



h_da

HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Aspects of Linguistic Complexity

Summary

Skype-Call (09.05.2018)

project structure

- three main subjects:
 1. corpus
 2. “LanguageTool”
 3. verb extraction

- project meetings every week

implementing a text corpus

- research of texts written in simple language / „normal“ language
- built a database (SQLite)
- Topics: Bible, fairytale, news, novel, political party programs
- Structure: parallel corpus (simple – „normal“ language)
- Dimension: 100.000 words „normal“ language / 10.000 simple language

„LanguageTool“

- **research**

programming Language-Tool:

- <http://www.pro-linux.de/artikel/1/62/languagetool.html>

Language-Tool rules:

- <https://multisprech.org/einfache-sprache/einfach-schreiben/languagetool-leichte-sprache/>

Language-Tool rules (code):

- <https://github.com/languagetool-org/languagetool/blob/master/languagetool-language-modules/de-DE-x-simple-language/src/main/resources/org/languagetool/rules/de-DE-x-simple-language/grammar.xml>

Language-Tool wiki:

- <http://wiki.languagetool.org/make-languagetool-better>

Language-Tool community:

- <https://community.languagetool.org/>

„LanguageTool“

- rules
 - analysing existing rules
 - think of new rules
 - syntax of programming language

„Flesch-index“

work with:

-Python

-Pyphen

-“Flesch-index“ formel

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

verb extraction

- research of algorithm (rules) for extraction
- NLTK import
- textblob
<http://textblob.readthedocs.io/en/dev/>
- pyphen
<http://pyphen.org/>