DOI: 10.33896/PorJ.2020.8.4

Magdalena Derwojedowa
(Uniwersytet Warszawski,
e-mail: derwojed@uw.edu.pl)
ORCID: 0000-0002-6515-2940

MIKROKORPUS GRONOWY POLSZCZYZNY 1830–1918

1. OGÓLNA CHARAKTERYSTYKA KORPUSU

Korpus, który jest przedmiotem niniejszego omówienia, to korpus a) historyczny; b) mały, a nawet – jak na dzisiejsze standardy, również korpusów historycznych – bardzo mały; c) gronowy – czyli składający się z wielu próbek o określonej wielkości; d) złożony z równych objętościowo podkorpusów.

Ze względu na dystans czasowy – prawie dwieście lat dla daty początkowej (1830), sto dla końcowej (1918) – jest to korpus historyczny. Powstał on na potrzeby projektu „Automatyczna analiza fleksyjna tekstów polskich z lat 1830–1918 z uwzględnieniem zmian w odmianie i pisowni",[1] którego częścią były badania zmian fleksyjnych i ortograficznych, zachodzących w tym okresie. Celem korpusu było zatem dostarczenie wiedzy o języku określonego przedziału czasu w zakresie, który dotyczył zjawisk gramatycznych i ortograficznych. Do takich celów – ale również innych, w których zróżnicowanie jednostek jest niewielkie, np. typów konstrukcji składniowych [por. Sambor, Hammerl 1990; Świdziński 1996] – nie są potrzebne korpusy duże czy bardzo duże, niezbędne przede wszystkim na potrzeby słowników. Korpus polszczyzny 1830–1918 liczy 1 067 891 słów rozumianych jako napisy rozdzielone spacjami lub znakiem równoważnym, co odpowiada ok. 1,3 miliona segmentów w rozumieniu NKJP [Przepiórkowski i in. 2012, 60–62].

Korpus składa się z tysiąca próbek o przeciętnej długości 1027 segmentów (z wyłączeniem powtarzających się postaci dramatu i didaskaliów). Jako zasadę zespół kompilujący korpus przyjął nieurywanie ostatniego zdania w próbce – co w większości wypadków (84,6% próbek)

[1] Projekt finansowany ze środków Narodowego Centrum Nauki, DEC-2012/07/B/HS2/00570. Zespół projektowy tworzyli, oprócz autorki niniejszego tekstu, Joanna Bilińska, Witold Kieraś, Monika Kwiecień, Danuta Skowrońska i Robert Wołosz. Prace informatyczne związane z tworzeniem słownika w środowisku Kuźnia wykonała Dorota Komosińska, konsultacji udzielił Marcin Woliński. Udostępnienia przez serwis Marasca [Bień 2011, 2012] dokonał Michał Rudolf.

powoduje niewielkie przekroczenie założonej długości (por. p. 4.). Każda próbka pochodzi z innego tekstu (por. p. 2. i 4.). Ponieważ pozyskany materiał językowy został równomiernie rozłożony w czasie – w każdym podkorpusie dla każdego roku znajduje się przynajmniej jedna próbka – mimo skromnej objętości korpusu i stosunkowo dużych próbek jest bardzo zróżnicowany [por. Kurcz i in. 1992, dalej SFPW]. W korpusie zostało wyodrębnionych pięć podkorpusów. W wypadku korpusów współczesnych oddają one zwykle strukturę ilościową czytelnictwa [Przepiórkowski i in. 2012, 11–37], w wypadku korpusów dawniejszych takich kryteriów nie można użyć ze względu na brak danych. Dlatego tworząc korpus, postanowiliśmy odzwierciedlić zróżnicowanie korpusu SFPW [Kurcz i in. 1992], tzn. pięć stylów funkcjonalnych w równych proporcjach. W efekcie powstały liczące po 200 próbek o łącznej objętości około 200 tys. segmentów podkorpusy 1) stylu popularno-naukowego (PN), 2) drobnych wiadomości prasowych (DWP), 3) publicystyki (PUB), 4) prozy artystycznej (PA) i 5) dramatu artystycznego (DR). Taki podział okazał się uzasadniony – drzewko konsensusowe pokazuje wyraźne zgrupowania tekstów należących do tego samego stylu (rys. 1.[2]).

Rysunek 1. Rozrzut stylów (1 – popularnonaukowy, 2 – drobne wiadomości prasowe, 3 – publicystyka, 4 – proza artystyczna, 5 – dramat).

[2] Analiza została przeprowadzona za pomocą popularnego pakietu *stylo* języka R [Eder, Kestemont, Rybicki 2013]. Dwa wyraźne niedopasowania próbek stylu popularnonaukowego (lata 1840 i 1850) zostały omówione w punkcie 4.

2. ŹRÓDŁA

Źródła do korpusu polszczyzny XIX w. (a nawet szerzej – do polszczyzny doby nowopolskiej) są liczne i łatwo dostępne. Olbrzymie ilości tekstów zostały udostępnione przez biblioteki cyfrowe i inicjatywy takie jak Wolne Lektury, Projekt Gutenberg czy WikiSource. Można zatem zgromadzić nawet dość rozległy korpus niespecjalistyczny niemal całkowicie automatycznie. Jeśli jednak ma on być zrównoważony (czasowo, stylistycznie, geograficznie lub pod innymi względami), zadanie jest trudniejsze, niektórych tekstów jest bowiem niewiele, a do tego nierzadko są one złej jakości technicznej. Gromadząc teksty z lat 1830–1918, przyjęliśmy założenie, że podstawą są pierwodruki w oryginalnej postaci ortograficznej, fleksyjnej i składniowej, by uniknąć zmian związanych na przykład ze standaryzacją lub uwspółcześnieniem ortografii czy konkurujących form. Konieczne okazało się przepisanie kilku próbek, głównie z najwcześniejszego okresu. Jednak gros pracy polegało na dotarciu do pierwodruku, wyborze próbek i korekcie błędów powstałych podczas automatycznego rozpoznawania pisma (OCR), a także znakowaniu pewnych elementów tekstu, takich jak np. postaci dramatu, i opracowaniu danych metatekstowych. Dodatkowym utrudnieniem może być zachowanie pisowni z „ſ ”, niektóre rozwiązania zecerskie, np. reprezentowanie [ɔw̃] przez „ǫ”, czy „š” i „ž” zamiast dwuznaku, a także stosowanie fraktury [próbka 1896_2.3,[3] „Gazeta Ludowa”] do zapisu polszczyzny, uzupełnianie składu polskiego literami cyrylickimi, obecność symboli nieliterowych (№, ½ itp.), a także liter greckich, por. przykłady (1)–(3). Zgodnie z przyjętymi zasadami zostały wykluczone teksty z wyrażeniami matematycznymi i większymi partiami obcojęzycznymi, wyjątek stanowiły powieści z pierwszego trzydziestolecia, których jest mało, a partie dialogowe po francusku można uznać za ich cechę typową.

Choć na okres 1830–1918 przypada wydanie wielu dzieł wybitnych, należących do kanonu literatury polskiej, zostały one w korpusie pominięte, ponieważ 1) w dużej części zostały one już uwzględnione w kartotece SJPDor, a co za tym idzie – w samym słowniku i SGJP3 [Saloni i in. 2015], który stanowi podstawę analizatora Morfeusz [Woliński 2006; 2014], czyli programu, który po adaptacji do pisowni i fleksji XIX w. został wykorzystany w znakowaniu korpusu; 2) w niewielkim pod względem liczby próbek korpusie, szczególnie w podkorpusach prozy i dramatu, uwzględnienie wszystkich najważniejszych dzieł z epoki zachwiałoby proporcjami na korzyść autorów najwybitniejszych; 3) korpus w zasadzie nie zawiera tekstów wierszowanych (poza dramatem, który przez długi czas właściwie innej formy nie przybierał, i podkorpusem PA,

[3] Dla cytowanych fragmentów podajemy numer próbki w korpusie. Pierwsze cztery cyfry oznaczają rok, kolejna – numer stylu, ostatnia – numer kolejny w danym roku i stylu.

w którym zostały uwzględnione trzy dzieła wierszowane J. Słowackiego (*Lambro*, *Beniowski*, *Jan Bielecki*), jedno C.K. Norwida (*Szczesna*) i po jednej H. Kajsiewicza, W. Dawida oraz K. Bujnickiego.

Nieco inaczej wygląda sytuacja z wykorzystaniem źródeł prasowych. Oprócz oczywistej kompilacji podkorpusu drobnych wiadomości prasowych z fragmentów notatek z prasy o dużej częstotliwości (od dzienników po dwutygodniki), źródła te zostały wykorzystane w stylu popularnonaukowym i publicystycznym. W wypadku prasy tytuły się powtarzają, próbki pochodzą jednak zawsze z różnych numerów. Po kilkanaście fragmentów pochodzi z „Gazety Lwowskiej" i „Gazety Warszawskiej", po kilka – z „Kurjera Warszawskiego", „Tygodnika Petersburskiego", „Czasu", „Gazety Wielkiego Xięstwa Poznańskiego", „Gazety Polskiej" i „Dziennika Poznańskiego". O ile publicystyka gościła zarówno w prasie codziennej, jak i w periodykach, o tyle próbką powstającego stylu popularnonaukowego były rodzące się czasopisma, takie jak „Biblioteka Warszawska" (17 próbek), „Kosmos" (11), „Ateneum" (7), „Wszechświat" (6), uzupełniające podręczniki i monografie z dzisiejszego punktu widzenia też raczej popularne niż ściśle naukowe.

3. MIKROSTRUKTURA KORPUSU

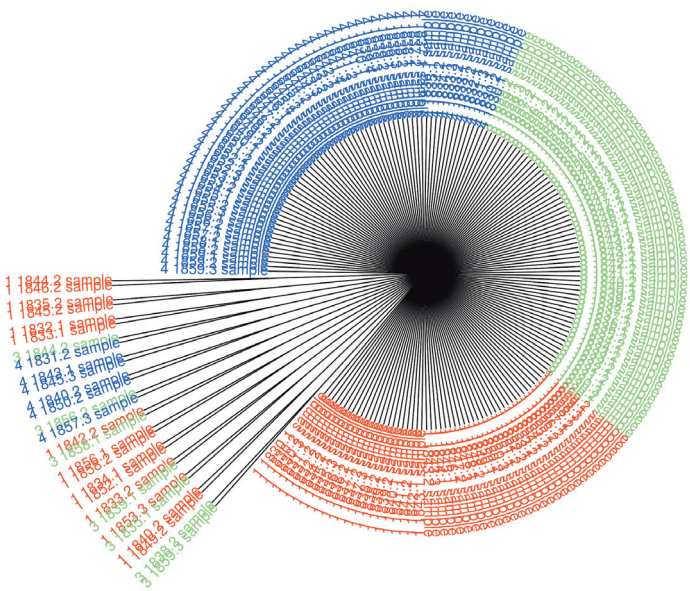
Każda próbka korpusu przechowywana jest w czterech plikach: 1) wybrany fragment tekstu (przykład (4)); 2) metadane, czyli informacja o autorze, tytule, miejscu i czasie wydania itd. (przykład (5)) oraz przynależności do podkorpusu; 3) plik źródłowy, czyli pełny skan dzieła (o ile był dostępny) lub – wyjątkowo – skany wykorzystanych stron; 4) plik korekty tekstu (nieudostępniany). Pliki próbek zostały zapisane w formacie tekstowym, co umożliwia ich wygodne przeszukiwanie i upraszcza import do programów do przetwarzania tekstu. Teksty oryginalne pozostały bez uwspółcześniania pisowni. Wprowadzone zostały oznaczenia elementów obcych (cytatów obcojęzycznych, wzorów itp.) i niemożliwych do odczytania (przykłady (6) i (7)), a w dramacie także didaskaliów i osób dramatu, dzięki czemu można je wyłączyć ze statystyk tekstu.[4]

[4] Na przykład pisownia z „x" jest w zgromadzonych tekstach stosunkowo rzadka, ale dwie trzecie jej wystąpień to postaci *Xiężnych*, *Xiążąt* i *Xięży* w dramatach, przy czym liczba wystąpień metatekstowych może być nawet trzykrotnie większa niż liczba użyć tekstowych (jak jest np. w *Posłannikach* S. Buszczyńskiego, 1876_5.2). Również wyrażenia anaforyczne mogą wpływać na wyniki użyć pisowni, jest ona bowiem stosowana konsekwentnie.

4. MAKROSTRUKTURA

Podział według stylów SFPW okazał się zaskakująco trafny (por. rys. 1.). Ponieważ próbki są stosunkowo krótkie, na potrzeby analizy zostały one pogrupowane w poszczególnych stylach według dat o tej samej liczbie setek i dziesiątek, tak więc „1_1830_sampl" oznacza próbki stylu „1" (PN) z lat 1830–1839, a „5_1910_sampl" – „5" (DR) z lat 1910–1918. Na analizę został nałożony warunek, by słowa powtarzały się przynajmniej w 20% tekstów (warunki ściślejsze przy sporej rozpiętości czasowej, zróżnicowaniu ortograficznym i fleksyjnym oraz niedużej objętości mogą zacierać różnice [por. Rybicki 2013]). Dwa fragmenty – lata 40. i 50. – zostały poddane badaniu na podkorpusie trzech stylów z lat 1830–1850 (204 próbki, 212 151 segmentów), na którym wyraźnie widać, że podkorpusy wykazują spójność stylistyczną, z wyjątkiem 28 próbek, które cechuje wzajemne podobieństwo i zarysowujące się mniejsze jednolite stylistycznie skupienia (por. rys. 2.).

Rysunek 2. Drzewko konsensusu PN, PUB, PA 1830-1859.

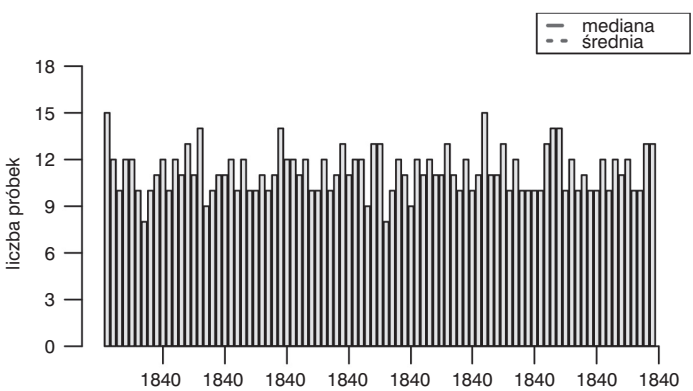


Ze względu na tematykę urywki PN z lat 30. dotyczą pieniądza (jako numizmatu i elementu operacji giełdowych), ponadto dwa teksty – anonimowy i Franciszka Grzymały – dotyczą kwestii politycznych. Próbki z lat 40. pochodzą z dzieł naukowych, w większości poświęconych naukom ścisłym (geometrii, biologii) i zagadnieniom inżynierskim, podobnie jak próbki z lat 50. (botanika, fizyka). Fragmenty powieściowe powstały

jako dzieła trzech autorek (A. Nakwaska,[5] Z. Klimańska, Ł. Rautenstrauchowa[6]) i dwóch autorów (F. Kozłowski i M. Czaykowski); w PUB trzy fragmenty są anonimowe, jeden pochodzi z tekstu Ł. Rautenstrauchowej,[7] wśród autorów pozostałych są ks. A. Malczewski i J. Klaczko. Właściwie jedynym wspólnym mianownikiem – jeśli nie liczyć przewagi liczebnej autorek powieści i tego, że jedna trzecia utworów została wydana za granicą, przede wszystkim w Paryżu – jest to, że większość autorów pochodziła z Kresów. Prawdopodobnym wyjaśnieniem tego odchylenia wydaje się jednak to, że nieliteracki styl pisania o zagadnieniach naukowych jeszcze nie ma wyraźnych cech, choć trzeba by dokładniej się przyjrzeć temu, dlaczego teksty z lat 30. takiej charakterystyki nie mają.

Zasada, że próbki nie są ucinane na tysięcznym słowie, ale na ostatnim wyrazie zdania, poskutkowała tym, że niektóre próbki są nieco krótsze, a niektóre nieco dłuższe niż 1000 segmentów. Odbija się to na długości podkorpusów liczonej w segmentach – DWP liczy ich niespełna 204 tys., ale też zdania z notatek prasowych są niedługie i stosunkowo łatwo jest uzyskać próbki oscylujące wokół zamierzonego tysiąca, w PA (nieco ponad 211 tys. segmentów) zdania są bardziej rozbudowane, a liczba składników dopełniających większa. Próbki dość równomiernie pokrywają cały uwzględniony przedział czasowy (por. rys. 3.).

Rysunek 3. Liczba próbek w poszczególnych latach.



[5] 1843_4.1 i 1845_4.3.
[6] 1831_4.3.
[7] 1839_3.3.

5. ZRÓŻNICOWANIE

Oprócz wspomnianej już preferencji dla autorów mniej znanych zespół kompilatorski starał się, by byli oni jak najbardziej różnorodni. Trudno podać precyzyjną liczbę autorów, ponieważ wiele tekstów, zwłaszcza prasowych (szczególnie DWP), nie jest w żaden sposób sygnowanych, jest sygnowanych inicjałami lub jako teksty redakcyjne, lub ma autora zbiorowego. Ekscerptów anonimowych jest 28%, pozostałe próbki zostały pobrane z dzieł nieco ponad 500 osób. W 39% wypadków autorzy są reprezentowani przez jeden tekst, w 13% – przez dwa, w 9% – przez trzy. Dziesięć cytatów pochodzi z utworów J. Słowackiego, po dziewięć – z tekstów E. Orzeszkowej i J.I. Kraszewskiego, każde z tych dwojga jest też współautorem jednej publikacji. Wśród często powtarzających się autorów są między innymi popularni dramatopisarze: J. Bliziński, W. Koziebrodzki, a także J. Dzierzkowski, publicysta i powieściopisarz, oraz autorki powieści: A. Nakwaska i Ł. Rautenstrauchowa, podobnie jak on aktywne jeszcze przed rokiem 1850. Autorzy cytowani kilkakrotnie byli na ogół aktywni na wielu polach w późniejszym okresie, próbki z dzieł E. Orzeszkowej, S. Przybyszewskiego i J.I. Kraszewskiego (ale też J. Słowackiego) znajdują się w podkorpusach prozy artystycznej, dramatu, i publicystyki. 29 autorów jest obecnych w podkorpusach prozy artystycznej i dramatu. Wydaje się, że zarówno wielokrotne cytowanie tych samych autorów, jak i ich obecność w różnych podkorpusach oddaje w jakimś stopniu ich społeczne oddziaływanie. Wybitna pozycja E. Orzeszkowej, słabsza (3 ekscerpty) M. Konopnickiej, a także generalnie nieduży udział autorek (32 na nieco ponad 500 nazwisk i sygnatur) odzwierciedlają skromny udział piszących kobiet w całej produkcji literackiej, obszary ich aktywności i wpływ na odbiorców.[8] Z 49 dzieł podpisanych przez autorki 28 to proza artystyczna, 13 – dramaty, a 5 – publicystyka.

Oprócz dużych ośrodków, w których działały wydawnictwa książkowe i prasa, staraliśmy się uwzględnić ośrodki mniejsze oraz skupiska emigracyjne. Z przyczyn technicznych (niedostępność tekstów w postaci elektronicznej) niedoreprezentowane jest Wilno, a Grodno jest praktycznie nieobecne. Proporcje publikacji z pozostałych miast przedstawia tabela 1. Ogółem teksty pochodzą ze 166 różnych ośrodków, niekiedy niewielkich, jak Tłuszcz, Racibórz, Chełmno, Bytom czy Striegau (Strzegom). Najbardziej zróżnicowany regionalnie jest podkorpus drobnych wiadomości prasowych, reprezentowanych jest aż 37 ośrodków, w tym emigracyjne, jak Detroit, Paryż i Zürich; najmniejsze zróżnicowanie regionalne wykazują teksty popularnonaukowe, ponad połowa próbek pochodzi z wydawnictw warszawskich, 15% ze Lwowa, 12% z Krakowa, 10% z Poznania.

---

[8] Proporcja ta może się zmienić, jeśli w prozie w większym stopniu uwzględni się mniejsze formy obecne tylko w prasie.

Tabela 1. Liczba próbek według miejsca wydania źródła.

miejsce wyd.	łącznie	PN	DWP	PUB	PA	DR
Warszawa	390	116	75	53	90	56
Lwów	156	30	21	37	28	40
Kraków	114	23	12	29	14	36
Poznań	71	11	24	16	6	14
Paryż	60	6	–	32	10	12
Wilno	50	2	6	1	24	17

6. UDOSTĘPNIENIE I ZASTOSOWANIE

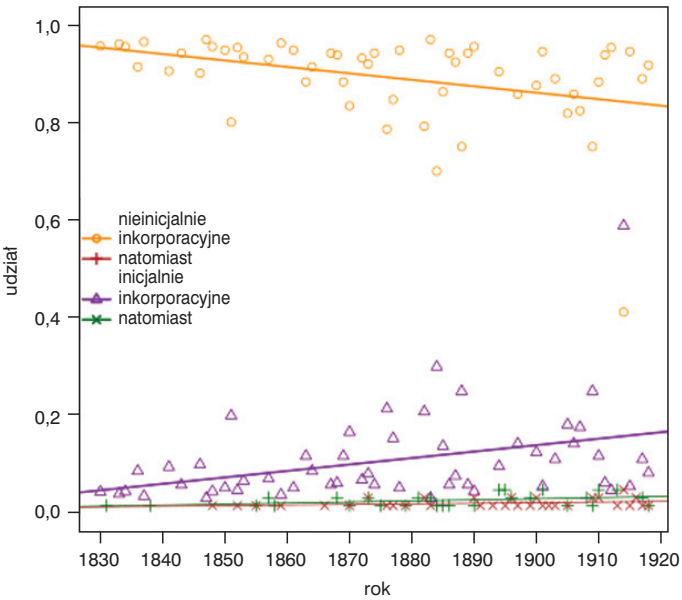
Korpus został udostępniony w postaci tekstowej, jako archiwa źródeł oraz skompilowany do przeglądarki Poliqarp-Marasca [por. Bień 2011; 2012] do użytku z programem off-line oraz jako serwis www pod adresem https://szukajwslownikach.uw.edu.pl/f19/ na licencji umożliwiającej całkowicie swobodne zastosowanie. Wersja udostępniona on-line i jako zasób Poliqarpa[9] została oznakowana z wykorzystaniem analizatora Morfeusz 2 [Woliński 2006; 2014] ze słownikiem dostosowanym do ortograficznej i fleksyjnej specyfiki korpusu.

Sposób udostępnienia korpusu umożliwia wykorzystanie go w wielu badaniach, które nie wymagają bardzo dużej liczby danych, np. obserwacji zmian ortograficznych, fleksyjnych czy składniowych. Jako przykład niech posłuży pozycja linearna spójników inkorporacyjnych BOWIEM, ZATEM i ZAŚ oraz porównawczo NATOMIAST. Spójniki inkorporacyjne zajmują pozycję po pierwszym segmencie wprowadzanego zdania, takie umieszczenie NATOMIAST jest uważane za błąd (NSPP). W korpusie wszystkich poświadczeń dla użyć BOWIEM, ZATEM (także w wariancie ZATÉM) i ZAŚ jest 1777 (dwie trzecie z nich przypada na ten ostatni), w ledwie 6% wypadków spójnik zajmuje pozycję inicjalną. Liczba użyć narasta wraz z upływem czasu. 89 poświadczeń NATOMIAST rozkłada się niemal po równo (42 po segmencie inicjalnym, 47 jako segment inicjalny, por. rys. 4.).

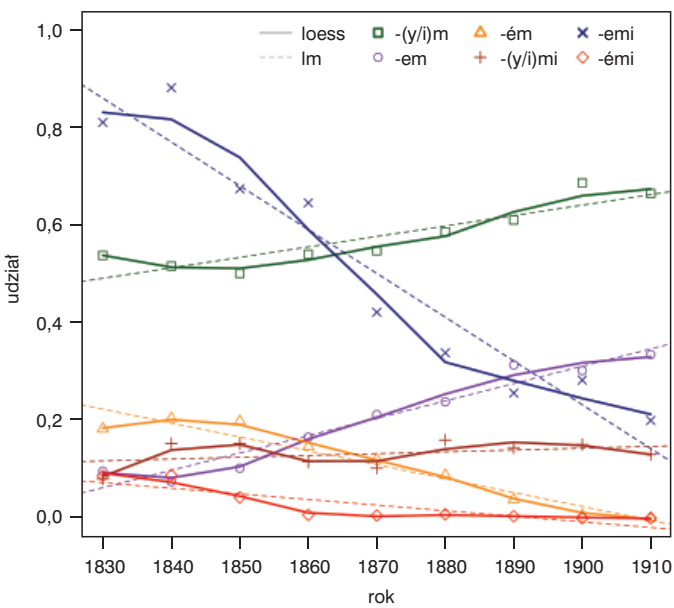
W porównawczym współczesnym korpusie milionowym NKJP [Przepiórkowski i in. 2012, 51, 58) z 508 wystąpień spójników inkorporacyjnych 128, czyli jedna czwarta, zajmuje pozycję inicjalną. Pozycje, które zajmuje NATOMIAST, pozostają w takich samych proporcjach, tzn. z lekką

[9] Jest to program stary i niezbyt przyjazny dla użytkownika, jednak oferujący subtelne wyszukiwania, możliwości obliczeń statystycznych według wybranych przez użytkownika parametrów oraz pobranie danych do dalszej obróbki, co sprawia, że ciągle jest narzędziem niemającym realnej konkurencji.

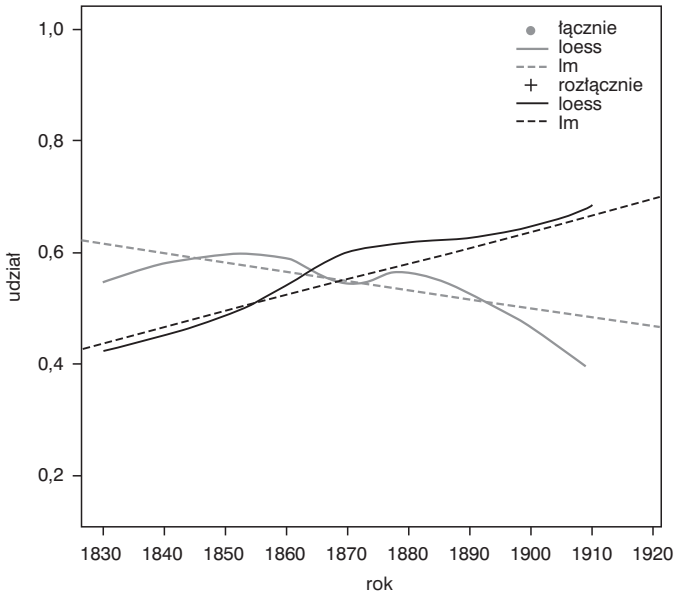
Rysunek 4. Pozycja linearna spójników inkorporacyjnych i NATOMIAST.



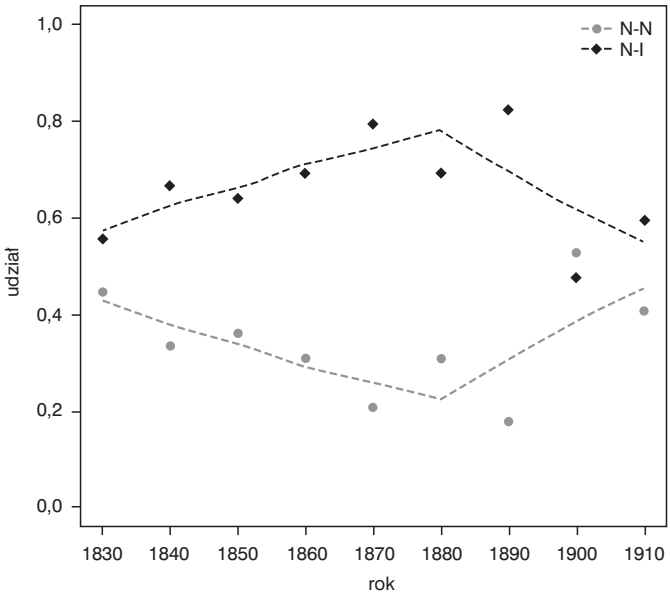
Rysunek 5. Końcówki N. i Msc. rodzaju męskiego i nijakiego przymiotników.



Rysunek 6. Pisownia łączna i rozdzielna Z DALA.



Rysunek 7. Przypadek orzecznika przymiotnikowego (N-N – mianownik, N-I – narzędnik).



przewagą dla pozycji inicjalnej (165 : 157). Rysunek 6. i rysunek 7. przedstawiają inne przykładowe badania: końcówek narzędnika i miejscownika rzeczowników męskich i nijakich, pisowni łącznej i rozdzielnej wyrażenia Z DALA, przypadka przymiotnikowego w orzeczeniach imiennych [por. Bajerowa 1992; Klemensiewicz 2002; Derwojedowa, Skowrońska, Kieraś 2014; Derwojedowa i in. 2014, 2016; Bilińska i in. 2016; Bilińska, Derwojedowa, Kwiecień 2018]. Dane dla większej przejrzystości zostały pogrupowane dziesięcioleciami.

7. ZMIANY I KIERUNKI ROZWOJU

Ponieważ każda próbka opatrzona jest adresem źródła, najbardziej oczywistym wykorzystaniem wydaje się pobranie pełnych źródeł i skonstruowanie na tej podstawie dość dużego, wstępnie zrównoważonego korpusu polszczyzny lat 1830–1919. Z pewnej perspektywy czasowej myślę, że znacznie bardziej potrzebne są narzędzia do opracowania tekstów, przede wszystkim do lematyzacji i ujednoznaczniania oraz generowania korpusów w formacie nadającym się do użycia w dostępnych już narzędziach. Celowe jest natomiast rozszerzenie korpusu gronowego o pierwsze trzydziestolecie i o teksty do 1945 r. Przy zachowaniu obecnej średniej długości tekstu, który przypada na każdy rok, zwiększyłoby to objętość zasobu o co najmniej 625 tysięcy segmentów. Wydaje się jednak, że nowe próbki – a równolegle lub w niedługim czasie również cały korpus – należałoby przebudować, zachowując objętość całości, w stylach i kolejnych latach, ale zwiększając liczbę próbek, czyli skracając istniejące i włączając krótsze, ale liczniejsze, nowe. Skrócenie próbek do 300–400 segmentów przy jednoczesnym zwiększeniu ich liczby nie odbije się ujemnie na jakości danych, a pozwoli włączyć 2,5–3-krotnie większą liczbę źródeł. Ułatwiłoby to też dołączenie – zapewne jako osobnych podkorpusów – liryki artystycznej, której obecność w ówczesnym obiegu czytelniczym trudno kwestionować, oraz przekładów – ciekawych ze względu na oddziaływanie na język i kształtowanie stylów [por. Even-Zohar 1990, 45–52, 73–78]. Częścią rozbudowy powinno też być dołączenie próbek najbardziej znanych tekstów, w pierwszej fazie pominiętych.

Uzyskany w ten sposób zasób niewielką objętość kompensowałby zróżnicowaniem tekstów, co ma istotne znaczenie dla ustalania powszechności zjawisk językowych (lub ich ograniczenia do pewnego typu tekstów). Taka długość próbki pozwala w zasadzie swobodnie rozszerzać go o źródła spoza domeny publicznej. Skonstruowany w ten sposób korpus mógłby też służyć jako materiał do trenowania różnych narzędzi automatycznych.

Przykłady

(1) Francuzi mówią wyraźnie, że ta wojna nie jest przeciw Cesarzowi marokańskiemu wymierzona, aczkolwiek jego kraj przez obie partyje naruszonym zostanie; zamiarem tych nieprzyjacielskich kroków nie może być, jak tylko schwytanie Abd-el-Kadera [„Gazeta Lwowska" 4 listopada 1845, 1845_2.2].

(2) (...) odebrano ważne nowiny z Mexyku, 51 (2 lipca 1839) [„Tygodnik Petersburski. Gazeta Urzędowa Królestwa Polskiego", 1839_2.2].

(3) [[Gom.]] Marysia nie chciała zaraz zemnǫ mówić, wybiegła za zamek i jak tylko mię spostrzegła, zaczęła zaraz do zamku uciekać. Ja daléj za nię i wołam: Marysiu! Marysiu! kochana Marysiu! serdeczna Marysiu! nareście Marysia stanęła [Michał Suchorowski, *Prawo litewskje czyli kara za całusek. Komedja we trzech oddziałach*, 1831_5.2].

(4) 54. Gdy wszystkie części arkady, będąc jak najlepiéj z sobą połączone, stanowią jednę całość; wypada ztąd, że dla utrzymania jéj w czasie roboty, nie ma potrzeby wznoszenia ciągłych, w łuki prowadzonych rusztowań, czyli bukszteli, jak pod sklepienia kamienne; lecz dosyć jest oprzéć ją w niewielu punktach na niewzruszonych podporach, między któremi można bez wielkich trudności składać zworniki, przy pomocy rusztowań wiszących (jak niżej się okaże), utrzymanych przez ukończone już części arkady, która tym sposobem sama sobie służyłaby za główne rusztowanie. Tak zaczynając od przyczołków, z któremi taż arkada jest mocno połączona, należy, w miarę postępowania roboty, podpierać ją w punktach, z początku bliższych A, B, C, (fig. 14), następnie zaś w punktach D i E, na taką jeden od drugiego odległość, na jaką części arkady CD, DE, bez niebezpieczeństwa utrzymać się mogą. Gdy w samym środku czyli w wierzchołku arkady, ma się kończyć składanie téjże, w tém miejscu zatém nie ma potrzeby stawiania nowéj podpory, i podpora EI, od odpowiadającéj sobie w drugiéj połowie arkady, może być dwa razy więcéj oddalona, niż od podpory DK. (...) [1830_1.1, F. Pancer, *Wiadomość o nowym rodzaju mostów żelaznych na wielką otwartość...*].

(5) autor: Pancer, Feliks
tytuł: Wiadomość o nowym rodzaju mostów żelaznych na wielką otwartość, wynalazku F. Pancera Porucznika Inżynierów Wojsk pol., Prof. Budownictwa w Szkole Wojskowej Applikacyjnej z zastosowaniem do rzeki Wisły pod Warszawą
data wydania: 1830
miejsce wydania: Warszawa
redaktor: M.A. Pawłowicz, S.F. Janicki
tytuł książki:
tytuł gazety, czasopisma, serii wydawniczej: Pamiętnik Fizycznych, Matematycznych i Statystycznych Umiejętności nr: 5 (maj)
wydawnictwo: S. Gałęzowski i komp.

numery stron: 258–262
styl: popularnonaukowy
źródło: e-Biblioteka Uniwersytetu Warszawskiego
link: http://ebuw.uw.edu.pl/publication/17406,
http://hint.org.pl/hid=AB085
uwagi: hint: budownictwo; -- ryciny po s. 301 [Metryczka próbki 1830_1.1., F. Pancer *Wiadomość o nowym rodzaju mostów żelaznych na wielką otwartość…*]

(6) Jak ja powiem za panią. Paniby wolała, coby wasz {{muż nie tolko}} nie był {{wowsie}} skazany, no coby on był wolny, {{sawsiem}} wolny, ha?

(7) Także {{…}} nocy będzie tam czuwał oficer a obcym osobom, nie należącym do fabryki, zostanie wstęp zupełnie wzbroniony [„Gazeta Ludowa. Pismo poświęcone ludowi ewangelickiemu" 30 września 1896, 1896_2.3].

Bibliografia

I. Bajerowa, 1992, *Polski język ogólny XIX wieku. Stan i ewolucja. Fleksja*, t. II, Kraków.

J.S. Bień, 2011, *Efficient search in hidden text of large DjVu documents* [w:] *Advanced Language Technologies for Digital Libraries*, *Lecture Notes in Computer Science (Theoretical Computer Science and General Issues)* 6699, s. 1–14.

J.S. Bień, 2012, *Skanowane teksty jako korpusy*, „Prace Filologiczne" LXIII, s. 25–36.

J. Bilińska, M. Derwojedowa, M. Kwiecień, 2018, *Microcorpus of Nineteenth-Century Polish* [w:] E. Fuß i in. (red.), *Grammar and corpora 2016*, Heidelberg, doi: http://dx.doi.org/10.17885/heiup.361.509.

J. Bilińska i in., 2016, *Dynamika zmian ortograficznych 1830–1918. Badanie korpusowe* [w:] M. Gębka-Wolak, J. Kamper-Warejko, I. Kaproń-Charzyńska (red.), *System – tekst – człowiek. Studia nad dawnymi i współczesnymi językami słowiańskimi*, Toruń, s. 15–33.

M. Derwojedowa, D. Skowrońska, W. Kieraś, 2014, *Korpus polszczyzny XIX wieku – od mikrokorpusu do korpusu średniej wielkości*, „Prace Filologiczne", s. 249–254.

M. Derwojedowa i in., 2014, *Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych*, „Polonica" XXXIV, s. 21–27.

M. Derwojedowa i in., 2016, *Dynamika zmian między reformami 1830–1918*, „Język Polski" z. 1 (96), s. 24–35.

M. Eder, M. Kestemont, J. Rybicki, 2013, *Stylometry with R: a suite of tools* [w:] *Digital Humanities 2013: Conference Abstracts*, Lincoln, s. 487–489; url: pdf:http://dh2013.unl.edu/abstracts/

I. Even-Zohar, 1990, *Polysystem studies*, „Poetics Today. International Journal for Theory and Analysis of Literature and Communication" t. 11, nr 1.

Z. Klemensiewicz, 2002, *Historia języka polskiego*, Warszawa.

NSPP: A. Markowski (red.), 2000, *Nowy słownik poprawnej polszczyzny*, Warszawa.
A. Przepiórkowski i in. (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.
J. Rybicki, 2013, *Stylometryczna niewidzialność tłumacza*, „Przekładaniec" 27, s. 61– 87; doi: 10.4467/16891864PC.13.004.1286.
SGJP3: Z. Saloni i in., 2012, *Słownik gramatyczny języka* polskiego, wyd. III; http://sgjp.pl
J. Sambor, R. Hammerl, 1990, *Statystyka dla językoznawców*, Warszawa.
SFPW: I. Kurcz i in., 1992, *Słownik frekwencyjny polszczyzny współczesnej*, Warszawa.
SJPDor: W. Doroszewski (red.), 1958–1969, *Słownik języka polskiego*, t. I–XI, Warszawa.
J. Szejko, 2016, *Aplikacja webowa do opracowywania słowników fleksyjnych*, praca magisterska napisana pod kierunkiem dra Marcina Wolińskiego, MIM Uniwersytet Warszawski.
M. Świdziński, 1996, Własności składniowe wypowiedników polskich, Warszawa.
M. Woliński, 2006, *Morfeusz – a Practical Tool for the Morphological Analysis of Polish* [w:] M.A. Kłopotek, S.T. Wierzchoń, K. Trojanowski (red.), *Intelligent Information Processing and Web Mining*, „Advances in Soft Computing", Berlin, s. 503–512.
M. Woliński, 2014, *Morfeusz Reloaded* [w:] N. Calzolari i in., *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland; url: http://www.lrecconf. org/proceedings/lrec2014/index.html

Cluster Micro-Corpus of the Polish Language of 1830–1918

Summary

This paper is dedicated to the construction of a small cluster corpus of Polish texts from the period 1830–1918. The assumptions of the corpus, its micro- and macro-structure, as well as stylistic, regional and author diversity, and method of making it available are presented. Its application capabilities are illustrated on the example of orthographic, inflectional, and syntactic studies.

Keywords: corpus – diachrony – inflection – syntax – orthography – stylometry

Trans. Monika Czarnecka