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

**Project Goals**
- Word prevalence is defined as a fraction of individuals who know a word and is an important metric in linguistics (Brysbaert et al. 2016)
- Existing research evaluates prevalence based on crowd-sourcing and corpora (Johns et al. 2020, Feltgen 2024)

**Motivation:** Prevalence measurements by creating a diachronic blog corpus with authorship information

Correlating data with Brysbaert et al.’s (2019) prevalence study of 62,000 Lemmas

**Creating the Blogspot Corpus**
- Custom-built web-scraper
- Scrapy-library via Python
- over 80 million tokens
- 104 authors
- spanning 20 years
- data cleaning (stopword-removal, removing named entities, PoS-tagging)

**Analysis & Results**
- Diachronic and overall prevalence calculation and comparison
- Significant correlation with Brysbaert’s data
- Emerging trends are visible in word usage patterns over time (e.g. smartphone)
- Text corpus available for future research