Web-Scraping Blogspot Data as an Individual-Based Diachronic Text Corpus for Measuring Linguistic Prevalence

Project Goals

Diachronic prevalence calculations

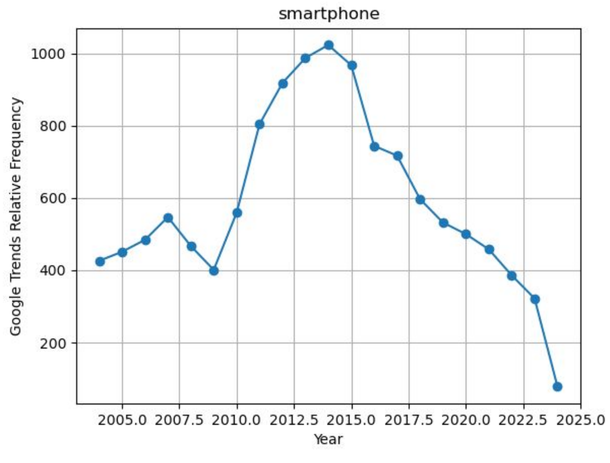
Comparison with Brysbaert et al. (2019)

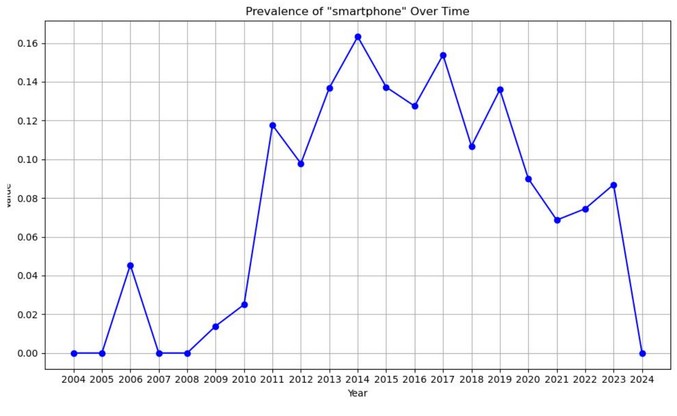
The Creation of the Blogspot Corpus

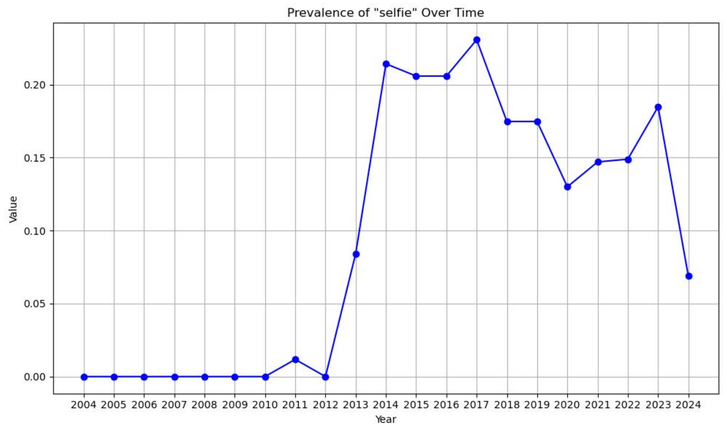
- Custom built web-scraper
- Scrapy-library via Python

Results

- Prevalence calculations for every year
- Results compared with Brysbaert et al. (2019)
- Word usage patterns; similar to Google Trends data
- Prevalence for written text often diverges from Brysbaert et al.'s data
- Text corpus available for future research

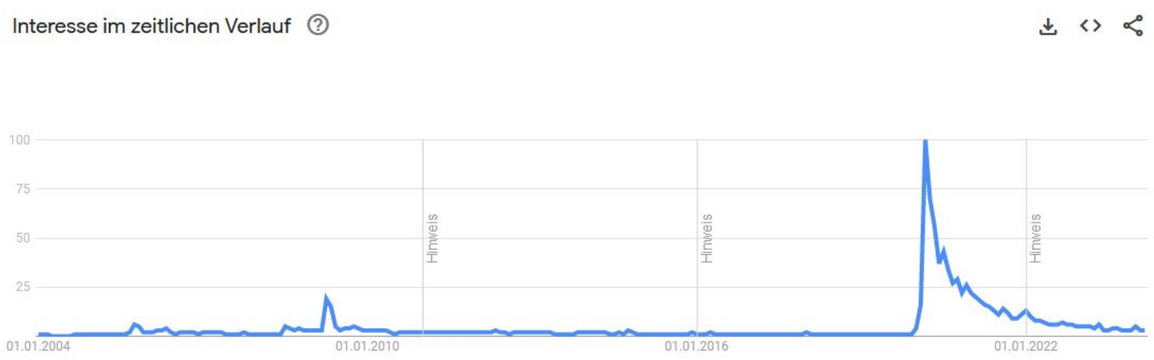


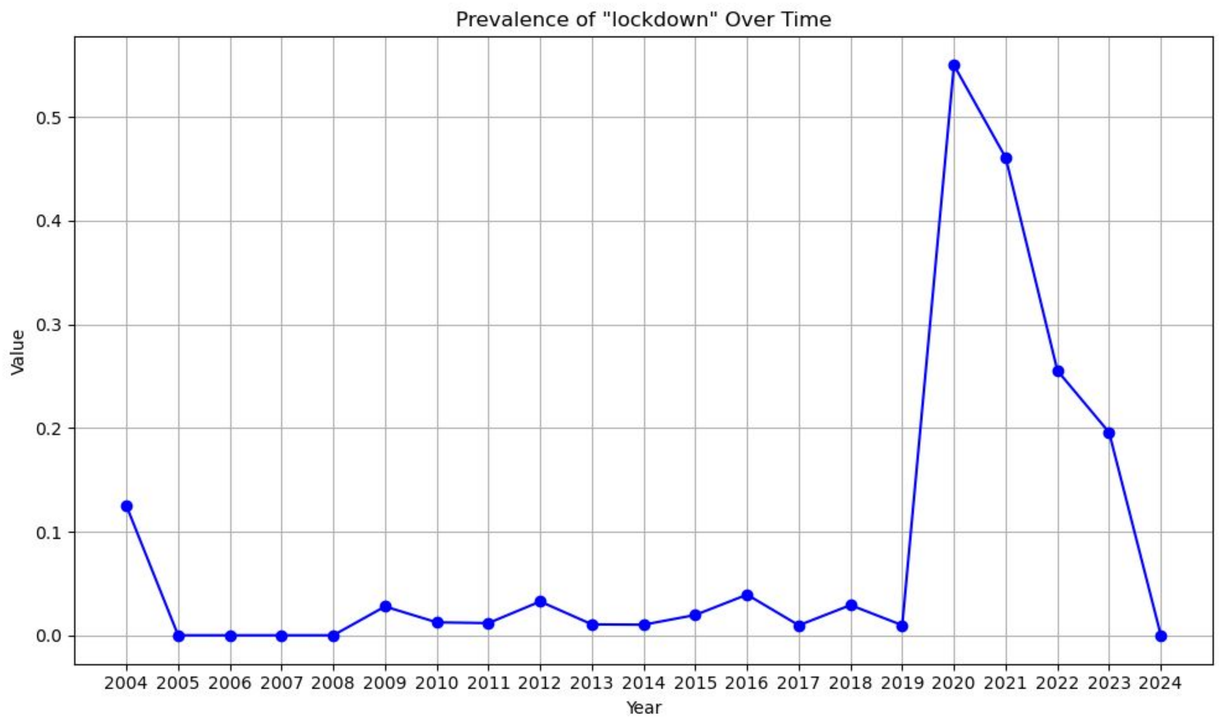


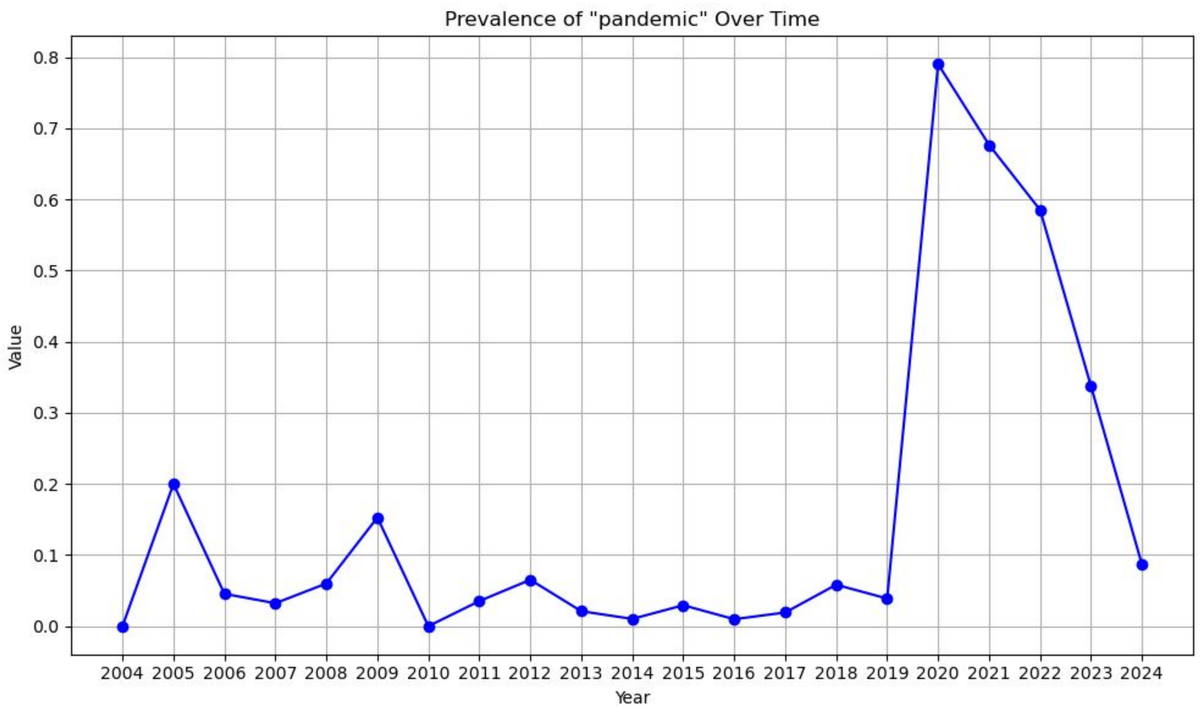


Facts about the Dataset

- over 80 million tokens
- 104 authors
- spanning 20 years (2004-2024)
- data cleaning (stopword-removal, NER, PoS-tagging)







Marja Nikolčič
Sebnem Yayla

universität wien

Patrick Konrad
Maximilian Berens

University of Vienna, Department of Data Science and Department of Digital Humanities