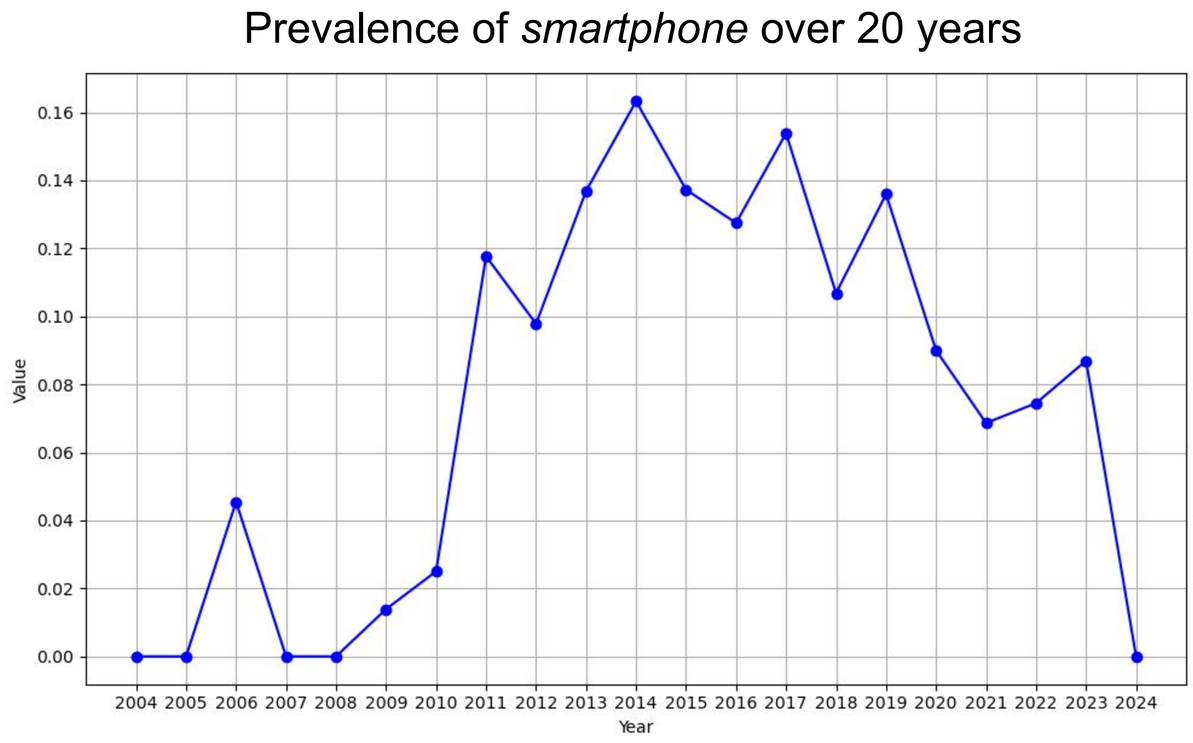
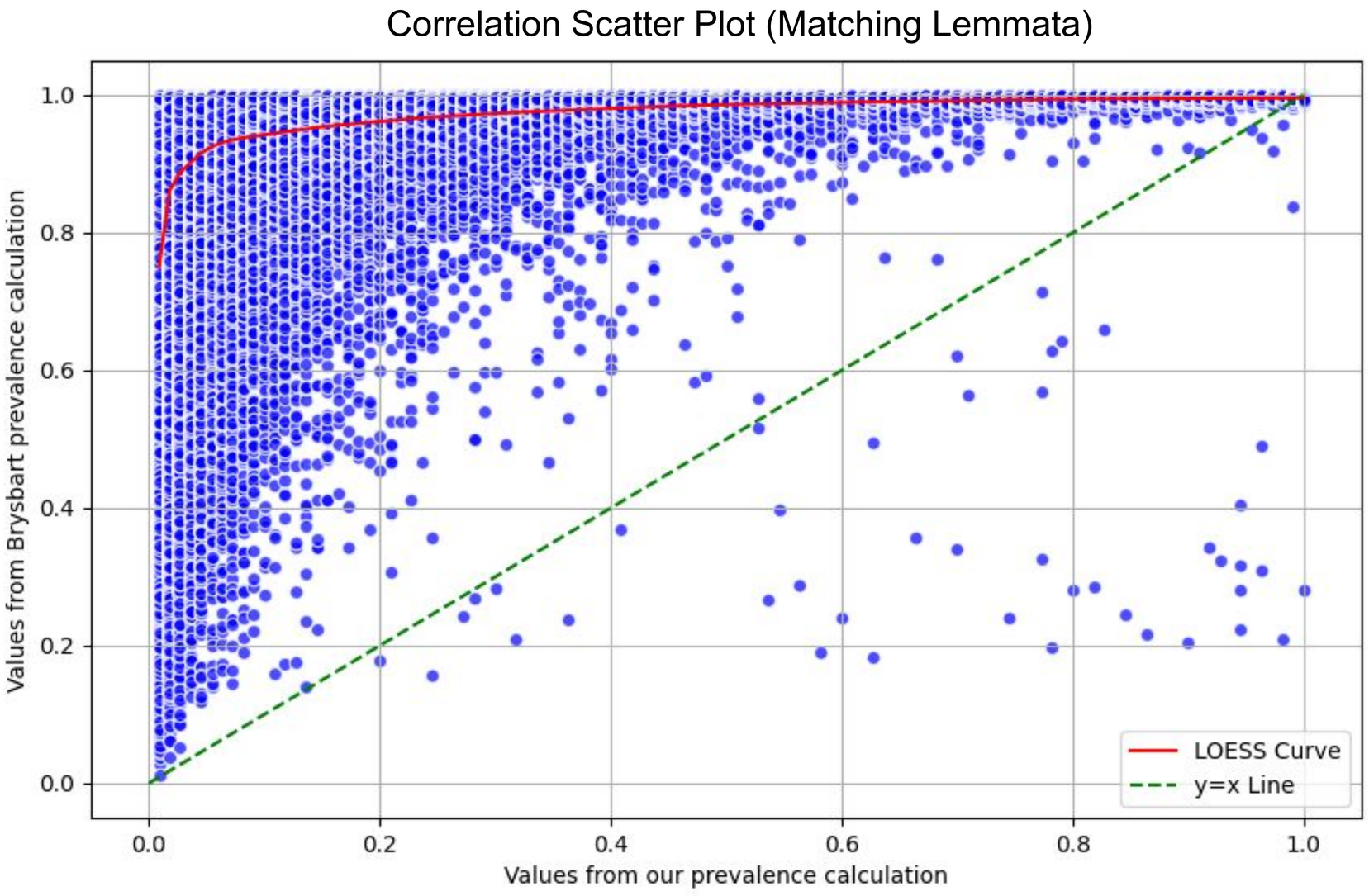
Web-Scraping Blogspot Data as an Individual-Based Diachronic Text Corpus for Measuring Linguistic Prevalence

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



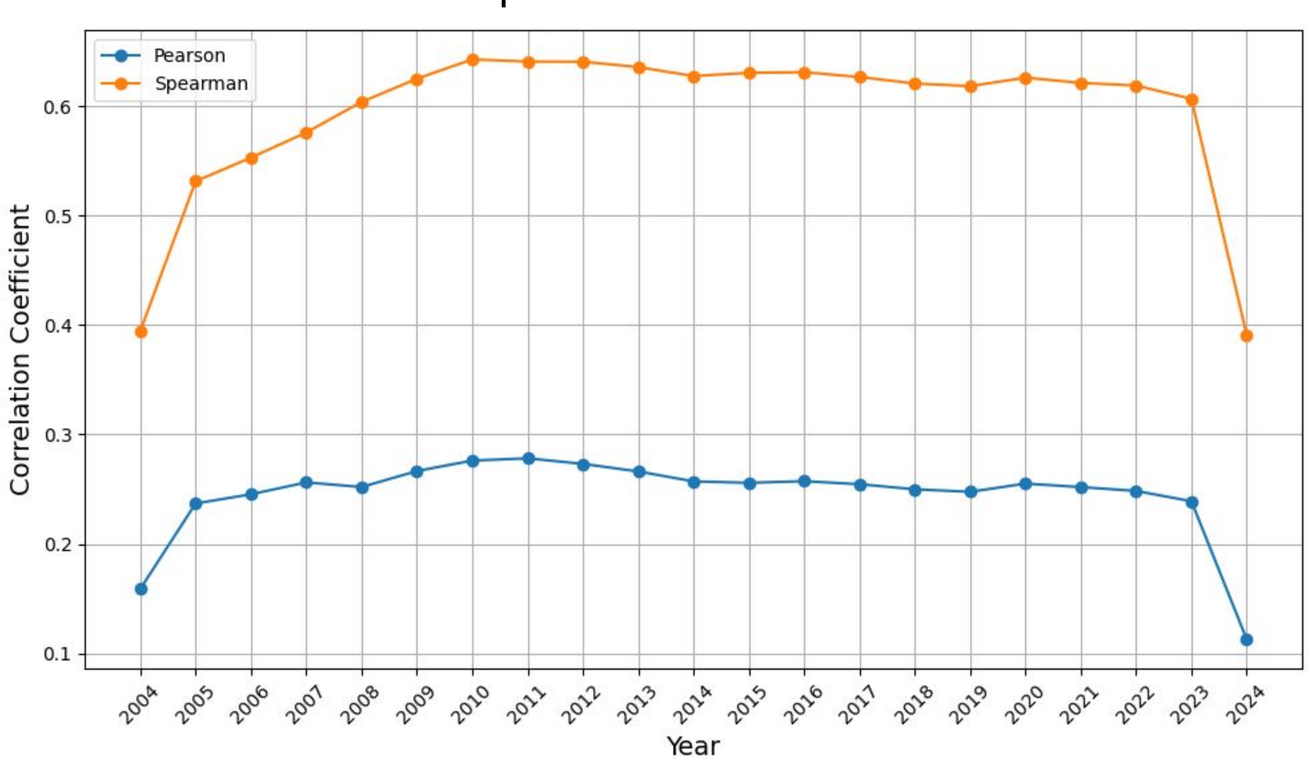


Creating the Blogspot Corpus

- 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

Patrick Konrad, Maximilian Berens,
Sebnem Yayla, Marja Nikolčič

universität wien
Master Digital Humanities, Data Science

References:
Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. Journal of Experimental Psychology: Human Perception and Performance, 42(3), 441.
Brysbaert, M., Mandera, P., Keuleers, E., & New, B. (2019). Word prevalence norms for 62,000 English lemmas. Behavior Research Methods, 51(2), 467–479.
Johns, B. T., Jones, M. N., & Mirman, D. (2020). Estimating the prevalence and diversity of words in written language.
Feltgen Q (2024) Is language change chiefly a social diffusion affair? The role of entrenchment in frequency increase and in the emergence of complex structural patterns. *Front. Complex Syst.* 2:1327425. doi: 10.3389/fcpxs.2024.1327425.