

# Vorlesung “Einführung in die Computerlinguistik”

Prof. Dr. D. Klakow

## Übung zur Statistischen Sprachverarbeitung

Ausgabe: 22.1.2008

Abgabe: 29.1.2008

### Aufgabe 1.1 (aus Manning+Schütze Seite 59)

Berechnen Sie die Wahrscheinlichkeit dafür, dass “Ein Punkt, der nach einem Wort aus drei Buchstaben folgt, eine Abkürzung kennzeichnet und nicht das Ende eines Satzes”. Die Wahrscheinlichkeiten sind gegeben durch:

$$P(Ist - Abk. | drei - Buchstaben - Wort) = 0.8$$

$$P(drei - Buchstaben - Wort) = 0.0003$$

### Aufgabe 1.2 (aus Manning+Schütze Seite 59)

Seien X und Y Zufallsvariablen für die gilt:

x	0	0	1	1
y	0	1	0	1
P(X=x,Y=y)	0.32	0.08	0.48	0.12

Sind die beiden Zufallsvariablen statistisch unabhängig?

### Aufgabe 1.3

Auf der Kurs-Webseite finden Sie ein Perl-Skript, mit dem Sie die Wahrscheinlichkeit schätzen könne, dass zuerst Wort w1 und d-Wörter später Wort w2 auftritt. Nutzen Sie dieses Skript um festzustellen, wie stark zwei Wörter in einem Corpus korreliert sind bzw. ob sie vielleicht für diesen Abstand sogar statistisch unabhängig sind. Benutzen Sie dazu den Quotienten

$$\frac{P(\text{zuerst Wort } w1 \text{ und d - Woerter spaeter Wort } w2)}{P(\text{Wort } w1)P(\text{Wort } w2)}$$

Als Corpus nehmen Sie die Englische Version von Tolstois “Krieg und Frieden” die Sie auf der Kurs-Webseite finden.

(Quelle: Projekt Gutenberg [http://www.gutenberg.org/wiki/Main\\_Page](http://www.gutenberg.org/wiki/Main_Page))

Hier ein Beispiel, wie Sie das Skript aufrufen müssen:

```
cat tolstoy_war_and_peace.dat | joint_prob.pl -w1="he" -w2="that" -d=3
```

Betrachten Sie folgende Wortpaare und Abstände

w1	w2	d
he	said	2
he	said	3
he	said	4
he	said	10
he	said	100

Wie interpretieren Sie Ihre Beobachtung?

w1	w2	d
Pierre	he	20
Pierre	she	20
Natasha	he	20
Natasha	she	20

Wie interpretieren Sie Ihre Beobachtung in diesem Fall?