UNIVERSITÄT
ZU KÖLN

CORPUS STATISTICS

Sprachverarbeitung: Übung

Janis Pagel

Institut für Digital Humanities

22 April 2025

01

REMINDER OF LECTURE

Relevant topics for today

- Absolute and relative counts
- Zipf distribution
- Type-Token-Ratio

02

TOKENIZATION WITH PYTHON

Tokenization with NLTK

- NLTK: Natural Language Tool Kit
- Provides functions for simple tokenization (among many other functionalities)

```
import nltk
nltk.download("punkt_tab") # Install tokenizer data
print(nltk.tokenize.word_tokenize("This text should be tokenized. Hyphen-words should be a single token!"))
~> ['This', 'text', 'should', 'be', 'tokenized', '.', 'Hyphen-words', 'should', 'be', 'a', 'single', 'token', '!']
```

03

EXERCISE 03

Exercise 03

- https://lehre.idh.uni-koeln.de/site/assets/files/5615/exercise03.pdf

UNIVERSITY
OF COLOGNE

Janis Pagel
Institut für Digital Humanities

eMail janis.pagel@uni-koeln.de
Homepage https://janispagel.de
Phone +49 221 470 5749

Foto: Gregor Hübl