






## Einführung in die Computerlinguistik – 7. Übungsblatt

### Aufgabe 7.1 - Sprachsynthese

- a) • *Der Staatsanwalt ist ein dem Gericht gleichgeordnetes Organ der Strafrechtspflege.*  
*Staatsanwalt* muss in die Komponenten *Staat*, *s* und *Anwalt* zerlegt werden, das Fugenelement *s* muss deutlich ausgesprochen werden. Weiterhin in *Staat* muss der Vokal *a* lang ausgesprochen werden, im Unterschied zu den Wort *Stadt*. Auch die Wörter *gleichgeordnetes* und *Strafrechtspflege* haben viele Komponenten, die korrekt prononciert werden müssen. 
- *Die Omikron-Variante dominiert inzwischen das Infektionsgeschehen in den USA.*  
*USA* ist die Abkürzung von United States of America, muss in entsprechend den Buchstaben ausgesprochen und groß geschrieben werden. Der Bindestrich zwischen den Wörtern *Omikron* und *Variante* ist schwer zu erkennen.
- b) Die Schwierigkeit der Sprachsynthese besteht in zusammengesetzter Wörtern, also Komposita. In diesem Satz bereitet das Wort *Waldecke* ein Problem. *Waldecke* kann als *Wal* + *decke* oder *Wald* + *ecke* analysiert werden. Eine Analyse des syntaktischen Kontextes ist nötig, um die richtige Aussprache dieses Wortes zu finden. 
- c) Dieser Satz wird bei der Beantwortung jeder Frage unterschiedlich betont. Bei der ersten Frage wird in dem Satz als Antwort das Wort *Peter* betont. Für den zweiten Frage ist es das Wort *Schild* und für den dritten Frage ist es das Wort *gesehen*. Die Regel der Satzbetonung ist, dass neue Information betont wird und gegebene Information nicht. 
- d) Die meisten Aspekte vom System werden gut behandelt. 
- e) Das System funktioniert gut, es gibt fast keine Aussprachefehler bei deutschen Wörtern. Aber es wurde *Waldecke* falsch als *Wal* + *decke* ausgesprochen. Sätze haben eine vernünftige Intonation und Pausen basierend auf Punkten, Kommas. Es macht auch hoch steigende Intonation am Ende jeder Frage. Aber die Antworten auf alle drei Fragen werden gleich betont. 


### Aufgabe 7.2 - n-Gramme

- a) •  $P(\text{Berlin ist eine Hauptstadt})$   
 $= P(\text{Berlin} \cap \text{ist} \cap \text{eine} \cap \text{Hauptstadt})$   
 $= P(\text{Berlin}) \cdot P(\text{ist}|\text{Berlin}) \cdot P(\text{eine}|\text{ist}) \cdot P(\text{Hauptstadt}|\text{eine})$   
 $= \frac{Fr(\text{Berlin})}{N} \cdot \frac{Fr(\text{Berlin} \cap \text{ist})}{Fr(\text{Berlin})} \cdot \frac{Fr(\text{ist} \cap \text{eine})}{Fr(\text{ist})} \cdot \frac{Fr(\text{eine} \cap \text{Hauptstadt})}{Fr(\text{eine})}$   
 $= \frac{516.000.000}{20.000.000.000} \cdot \frac{7.910.000}{516.000.000} \cdot \frac{982.000.000}{13.760.000.000} \cdot \frac{129.000}{3.430.000.000}$

$$= 1,06 \cdot 10^{-9}$$

- $$\begin{aligned}
 &P(\text{Berlin ist eine Frage}) \\
 &= P(\text{Berlin} \cap \text{ist} \cap \text{eine} \cap \text{Frage}) \\
 &= P(\text{Berlin}) \cdot P(\text{ist}|\text{Berlin}) \cdot P(\text{eine}|\text{ist}) \cdot P(\text{Frage}|\text{eine}) \\
 &= \frac{Fr(\text{Berlin})}{N} \cdot \frac{Fr(\text{Berlin} \cap \text{ist})}{Fr(\text{Berlin})} \cdot \frac{Fr(\text{ist} \cap \text{eine})}{Fr(\text{ist})} \cdot \frac{Fr(\text{eine} \cap \text{Frage})}{Fr(\text{eine})} \\
 &= \frac{516.000.000}{20.000.000.000} \cdot \frac{7.910.000}{516.000.000} \cdot \frac{982.000.000}{13.760.000.000} \cdot \frac{64.300.000}{3.430.000.000} \\
 &= 5,29 \cdot 10^{-7}
 \end{aligned}$$

"Berlin ist eine Frage" hat die höhere Bigramm-Wahrscheinlichkeit. 

- b) Das Ergebnis entspricht nicht meiner Intuition. "Berlin ist eine Hauptstadt" sollte höhere Bigramm-Wahrscheinlichkeit haben. Die linguistische Plausibilität eines Satzes kann durch Bigramme nicht abgedeckt werden. 

c) • 
$$\begin{aligned}
 &P(\text{Ich esse einen Apfel}) \\
 &= \frac{Fr(\text{Ich})}{N} \cdot \frac{Fr(\text{Ich} \cap \text{esse})}{Fr(\text{Ich})} \cdot \frac{Fr(\text{esse} \cap \text{einen})}{Fr(\text{esse})} \cdot \frac{Fr(\text{einen} \cap \text{Apfel})}{Fr(\text{einen})} \\
 &= \frac{5.950.000.000}{20.000.000.000} \cdot \frac{3.250.000}{5.950.000.000} \cdot \frac{53.900}{12.700.000} \cdot \frac{555.000}{7.420.000.000} \\
 &= 5,16 \cdot 10^{-11}
 \end{aligned}$$

$$\begin{aligned}
 &P(\text{Ich mache einen Apfel}) \\
 &= \frac{Fr(\text{Ich})}{N} \cdot \frac{Fr(\text{Ich} \cap \text{mache})}{Fr(\text{Ich})} \cdot \frac{Fr(\text{mache} \cap \text{einen})}{Fr(\text{mache})} \cdot \frac{Fr(\text{einen} \cap \text{Apfel})}{Fr(\text{einen})} \\
 &= \frac{5.950.000.000}{20.000.000.000} \cdot \frac{15.100.000}{5.950.000.000} \cdot \frac{552.000}{51.900.000} \cdot \frac{555.000}{7.420.000.000} \\
 &= 6,01 \cdot 10^{-10}
 \end{aligned}$$

- $$P(\text{Ich esse einen Apfel}) = 5,16 \cdot 10^{-11}$$

$$\begin{aligned}
 &P(\text{Ich esse einen Tisch}) \\
 &= \frac{Fr(\text{Ich})}{N} \cdot \frac{Fr(\text{Ich} \cap \text{esse})}{Fr(\text{Ich})} \cdot \frac{Fr(\text{esse} \cap \text{einen})}{Fr(\text{esse})} \cdot \frac{Fr(\text{einen} \cap \text{Tisch})}{Fr(\text{einen})} \\
 &= \frac{5.950.000.000}{20.000.000.000} \cdot \frac{3.250.000}{5.950.000.000} \cdot \frac{53.900}{12.700.000} \cdot \frac{7.080.000}{7.420.000.000} \\
 &= 6,58 \cdot 10^{-10}
 \end{aligned}$$

d) 

e) 