Joanna Bilińska, Monika Kwiecień, Magdalena Derwojedowa

Microcorpus of Nineteenth-Century Polish

Abstract In the paper, a 1M word corpus of Polish texts from the period 1830–1918 is described. The corpus was compiled to provide diversified linguistic data for morphological analysis, however several tests proved that it can be used as a versatile resource to identify various linguistic phenomena and trace their dynamics in regard to inflection, spelling or even syntax. It is divided into five equal subcorpora to provide stylistic variety: scientific texts for general public, news, feuilletons, fiction and drama. In order to conduct morphological analysis an analyzer made for contemporary texts was adapted, which can, therefore, process word forms that differ from contemporary inflection and spelling. In the paper, several experiments made with the use of the corpus are discussed.

Keywords Morphological analysis, spelling, 19th century Polish, corpus

1 Introduction

The aim of this paper is to present a 1M word corpus of Polish texts from the period 1830–1918, available as text samples and metadata files (http://www.f19.uw.edu.pl/download/korpus-f19-v1-0/).[1] A browsable version, using the Polish National Corpus Poliqarp engine (Przepiórkowski et al. 2012), is available at https://szukajwslownikach.uw.edu.pl.[2] Originally the corpus was compiled to deliver as much diversified data as possible for morphological analysis, any other research in diachrony or history of language being just an additional possibility facilitated by this project (cf. Derwojedowa et al. 2014a, b). The paper is organized as follows: in the first part we present the overall design of the corpus

1 This research was funded in the years 2013–2017 by the Polish National Science Centre grant DEC-2012/07/8/HS2/00570.

2 The instance of the corpus compiled for Poliqarp browser off-line is available at http://www.f19.uw.edu.pl/download/obraz-korpusu-1830-1918/.

Published in: Eric Fuß, Marek Konopka, Beata Trawiński, Ulrich H. Waßner (eds.): *Grammar and Corpora 2016.* Heidelberg: Heidelberg University Publishing, 2018.
DOI: https://doi.org/10.17885/heiup.361.509

(macrostructure), then we present the design of a sample (microstructure). In the next part, there is a discussion of some experiments conducted with the use of the corpus.

2 Corpus' structure

The corpus consists of 1000 samples of 1000 tokens each. The samples were divided equally into 5 subcorpora: scientific texts for the general public (1), news (2), feuilletons (3), fiction (4) and drama (5). This method differs from the choice of texts made for the Polish National Corpus (PNC, Przepiórkowski et al. 2012) and the corpus of Baroque-period Polish — KorBa (Gruszczyński et al. 2013; under construction), but such a division was well-tested on the small-scale corpus of *the Frequency Dictionary of Polish* (Kurcz et al. 1990). In the time span of our research, drama seems the best approximation of speech, but also the burgeoning vocabulary of emerging science, engineering and fast-changing social reality need to be taken into account. Tests such as cluster analysis and multidimensional scaling (Eder et al. 2013, cf. R-manual 2015) concluded that the texts are distinctively spread between styles (cf. Figure 1).

The overriding principle of the project was that first printed editions of texts written originally in Polish were included in the corpus. Some exceptions were applied in special cases (e.g. literary works first issued in episodes in a newspaper or a magazine; cf. Bilińska et al. 2016).

Most texts were acquired from digital libraries. Despite the rule of at least one sample per year in each subcorpus, the acquisition was a result of rather opportunistic guidelines: we searched sources with a text layer (e.g. plain text and/or layered djvu). If such a source was not available, which was the standard case for the earlier quarter of the period, we decided to OCR files in graphic formats (.jpg or .png).

The number of samples in a style for a given year never exceeds four. In the whole corpus, each year is represented by at least five samples but no more than twenty. The majority of years is represented by 10–13 samples with an average of 11 samples per year (cf. Figure 2).

3 Sampling the corpus

A sample comprises a couple of files: a fragment of continuous text, its metadata and a source graphic file (.png, .jpg, .djvu, .pdf, .tiff; cf. Figure 3). The excerpt — a proper text sample for research — is the most accurate representation of the source text. The footnotes, incomprehensible fragments, stage directions and

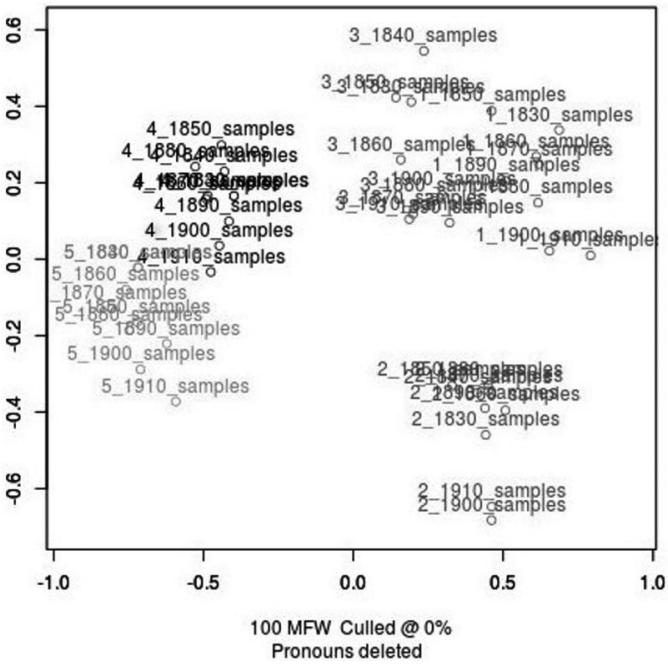


Figure 1. MDS grouping of styles (samples of each style merged by decades).

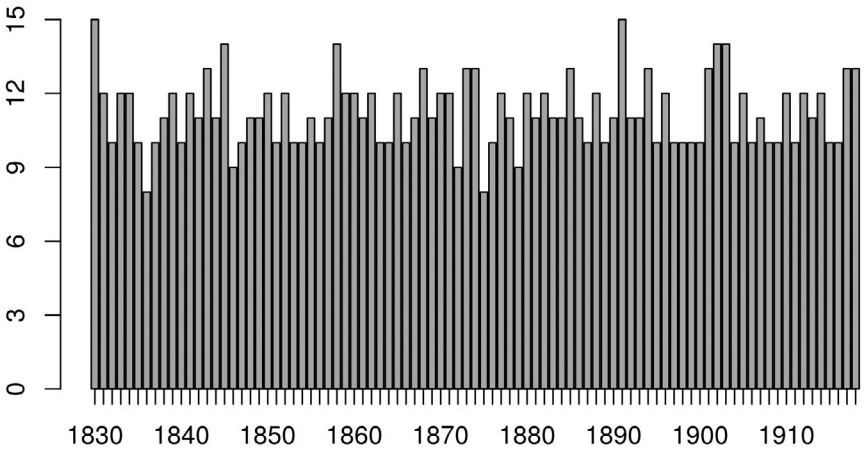


Figure 2. Number of samples per year.

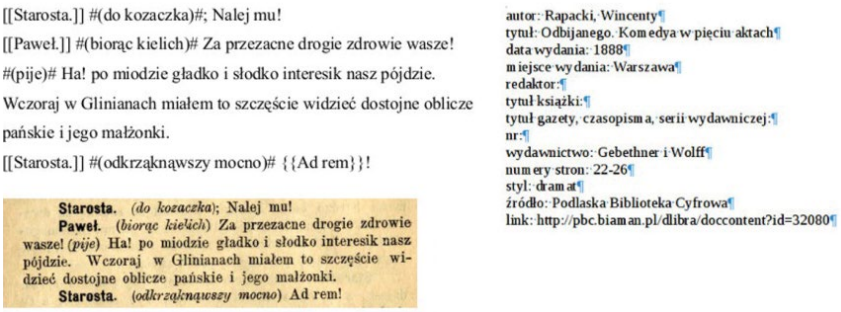


Figure 3. Text, metadata and source file of sample 1888_5.1.

small fragments in foreign languages, even misspellings were marked, but left unedited.

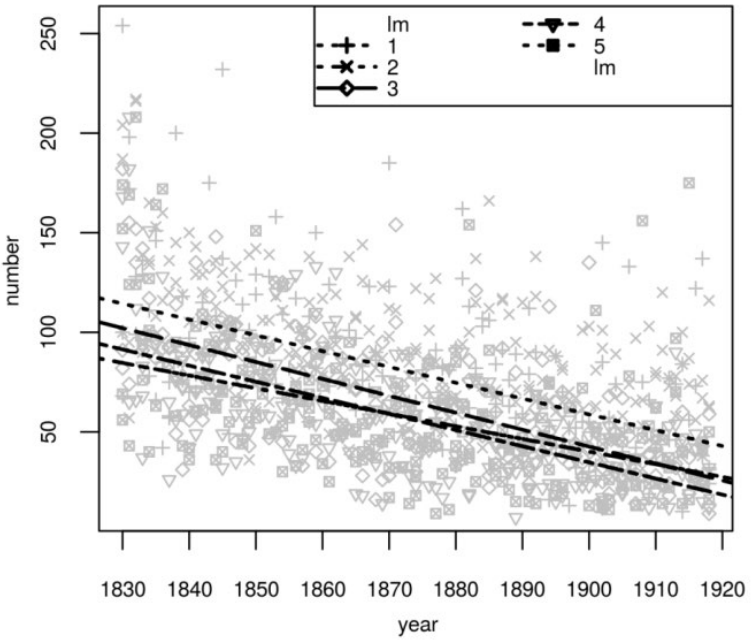
4 Diversity of the corpus

It is difficult to ascertain the exact number of authors without in-depth research (newspaper texts are often signed with initials or left unsigned; in the whole corpus there are 270 such samples), however there are circa 650 individual writers. Some are represented in more than one sample, but never in more than one style per year. In total there are 106 writers cited more than once.

Even though we struggled to create as diversified a collection of texts as possible, we did not select texts with respect to regional linguistic features. In effect, almost 2/3 of the texts were printed in Warsaw (almost 40 %), Lviv and Cracow. Together with texts issued in Paris, Vilnius, St. Petersburg and Leipzig they comprise almost 90 % of the corpus. The remaining 68 printing centers are represented several times and 39 of them – just once.

The majority of sources comes from big academic centers that undertook substantial projects of digitizing library archives. We used 43 such archives but 54 % of samples were excerpted from just three of them (Polish National Library on-line Polona, Warsaw University Digital Library, Digital Library of Wielkopolska).

The corpus is a resource of nineteenth-century Polish language indispensable for modifying a morphological analyzer in order to enhance its capabilities to analyze older texts. For this reason, we initially analyzed each sample and each subcorpus with an unmodified (i.e. trained on contemporary Polish) analyzer. Generally speaking, the number of unrecognized segments decreases with every newer sample and differs between circa 5 % and 15 % for a style and between 2 % to circa 25 % in case, respectively, of the best and the poorest sample in a given style (cf. Figure 4). The best results come from analyzing fiction, which can be



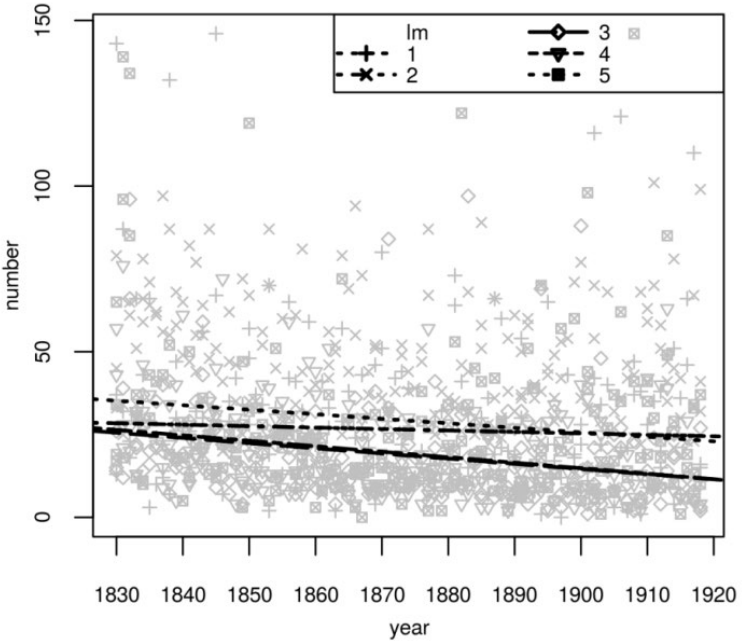


Figure 4. Unrecognized tokens in the 5 styles (1 is for science for general public, 2 —for press news, 3 — for feuilletons/journalism, 4 — for fiction, 5 — for drama), not modified analyzer. / Figure 5. Unrecognized tokens in the 5 styles (1 is for science for general public, 2 —for press news, 3 — for feuilletons/journalism, 4 — for fiction, 5 — for drama), modified analyzer (http://www.f19.uw.edu.pl/download-category/analizator/).

attributed to the fact that this type of language is mostly represented in dictionaries that constitute a base for any NLP device (cf. Saloni et al. 2015, Woliński 2014). For the same reason, the outcome of journalistic subcorpus' analysis is quite similar because this style is also included by lexicographers in the material base of their works. The poorest result comes from analyzing subcorpus of drama — in these texts there are, seen relatively, a large number of proper names, colloquial expressions, interjections etc.

5 Subcorpora

We will characterize each subcorpus in brief. The subcorpus containing scientific texts for the general public is comprised of samples excerpted from monographies, textbooks as well as scientific papers and popular science articles in the magazines. These were foremost the emerging Polish periodicals (written in Polish) aimed at popularizing current scientific achievements and discoveries especially in the life sciences. Magazines and books are almost equally represented.

In this subcorpus the morphological analysis gave results spanning from 1.3 % (sample from 1897) unrecognized segments to almost 25 % (sample from 1830). The reason for such a high percentage of unrecognized forms is not just spelling that was different from contemporary orthography but also foreign words in different stages of assimilation (e.g. *feldspat* 'feldspar'), technical terms and suggested Polish equivalents that were not accepted in the end (e.g. *błyszcz* 'stibnite (antimonite)').

The second subcorpus — containing short press texts — mainly consists of short relations from daily newspapers published in the biggest Polish cities. Apart from the daily press, newspapers issued twice or once a week and every two weeks were also considered, which was common for places with no daily press. The language of press notes did not differ from the language of scientific texts for the general public (2.3 % in the most recognizable sample, 25.3 % in the least recognizable one), however the main source of unidentified parts are different spelling or older forms of inflection.

The journalistic subcorpus includes texts published in newspapers, journals and books. The most characteristic feature of the style is the anonymity of texts — almost half of them are signed only by initials, a pseudonym or collective author. On the other hand, these excerpts are almost fully recognizable (0.9 % to 18.2 %, about 6 % on average), possibly because of the style's closeness to general language, the small number of foreign words and/or professional vocabulary.

The fiction subcorpus contains mainly samples of novels and stories. Seven samples of verse novels and epic poems may be treated as an exception, however they are typical for the earliest 25 years of the period. In metadata they

are marked as verse prose because this information may be useful for natural language processing. In novels and stories (mainly romances) from the earlier period there are many fragments in French, on the other hand there was very limited availability of prose texts at that time, so they cannot be replaced with other material. In the later samples, mainly older inflectional forms are not recognizable — the average is about 5.5 % with a range from 0.5 % to 20 %.

The drama subcorpus contains samples of different kinds of dramatic works — from the masterpieces of Polish playwriting to the libretti of operettas and vaudevilles. As stated before, the analysis of these texts gave the weakest results (1 % to 28 %, 8 % on average). It is most unlikely that these results can be improved because there are a lot of interjections, dialect words etc., even though the utmost care was taken to avoid texts with strong dialect, historical or parodic stylization.

6 Processes of linguistic change through the corpus' lens

In spite of its small size, the corpus may be used not only as a source of data for an analyzer but also as material for research on the linguistic processes of change in regard to inflection, spelling (cf. Derwojedowa et al. 2016) or, to some extent, syntax (it consists of more than 11,000 sentences). Clear distribution of texts between styles (cf. Figure 1) allows even the formulation of tentative hypotheses concerning the differences between the subcorpora. First of all, changes listed in grammar books (cf. Bajerowa 1986, 1992, Klemensiewicz 2001) were looked at more closely. There are about 20 features of that period that may be verified on small-scale datasets. Figures 6 to 9 provide some examples. Figure 6 presents an overall picture of the evolution of adjective endings in the nineteenth century—*-em(i)/ém(i)* and *-éj* made by Kopczyński (1817) and those inherited from earlier stages of Polish.

Figure 7 presents the dynamics of change in adjective endings in instrumental and locative singular and the instrumental plural of both masculine and neuter from late Middle Polish -ym(i)/-im(i) to nineteenth century. -ém(i)/-em(i) and earlier.

In Figure 8 contraction [ij]/[ij] → [j]/[i] in loanwords is shown. Bajerowa's (1986) claim that the process was almost finished at the time is generally right, however it seems that it is still active (even if only simmering) in a wider class of left context consonants than in her research. It can be clearly observed that mostly stem-syllables are affected, long syllables in the stem being rare (circa 70 wordforms of 30 lexemes in 650 wordforms altogether), with *austryj~* ('Austrian', presently *austri~*) being most frequent (cf. Figure 7).

When compared with the frequency of *Ross(y)ja*, *rossyjsk-* ('Russia', 'Russian' *Rosja*, *rosyjski*) and *Prussy*, *prussk-* ('Prussia', 'Prussian' *Prusy*, *pruski*) with respect

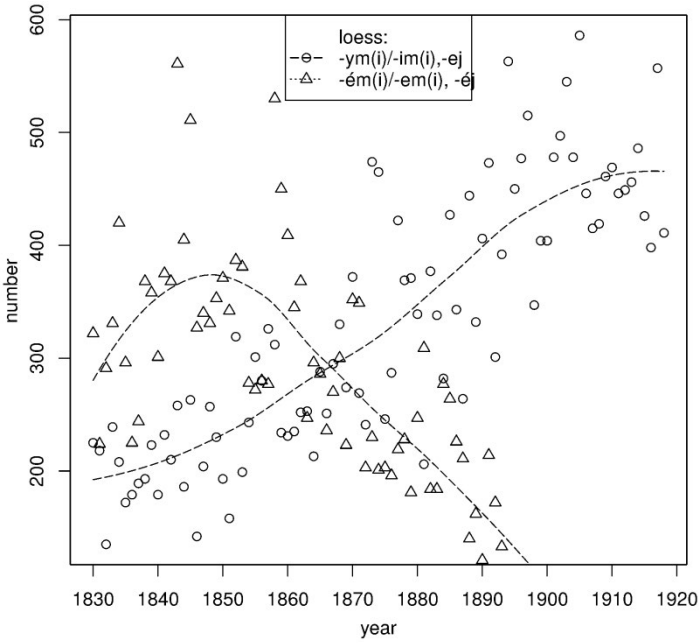


Figure 6. Innovative and historically developed endings of adjective-altering between 1830 and 1918. The dotted line represents innovative endings in total, i.e. any endings with é and e (loess = locally weighted scatterplot smoothing, cf. Cleveland et al. 1988).

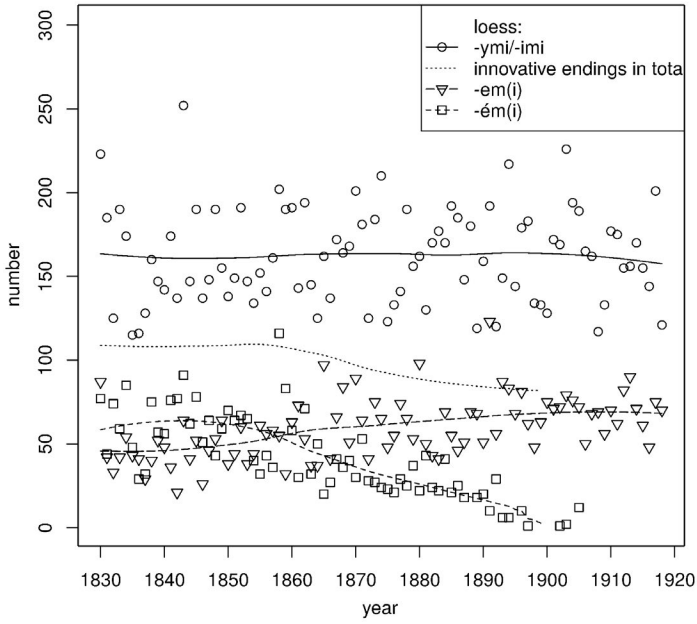


Figure 7. Innovative and inherited masculine and neuter endings of *adjectives* in instrumental and locative singular, instrumental plural, all genders.

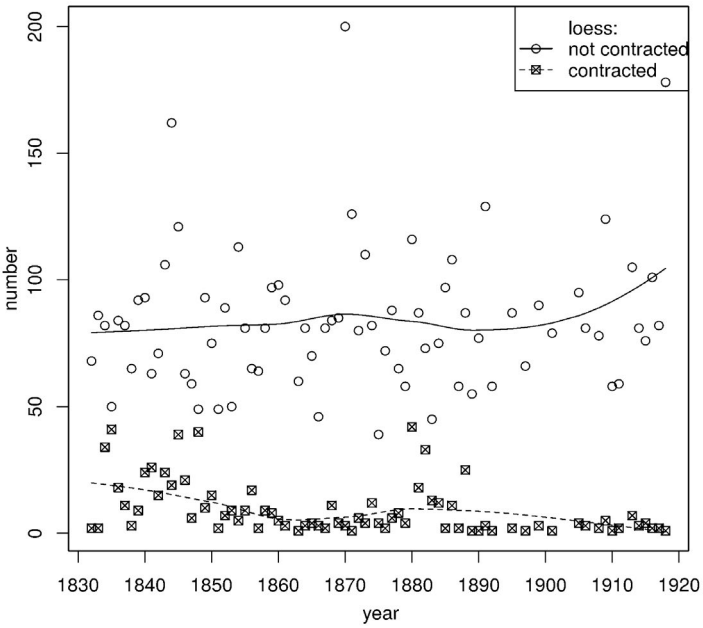


Figure 8. Words with contracted and uncontracted syllable [ɨj]/[ij].

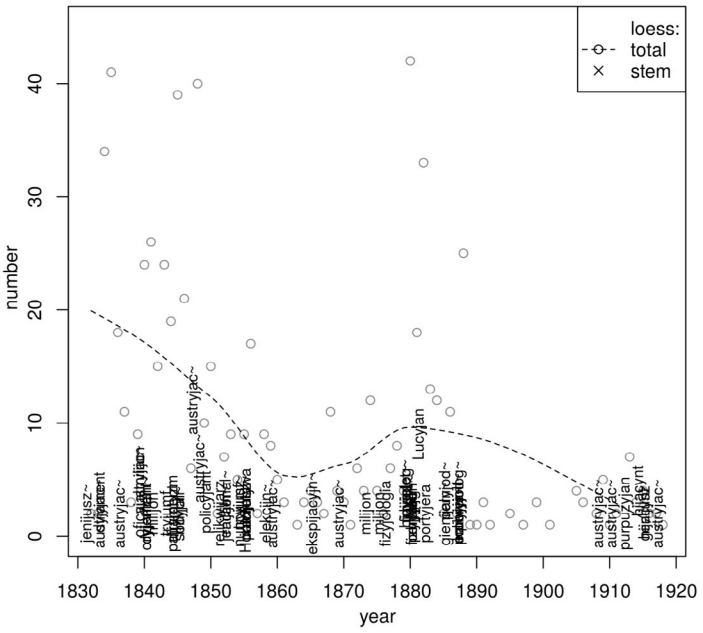


Figure 9. Uncontracted syllables in total, stems with not contracted syllables.

to the usage of doubled letters in loanwords, we clearly see that instead of processes, we rather observe lexical phenomena – all three are stems used in names of offices and institutions. Figure 9 shows the number of all contracted forms and points to individual uncontracted stems over the time span of 1830–1918.

The last example is the spelling of the (orthographic) string *ge* in loanwords. It is well attested that over time, the string became depalatalized in the period in question, being pronounced (and in consequence spelled) with *je*, *gie* and (innovative) *ge* by no other rule than according to a writer's belief or habit, e.g. spelling *jenerał* ('general') is almost three times more frequent than *generał*, no evidence of *gienerał*, whilst in the case of *geografia* and *jeografia* ('geography'), the spelling is exactly the opposite, with just one *gieografia*. The Dictionary of Polish by Niedźwiecki, Karłowicz and Kryński (1900–1927) quotes over 1,300 entries with *gie*, while there are less than 30 words with *gie* in the corpus. Some of them are lexical derivates (e.g. *Giermanie* 'Germans' and *giermański* 'German, adj'), and are present only in 5 % of samples. All others are spelled with an original *ge*.

7 Conclusion

Until now, neither a balanced, tagged and verified corpus of nineteenth century Polish nor an analyzer able to process older Polish texts have been available. Because of relatively small samples, the diversity of the corpus in many respects (places, authors, printed sources etc.) is quite satisfactory. Several tests passed on the corpus have proved that it can be used as a versatile resource to identify linguistic phenomena, trace their dynamics (cf. Figures 4–7) and turning points or to confront the emerging rules of orthography and good usage from the grammar handbooks with everyday practice. The corpus may be treated as an independent resource for research in inflection, morphonology and, to some extent, syntax. The considerable differentiation of samples makes it useful as an initial resource for research in new vocabulary and lexical changes as well.

References

Bajerowa, Irena. 1986. *Polski język ogólny XIX wieku. Stan i ewolucja.* T. I. *Ortografia, fonologia z fonetyką, morfonologia.* Katowice: Uniwersytet Śląski.

Bajerowa, Irena. 1992. *Polski język ogólny XIX wieku. Stan i ewolucja. T. II. Fleksja.* Katowice: Uniwersytet Śląski.

Bilińska, Joanna, Magdalena Derwojedowa, Monika Kwiecień and Witold Kieraś. 2016. Mikrokorpus polszczyzny 1830–1918. *Komunikacja Specjalistyczna*, in print.

Cleveland, William S. and Susan J. Devlin. 1988. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* 83 (403): 596–610. doi:10.2307/2289282. JSTOR 2289282 Derwojedowa, Magdalena, Witold Kieraś, Danuta Skowrońska and Robert Wołosz. 2014a. Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych. *Polonica* XXXIV: 21–27.

Derwojedowa, Magdalena, Witold Kieraś, Danuta Skowrońska and Robert Wołosz. 2014b. Zasób leksykalny polszczyzny II połowy XIX wieku a możliwość automatycznej analizy morfologicznej tekstów z tego okresu. In Małgorzata Gębka-Wolak, Joanna Kamper-Warejko and Andrzej Moroz (eds.), *Leksyka języków słowiańskich w badaniach synchronicznych i diachronicznych*, 183–196. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Derwojedowa, Magdalena, Witold Kieraś, Joanna Bilińska and Monika Kwiecień. 2016. Dynamika zmian między reformami 1830–1918. *Język Polski* z. 1 XCVI: 24–35.

Eder, Maciej, Mike Kestemont and Jan Rybicki. *Stylometry with R: a suite of tools*, *leksykalny polszczyzny II poł. XIX wieku a możliwość automatycznej analizy morfologicznej* Digital Humanities 2013: Conference Abstracts.

Gruszczyński, Włodzimierz, Dorota Adamiec and Maciej Ogrodniczuk. 2013. Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu badawczego. *Polonica* XXXIII: 309–316.

Klemensiewicz, Zenon. 2002. *Historia języka polskiego*, Warszawa: Wydawnictwo Naukowe PWN.

Kurcz, Ida, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran and Jerzy Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej*. T. 1–2. Kraków: Polska Akademia Nauk.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski and Barbara Lewandowska-Tomaszczyk (eds.). 2012. *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

R-manual 2015. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Saloni, Zygmunt, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński and Danuta Skowrońska. 2015. *Słownik gramatyczny języka polskiego*. http://sgjp.pl/.

Woliński, Marcin. 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1106–1111. Reykjavik: European Language Resources Association (ELRA).