NLP for Digital Humanities: Processing Chronological Text Corpora

Adam Pawłowski
University of Wrocław
pl. Uniwersytecki 1
50-137 Wrocław, Poland
adam.pawlowski@uwr.edu.pl

Tomasz Walkowiak
Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego St.
50-370 Wrocław, Poland
tomasz.walkowiak@pwr.edu.pl

Abstract

The paper focuses on the integration of Natural Language Processing (NLP) techniques to analyze extensive chronological text corpora. This research underscores the synergy between humanistic inquiry and computational methods, especially in the processing and analysis of sequential textual data known as lexical series. A reference workflow for chronological corpus analysis is introduced, outlining the methodologies applicable to the ChronoPress corpus, a data set that encompasses 22 years of Polish press from 1945 to 1966. The study showcases the potential of this approach in uncovering cultural and historical patterns through the analysis of lexical series. The findings highlight both the challenges and opportunities present in leveraging lexical series analysis within Digital Humanities, emphasizing the necessity for advanced data filtering and anomaly detection algorithms to effectively manage the vast and intricate datasets characteristic of this field.

1 About Digital Humanities

Digital humanities (DH) today is a broad domain of research and practical applications of various techniques for automatic processing of data, representing linguistics, literary studies, history, art history, cultural anthropology, or archaeology, among others. It can be defined as a system of interrelated resources, functionalities, and cognitive practices created by transferring to the digital realm and creatively expanding the heritage of the humanities that grew out of print culture (Terras et al., 2013; Schreibman et al., 2008; Sinatra and Vitali Rosati, 2014). However, behind the apparent plethora of different descriptions of digital humanities lies the same recurring set of characteristics, among which the most important are: the study of large data sets, the linked open data approach, the extensive use of metadata, the fusion of natural and artificial intelligence, transdisciplinarity, multimedia, an almost radical empiricism, and last but not least, the dominance of interactive, dynamic infrastructures over the static products of the Gutenberg era, such as articles, chapters, or monographs.

Describing the current state-of-the-art of digital humanities is therefore a task that is all the more difficult because the discipline is undergoing a phase of rapid, dynamic expansion today. Moreover, the efforts of many DH centers focus on digitizing the resources produced by humanity in the past centuries and enriching them with the tools of computational intelligence. The obsolescence of DH definitions, theories, and research practices is hence rapid. Here, we adopt a research perspective in which digital humanities is a strand at the intersection of applied computer science and natural language processing (NLP). This assumption does not contradict the principles of HC, but it exposes the methodological aspects on which NLP's attention is focused, rather than the problem of digitizing and/or sharing resources.

2 Language and NLP vs Digital Humanities

Language has not been a major area of digital humanities in recent years. As the topics of papers submitted to major DH conferences[1] show, interdisciplinary topics dominated there, combining history, art history, cultural anthropology or archival science. The formats processed in DH were, in varying proportions, text, but also image, sound, geographic coordinates (used to formally represent geolocation of objects) and time (dates, hours). However, on a more general level of reflection, natural language preserves its privileged role in the humanities, and this is something that the digital world will not change. After all, language is the primary and universal tool of communication, shaping

[1] See https://adho.org/conference/, section 'Past Conferences'.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 105–112
November 16, 2024.

in the human mind a system of representation of the world and modeling the processes of thinking.

Taking the above statement about the privileged role of language in the digital humanities as correct, the potential for collaboration between humanists and language engineers is enormous. NLP applications start with simple text string processing tasks, including segmentation, disambiguation, and morphosyntactic annotation of lexemes, followed by calculating the frequency of any segments. A more advanced level of NLP applications involves generating segment relationships (lexemes or multiword units) that take the form of networks (e.g., wordnets), hierarchical structures (e.g., dendrograms) or point clouds. The next level of NLP application is semantics analysis, which is usually based on distributional relationships of segments (e.g., topic modeling) and/or uses neural networks. A specific class of textual resources is the so-called chronological corpora, which contain samples arranged on a timeline. Below is an example of the application of NLP methods and mathematical statistics to such a corpus.

3 Sequential Data and Chronological Corpora

By textual sequential data, we mean data representing consecutive segments of a corpus on a linear time axis. Representation here means the moment of text production, but is not related to style, genre, age of the writer, or context. Out of the wide spectrum of discourses, the best material to construct chronological corpora are media texts generated by the press, radio, television and all Internet formats. The press is of particular importance here, since it covers long time spans: media institutions in the modern sense have existed in various countries of the world for at least two hundred years. The granularity of the existing resources varies. The earliest texts may have daily or even monthly dates, whereas contemporary news is marked with minute or second accuracy. The chronology of texts can also be discovered in corpora of literary and applied texts, but the granularity in such a situation is at least annual, which makes it necessary to have really large volumes of evenly distributed data.

Sequential text data, which by analogy with time series can be referred to as 'lexical series', obviously have their weaknesses. When extracting electronic text from print, acquiring clean OCR output is a problem. Recordings from radio, television, and the Internet require transcription supervised by a human. Cleaning and curing the data thus adds significantly to the cost of producing press corpora suitable for chronological analysis. As a result, research on such corpora is not as developed as on large corpora of general language.

An important issue here is to distinguish chronological analysis from diachronic studies. The former deals with changes in the frequency of relevant lexemes at fixed intervals of time, while the latter describes the evolution of language forms over time. Diachronic research can be carried out with quantitative methods, but the nature of the phenomena under study is quite different: it is about change, disappearance and/or appearance of new lexical (rarely morphological or syntactic) forms. An example of a quantitative diachronic research is a study of lexical changes in the Polish language over a period of 600 years (Górski and Eder, 2023). One of the methods for quantitative modeling the dynamics of language change is Piotrovsky's law. This Soviet linguist with Polish roots noted that linguistic changes usually have non-linear patterns, resembling a logistic function (Leopold, 2005; Górski and Eder, 2023).

For the above reasons both approaches should not be confused: they deal with completely different problems, and the reference to time is only an apparent similarity. Bringing them together is all the more difficult because chronological research by NLP methods assumes stable orthography and does not require corpora, derived from long (preferably several hundred years) periods. In contrast, the object of chronological studies are corpora covering shorter periods with stable orthography.

4 Purpose of the Project and Test Data

The purpose of the project is to create a reference processing workflow for textual chronological data and to prepare NLP tools that would serve for lexical series analysis. The first step of the processing flow is to determine the characteristics of the data that are suitable for such analyses and to prepare the test material. The second stage involves the definition of patterns of lexical series, relevant to the needs of the humanities (in particular, linguistics and cultural anthropology), and the preparation of algorithms for extracting such patterns (a priori approach). For some lexemes (represented by lexical series), this stage may include the estimation of trend models and/or stochastic processes if the se-

ries contains periodic oscillations. The third stage involves conducting an unsupervised taxonomy of lexical series, leading to the empirical extraction of semantic classes (a posteriori approach).

As to the first issue, sequences of text samples of the same volume, produced at equal intervals of time (the benchmark example of such a source is the press) should be considered the data best suited for chronological analysis. Additionally, the volume of such samples must be large enough to generate statistically significant lexeme frequencies. Equal sample lengths eliminate in the simplest possible way a troublesome feature of linguistic systems, which is that the dependence of the frequency of the vast majority of lexemes and the volume of the sample is not linear, and, in addition, for each lexeme follows a slightly different curve (in simple terms, if the frequency of a lexeme L in a sample of volume N is L_i, it does not mean that in a sample of $10 * N$ it will be $10 * L_i$ - actually it will be lower due to the constant appearance of new words). Thus, only comparing samples of the same length gives reliable and indisputable results.

As for the sought-after patterns in lexical series, they duplicate and minimally extend the patterns identified in classical time-series analysis, that is, stable trend and periodic oscillations (Box and Jenkins, 1976). In addition, random series and anomalies ("catastrophes") as non-deterministic patterns are also important in lexical series analysis. The first two types of series express either long-term processes of cultural change (trends) or periodic phenomena, driven by natural cycles (weather, agricultural work, seasonal diseases, etc.) and rituals of culture (anniversaries, holidays, cyclical political events). A non-deterministic pattern with significant informational value is an anomaly, i.e., a sudden jump in the value of a series, caused by some sudden event (the death of a well-known figure, a natural disaster, a change in the name of a great city, etc.). Interestingly enough, in the context of language and culture, researchers may also be interested in random lexical series.

The third stage of the lexical series processing flow potentially includes two modules. If periodicity of the series is detected, a process model (AR, MA, ARMA, ARIMA) can be estimated. However, in the digital humanities, the usefulness of such models is low and it is also difficult to interpret their parameters. As for trend estimation, only working on large datasets covering long periods gives reliable results. On the other hand, a much more interesting and so far unused approach in this respect is the taxonomy of series. Each lexical series can be treated as a vector representing a point in some multidimensional space. Thus, all lexical series can be projected into this space and can, after dimension reduction, generate human-readable dendrograms or point clouds. It can be expected that such a taxonomy will not be very transparent since placing tens or hundreds of thousands of points in a single space produces a result that is opaque and difficult to interpret. Nevertheless, lexical series (and thus lexemes) with similar frequency characteristics are likely to form at least some visible clusters.

The above processing workflow was tested on the ChronoPress corpus, which represents 22 years of Polish press from the period 1945-1966 [2]. This corpus has a volume of ca. 24 milions of segments, evenly distributed by year and month. The texts are lemmatized and chronologically annotated. One year is represented on average by 1,098,526 segments with a standard deviation of 69,843, and the average volume of monthly segments is 91,544 with a standard deviation of 6,201. The volume of the corpus thus allows for annual and monthly granulation. Here, monthly granulation was used, which made it possible to generate lexical series (or vectors) of 264 (=22*12) units in length.

5 Research Methods

In the case of linguistic data, the study of time series consists of two modules. The first comprises NLP tools necessary to generate the input data from the corpus, and the second includes numerical methods of time-series analysis and taxonomy. As for the NLP methods module, the corpus was lemmatized using the WCRFT2 tagger for Polish (Radziszewski, 2013). Each sample includes publication data, allowing us to calculate the number of lexeme occurrences in each month by summing the lexeme numbers obtained from samples published in that month, obtaining time series.

In terms of time series analysis, we rely on the standard model of Box and Jenkins (Box and Jenkins, 1976), exposing, however, the specifics of text corpora. The standard model assumes that any time series consists of a trend, periodic oscillations, and noise. The processing flow includes: (1) identifying the trend and cutting it off from the data,

[2] https://chronopress.clarin-pl.eu/

thus making the series stationary; (2) identifying periodic oscillations by calculating the autocorrelation function (ACF) and partial autocorrelation function (PACF). The shape of the ACF and PACF functions allows one to choose the optimal type and order of the model - autoregressive, moving average, or mixed (ARMA, ARIMA). After cutting off the trend and periodic oscillations from the time series, the residual series that remains should meet the criteria of white noise, and the percentage of explained variance indicates the contribution of the deterministic component to the series.

However, from the perspective of the digital humanities and text processing, some of the functionalities of the standard Box and Jenkins model are not as useful as in economics or engineering. This model, as the title of the cited work indicates (Time series analysis: forecasting and control), was created to predict and/or steer processes. So, while one can understand an economist trying to predict in advance the price of some raw material or the exchange rate of a currency, a humanist does not wonder what the frequency of some lexeme in a stream of media texts will be next month. Therefore, the module for modeling stochastic processes in lexical series of the type described here is not particularly important. An autoregressive model identified in a lexical series would at most show the depth of cultural or societal 'memory'. The previous application of this method in linguistics or textual studies confirms this statement, since the object was to explain some linguistic phenomena, not predict them (Pawłowski, 1997; Pawlowski and Eder, 2001; Mikros and Macutek, 2015). For the above reasons, the ACF and PACF functions should be considered key tools for the humanities to identify periodic phenomena in the great mass of data. This situation raises a fundamental challenge for NLP, namely the need to generate and filter data from a corpus. Chronological analysis of a corpus of texts assumes that there are as many time series as there are different lexemes in the corpus, and the task of the researcher is to identify among them those that are for some reason significant. Despite the great cognitive capabilities of human mind, this task is not feasible without the support of NLP. For example, in the ChronoPress test corpus processed here, the initial number of series was close to 100,000, and this is more likely to be the lower limit, since we are talking about a corpus with an average volume. An additional difficulty is that the anomaly (catastrophe) pattern can involve series that have subsequent values close to zero almost throughout the run but once their value unexpectedly spikes. An example of such a lexeme is 'comet'. Normally, the press does not write about comets, so the word is almost absent from the media discourse. But, like a real comet, it suddenly appears every few or a dozen years and has higher frequencies. Similarly, lexemes with relatively low frequency, and therefore irrelevant in the perspective of big data, can appear rhythmically. Therefore, the low average frequency of a lexeme is not a sufficient criterion for its elimination. The same applies to the variance of the series, which may be too low to indicate interesting cases of anomalies. To overcome these difficulties, we have developed algorithms for automatic series filtering and anomaly detection.

6 Results

6.1 Overview

From the ChronoPress corpus, a total of 99,528 lexical series with monthly granularity were generated, each with a length of 264 units. The initial issue that required attention was the normalization of the data in order to facilitate comparison. We divided each occurrence of a lexeme by the total occurrences of that lexeme. This allowed us to obtain the probability density function of lexeme occurrences over the analyzed time period.

6.2 Linear Regression

In the first set of experiments, we applied linear regression to normalized time series to calculate the slope and the coefficient of determination R^2 for each lexeme. Next, we have planned to identify the lexemes with the highest, closest to zero, and lowest slopes, representing those that were the most rising, flat, and descending, respectively. However, a decision needed to be made regarding the inclusion of all series in the analysis. Empirical evidence suggested that many series deviated significantly from linearity. Therefore, we excluded series with R^2 values less than 0.5, resulting in the retention of 138 lexical series. The results, including the 10 lexemes with the largest slope, the slope closest to zero, and the smallest one, are presented in Table 1. First of all, we can observe that the majority of lexemes (99.86%) are non-linear, with a coefficient of determination smaller than 0.5. Among those assumed to be linear, functional words, such as

rising	flat	descending
West German	my	self-help
specialized	such	rebuilding
television set	just	fascism
currently	a few	rebuild
set	very	fascist

Table 1: Lexemes with the most rapidly increasing, consistently flat, and sharply declining time series among linear ones (i.e., time series with an R^2 greater than 0.5).

old, at the same time, valuable, date, prove, put in, get to know, upbringing, beginning, leave

Table 2: The 10th most stationary lexemes. Those for which their time series yield the smallest p-values for the Augmented Dickey-Fuller test (the p-values for presented lexemes are approximately e^{-30}).

pronouns and adverbs, tend to have a flat shape, with a slope close to 0. The most rapidly declining words are those associated with World War II and the process of rebuilding after the damage of war.

The interpretation of this result is very positive from the point of view of the efficiency of the model. The ChronoPress corpus reflects the events of post-war Poland, where the trauma of World War 2 is very strong right after 1945, but is gradually dying out, replaced in the official propaganda by Cold War events (e.g., the Korean War). In contrast, the behavior of function words is completely different. Their frequency is largely independent of the sample size, which is the reason why they are massively used in stylometry. Our research has shown that they are also immune to the time factor: successive samples of texts on the timeline are stable in this respect. Finally, technology-related lexemes (e.g., television set) are trending upward.

6.3 Autoregresive Model

In order to detect lexemes with seasonal patterns, we applied the Partial Autocorrelation Function (PACF), which helps to determine the order of an autoregressive model. PACF assumes the stationarity of the underlying time series. We used the Augmented Dickey-Fuller (ADF) test to check the stationarity of the time series with a periodic component. Table 2 shows the most stationary lexemes, where the ADF test p-value is the smallest. These lexemes express cultural rituals (such as religious and national celebrations) and cycles of nature.

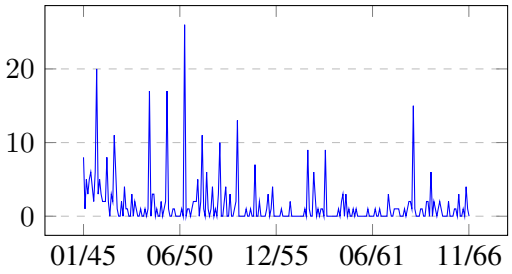

Figure 1: Occurrence of the lexeme 'Lenino' as a time series.

Christmas Eve, New Year, Lenino, August, Barbour, Christmas, April, May, September, Christmas tree

Table 3: The lexemes with the highest values of partial autocorrelation at lag 12 among those assumed to be stationary. The partial autocorrelation values range from 0.66 (for Christmas Eve) to 0.54 (for Christmas tree).

To choose stationary time series, we only consider lexemes with a p-value of <= 0.05. This allows us to reject the null hypothesis, indicating that the time series does not have a unit root and is stationary. Resulting in the retention of 93,039 lexical series. The lexemes with the highest PACF values for lag 12 are shown in Table 3. We can notice some expected events occurring once a year like Christmas Eve, New Year or name of months. However, Lenino looks at first surprisingly (refer to Figure 1 for Lenino time series). But, as the battle of Lenino (12.10.1943) was an important event for communist propaganda, the algorithm found it to be as oscillating as Christmas. The battle of Lenino was the the baptism of fire of Polish troops organized in the USSR, and from 1950 to 1991 it was celebrated as the Polish Army Day[3]. Of course the event had an important press coverage.

Table 4 presents the lexemes with the highest Partial Autocorrelation Function (PACF) values for a lag of 1. These lexemes are associated with the theme of postwar reconstruction in Poland. The high PACF values, ranging from 0.85 to 0.91, may indicate that these lexeme time series likely follow an AR (1) (AutoRegressive) process. This means that the present value of each series is primarily influenced by its immediate past value.

[3] https://en.wikipedia.org/wiki/Battle_of_Lenino

reconstruction, occupation, Poland, fascism, Polish, democracy, war, UNRRA, destroyed, allied

Table 4: The lexemes with the highest values of partial autocorrelation at lag 1 among those assumed to be stationary. The partial autocorrelation values range from 0.91 (for reconstruction) to 0.85 (for allied).

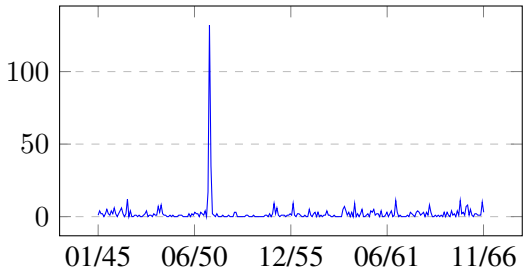

Figure 2: Occurrence of the lexeme 'plebiscite' as a time series.

plebiscite, Dzerzhinsky, senate, ratification
Communist Party of Poland, coffin, referendum
rally, constitution, pre-convention, Stalin
Indochina, Grunwald, Potsdam, capitulation

Table 5: The lexemes with the highest values of the catastrophe index.

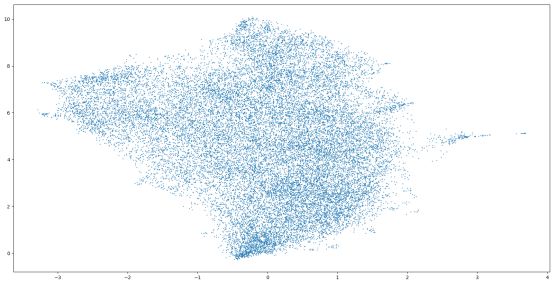

Figure 3: UMAP projection of lexems - represented by 264 dimensional vectors (default parameters of UMAP method).

6.4 Catastrophe Detection

In order to identify sudden spikes in a series, we calculate the difference between the current value of a normalized time series and a moving average with a length of 5. We have defined a catastrophe (anomaly) index as the maximum absolute value of this difference. An issue we needed to address was determining which series should be included in the catastrophe (anomaly) analysis. We decided to exclude series with a low sum (less than 500 occurrences in the entire series). After this exclusion, we were left with 6612 lexical series, which we then analyzed for anomalies. The lexemes with the highest values of the 'catastrophe index' are displayed in Table 5. The values for these lexemes range from 0.17 to 0.07.

An analysis (refer to Figure 2) of the time series for the anomaly of the lexeme 'plebiscite' reveals that it originates from the National Plebiscite for Peace held in Poland in May 1951 under the auspices of the Polish Committee of the Defenders of Peace (Dawid, 2018). The lexeme 'Dzerzhinsky' (peaking in June 1951) is linked to the 25th anniversary of his death and the unveiling of a monument in his honor in Warsaw.

6.5 Clustering and Dimension Reduction

In the next step, we cluster the analyzed lexemes using the normalized time series as vectors. We have utilized the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm (Campello et al., 2015). However, it did not detect any distinct groups (only one, large group was identified). Modifying the default values of the HDBSCAN parameters and trying different metrics (the default being Euclidean) did not yield different results. To better understand the issue, we employed the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) technique to reduce dimensions and visualize the similarities in the data. The results are presented in Figure 3. It is clear that UMAP does not reveal any distinct groups in the data, suggesting a relatively equal distribution of vectors in multidimensional space. This is probably why HDBSCAN was unable to identify any well-defined clusters in the data.

Should this result be evaluated entirely negatively? In our opinion, the method described above is not inefficient per se, but it may well be that it performs well when processing data are structured differently (smaller granularity, higher volume of samples, representing years or even decades).

6.6 Similarity of Lexems

A distance metric between vectors defined by normalized time series could be used to identify lexemes that have a similar 'longitudinal shape', i.e. the pattern of their distribution over time line. The results for exemplary lexemes are displayed in Table 6. Some of the results are predictable, such as the similarity between 'harvest' and semantically related terms. However, others, such as the pairing of 'worker' with 'fight' or 'forefront' reveal

war	**harvest**	**flood**
Poland	harvesting	simple
government	July	flooded
representative	harvest like	touch
series	July	loss
camp	harvester	water
association	rye	dear
Polish	June	last
area	grain	height
huge	barley	bridge
state	cartload	none

Stalin	**Gomułka**	**worker**
leadership	Władysław	female worker
Stalinist	tendency	fight
brilliant	self-gover.	mass
successor	common sense	association
proletarian	of course	fighting
Leninism	environment	people
invincible	opinion	working
Leninist	surely	forefront
leader	demand	factory
generalissimus	view	segment

Table 6: Lexeme similarity results. The lexemes most similar to the exemplar lexemes (those in bold) in order of increasing distance defined over normalized time series vectors.

specific features of the communist propaganda discourse of the post-war period not present in the general language.

7 Conclusions

While traditional time series analysis seems to be a task of fortune-telling (who wouldn't want to know next week's stock prices...), lexical series analysis is akin to looking for a needle in a haystack. In the research presented here, the 'needle' was the time series, containing the trend, periodic oscillations and anomalies, and the haystack was the corpus of 24 million words, divided into 264 sections. In addition, we were looking for lexemes that display similar shapes of time series.

The techniques elaborated during our research allowed us to identify automatically a set of lexemes from the corpus of nearly 100,000 that we found relevant in some way. In particular anomaly / catastrophe detection helps us to pinpoint lexemes that undergo rapid changes in occurrence, with some unexpected cases like "plebiscite" or "Dzerzhinsky". Another promising technique involves detecting lexemes with similar time-series patterns. The semantic similarity between words has been a foundational concept in modern NLP, based on deep neural networks and generative models. The Transformers architecture (Vaswani et al., 2017; Devlin et al., 2019) originates from the word2vec method (Mikolov et al., 2013), which creates word representations using large data sets and word occurrences in similar contexts. The method presented here, utilizing normalized time series, shares similarities with word2vec, as it constructs vectors from a large corpus. However, the novelty of our approach is the reliance on co-occurrence in recurring units of time (e.g., years) rather than cooccurrence in the text. The results shown in Table 6 reveal natural recurrence patterns, such as "harvest" and "July," but also provide insights into the communist perspective on the world, identifying semantic clusters typical for the totalitarian propaganda of the communist period. Last but not least, the great advantage of the method developed is that it is language-independent – any chronological (longitudinal) corpora can be processed in this way.

Limitations

The processing workflow presented here was developed and tested on a single corpus of Polish. However, linguistic aspects seem to be the easiest to overcome, as the workflow can be easily extended to other languages by using appropriate, language-specific taggers, such as those available from the spaCy framework (Honnibal et al., 2020). The main limitation of a reliable analysis of chronological corpora is their volume and time distribution. To derive statistical patterns, the corpus must be sufficiently large and balanced in terms of the distribution of analyzed time slots over time. While large corpora of contemporary language are numerous, balanced coverage of long time periods is rare. And the essence of chronological analysis is precisely to describe the "long duration" – far longer than the Internet era – which allows the correct identification of events, processes, cultural phenomena, etc. Practice shows that the most effective method of expanding chronological corpora is to scan the press, which significantly increases the cost of such an activity. Filling in data gaps could be achieved by subsampling underrepresented time periods, but this would result in data loss.

Another limitation is the number of parameters that must be set to use the proposed methods, which can influence the results obtained. These parameters include the minimum value of the coefficient of determination in linear regression analysis and the minimum size of the analyzed time series in the case of anomaly (catastrophe) detection.

Ethics Statement

The data and resources used in this study contain no sensitive data, they are publicly available and have been used in other researches.

Acknowledgements

The work was financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure" (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01.

References

George Box and Gwilym Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day.

Ricardo Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1).

Adriana Dawid. 2018. Organization and running of the national plebiscite for peace in opole voivodeship. *Rocznik Ziem Zachodnich*, 2018(2).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rafał L. Górski and Maciej Eder. 2023. Modelling the dynamics of language change: Logistic regression, piotrowski's law, and a handful of examples in polish. *Journal of Quantitative Linguistics*, 30(1):125–151.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Edda Leopold. 2005. Das piotrowski-gesetz. In *Quantitative linguistik – quantitative linguistics. Ein Internationales Handbuch*, page 627–633. de Gruyter.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv: 1802.03426*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

George K. Mikros and Ján Macutek, editors. 2015. *Sequences in Language and Text*. De Gruyter Mouton, Berlin, München, Boston.

Adam Pawlowski and Maciej Eder. 2001. Quantity or stress? sequential analysis of latin prosody. *Journal of Quantitative Linguistics*, 8(1):81–97.

Adam Pawłowski. 1997. Time-series analysis in linguistics: Application of the arima method to cases of spoken polish. *Journal of Quantitative Linguistics*, 4(1-3):203–221.

Adam Radziszewski. 2013. A tiered crf tagger for polish. In *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, pages 215–230. Springer Berlin Heidelberg.

Susan Schreibman, Ray Siemens, and John Unsworth. 2008. *A Companion to Digital Humanities*. Wiley Publishing.

Michael Sinatra and Marcello Vitali Rosati, editors. 2014. *Pratiques de l'édition numérique*. Parcours numériques. Les Presses de l'Université de Montréal, Montréal. OCLC: 870269451.

Melissa Terras, Julianne Nyhan, and Edward Vanhoutte. 2013. *Defining Digital Humanities: A Reader*. Ashgate Publishing Company, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.