Lösungen für Aufgabenblatt 9 — 19.12.2019

Einführung in die Computerlinguistik (WS19/20) — Abgabe bis 11.01.2020

Aufgabe 1) Normalform von Grammatiken **Punkte: 3**

Gegeben ist eine probabilistische kontextfreie Grammatik, bestehend aus folgenden Regeln (Terminale in *kursiver Schrift*):

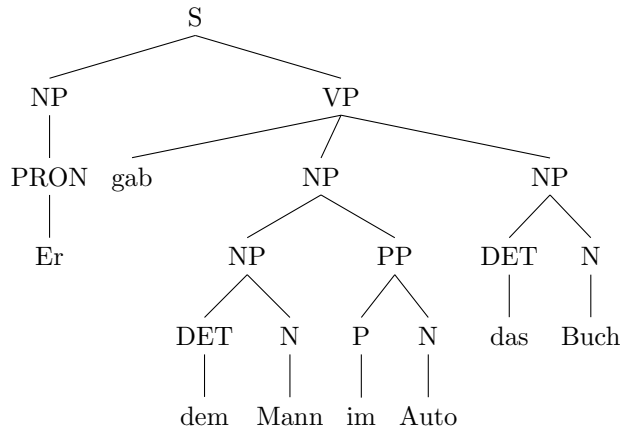
S → NP VP [1.0]
NP → PRON [0.4]
NP → DET N [0.3]
NP → NP PP [0.3]
VP → gab NP NP [0.4]
VP → gab NP PP NP [0.6]
PP → P N [1.0]
PRON → *Er* [1.0]
DET → *dem, das* [0.5, 0.5]
N → *Mann, Auto, Buch* [0.4, 0.4, 0.2]
P → *im* [1.0]

Bringen Sie die obige Grammatik in Chomsky Normalform.

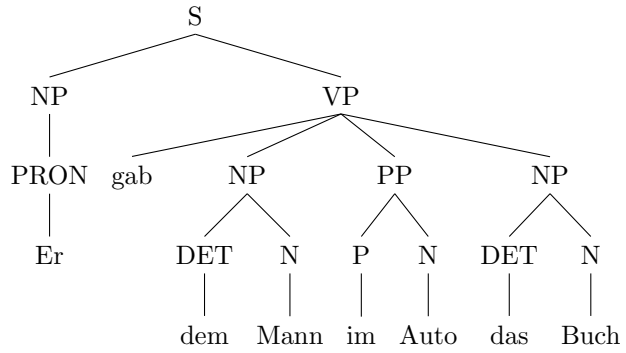
Lösung zu 1)

Die Grammatik soll die folgenden beiden Lesarten widerspiegeln (nicht gefragt, nur zur Extrainfo):

1.



2.



Umwandlung:

1. unit productions entfernen:
   NP $\rightarrow$ *Er* [0.4]

2. Ersetze Terminale, die in "gemischten" Regeln vorkommen:
   V $\rightarrow$ *gab* [1.0]

3. längere Verkettungen durch Einführen zusätzlicher Nichtterminale beheben[1]:

   a) VP $\rightarrow$ X_1 NP [0.4]
      X_1 $\rightarrow$ V NP [1.0]

   b) *VP $\rightarrow$ X_2 PP NP [0.6]
      X_2 $\rightarrow$ V NP [1.0]

   c) VP $\rightarrow$ X_3 NP [0.6]
      X_3 $\rightarrow$ X_2 PP [1.0]

Finale Grammatik:

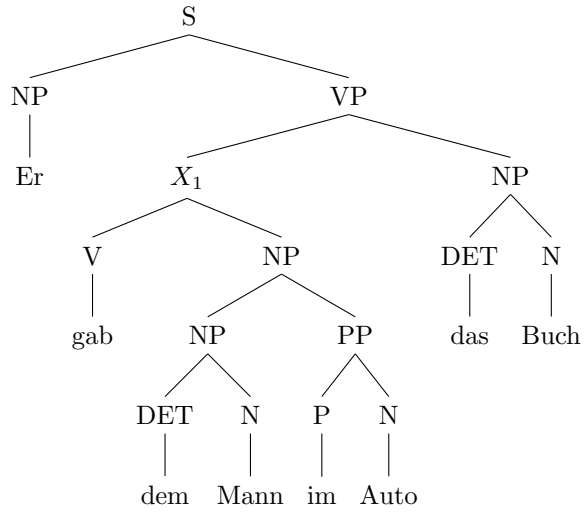
S $\rightarrow$ NP VP [1.0]
NP $\rightarrow$ DET N [0.3]
NP $\rightarrow$ NP PP [0.3]
NP $\rightarrow$ *Er* [0.4]
VP $\rightarrow$ X_1 NP [0.4]
VP $\rightarrow$ X_3 NP [0.6]
X_1 $\rightarrow$ V NP [1.0]
X_2 $\rightarrow$ V NP [1.0]

[1] Asterisk "*" bedeutet, dass die Regel zwar schon angepasst wurde, allerdings immer noch nonkonform ist.

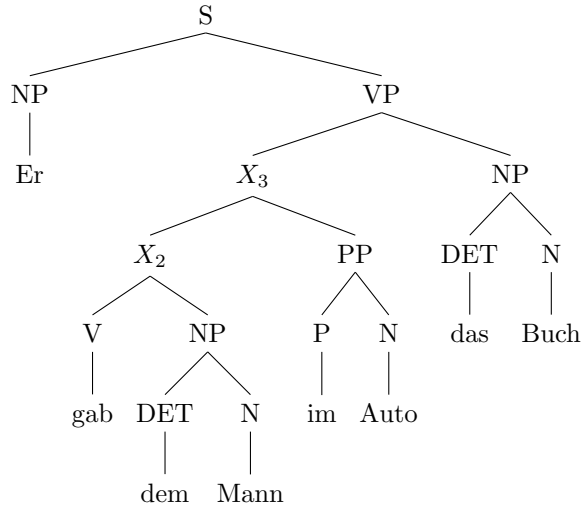
$X_3 \rightarrow X_2$ PP [1.0]
PP $\rightarrow$ P N [1.0]
DET $\rightarrow$ *dem, das* [0.5, 0.5]
N $\rightarrow$ *Mann, Auto, Buch* [0.4, 0.4, 0.2]
P $\rightarrow$ *im* [1.0]
V $\rightarrow$ *gab* [1.0]

Die beiden möglichen Bäume nach CNF-Konvertierung sollte also so aussehen (nicht gefragt, nur zur Extrainfo angegeben):

1.



2.



Aufgabe 2) Probabilistischer CKY-Algorithmus **Punkte: 7**

Gegeben ist eine PCFG, in Chomsky Normalform, bestehend aus folgenden Regeln:

S → PRON VP [1.0]
NP → DET N [0.7]
NP → NP NP [0.3]
VP → V VP′ [1.0]
VP′ → NP PART [0.4]
VP′ → NP VP′ [0.6]
PRON → *Er* [1.0]
V → *liest* [1.0]
DET → *das, seiner* [0.6, 0.4]
N → *Buch, Schwester* [0.5, 0.5]
PART → *vor* [1.0]

Gegeben sei auch der folgende Satz S:

```
Er liest das Buch seiner Schwester vor
```

1. Parsen Sie S mittels des probabilistischen CYK Algorithmus. Berechnen Sie nur die wahrscheinlichste Lesart (4 Punkte).

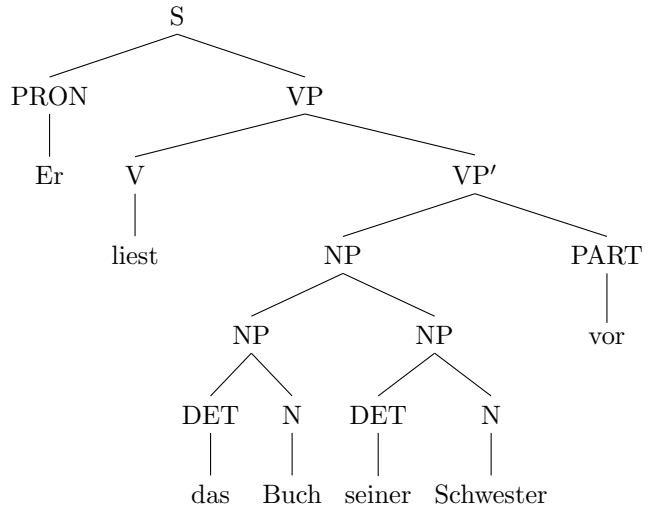
2. Welcher Lesart von S entspricht die wahrscheinlichste Lesart, die Sie berechnet haben? Bitte zeichnen Sie den entsprechenden Baum. (2 Punkte)

3. In welcher Zelle spiegelt sich die Ambiguität des Satzes wider? Wie müssten Sie vorgehen, um alle Lesarten sowie deren Wahrscheinlichkeiten zu berechnen? (1 Punkt)

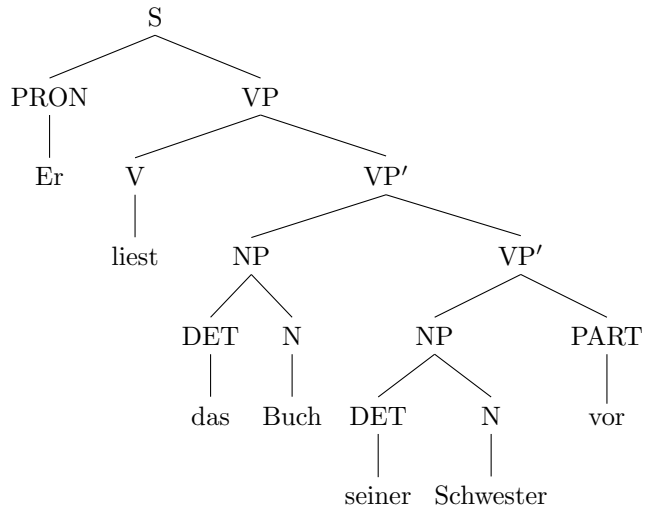
Lösung zu 2)

Die Grammatik soll zwei Lesarten widerspiegeln:

1.


S
PRON
VP
Er
V
VP′
liest
NP
PART
NP
NP
vor
DET
N
DET
N
das
Buch
seiner
Schwester

2.


S
PRON
VP
Er
V
VP′
liest
NP
VP′
DET
N
NP
PART
das
Buch
DET
N
vor
seiner
Schwester

Parsing:

0 *Er* 1	*liest* 2	*das* 3	*Buch* 4	*seiner* 5	*Schwester* 6	*vor* 7
PRON (1.0) [0, 1]	- [0, 2]	- [0, 3]	- [0, 4]	- [0, 5]	- [0, 6]	S ($1.0 \cdot 1.0 \cdot 0.0106 = 0.0071$) [0, 7]
	V (1.0) [1, 2]	- [1, 3]	- [1, 4]	- [1, 5]	- [1, 6]	VP ($0.0071 \cdot 1.0 \cdot 1.0 = 0.0071$) [1, 7]
		DET (0.6) [2, 3]	NP ($0.5 \cdot 0.6 \cdot 0.7 = 0.21$) [2, 4]	- [2, 5]	NP ($0.3 \cdot 0.14 \cdot 0.21 = 0.0088$) [2, 6]	VP'_1 ($0.0088 \cdot 1.0 \cdot 0.4 = 0.0035$), VP'_2 ($0.21 \cdot 0.056 \cdot 0.6 =$ **0.0071**) [2, 7]
			N (0.5) [3, 4]	- [3, 5]	- [3, 6]	- [3, 7]
				DET (0.4) [4, 5]	NP ($0.4 \cdot 0.5 \cdot 0.7 = 0.14$) [4, 6]	VP′ ($0.4 \cdot 0.14 \cdot 1.0 = 0.056$) [4, 7]
					N (0.5) [5, 6]	- [5, 7]
						PART (1.0) [6, 7]

Die wahrscheinlichste Lesart ist die indirekte Objekt Lesart (zweiter Baum siehe oben).

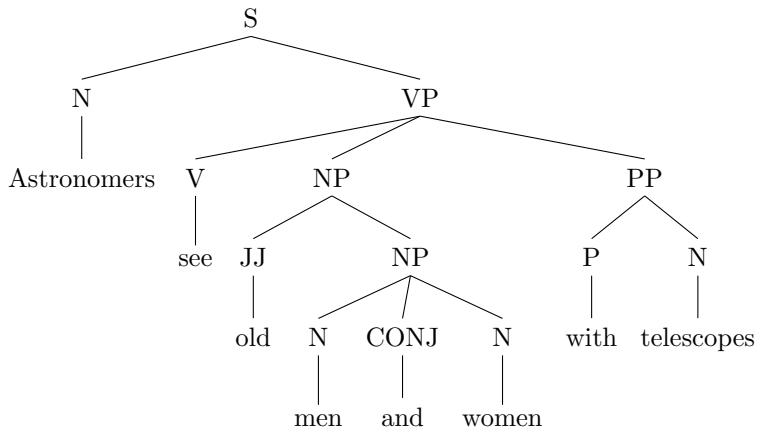
Hier spiegeln die Farben die Ambiguität in Zelle [1,7] wieder. Wollte man alle Lesarten berechnen, so dürfte man in Zelle [1,7] nicht maximieren, sondern müsste mit beiden VP' weiterrechnen.

Aufgabe 3) Parsing Evaluation **Punkte: 6**

Gegeben ist der Satz:

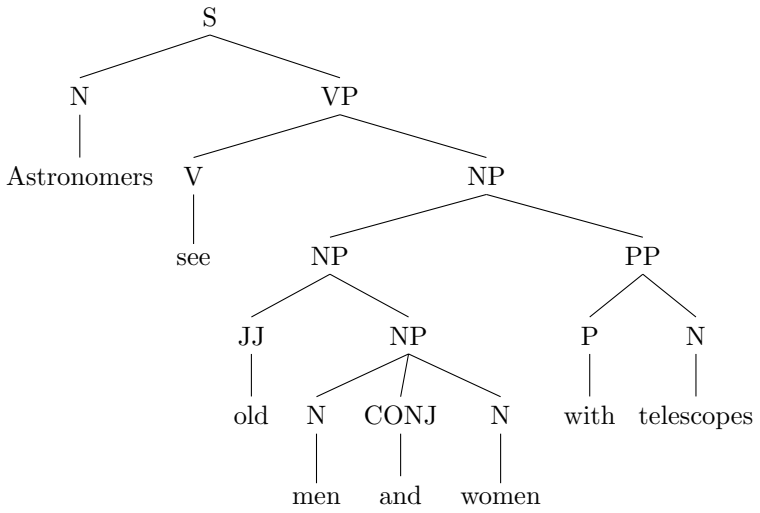
```
Astronomers see old men and women with telescopes
```

Der Satz ist in einem Korpus mit folgendem Gold-Parse (GP) annotiert:

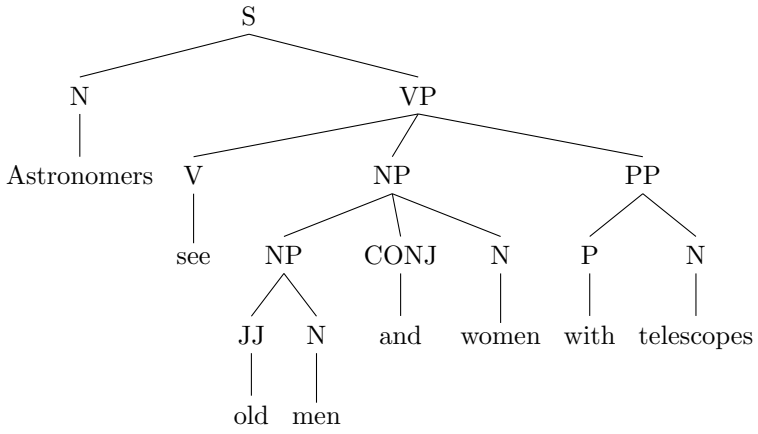


Sie haben zwei Parser, von denen einer den untigen Parsebaum SP_1 und der andere den Parsebaum SP_2 ausgibt:

1. SP1:]



2. SP2:



a) Precision / Recall *Punkte: 4*

Berechnen Sie **labelled Precision** und **labelled Recall** für SP1 und SP2.

Lösung zu 3a)

Span	Gold-Parse	System Parse 1	System Parse 2
(0:8)	S	✓	✓
(1:8)	VP	✓	✓
(2:6)	NP	✓	✓
(2:8)		NP	
(3:6)	NP	✓	
(2:4)			NP
(6:8)	PP	✓	✓

Precision_{SP1}: $\frac{5}{6} = 0.83$
Recall_{SP1}: $\frac{5}{5} = 1.00$

Precision_{SP2}: $\frac{4}{5} = 0.80$
Recall_{SP2}: $\frac{4}{5} = 0.80$

b) Interpretation *Punkte: 2*

Obwohl beide Parses jeweils in nur einer attachment Entscheidung vom Goldparse abweichen, erhalten Sie unterschiedliche Werte für Precision_{SP1} und Precision_{SP2} bzw. Recall_{SP1} und Recall_{SP2}. Begründen Sie dies anhand der verwendeten Metriken bzw Grammatik.

Lösung zu 3b)
Was Precision betrifft, so kann diese immer noch sehr hoch sein, so lange man "nur" zusätzliche

Nichtterminale “zwischenschaltet”, aber immer noch alle Gold-Verzweigungen abbildet. Dies ist bei SP1 der Fall, während in SP eine NP “nach links wandert”.

Dementsprechend ist der Recall für SP1 vollständig, für SP2 fehlt aber die veränderte NP.

Man sieht auch, dass die Bestrafung/Belohnung grammatikabhängig ist. Hätte man das Verb-attachment der PP mit Chomsky/Normalform dargestellt, anstatt flach mit V NP PP, so würde sich auch die Bestrafung des veränderten PP-attachment ändern. Ebenso bekommt man in der gegebenen Grammatik keine Belohnung für das Erkennung von *Astronomers* als Subjekt, da direkt als Präterminal kodiert.

Aufgabe 4) Wortsemantik **Punkte: 4**

Sie sollen die Bedeutungen des deutschen Nomens *Gericht* bestimmen. Hierzu bekommen Sie Korpusvorkommen aus dem Deutschen Kernkorpus. Gehen Sie auf `www.dwds.de/r`. Hier können Sie nach der genauen Wortform *Gericht* mittels *@Gericht* suchen. Schauen Sie sich die ersten 50-100 Beispiele aus dem DWDS Kernkorpus (2000-2010) an. Am besten stellen Sie die KWIC (Keyword in Context)-Anzeige auf maximal und verändern sonst die Grundeinstellungen nicht. Weitere Details zur Suche finden Sie unter `https://www.dwds.de/d/suche#korpussuche`.

1. Finden Sie die homonymen Kernbedeutungen von *Gericht* und geben Sie Evidenz dafür an, dass diese Bedeutungen honomym sind (2 Punkte)

2. Welche polysemen Bedeutungen finden Sie (den homonymen Kernbedeutungen zugeordnet) noch? Belegen Sie diese mit den Beispielen aus dem Korpus —geben Sie die Beispiele vollständig an, nicht nur die Nummern, um die Korrektur zu erleichtern. (2 Punkte).

Lösung zu 4)

1. Man sollte zwei Homonyme finden (Essen bzw rechtliche Institution). Evidenz: verschiedene Kollokationen, kein Zeugma möglich.

2. Man findet unter *Gericht* als Instiution mindestens noch:

   - Das Richterkollegium
   - die abstrakte Institution (vor Gericht gestellt werden)
   - Das Gebäude (Diskussion um Anwesendheitspflicht von Richtern im Gebäude)

   mit entsprechenden Belegen. Diese sind durch systematische Polysemien miteinander relationiert.

Aufgabe 5) Bonusaufgaben **Punkte: 7**

1. Fortsetzung der Aufgabe 4 (Lexikalische Semantik). Nehmen Sie an, Sie wollten einen automatischen Algorithmus zur Erkennung der verschiedenen Bedeutungen von *Gericht* schreiben. Dieser könnte zum Beispiel auf einem Naive Bayes Verfahren mit Bag of Words beruhen. Was wären dann die beiden wichtigsten Informationsquellen für die Bestimmung der Bedeutung in einem gegebenen Kontext? Geben Sie jeweils ein Beispiel für die Aussagekraft dieser Informationsquellen. (2 Punkte).

2. Wir haben in der Vorlesung Beispiele dafür gesehen, in denen die Intension über die Zeit gleichbleibt, die Extension aber nicht. So ändert sich die Intension/Denotation des Begriffes *Junggeselle* nicht, wenn ein einziger Junggeselle heiratet und damit die Extension sich ändert. Die Intension von *Bundeskanzler(in)* ändert sich nicht, auch wenn das Amt je nach Wahlausgang von verschiedenen Menschen (Extension) ausgefüllt wird. Nehmen wir nun an, die Extension zweier Lemmata sei gleich, und dies zu *allen* Zeitpunkten. Heißt dies, dass dann auch die Intension der beiden Lemmata (abgesehen von verschiedenen Konnotationen) gleich ist? (2 Punkte)

3. Geben Sie drei verschiedene Ursachen für die Entstehung von Synonymen an. Belegen Sie jede mit einem Beispiel (aus dem Deutschen oder Englischen). Benutzen Sie nicht die Beispiele aus den Vorlesungsfolien. (3 Punkte)

Lösung zu 5)

1. Die Häufigkeit einzelner Bedeutungen (prior) sowie die anderen Worte, die in einem Kontextfenster um *Gericht* herum vorkommen. Der Prior würde uns beim DWDS Korpus sagen, dass die rechtliche Bedeutung häufiger ist als die Essensbedeutung. Kontextworte wie *lecker* würden für die Essensbedeutung sprechen.

2. Nein, dies würde nicht gelten. Einfachstes Gegenbeispiel: zwei Lemmata, die jeweils leere Extension haben (*Einhorn* vs. *Hypogriff*) hätten dennoch nicht die gleiche Intension.

3. a) Fremdspracheneinfluss: *snake, serpent*

   b) politische Lenkung: *Engel, Jahresendfigur*

   c) Regionale Unterschiede: *Sonnabend, Samstag*