerkategorie*
- kleinere Regelmenge durch **selektive parent annotation**
→ *nur Splitten, wenn accuracy erhöht wird*