# Lab Class NLP:II

Exercise 1 : Zipf Distribution

The lecture introduced Zipf's Law as a statistical law that describes the frequency of words in a language.

(a) What does Zipf's law state?

(b) What implications does it have for developing statistical models in NLP?

(c) Consider the following table of words that appeared in a corpus:

| Rank | Word | Frequency | Expected Frequency |
|---|---|---|---|
| 1 | the | 36000 | |
| 2 | that | 18000 | |
| 3 | for | 12000 | |
| 4 | is | 9000 | |
| 5 | said | 7200 | |
| 6 | on | 6000 | |
| 8 | in | 4500 | |
| 9 | it | 4000 | |
| 10 | by | 3600 | |
| 12 | from | 3000 | |
| 15 | million | 2400 | |
| 16 | at | 2250 | |
| 18 | as | 2000 | |
| 20 | with | 1800 | |
| 24 | a | 1500 | |
| 25 | was | 1440 | |

Assuming that there are a total of 500,000 words out of which 50,000 are unique, do the word frequencies satisfy Zipf's law in this case? Explain why or why not.

Exercise 2 : Data Acquisition

You decide to study the problem of computational argumentation.

(a) Name two possible sources of argumentative text like arguments, debates, or claims as text or speech transcripts.

(b) For each source from (a), describe how you would collect this data and what problems you could face collecting them.

(c) Inform yourself about common sources of bias in language corpora or datasets. Name the biases you found and how they can be mitigated during data acquisition.

Exercise 3 : Corpora and Annotation

In this exercise, you will explore the challenges involved in annotating, both for humans and automated systems. Consider the following corpus:

1. Paris Hilton stayed at the Hilton in Paris.
2. Quentin Bloody Tarantino.
3. Max works for Berlin & Brandenburg Post.

(a) Identify and annotate all named entities in the sentences above. For this task, a named entity is any word or phrase that refers to

- a person (PER),
- a location (LOC), or
- an organisation (ORG).

Use the following format for your annotations: [Mary Smith]$_{\text{PER}}$ went to [Tokyo]$_{\text{LOC}}$.

(b) Discuss any instances you found challenging to annotate. Which cases might be particularly difficult for an automated system to annotate? Why?