Übungsblatt 7 – Musterlösung

7.1 Sprachsynthese

(a) 1. beinhalten

Das Wort *beinhalten* kann morphologisch sowohl als *be-in-halt-en* als auch als *bein-halt-en* interpretiert werden. Entsprechend ist die standarddeutsche Aussprache [bəˈʔɪnhaltn̩] bzw. [ˈbaɪ̯nhaltn̩].

2. Häuschen

Das Substantiv *Häuschen* sollte normalerweise morphologisch als *Häus-chen* (also *Haus*, Diminutiv-Suffix *chen* und Umlaut) interpretiert und dementsprechend [ˈhɔʏ̯sçən] ausgesprochen werden. Allerdings wäre auch eine Interpretation als Nominalisierung des fiktiven Verbes *häuschen* denkbar, was zur morphologischen Zerlegung *Häusch-en* und damit zur Aussprache [ˈhɔʏ̯ʃn̩] führen würde.

(b) Das Verb *umfahren* hat 2 verschiedene Bedeutungen, die sich phonetisch nur auf suprasegmentaler Ebene, nämlich der Betonung, unterscheiden:

#	Aussprache	Bedeutung
1	[ʔʊmˈfaːʁən]	Um etwas herumfahren
2	[ˈʔʊmfaːʁən]	Über etwas fahren

Da die beiden Varianten semantisch gesehen Antonyme sind, ist es hier von besonderer Wichtigkeit, die korrekte Aussprache zu wählen. In den meisten Fällen wird Variante 1 gemeint sein, da ein TTS-System wohl kaum jemanden bitten wird, vorsätzlich eine Baustelle zu zerstören. Generiert das TTS-System die Aussprache Wort für Wort, hat es allerdings kein Wissen darüber, in welchem Kontext das Wort *umfahren* vorkommt, womit theoretisch beide Varianten möglich wären. Verfügt es aber über Wissen über den Satzkontext, kann es durch die Wörter *Baustelle* und *großräumig* darauf schließen, dass die intendierte Bedeutung und damit auch Aussprache die von Variante 1 ist.

(c) Das betonte Wort in der Antwort ist jeweils kursiv geschrieben.

- Wer hat den Kuchen gegessen? *Peter* hat den Kuchen gegessen.
- Was hat Peter gegessen? Peter hat den *Kuchen* gegessen.
- Isst Peter gerade den Kuchen? Peter *hat* den Kuchen gegessen.

Die jeweils neue Information (*Peter*, *Kuchen*, *hat*), nach der in der Frage gefragt wurde, wird betont.

(d) *Die Lösungen dieser Teilaufgabe beziehen sich auf* iSpeech*, da der* Vocalizer *von Nuance immer noch nicht funktioniert.*

Folgendes funktioniert gut:

- Abkürzungen (*CDU*, *SPD*, *ca.*, *bzw.*)
- Datum / Uhrzeit (*24.02.2014, 11:46*)
- Aussprache von Homographen im Satzkontext (*Die vielen Staubecken* vs. *Das große Staubecken*)

Folgendes funktioniert weniger gut:

- Intonation bei Fragesätzen (*Wer bist du?* vs. *Wer bist du.*)
- Anglizismen / Fremdwörter (*Software*, *Donald Trump*, *Renaissance*)
- der eigenen Name (*iSpeech*)

(e) *Die Lösungen dieser Teilaufgabe beziehen sich auf* iSpeech*, da der* Vocalizer *von Nuance immer noch nicht funktioniert.*

- Die Wörter *beinhalten* und *Häuschen* werden korrekt ausgesprochen
- Der Satz *Die Baustelle bitte großräumig umfahren* wird ebenfalls korrekt ausgesprochen
- Bei den Frage-Antwort-Paaren ist keine Änderung der Intonation feststellbar

7.2 n-Gramme

Im Folgenden gilt $P(A \cap B) = P(A\ B)$

(a) *Saarbrücken liegt im Saarland* vs. *Saarbrücken liegt im Garten*

Bigramm-Satzwahrscheinlichkeiten (nach (8) der Bigramm-WK-Erklärung):
$P(\text{SB liegt im Saarland}) = P(\text{SB}) \cdot P(\text{liegt}|\text{SB}) \cdot P(\text{im}|\text{liegt}) \cdot P(\text{Saarland}|\text{im})$
$P(\text{SB liegt im Garten}) = P(\text{SB}) \cdot P(\text{liegt}|\text{SB}) \cdot P(\text{im}|\text{liegt}) \cdot P(\text{Garten}|\text{im})$

Da die ersten 3 Faktoren der beiden Satzwahrscheinlichkeiten gleich sind und nur nach dem wahrscheinlicheren Satz gefragt ist, reicht es, im Folgenden die Terme $P(\text{Saarland}|\text{im})$ und $P(\text{Garten}|\text{im})$ zu betrachten.

$$P(\text{Saarland}|\text{im}) = \frac{P(\text{im Saarland})}{P(\text{im})} = \frac{Fr(\text{im Saarland})}{Fr(\text{im})} = \frac{599.000}{3.540.000.000} = 1 \cdot 10^{-4}$$

$$P(\text{Garten}|\text{im}) = \frac{P(\text{im Garten})}{P(\text{im})} = \frac{Fr(\text{im Garten})}{Fr(\text{im})} = \frac{20.200.000}{3.540.000.000} = 5 \cdot 10^{-3}$$

Es gilt also:

$$P(\text{Saarbrücken liegt im Garten}) \gg P(\text{Saarbrücken liegt im Saarland})$$

(b) Intuitiv wäre zu erwarten gewesen, dass *Saarbrücken liegt im Saarland* eine höhere Wahrscheinlichkeit hat, da dieser Satz semantisch wohlgeformt bzw. logisch wahr ist. Wie in (b) gezeigt, beruht unter der Bigramm-Annahme der Unterschied zwischen den beiden Sätzen einzig und allein auf der Häufigkeit des letzten Bigramms (*im Garten* bzw. *im Saarland*). Da *im Garten* wesentlich häufiger vorkommt, ist so auch die Bigramm-Satzwahrscheinlichkeit des semantisch unplausibleren Satzes wesentlich höher.

Mit der Bigramm-Satzwahrscheinlichkeit werden immer nur Bigramme erfasst, so dass Satzkontext nur selten zum Tragen kommt. Dass *Saarbrücken* hier das Subjekt des Satzes ist und somit nicht im *Garten* liegen kann, wird also nicht berücksichtigt.

(c) 1. *Ich beherrsche Deutsch* vs. *Ich nicht Deutsch* (fehlerhafte Syntax)

Satzwahrscheinlichkeiten:
$P(\text{Ich beherrsche Deutsch}) = P(\text{ich}) \cdot P(\text{beherrsche}|\text{ich}) \cdot P(\text{Deutsch}|\text{beherrsche})$
$P(\text{Ich nicht Deutsch}) = P(\text{ich}) \cdot P(\text{nicht}|\text{ich}) \cdot P(\text{Deutsch}|\text{nicht})$

Relevant sind also $P(\text{beherrsche}|\text{ich}) \cdot P(\text{Deutsch}|\text{beherrsche})$ und $P(\text{beherrsche}|\text{ich}) \cdot P(\text{Deutsch}|\text{beherrsche})$.

Für *beherrsche* ergibt sich:

$$P(\text{beherrsche}|\text{ich}) = \frac{Fr(\text{ich beherrsche})}{Fr(\text{ich})} = \frac{92.500}{1.120.000.000} = 8 \cdot 10^{-5}$$
$$P(\text{Deutsch}|\text{beherrsche}) = \frac{Fr(\text{beherrsche Deutsch})}{Fr(\text{beherrsche})} = \frac{3.750}{517.000} = 7 \cdot 10^{-3}$$
$$P(\text{beherrsche}|\text{ich}) \cdot P(\text{Deutsch}|\text{beherrsche}) = 6 \cdot 10^{-7}$$

Für *nicht* ergibt sich:

$$P(\text{nicht}|\text{ich}) = \frac{Fr(\text{ich nicht})}{Fr(\text{ich})} = \frac{50.300.000}{1.120.000.000} = 4 \cdot 10^{-2}$$
$$P(\text{Deutsch}|\text{nicht}) = \frac{Fr(\text{nicht Deutsch})}{Fr(\text{nicht})} = \frac{385.000}{1.450.000.000} = 2 \cdot 10^{-4}$$
$$P(\text{nicht}|\text{ich}) \cdot P(\text{Deutsch}|\text{nicht}) = 1 \cdot 10^{-5}$$

Es gilt also:

$$P(\text{Ich nicht Deutsch}) \gg P(\text{Ich beherrsche Deutsch})$$

2. *Ich esse Kuchen* vs. *Ich essen Kuchen* (Fehlerhafte Morphologie)

Satzwahrscheinlichkeiten:
$P(\text{Ich esse Kuchen}) = P(\text{ich}) \cdot P(\text{esse}|\text{ich}) \cdot P(\text{Kuchen}|\text{esse})$
$P(\text{Ich essen Kuchen}) = P(\text{ich}) \cdot P(\text{essen}|\text{ich}) \cdot P(\text{Kuchen}|\text{essen})$

Relevant sind also $P(\text{esse}|\text{ich}) \cdot P(\text{Kuchen}|\text{esse})$ und $P(\text{esse}|\text{ich}) \cdot P(\text{Kuchen}|\text{esse})$.
Für *esse* ergibt sich:

$$P(\text{esse}|\text{ich}) = \frac{Fr(\text{ich esse})}{Fr(\text{ich})} = \frac{466.000}{1.120.000.000} = 4 \cdot 10^{-4}$$
$$P(\text{Kuchen}|\text{esse}) = \frac{Fr(\text{esse Kuchen})}{Fr(\text{esse})} = \frac{4.010}{549.000.000} = 4 \cdot 10^{-6}$$
$$P(\text{esse}|\text{ich}) \cdot P(\text{Kuchen}|\text{esse}) = 3 \cdot 10^{-9}$$

Für *essen* ergibt sich:

$$P(\text{essen}|\text{ich}) = \frac{Fr(\text{ich essen})}{Fr(\text{ich})} = \frac{572.000}{1.120.000.000} = 5 \cdot 10^{-4}$$
$$P(\text{Kuchen}|\text{essen}) = \frac{Fr(\text{essen Kuchen})}{Fr(\text{essen})} = \frac{345.000}{247.000.000} = 1 \cdot 10^{-3}$$
$$P(\text{essen}|\text{ich}) \cdot P(\text{Kuchen}|\text{essen}) = 7 \cdot 10^{-7}$$

Es gilt also:

$$P(\text{Ich essen Kuchen}) \gg P(\text{Ich esse Kuchen})$$

(d) *Die EU will die EU*

1. Mögliche POS-Strukturen eines Satzes mit 5 Wörtern und deren Häufigkeit im Negra-Korpus:
   - ART NN VV ART NN (178 mal)
   - NN VV ART ADJA NN (141 mal)
   - ART ADJA NN VV NN (18 mal)
   - ...
2. ART-NN-Kombination mit hoher Trefferzahl in Google: *Die EU*
3. *EU*-VV-Kombination mit hoher Trefferzahl in Google: *EU will*
4. *will*-ART-Kombination mit hoher Trefferzahl in Google: *will die*
5. Eine gute *die*-NN-Kombination ist ja bereits bekannt: *die EU*

$$\begin{aligned} P(\text{Die EU will die EU}) &= P(\text{Die}) \cdot P(\text{EU}|\text{Die}) \cdot P(\text{will}|\text{EU}) \cdot P(\text{die}|\text{will}) \cdot P(\text{EU}|\text{die}) \\ &= \frac{Fr(\text{Die})}{N} \cdot \frac{Fr(\text{Die EU})}{Fr(\text{die})} \cdot \frac{Fr(\text{EU will})}{Fr(\text{EU})} \cdot \frac{Fr(\text{will die})}{Fr(\text{will})} \cdot \frac{Fr(\text{die EU})}{Fr(\text{die})} \\ &= \frac{3,8 \cdot 10^9}{2 \cdot 10^{10}} \cdot \frac{1,4 \cdot 10^7}{3,8 \cdot 10^9} \cdot \frac{2,8 \cdot 10^6}{2,8 \cdot 10^9} \cdot \frac{4,3 \cdot 10^7}{1 \cdot 10^{10}} \cdot \frac{1,4 \cdot 10^7}{3,8 \cdot 10^9} \\ &= 1.1 \cdot 10^{-11} \end{aligned}$$

(e) *die der von der von*

$$\begin{aligned} P(\text{die der von der von}) &= \frac{Fr(\text{die})}{N} \cdot \frac{Fr(\text{die der})}{Fr(\text{die})} \cdot \frac{Fr(\text{der von})}{Fr(\text{der})} \cdot \frac{Fr(\text{von der})}{Fr(\text{von})} \cdot \frac{Fr(\text{der von})}{Fr(\text{der})} \\ &= \frac{3,9 \cdot 10^9}{2 \cdot 10^{10}} \cdot \frac{5,3 \cdot 10^7}{3,9 \cdot 10^9} \cdot \frac{4,9 \cdot 10^7}{3,4 \cdot 10^9} \cdot \frac{4,3 \cdot 10^8}{2,8 \cdot 10^9} \cdot \frac{4,9 \cdot 10^7}{3,4 \cdot 10^9} \\ &= 9,6 \cdot 10^{-8} \end{aligned}$$

7.3 Unix-Tools

(a)

```
# Nur den Anfang der ersten Zeile durch einen Zeilenumbruch
    ersetzen
$ sed '1s/^/\n/' tiger.txt > tiger_2.txt
$ mv tiger.txt tiger_1.txt
# "tiger*" expandiert zu "tiger_1.txt tiger_2.txt"
$ paste tiger* > bigrams.tsv
$ head bigrams.tsv
             ``
``           Ross
Ross         Perot
Perot        wäre
wäre         vielleicht
vielleicht   ein
ein          prächtiger
prächtiger   Diktator
Diktator     ''
''           Konzernchefs
```

(b)

Unigramme

```
$ grep "^Saarbrücken$" tiger_2.txt | wc -l
11
$ grep "^liegt$" tiger_2.txt | wc -l
210
$ grep "^im$" tiger_2.txt | wc -l
5138
$ grep "^Saarland$" tiger_2.txt | wc -l
20
$ grep "^Garten$" tiger_2.txt | wc -l
9
```

Bigramme

```
$ grep "^Saarbrücken liegt$" bigrams.tsv | wc -l
0
$ grep "^liegt im$" bigrams.tsv | wc -l
11
$ grep "^im Garten$" bigrams.tsv | wc -l
2
$ grep "^im Saarland$" bigrams.tsv | wc -l
6
```

Anzahl der Wörter im Tiger-Korpus

```
$ wc -l tiger_2.txt
888238 tiger_2.txt
```

Das Problem ist, dass das Bigramm *Saarbrücken liegt* nicht im Korpus vorkommt. Damit ist dessen Frequenz und auch Wahrscheinlichkeit 0, was die ganze Satzwahrscheinlichkeit zu 0 werden lässt. Dieses Problem ist als *Sparse Data Problem* bekannt. Das zur Verfügung stehende Korpus bzw. die Wahrscheinlichkeit des Bigrammes ist zu klein. Daher bleibt uns nichts anderes übrig, als den Term mit $P(\text{liegt}|\text{Saarbrücken})$ im Zähler in beiden Sätzen zu ignorieren.

(c) Es bleibt also noch Folgendes zu berechnen:

$$\begin{aligned} P(\text{Saarbrücken liegt im Saarland}) &\approx \frac{Fr(\text{Saarbrücken})}{N} \cdot \frac{Fr(\text{liegt im})}{Fr(\text{liegt})} \cdot \frac{Fr(\text{im Saarland})}{Fr(\text{im})} \\ &= \frac{11}{888.238} \cdot \frac{11}{210} \cdot \frac{6}{5138} \\ &= 2,5 \cdot 10^{-10} \end{aligned}$$

$$\begin{aligned} P(\text{Saarbrücken liegt im Garten}) &\approx \frac{Fr(\text{Saarbrücken})}{N} \cdot \frac{Fr(\text{liegt im})}{Fr(\text{liegt})} \cdot \frac{Fr(\text{im Garten})}{Fr(\text{im})} \\ &= \frac{11}{888.238} \cdot \frac{11}{210} \cdot \frac{2}{5138} \\ &= 7,6 \cdot 10^{-10} \end{aligned}$$

Es gilt also

$$P(\text{Saarbrücken liegt im Saarland}) > P(\text{Saarbrücken liegt im Garten})$$

Da das Tiger-Korpus auf Artikeln der *Frankfurter Rundschau* basiert, ist hier die Wahrscheinlichkeit für *Garten* recht niedrig. Folglich ist die Wahrscheinlichkeit für den Satz mit *Saarland* höher. Bei der Entwicklung eines Sprachmodells muss man also darauf achten, dass man Daten aus der Textsorte nimmt, auf die das Modell angewendet werden soll.