Corpus REDEWIEDERGABE

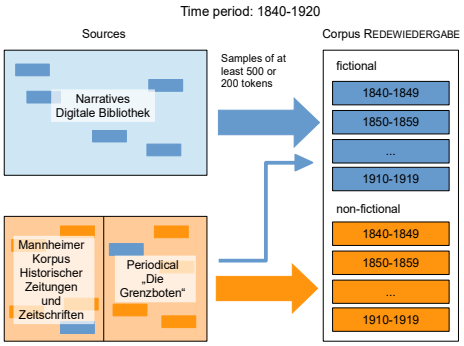
Lukas Weimer, Annelen Brunner, Stefan Engelberg,
Fotis Jannidis, Ngoc Duyen Tanja Tu

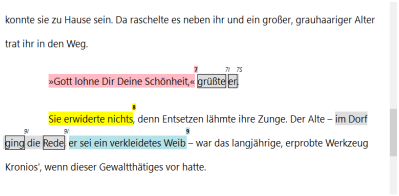
IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Julius-Maximilians-
UNIVERSITÄT WÜRZBURG

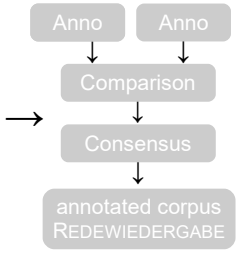
Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Annotation





Annotation environment ATHEN
https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen



Annotation guidelines

direct	He thought, “I'm hungry.”
free indirect	Where on earth should he get something to eat now?
indirect	He said he was hungry.
reported	He was talking about his hunger.

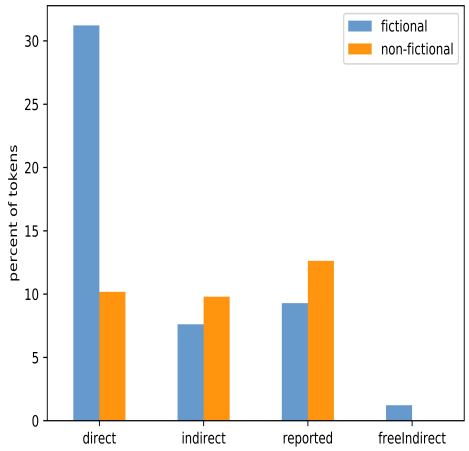
Corpus statistics

Type	Instances	Tokens
direct	2,929	77,246
free indirect	97	2,238
indirect	2,077	32,740
reported	4,204	40,647
all	9,307	152,871

Medium	Instances
speech	6,003
thought	2,241
writing	873
ambiguous	302

Overall:
619 samples with
630,974 tokens

download
the corpus at
GitHub
github.com/redewiedergabe



Conclusions

- *direct* seems to be the preferred form of representation
- *free indirect* and *direct* seem to be preferred in fictional literature, whereas *indirect* and *reported* are more common in non-fictional literature
- Speech is the medium that is most frequently represented, followed by thought, followed by writing
- Instances of *direct* representation are on average the longest (26 tokens), whereas *reported* is a rather short form of representation (10 tokens)

Next steps

- programming an automatic recognizer for the different types of representation
- second release of our corpus with additional text material such as single annotated samples and full texts

www.redewiedergabe.de