Universität Wien

SPL 13 - Finno-Ugristik, Nederlandistik, Skandinavistik und Vergleichende Literaturwissenschaft

Masterstudium Digital Humanities, Data Science, und Business Analytics

Data Analysis Project:

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

Submitted by:

Marja Nikolčič
marja.nikolcic@gmail.com, a12124067@unet.univie.ac.at

Sebnem Yayla
sebnem.sara.yayla@gmail.com, a12043536@unet.univie.ac.at

Patrick Konrad
konrad-patrick@gmx.at, a01447066@unet.univie.ac.at

Maximilian Berens
maximilian.berens@email.de, a12250229@unet.univie.ac.at

Course: 053631 LP Data Analysis Project (2023WS)

Supervisor: Ass.-Prof. Mag. Mag. Dr. Andreas Baumann

Inhaltsverzeichnis

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

Introduction

In recent years, the notion of prevalence has risen notably in linguistic research. In response to its growing significance, this project aims to contribute valuable insights to the field by developing and analyzing an individual-based diachronic text corpus for measuring linguistic prevalence.

Our methodology involves scraping blog data spanning roughly two decades, to build a comprehensive database of users' writings across various thematically different fields of interest. Through this approach, we address some of the potential challenges when measuring prevalence, namely e.g. the lack of essential speaker information when analyzing corpus data, as well as providing an adequate baseline regarding the number of individuals analyzed. Through these measures we tend to provide sufficient data to create a successful blueprint which also ensures representativeness at the individual level by providing an adequate sample size. In this project report, we will present our findings throughout the roughly six month journey of developing this project.

First, since the main research we built our project upon was conducted by Brysbaert et al. (Brysbaert et al. 2019, Behavior Research Methods), it is necessary to present the paper and its research to establish a baseline for further discussion. This will also help when comparing the findings of our project with the results provided by Brysbaert et al. which is an important part of this report.

Afterwards, we will provide insights to the process of building the dataset for this project. We will discuss the initial approach and the various challenges during this project stage, as well as the reasons for taking one particular approach over the other. Namely, we build the text corpus by scraping openly accessible blog data on the web. We will also present the methodology of finding and scraping the selected blogs.

We will present our findings, as well as the prevalence scores from the collected data, while also looking at further interesting findings we discovered through our results. Additionally, as mentioned, we will discuss the comparisons with the previous research provided by Brysbaert et al (2019). Finally, the concluding section will discuss the outcomes of our findings and the implications for future research on the topic of linguistic prevalence.

Brysbaert et al. literature review

In their research paper from 2019, Brysbaert et al. estimate linguistic prevalence data for approximately 62.000 words in english. They define word prevalence as the indication of which number of people know a certain word. This term of word prevalence aims to solve the issue of differences regarding word knowledge that are unrelated to word frequency.

Methodologically, the study was conducted by online crowdsourcing involving over 220.000 people who indicate that they know a certain word. The data material used in the study consisted of 61.858 English words, together with a list of 329.851 pseudowords. Study participants were given a random sample for the vocabulary test which consisted of 67 words and 33 nonwords. “For each letter string, participants had to indicate whether or not they knew the stimulus. At the end of the test, participants received information about their performance, in the form of a vocabulary score based on the percentage of correctly identified words (Brysbaert et al. 2018, p. 468).” Additionally, metadata such as age, gender, the level of education and the question whether English was their native language was collected. On average, each word was judged by 388 participants and the rates of which they knew certain words ranged from 2% to 100%. The study showed that most words were known by 90% or more of the participants. Some other main differences were that certain words were known by more participants in the United Kingdom than in the United States of America, due to cultural differences, and vice versa. Similarly, there are also differences based on gender which Brysbaert et al contribute to gender differences in interests.

Word prevalence has a wide array of potential use-cases in research and beyond. For instance, according to the researchers, the discussed prevalence is a good general indicator for word difficulty. Words which are relatively unknown to a large number of people can be avoided by using words with the appropriate prevalence values. Further, prevalence is a potentially promising variable in natural language processing particularly considering algorithms to determine text difficulty. Since the value is not just reduced to differences in word frequency, word prevalence is likely an insightful measure for evaluating text difficulty.

Finally, all the computed results in the research were made accessible via an Excel sheet with the collected values. Further, the file also contains information regarding the discussed differences between gender and countries.

Attempting to Crowdsource Prevalence Data

One of the initial goals of the Data Analysis Project was to crowdsource prevalence data and to compare our data with Brysbaert et al. (2019). Crowdsourcing is a process to combine the inputs or activities of many individuals who are contributing to scientific goals,one of which is data collection (Sari et al. 2019). Within the last five years there have been many interdisciplinary papers which have been embracing research in crowdsourcing, but none have been published yet with the explicit intent to crowdsource large scale text data for prevalence research within the computational humanities. At the start of the project, our team has attempted to pioneer in this area of research, but ran into multiple issues along the way, which prompted us to change our method of data collection. However, it might still prove to be useful for future researchers to know what did not work for our team.

The initial plan was to crowdsource prevalence data from at least 100 participants from within the USA, preferably. To accomplish this we thought about deploying a survey, using SoSci-Survey (c.f. www1), on Prolific Academic (c.f. www2) which is a platform where people are paid to be participants in studies. On Prolific Academic there would have been options available to only display our survey to native English speakers, coupled with launching the survey during the night time in Europe would reduce the amount of Europeans and people of other geolocations participating to ensure our data would match as closely as possible Brysbaerts.

The most important problems that occurred were the requirement of a huge size of individually written data of at least 100.000 words per participant, insufficient funding[1], presumably heavily biased data towards academic writing – which does not represent orally conceptualized or everyday language in any way, neither in regards to the vocabulary nor to its topics – privacy concerns, potentially with bad effects on the amount of provided data, its quality, and relevance to our research, but also a lot of unclean data with little structure. In conclusion, while crowdsourcing prevalence data might be very well possible with other approaches, we found conducting online-surveys to be not a fruitful method.

[1] Especially in regards to the estimated time it would take to remember where the data was saved, uploading it, entering basic information etc. Copy and pasting data of that size would already take roughly thirty minutes for proficient and organized enough computer users, let alone anything else required of the participants. On Prolific Academic it is important that the ratio of payment to time spent is right, otherwise the study might fail if not enough people are participating.

After coming to terms with these methodological problems our team had decided to create our own dataset by the means of web-scraping blog data.

The Creation of the Blogspot Corpus

Primarily, our method involved scraping content from existing online blogs. Some of which have been active for a long time, stretching up to two decades. So, over time, a substantial amount of textual content has been created which would be sufficient for our objective of analyzing a large amount of text data which could be traced to individual blog owners. After initial research, we found that constructing a text corpus from these sources was promising, so we decided to proceed with this idea.

However, one drawback was that this initial search for potential blogs to scrape could not be done automatically. The blogs had to be sufficient for our research criteria, meaning that they need to have a robust amount of text spanning throughout the blog's existence.
Another challenge was that it is hard to verify if the authors are English native speakers. Most blogs offer an "about" page containing some information about the author. So, we manually looked into the information to try and verify that the blogs are operated by native speakers. However, this process still carries potential uncertainties.

As a baseline, generally our approach focused on searching for long-running blogs managed by native English speakers which have written close to, or above 100.000 words to ensure a substantial amount of content. For the larger blogs, it would be ideal if they had written over 100.000 words per year. Further, the different blogs should represent a wide array of interests spanning over all kinds of different fields to represent a broadly diversified and well-rounded corpus.

Before starting the scraping process, we collected potential blogs to scrape in an Excel-file which was created manually. We found the blogs by using online searches and blog search engines over a period of approximately one month. We settled on scraping data specifically from blogs hosted by Blogspot (a subdomain of Blogger.com). As one of the oldest and most widely used blogging platforms, it hosts a huge amount of content with a wide range of topics, which proved to be great for creating our corpus.

One significant advantage of utilizing the Blogspot framework lies in its straightforward HTML structure and the absence of anti-scraping mechanisms. Unlike some other platforms, Blogspot often tends to maintain consistent naming conventions for HTML elements throughout its architecture. This consistency simplifies the scraping process, enabling the creation of automated scripts capable of efficiently extracting content from multiple blogs within the platform. We employed Python along with the Scrapy library—a powerful web scraping framework. Using Scrapy, we programmed the script to automatically locate and extract desired data within the targeted HTML elements. By utilizing CSS selectors, we iteratively targeted elements containing textual content and publication dates across every entry and page within the blog.

For each blog, the results of the scraping process were compiled into a structured JSON file with a title, publication date, and textual content for each post. This scraper then was used over all researched blogs and these files collectively formed the text corpus for this project. Afterwards, we proceeded to process the JSON files to clean the data by removing the stop words which were not relevant for our analysis. Also, we employed Part-of-speech tagging and Named-entity recognition to clean the data further. Through these post-processing steps, we created a cleaned dataset composed of multiple JSON files representing the different blogs.

All in all, our collected Blogspot text corpus contains approximately 81.000.000 words divided across 104 total blogs. Almost all of the blogs consist of well over 100.000 words. With this robust corpus, we are confident that there is enough reliable data to compute word prevalences. In order to evaluate the whole size of the dataset, a Python script was used to search through the gathered JSON files to determine the total amount of words per year for each blog. Therefore, we also have diachronic word prevalence data.

Problematic Aspects of Using Blog Data for Prevalence Calculations

It seems like most methods have methodological downsides to them when they are applied in the field. The study of Johns et al. (2019), in some regards similar to our approach, used the textual data of published authors to assign individuals textual data for word prevalence calculations. Looking at the process of publishing a book, the drawbacks of such an approach become apparent as not only the author, but also editors and publishers have a say in the creation of the opus. Therefore, the measured language itself is not necessarily representative of only the author itself, somewhat contrary to the modern author-opus paradigm. Also, the language within novels is edited to a high-degree and might offer a wider range of words used leading to a difference in word prevalence between English in literature and 'everyday' English where not every utterance is fine-tuned close to perfection.

By creating our Blogspot corpus we aimed to create a data set that offered more authentic and less edited textual data which did not go through an extensive editing process (other than from the author himself).

However in regards to co-authorship, and in general authorship attribution, blog data is still problematic. The author of a published novel is still most responsible for the writings in it, while a blog could be co-authored by multiple people, in the best case by e.g. an institution which makes it transparent, but this does not always have to be overt.

The scraped blogs in our corpus are all in English, but the location of the authors themselves is not always mentioned on the blog. While we tried to exclude blogs that most certainly were not from the US, ultimately there are a lot of blogs in the corpus where we do not know where they originated from. Time constraints also played a role in this circumstance. Given the specified constraint, the blog corpus should be considered as a dataset primarily comprising blogs written by English speakers, with a potential bias towards individuals located in the United States.

Since blogs are less edited, there is also a bigger need for data cleaning, but one of the biggest advantages of it is the more creative use of language.

Case Study: Creative Language Usage in Blogs

When plotting diachronic prevalence data, we found that some words that have gone viral or have been invented recently were not to be found in Brysbaert's study (2019), while our study could not only show that these words are known, but due to our diachronic data we could show when the interest in these words spiked, according to the expected contexts.

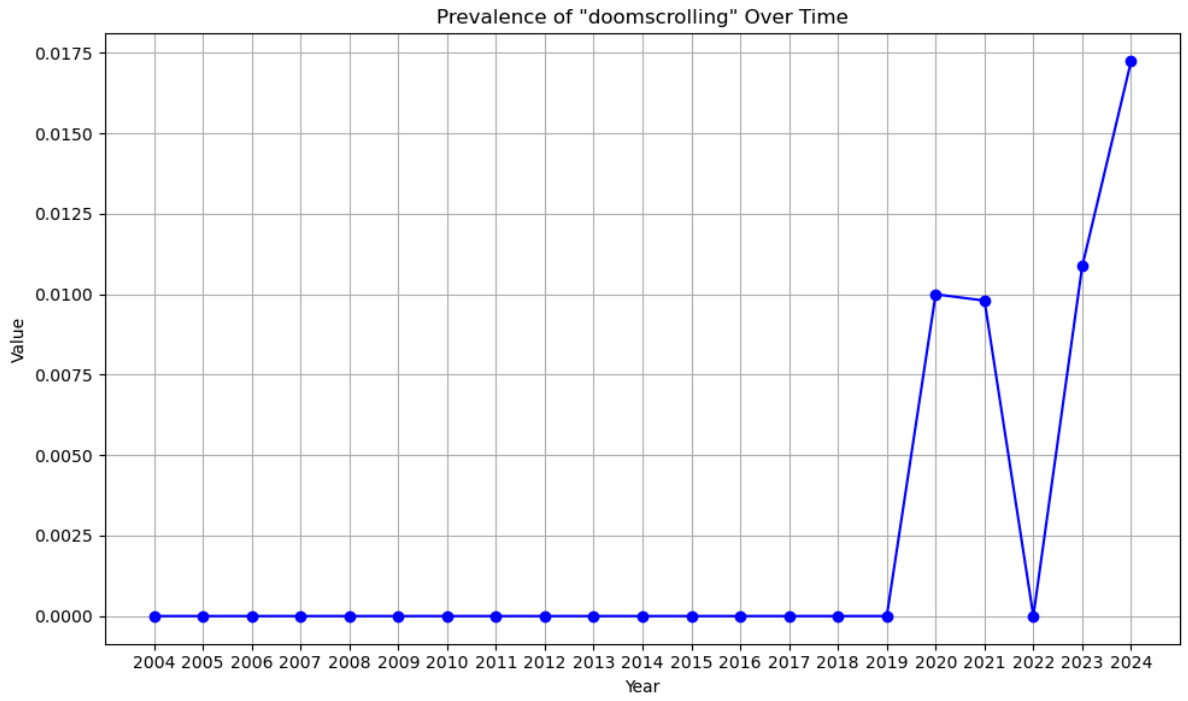

Table 1. Prevalence of *doomscrolling* over time.

Here is a chart of the prevalence of the word *doomscrolling* which is not part of the standard English vocabulary. This word has gone viral during the pandemic which is clearly depicted by this graph. Interestingly enough, in 2022 the prevalence score dropped down to zero, which we do not know how to explain, but it might be due to not having enough individual blogs to still measure when the interest in a certain word has come down a bit. The word is even relevant in 2024, even though 2004 and 2024 were the years which we actually excluded from our research because these ones did meet our data threshold requirements.

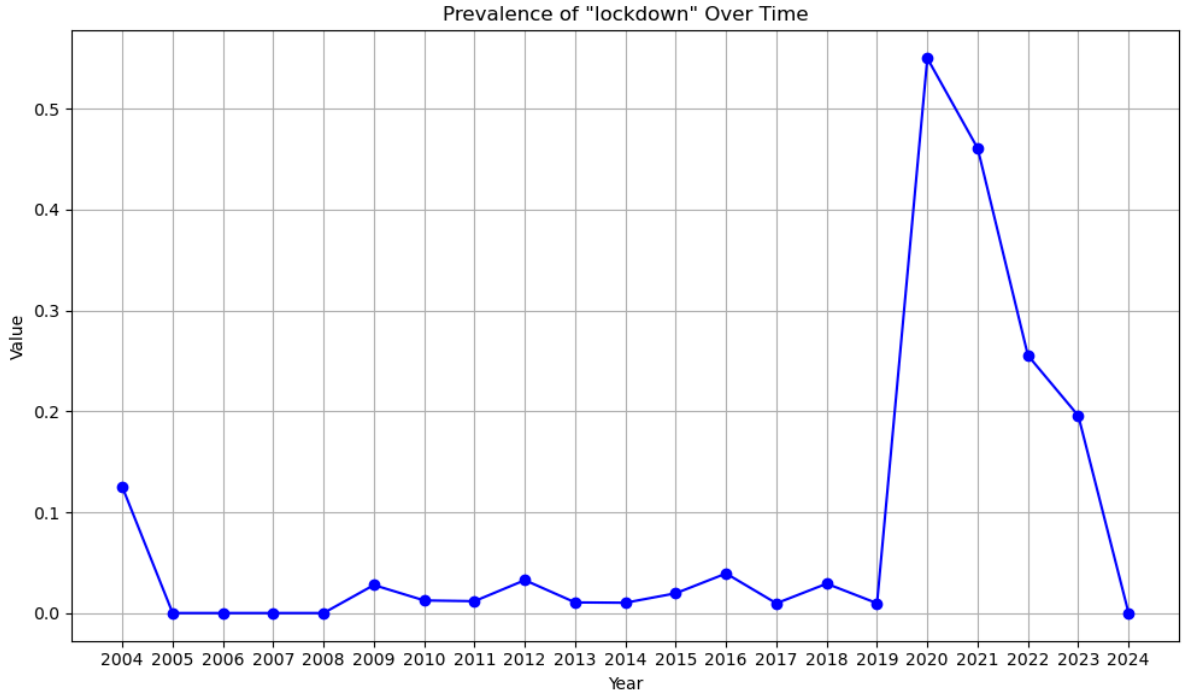

Table 2. Prevalence of *lockdown* over time

Another example is *lockdown* which always was somewhat prevalent in our data, but saw a drastic increase during the covid lockdowns. The same goes for a plot showing the prevalence of *pandemic*.

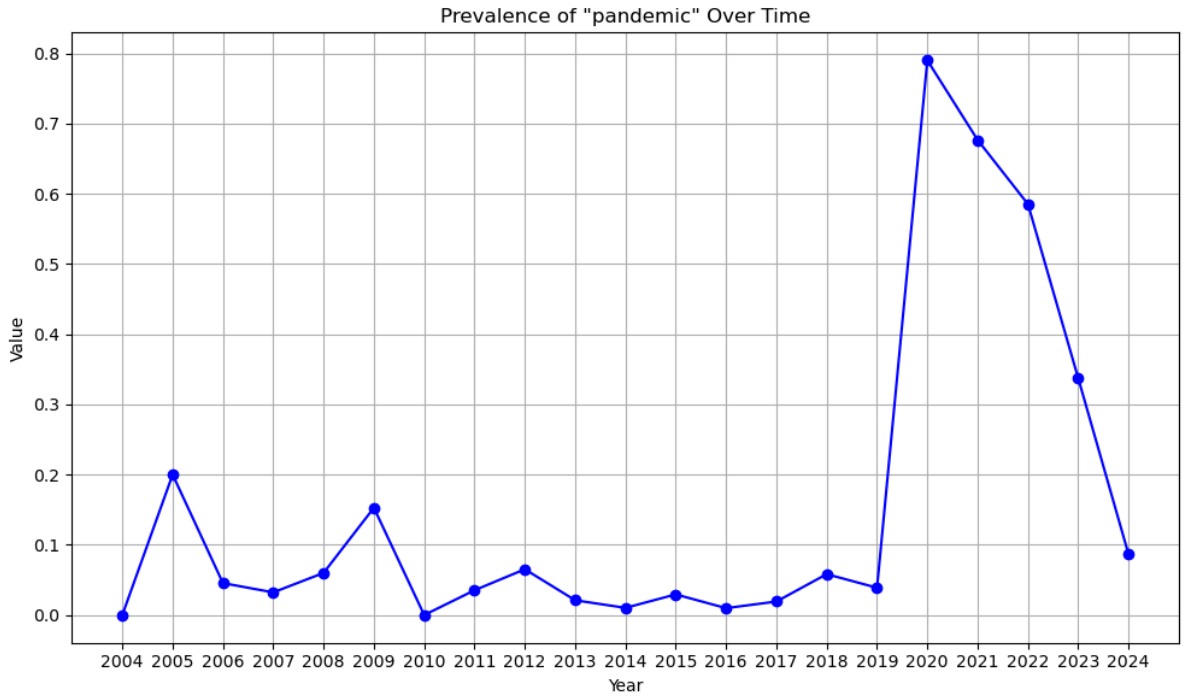

Table 3. Prevalence of *pandemic* over time.

Since the graphs of both *lockdown* and *pandemic* are so similar, it is safe to assume that there is a connection between the two words; as a matter of fact we do know that there is, of course – after the Covid-19 pandemic this is to be treated as a given – , but this shows that relationships between word prevalence scores of specific words could be potentially computed in further studies. One of such studies could e.g. cluster words based on prevalence values.

This shows that our Blogspot corpus has scientific merit:

Firstly, it contains creative colloquial English language data. One speculation as to why this is, is the nature of the text genre *blog*. Blogs can be conceptually oriented towards oral or written language, due to the wide variety of contexts blogs can have.

Secondly, our data seems to show when words became trendy and widely used as well as seen in the visualization of the prevalence of the words *lockdown* and *pandemic*. The measurement of their prevalence has drastically increased during 2019 and 2023.

The Meaning of Increased Measured Word Prevalences?

An interesting and not at all straightforward question is what it means if the prevalence value increases at specific points in time. Word prevalence has been defined as a measurement of how widely known a word is; the amount of people who know the word divided by the total number of people (e.g. Brysbaert et al. 2019).[2] However, if we are looking diachronically at the word *pandemic* has the value of measured prevalence truly increased because the word was more well known than before or was it due to the Covid-19 pandemic, a major rupture or turning point in many people's lives that warranted the increased usage of the word leading to a perceived increase in prevalence.

On the one hand, the word *pandemic* might have been more or less well known before, but there might not have been a reason to actively use the word, on the other hand, the knowledge of the meaning of the word *pandemic* might very well have increased drastically

[2] If a total of four people are asked whether they know the word *cicada* and only one person knows the word, then the calculated prevalence would be 0.25 or 25%. The only allowed values in this specific case would be 0.0, 0.25, 0.50, 0.75, and 1, which are 5 values total. The amount of allowed values is the number of people asked +1.

as well, too. This ambiguity that is displayed here leads to the question: What is actually measured here?

To further support the above claims that the actuality or trendiness of a word in the context of the living conditions matter in regards to our measurements of increased prevalence, the trajectory of the graph of Google Trends data shows that it is almost the same as our visualization.[3] The region that was looked at is the US in the timeframe from 2004 until today (roughly the end of March in 2024). The graph visualizes the relative frequency of Google searches in relationship to all other searches (for a more in-depth explanation c.f. the FAQ from Google: www3).

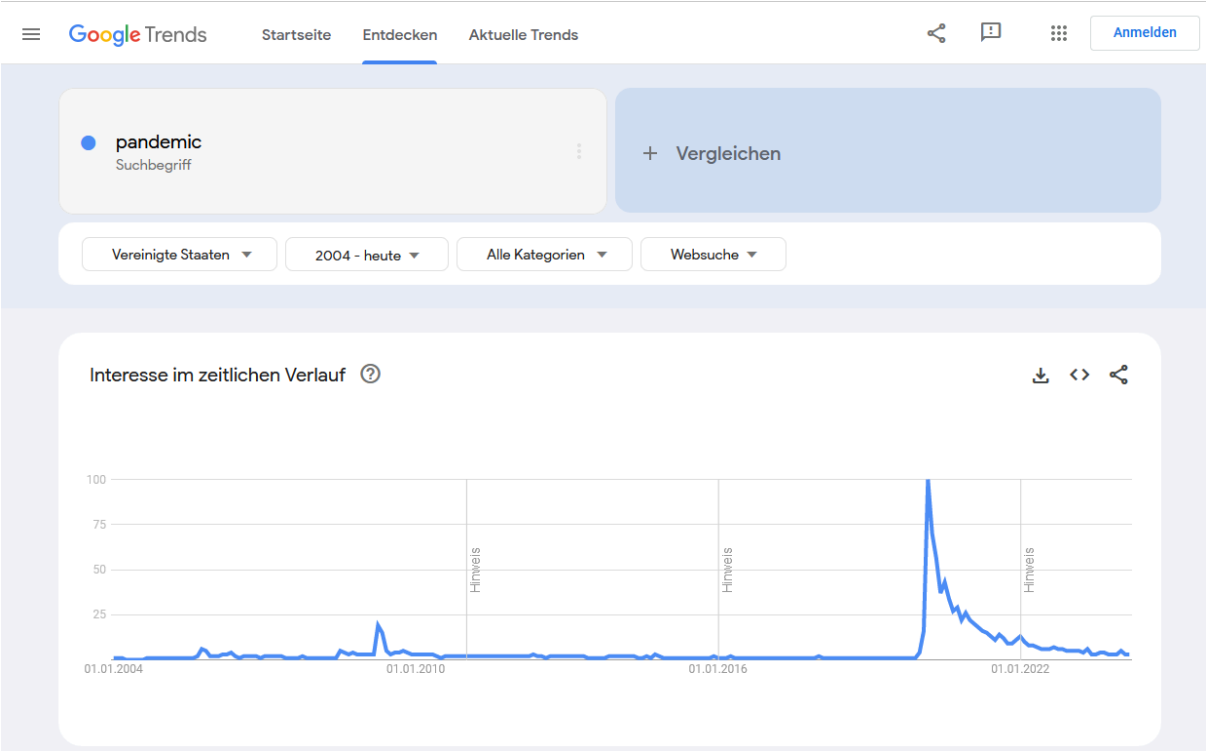

Table 4. Google Trends Data for *pandemic* from 2004 to March 2024.

Possible answers are only speculation at this point, and this situation shows that a lot more research is required, especially on the relationship of prevalence with other linguistic variables.

[3] This is not only the case for the word *pandemic*, but plotting Google Trends Data for the word *smartphone* in Python, using the Pandas and Matplotlib libraries, shows also a very similar graph trajectory, even though it is not as obvious as here.

Possibly, measuring word prevalence diachronically on the basis of a (blog)corpus seems to measure the frequency of when word prevalence can be measured, so to speak the **word prevalence frequency**.

To make this situation even more confusing, our data shows significant differences to Brysbaert et al. (2019)'s study in some instances, while in others that is not the case. Diachronic word prevalence for every blog in the range of 2004 and 2024 have been computed, but no overall word prevalence, due to time constraints.

Results and Comparison with Brysbaert et al.

All in all, our findings are available in two formats: One is the complete dataset of yearly prevalence calculations for all words contained within the text corpus post-processing and cleaning the original raw data. Secondly, we provide a comparison with the data collected by Brysbaert et al. (2019), spanning across 61.857 English words available in their dataset. Notably, some words are not present in our collected corpus, and thus do not appear in the file of the comparison. Further, certain words might only appear during specific years and not throughout the whole span of the collected data from 2004 to 2024 since these words might have not been used by the blog authors during those specific years. Lastly, words collected in our dataset but which are absent in Brysbaert et al.'s (2019) data are excluded from the Excel comparison file, but instead only available in the full dataset we collected since they are not needed for the direct comparison of the two datasets.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Word	Pknown	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
2	a	0,977169	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	aardvark	0,958525	-	-	-	-	-	-	0,0125	0,011765	-	-	0,010204	0,009804	-	-	0,009709	0,009709	-	-	-	-
4	aardwolf	0,212617	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	abaca	0,237374	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	aback	0,862974	-	-	0,136364	0,16129	0,06	0,138889	0,1125	0,129412	0,065217	0,073684	0,05102	0,068627	0,068627	0,115385	0,067961	0,07767	0,05	0,088235	0,042553	0,043478
7	abacus	0,927681	-	-	-	-	-	0,013889	-	0,011765	0,01087	0,010526	0,020408	0,009804	0,019608	-	0,019417	-	-	-	0,010638	0,021739
8	abaft	0,187328	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	abalone	0,691906	-	-	-	-	-	-	-	0,011765	0,01087	0,010526	-	0,009804	0,009804	0,019231	-	-	-	-	-	-
10	abandon	0,997354	0,625	0,6	0,545455	0,645161	0,52	0,527778	0,6375	0,611765	0,543478	0,547368	0,5	0,529412	0,529412	0,5	0,524272	0,446602	0,54	0,401961	0,414894	0,391304
11	abandoned	1	-	-	-	-	-	-	-	-	0,01087	0,010526	-	-	0,009804	-	-	-	-	-	-	-
12	abandonee	0,660221	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	abandoner	0,863861	-	-	-	-	-	-	-	-	-	-	0,010204	-	-	-	-	-	-	-	-	-
14	abandonment	0,990453	-	0,066667	0,136364	0,064516	0,08	0,027778	0,0875	0,082353	0,108696	0,073684	0,071429	0,088235	0,058824	0,057692	0,087379	0,07767	0,07	0,058824	0,053191	0,021739
15	abase	0,75	-	-	-	-	-	0,013889	-	-	-	-	0,010204	0,009804	0,009804	-	-	-	0,01	0,009804	0,010638	-
16	abased	0,786486	-	0,066667	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	abasement	0,811275	-	-	-	-	0,02	0,013889	-	0,011765	-	-	-	-	-	-	-	-	-	0,009804	-	-
18	abash	0,887446	-	-	0,045455	-	0,02	-	0,0125	0,011765	0,01087	-	-	0,009804	-	-	-	-	0,01	0,019608	0,010638	-

Table 5. A snippet of the computed comparison file.

At the end, it was possible to compare the results we achieved to the data which Brysbaert et al. collected during their research. The prevalence scores we computed for each word ranged

between 0 and 1, depending on the word usage per year. With that, there are certain fluctuations during the decades which also implicitly shows how word usage evolved over the years, as shown in the case study above. For the comparison, we utilized the *Pknown* values from the results gathered by Brysbaert et al. (2019) and compared them with our yearly results for the corresponding words. These comparisons are stored in an Excel-file which is made available alongside this project.

Some interesting details can be gathered from the comparison results. Certain results tend to vary quite drastically for some words, while others are more in line with the results of Brysbaert et al. (2019).

One example of that is the word *toothbrush* which is a very common and well-known word with a prevalence score of 0,993 as evaluated by Brysbaert et al. (2019). In our corpus, it also appears across all collected years apart from 2024, but with a considerably lower score ranging between 0,045 in 2006, and 0,167 in 2009. As always, one reason for that might be a small sample size. Even though the text corpus is considerably large, certain years have more collected data than others. This also coincides with the popularity of blogs and the internet as a whole which could be a possible explanation why the collected data is much lower for the early years, especially 2004, and most data is available from 2010 onwards. Additionally, it is harder to collect older blog data from the earlier days of the internet as much data from that era might have already disappeared over time.

Another possible explanation for the divergence between the results of the two datasets could be attributed to the different research methodologies used to collect the data. While Brysbaert et al. (2019) asked study participants if they knew a certain word, our approach encompasses blog users to actually have used the word in writing within their blog posts. Thus, while they might be familiar with a certain word (such as *toothbrush*) they might not have actively used such a word within the topics they write about as some of these well-known words do not appear often enough in writing. This is something to consider while looking at our research results. However, the data collected that way might also allow interesting insights in certain word usage patterns over time since the data is available with yearly fluctuations.

To further illustrate, another example of a very common English word, such as the word *want* which is ranked among the most used English words according to an analysis of the Oxford English Corpus (www4), has a high prevalence in both of the compared datasets.

In our case based on the yearly fluctuations it typically hovers around or well above 0,95. This reflects the appearance of *want* in written English.

The word *war* is also very well known (prevalence score of 0,995 according to Brysbaert et al.), but it demonstrates a slight divergence in our results, fluctuating between 0,478 in 2023 and 0,694 in 2011 (with the outlier being 2024 with 0,120, likely due to the lack of data for the current year).

Apart from the prevalence of individual words, our data seems to suggest that another important factor is if the word appears often enough in written English to gather a large enough sample size to adequately represent the results, especially comparing them with a dataset based solely on known words without taking their usage into account.

Conclusion

In our Data Analysis Project, we aimed for building a robust text corpus of blog data for measuring linguistic prevalence in written text. Through research and blog scraping we managed to achieve a robust size for our corpus, incorporating over 80 million words by more than 100 authors written over the span of 20 years.

Afterwards, we cleaned the collected corpus of raw data before we computed the prevalence scores using said data and compared them with the results gathered by the research of Brysbaert et al. (2019).

We have shown that diachronic prevalence data offers valuable insight to word usage over time and the emergence of words.

Furthermore, the ambiguous nature of diachronic word prevalence in contrast to the study conducted by Brysbaert et al. (2019) was discussed, and future research regarding the relationship between other linguistic variables and word prevalence is a necessity.

With all of this and our discussed results for this research project, we are confident that we have contributed to valuable insights regarding prevalence and potential reasons for disparities when comparing datasets, attributing them potentially to variances in research methodologies regarding the data collection and sample sizes.

Additionally, our text corpus offers a large sample size regarding the total word count by a substantial number of authors contributing to said corpus. With that, we hope that making it available aids future researchers who are interested in researching blog corpus data.

References

Articles

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.

Sari, A., & Tosun, A. (2019). A systematic literature review on crowdsourcing in software engineering. *Journal of Systems and Software*, 153, 200–219.

Internetressources

www1 = SoSci Survey. https://www.soscisurvey.de/de/index [Last Access: 03.26.2024].
www2 = Prolific Academic. https://www.prolific.com/ [Last Access: 03.26.2024].
www3 = FAQ about Google Trends Data. https://support.google.com/trends/answer/4365533?hl=en [Last Access: 03.26.2024].
www4 = The OEC: Facts about the language. https://web.archive.org/web/20111226085859/http://oxforddictionaries.com/words/the-oec-facts-about-the-language [Last Access (by WayBackMachine): 03.26.2024].