

Chronological Corpus of Polish Press Texts

Adam Pawłowski

IINiB, University of Wrocław, Poland

Abstract

The paper describes briefly the project of the chronological corpus of press text covering the communist period in the contemporary history of Central and Eastern Europe. The kernel of the *Wrocław Corpus* consists of Stalinist texts published in the years 1945-1958. The Corpus will apply selected methods of time-series analysis to lexical series composed of word frequencies and spanned over different periods of time. The future users of the Corpus are linguists, sociologists, historians of contemporary Europe, cultural anthropologists, political scientists and students in the history of totalitarianism.

Keywords: text corpus, press texts, sequential analysis, lexicometry, communism

1 Introduction

The practice in corpus linguistics has been to build large collections of textual data implemented with basic search tools designed mainly for linguists. The chronological corpus of Polish press texts, further referred to as the *Wrocław Corpus*, will incorporate these functions, but apart from that it will also support additional functionalities. The features distinguishing it from other corpora will be the following: 1) It will expand beyond rather tight boundaries of linguistic applications (the Corpus will be a knowledge base, not just a source of linguistic data); 2) It will introduce chronological (sequential) dimension into the analysis of civilization, cultural, economic and political phenomena reflected in the media in a given period of time; 3) As the project stems from the Clarin philosophy, it will offer the possibility of multilingual extensions (inter-operability of resources).

2 Goal and Description

The authors of the *Wrocław Corpus* aim at creating a tool facilitating the study of the historical period of communist Poland and other countries of the former Soviet Bloc (1944-1989). New generations know very little about the history of Central and Eastern Europe from that period and science should contribute to the process of the widely understood social education, which uses modern tools of language engineering and mass communication. The Corpus will help reveal and closely examine the mechanisms of verbal coercion of the society by the official propaganda of the totalitarian regime. Consequently the creation of the *Wrocław Corpus* is not only a scientific project but also an educational one.

From today's perspective communism is hopefully a closed chapter in the history of Central and Eastern Europe. Therefore this period of time can be treated as one coherent entity. The Corpus, when completed, will be a sort of photograph of the official press language in the communist Poland. It will be further expanded by analogous modules representing the media of other "sister countries" in the same period. The first phase of the project will consist in creating a corpus that covers Polish press texts in the years 1945-1958. Since this period was abundant with momentous political, social and cultural events it will serve as excellent material to test the performance of the implemented tools of information retrieval and processing. After this phase is completed, a much cheaper expansion of the Corpus will commence, which consists in adding data from the subsequent and/or preceding historical periods. In time parallel chronological data modules in other languages will be incorporated, as well as texts originating from earlier and later periods (before 1944 and after 1989).

Text fragments will be randomly sampled from newspapers and periodicals with the highest circulation and proportionally distributed onto particular titles, which represent the official information and propaganda policy of the state (I presented the preliminary list of titles in the paper Pawłowski, 2006). It was assumed that every month will be represented by 100000 text words (1200000 text words per year). Since the average sample text length is 150-300 words, each year will be represented by 4000 to 8000 excerpts. Based on their printed editions these fragments will be digitalized, processed by OCR software, verified, and finally annotated.

The Corpus will be annotated morphosyntactically in accordance with the IPI PAN standard. Semantic analysis will be based on the word profiles extracted from the Polish WordNet and, if there is a possibility, on the implementations of other lexical resources (Derwojedowa et al., 2007). This will allow automatic retrieval and processing of the entire semantic nests (for example the lexeme "church" would point to other lexemes of the same semantic node, such as "faith", "believer", "priest", "primate", "Vatican", etc.). The samples will also be annotated chronologically in order to measure the dynamics of change in the coverage of the events both in long and short time spans (years, months, weeks). Statistical analysis will be based on the chronologically sorted frequencies of lexemes or sets of lexemes, defined by the user, which will appear as time series. This will allow the application of the advanced tools, such as trend estimation, as well as stochastic processes identification and modeling (Pawłowski, 2001).

3 Previous Research or Projects

For the time being the project remains definitely innovative and unique. So far no one in Poland and the world has created such a wide and representative corpus, which would allow analysis of chronologically sorted textual data. The only contemporary example which resembles this idea is the corpus of the "Time" magazine covering the years 1923-2007 with division into decades and years but not months.¹ This corpus was created a few years ago and made public only in June 2007. In

¹<http://corpus.byu.edu/time/>

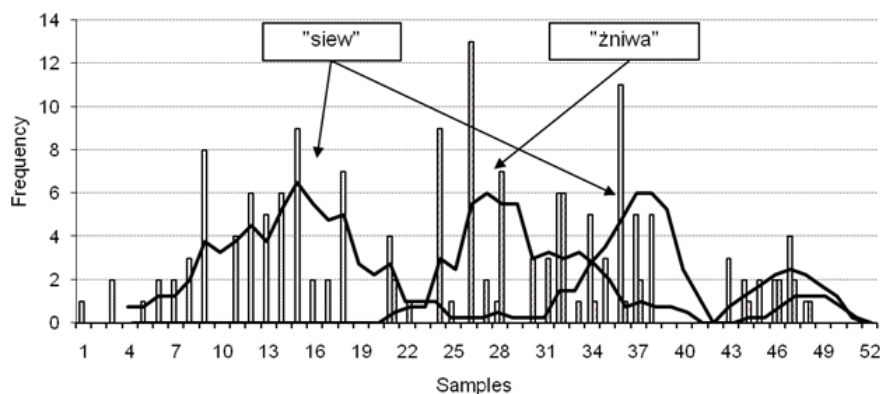


FIGURE 1: Frequencies of the lexemes *siew* (sowing) and *żniwa* (harvest), 1953, weekly samples

the case of the well known and time-honoured corpora, such as TLF, BNC and ARTFL, chronological tools were restrained to the visualisation of particular lexeme frequencies and in practice they were only an addition to the non-sequential search and processing tools designed to serve the needs of linguistic research (cf. Brunet, 1981).

4 Preliminary Tests

Preliminary analyses of chronological data, tested against a pilot sample from 1953 (“Trybuna Ludu”, “Sztandar Młodych”, agricultural press, a total of about 500000 text words) produced very good results and fully justify the need of creating this project (Pawłowski, 2006; Pawłowski, 2008). Below I quote an example graph of the dynamics of lexemes related to agricultural activity of sowing (*siew*) and harvest (*żniwa*) in 1953 in the journals mentioned above (Pawłowski, 2006, pages 22).

5 Target User

Given the scope and the amount of information included in the *Wrocław Corpus*, as well as the simplicity of the communication interface, the Corpus is aimed at various users including linguists, sociologists, historians of contemporary Europe, cultural anthropologists, political scientists and students in the history of totalitarianism.

6 Stages of the Project

The project will be carried out in several stages. The first is the OCR processing of excerpted text samples. Since the estimated average volume should be 6000 fragments per year (in form of digital photographs or miniscans), the generation of a common text format from graphics files will be automated and based on dedicated OCR software. In the next stage the digitalized samples will be verified. This will require proof reading of all texts by a human expert to correct possible scanning errors.

In the subsequent phase the samples will be annotated with meta-information tags and consolidated. In particular, they will be provided with metrical data, such as source name, as well as the publication year, month and week. These chronological details are quite important, as they will allow the generation, visualization and statistical analysis of what should be called **lexical time series**. The text prepared in this way will be annotated with the TaKIPI morphosyntactic tagger (Piasecki and Godlewski, 2006). When performing search tasks on the annotated text material, a user will have the possibility to freely apply the basic morphological forms – verb infinitives, nominative masculine form of nouns and adjectives. Once this stage is completed the Corpus will be provided with some basic semantic pre-processing tools, derived from the Polish WordNet. They will allow automatic identification and further processing of the entire lexical fields (determined by the relation of contextual proximity). The transformation of text into lexical time series, as described above, should be emphasized here, as it allows the application of various mathematical parameters and modeling tools, e.g. moving average function, series smoothing, autocovariance and autocorrelation functions, trend estimations and modeling of stationary stochastic processes. All these procedures, when implemented, may remarkably increase the explanatory power if the Corpus considered as a knowledge data base.

The final stage of the project will consist in creating a friendly and bilingual interface including, apart from the functionalities described above, help for beginners, description of the implemented procedures and references to works which use data from the *Wrocław Corpus*.

7 Extension Possibilities

The parallel political histories of the post-soviet Clarin member countries serve as a good reason to extend the *Wrocław Corpus* by materials extracted from their press texts. This would guarantee the interoperability of resources, as recommended by Clarin. One should also consider comparing the resources based on the data from the totalitarian countries with the analogous media information stream from the democratic European states. This would allow one to visualize similarities and differences between the two opposing halves of the post-Yaltian Europe, as well as the tendencies in entire regions of Europe, not only in individual countries. The experience of the Wrocław team with the creation of the kernel of the Corpus would place Poland in the lead of this extended project.

8 Corpus vs. Digital Library

It should be pointed out that text corpora are not digital libraries. The function of a digital library is to collect, store and make available digital documents, which originally existed in a printed form or as voice and image recordings. In these retrospective collections the stress is laid on the database or catalogue functionalities, while text content remains inaccessible for the search engines. A corpus, on the other hand, contains numerous selected excerpts of longer texts, which make up a coherent and most often very complex structure rather unsuitable for reading. They represent historical epochs, styles, functional types, press titles, literary works of art, authors, etc. and in the case of so-called national corpora, a certain abstract entity referred to as general language. In accordance with the principles of the representative method corpus data permit inductive inferences concerning language system and a hypothetic infinite population of texts (this is one of the objectives of the corpus linguistics).

Irrespective of the well known limitations of reliability of the inductive reasoning, observed language regularities refer indirectly to the extralinguistic reality. This is seemingly the point where the functions of the libraries and text corpora converge. The difference between them is, however, huge. Corpora help discover and characterize significant events, developmental trends, customs, internal structures and hierarchies in a given community or other reality represented by the available texts. This information usually remains hidden for an individual scientist, unless he approaches the data equipped with semantically oriented retrieval tools.

It should be finally underlined that full digitalization of printed press texts from the countries of Central and Eastern Europe will not be possible in the predictable future and most likely will not take place over the next few decades. The libraries in these countries are currently facing the problem of retrospective digitalization and compacting of catalogues and therefore are not ready to deal with such challenges. The *Wrocław Corpus* seems thus the only way to generate and share digital information representing the press of this period with a significant number of users. Since the corpus tools are unique and dedicated to this specific task, only renowned academic institutions with competent academic staff are capable of creating them and maintaining over a long period of time.

9 Summary

The Corpus in the proposed form meets all the requirements for advanced research projects:

1. The scientific quality of project is high and fits high European standards;
2. The addressee of the Corpus data is the wide scientific community of linguists, social scientists, political scientists, contemporary historians and students of the humanities;
3. The mission of the project corresponds to the educational mission of Central and Eastern European countries, which includes the postulate of debunking the false beliefs about the totalitarian period (1944-1989);

4. Once the *Wrocław Corpus* is created, it will be a durable and reusable piece of work designed for subsequent generations of users and scientists;
5. The Corpus will cover a well determined and finite period;
6. The formula of the Corpus will remain open to new functional and chronological modules;
7. The structure of the Corpus is transparent and allows integration with similar tools created for other languages, which corresponds to the Clarin idea of building and harmonizing the European information resources.

References

- Etienne BRUNET (1981), *Le vocabulaire français. De 1789 a nos jours*, Slatkine-Champion, Paris and Geneve.
- Magdalena DERWOJEDOWA, Maciej PIASECKI, Stanisław SZPAKOWICZ and Magdalena ZAWISŁAWSKA (2007), *Polish WordNet on a Shoestring* in: G. Rehm, A. Witt, L. Lemnitzer, editors, *Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, April 11–13 2007, Universität Tübingen, pp. 169-178.
- Maciej PIASECKI, Grzegorz GODLEWSKI (2006), Effective Architecture of the Polish Tagger, in: Petr Sojka, Ivan Kopecek, and Karel Pala, editors (2006), *Proceedings of the Text, Speech and Dialog 2006 Conference, LNAI*, Springer, pp. 213-220.
- Adam PAWŁOWSKI (2001), *Metody kwantytatywne w sekwencyjnej analizie tekstu* [Quantitative methods of sequential text analysis], Katedra Lingwistyki Formalnej Uniwersytetu Warszawskiego, Warszawa.
- Adam PAWŁOWSKI (2006), Chronological analysis of textual data from the “Wrocław Corpus of Polish”, *Poznań Studies in Contemporary Linguistics* 41, pp. 9-29.
- Adam PAWŁOWSKI (2008), Ewolucja dyskursu medialnego okresu stalinowskiego [Evolution of media discourse during the Stalinist period], in: *Teorie komunikacji i mediów* (in print).