ChronoPress – Chronological Corpus of Polish Press Texts (1945–1962)

Adam Pawłowski
Faculty of Philology
University of Wrocław,
Poland
adam.pawlowski@uwr.edu.pl

Abstract

This contribution aims to introduce the main characteristics and functionalities of the ChronoPress corpus. It consists of three parts. First will be presented some definitions and goals of chronological text analysis. Then selected web applications provided with tools of sequential analysis will be shortly reviewed. These theoretical and "state-of-the-art" introduction will be followed by the demonstration of the functionalities of the ChronoPress web service, such as: time series, quantitative analysis, semantic word profiles, lexical maps. At this stage, different case studies will be examined. The final part will include a discussion of the current state of the web service and its possible/potential future development.

1 Introduction and goal of the project

The goal of the Chronological Corpus of Polish Press Texts (henceforth ChronoPress) was to create an open-source, searchable text resource consisting of carefully selected samples of Polish daily press, structured according to the publication time of these samples, and annotated with relevant metadata. The introduction of the time variable has allowed the users to discover and explore the dynamics of the events and phenomena represented in daily press over long periods of time. The user of the ChronoPress web service has been also provided with effective statistical tools of text analysis "in the mass" and "in the line" (using Gustav Herdan's terminology). In order to facilitate the extraction of knowledge from the data, the Corpus has been annotated morphosyntactically in accordance with CLARIN standards and provided with a hinting user interface (Piasecki 2007).

At the moment, ChronoPress covers the period from 1945 to 1962, but it is intended to be extended to one hundred years, starting from 1918, when Poland regained independence. The minimum time spans visible from the level of user interface are subsequent months (there is a possibility to aggregate them into subsequent years). Every month is represented by approximately 100‘000 text words, i.e. 1.2–1.4 million words per year. The corpus has a strong orientation towards the needs of users in the humanities and social sciences; in other words, it is aimed not only at linguists, but also at historians, anthropologists and political scientists.

The project has been realised at the University of Wrocław in collaboration with the Technical University of Wrocław in the framework of CLARIN-PL consortium. Almost all the technological solutions (e.g. tagging, named entities recognition, frequency counts) have been developed by the Clarin team. Some modules of its actual version (http://chronopress-test.clarin-pl.eu/) are still under development, but are expected to be operational by August 2017 (including the English version).

2 Previous and actual research

Time variable and time-series tools have not yet been effectively implemented in/on? the existing text corpora, including the Polish ones. The first attempts at such systems were made in France in the 1970s on the material of the Trésor de la Langue Française (Brunet 1981). Some theoretical and em-

pirical research of serial phenomena in the analysis of political texts was conducted by A. Salem (1987). His works have shown a great potential of this approach, relevant to linguistics, political science, and information extraction techniques. However, no (or only few) technological solutions have followed his research. At the moment, there exist just a few chronological corpora of popular reviews or magazines, where electronic data are being used as a sort of a by-product of their actual or past commercial mainstream activities (e.g. Time Magazine Corpus); some new text corpora are provided with functionalities of basic sequential analysis (e.g. Deutsches Text Archiv, Hansard Corpus of British Parliament speeches from the period of 1803-2005). A new generation of innovative products, which take account of the time variable is represented by DiaCollo, developed by Bryan Jurish in the framework of Clarin (https://www.clarin.eu/showcase/diacollo). One of the latest, relatively innovative and publicly available resources equipped with chronological tools is the Google n-gram Viewer, which displays raw frequencies of n-grams in vaguely defined corpora of English or other languages (Google Books). As for the tools offered by Google, despite their huge scope they have numerous flaws:

– they do not rely on precisely preselected material (as the data underlying displayed histograms are uncertain, no reliable inference can be based upon them);
– representation of medium or "lesser-spoken" languages is weak and very imprecise, thus practically ineffective;
– no linguistic tools appropriate for inflectional languages are applied in text processing, which eliminates Polish and other Slavic languages as reliable research objects;
– no quantitative data is available for deeper analysis of trends, oscillations or other parameters.

In conclusion, none of the above mentioned resources can fully serve scientific purposes. Especially Polish press published before 1989 remains fairly inaccessible to scientists who prefer digital tools, capable of producing derived information, such as statistics, graphs etc. The existing tools of chronological analysis are dedicated to mathematical or economic applications (e.g. market trends or exchange currency rates analysis), where sequential data in a numerical form are generated independently.

3 Corpus description

The ChronoPress corpus has been created using the representational method, which means that it reflects the basic features of the political discourse of the time. It covers 18 years (1945–1962) of Polish press published during the post-war, so-called "communist" period [in the history of Poland]. Every year is represented by approximately 5700 text samples excerpted from a representative set of periodicals of large circulation. One sample is composed of ca 300 words (sentence boundaries are respected), so the size of the corpus is about 30 million words. The number of sampled newspapers ranges from about 20 titles right after the war, to 12 during the Stalinist period (from 1948 until mid 1956) and 15 afterwards. All samples are citations from press articles, so there is no need to negotiate copyright with countless individuals or institutions. Corpus users should be aware of all the fallacies of the "newspeak" of the communist propaganda. However, when processed statistically in great quantities, texts reveal processes and phenomena typical for any society undergoing technological and cultural change, corresponding to the development of other European nations, but shifted in time.

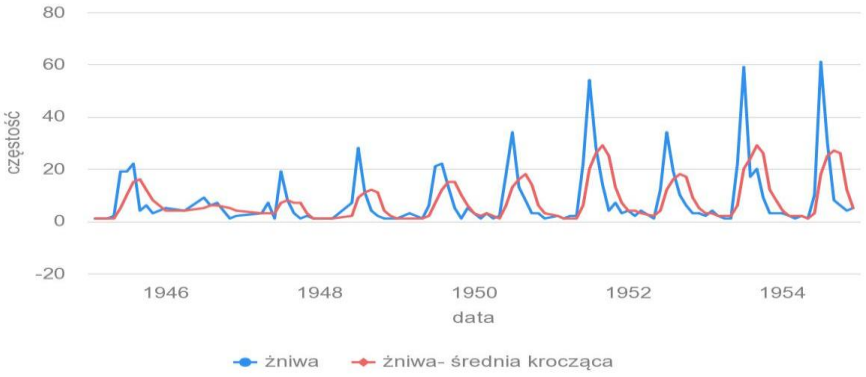
It needs to be said here that the notion of a chronological corpus is relatively new in corpus linguistics. It is defined as a sequence of text samples consistent in terms of their spelling and grammar, corresponding to subsequent points on the axis of time (e.g. weeks, months etc.) and annotated in terms of timing. A chronological corpus should not be confused with a diachronic one: in the former texts are evenly spread in time and word forms remain unchanged, while in the latter it is the opposite – word forms must evolve to become object of interest and time spans between measurements may be of any length (Pawłowski 2016).

4 Selected corpus functionalities

4.1 Time series

ChronoPress allows generation of time series of single word frequencies (Gottman 1981, Pawłowski 2001). It helps discover cultural trends in the data. In Figure 1 a cyclical time-series is displayed, representing natural sequence of field works (harvest follows sowing at more or less precise times of a year).

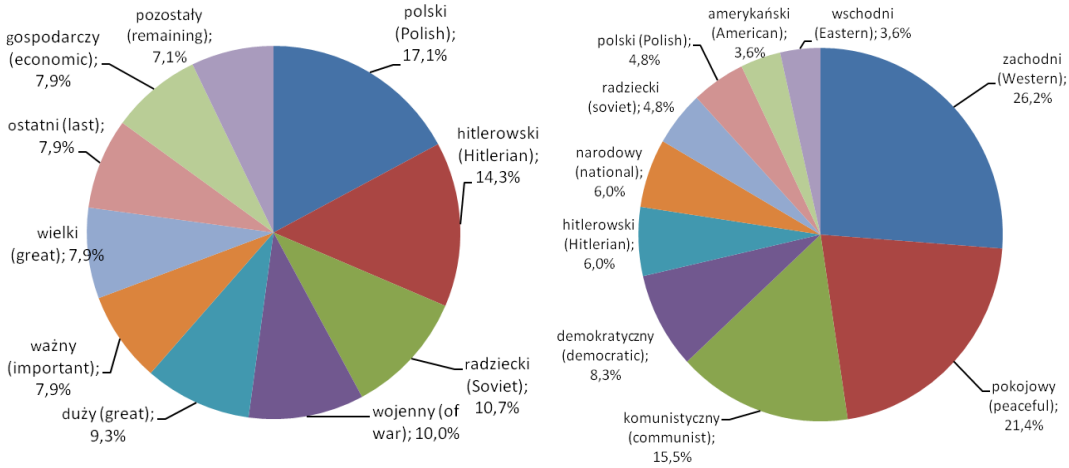
Fig. 1 Lexemes *siew* (‘sowing’) and *żniwa* (‘harvest’) in ChronoPress (1945–1954)



4.2 Word profiles

ChronoPress allows generating semantic word profiles. Below are two simple frequency-based profiles of the lexeme *Niemiec* (the noun ‘German’), which show different attitudes of the official discourse in Poland towards post-war Germany.

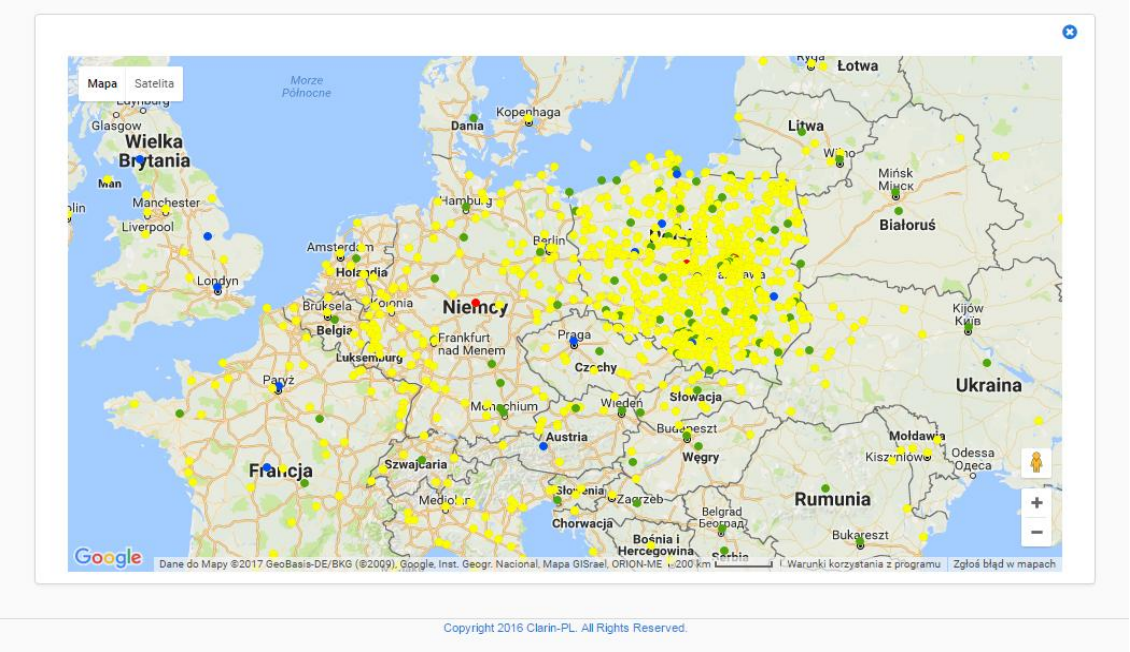
Fig. 1 Profiles of the lexemes *Niemiec* (‘German’) in 1945 and in 1954



4.3 Word maps

ChronoPress allows us to assign toponymes recognised in the corpus to points on a map (Google maps are used at the moment but more flexible and independent solution is planned for the future). Figure 3 shows a scan of a map generated for year 1945. The colourful points displayed on a map denote term frequency (yellow 1–10, green 11–100, blue 101–1000, red above 1000). They are mouse-sensitive: as hyperlinks they open word concordance of a given toponym which helps interpret the data.

Fig. 1 Map of toponyms from ChronoPress (1945)



5 Conclusion

ChronoPress is an ongoing project of a corpus and a webservice provided with analytic tools of sequential analysis and of text exploration of Polish press. It allows on-line extraction of linguistic patterns, as well as information regarding cultural phenomena (political campaigns, natural and/or cultural trends/cycles, sudden/catastrophic phenomena), if they are reflected in the daily press. Time series, quantitative parameters, word lists, concordances and lexical maps are generated to give users an easier access to the data found in libraries and archives and thus produce knowledge.

References

Brunet E. (1981), *Le vocabulaire français. De 1789 à nos jours*. Paris–Genève: Slatkine, Champion.

Gottman J.M. (1981), *Time-series analysis: a comprehensive introduction for social scientists*. Cambridge, London etc.: Cambridge University Press.

Piasecki M. (2007), Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *TASK Quarterly* 11, 151-167.

Pawłowski A. (2001), *Metody kwantytatywne w sekwencyjnej analizie tekstu* [Quantitative Methods in Sequential Text Analysis]. Warszawa: Uniwersytet Warszawski, Katedra Lingwistyki Formalnej.

Pawłowski A. (2016), *Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish*. In *Digital Humanities 2016*: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, 311-313.

Salem A. (1987), *Pratique des segments répétés. Essai de statistique textuelle*. Paris: Klincksieck.