LingVaria
Rok XVIII (2023) | 1 (35) | s. 125–134
https://doi.org/10.12797/LV.18.2023.35.09

Marek Łaziński
Uniwersytet Warszawski, Warszawa
m.lazinski@uw.edu.pl

Michał Woźniak
Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków
michal.wozniak@ijp.pan.pl

Rafał L. Górski
Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków
rafal.gorski@ijp.pan.pl

KORPUS XIX W. UNIWERSYTETU WARSZAWSKIEGO I IJP PAN[1]

Słowa klucze: korpus językowy, polszczyzna historyczna, okres nowopolski, językoznawstwo korpusowe
Keywords: language corpus**,** historical Polish, Modern Polish Period, corpus linguistics

1. Korpusy historyczne

Językoznawstwo historyczne, o ile zajmowało się epokami, które pozostawiły po sobie świadectwa pisane, było zawsze „językoznawstwem korpusowym". Tzw. metoda filologiczna polegała na poszukiwaniu interesujących badacza zjawisk w dawnych tekstach (Campbell 1999: 333). Nie znaczy to jednak, że zmiana technologiczna, a więc mniej czy bardziej zautomatyzowane przeszukiwanie (tzw. *distant*, a zwłaszcza *middle reading*) zdygitalizowanych tekstów, nie zmieniła oblicza tej subdyscypliny językoznawstwa.

1 Korpus i artykuł zostały sfinansowane z grantu BEETHOVEN 2 Narodowego Centrum Nauki (NCN; numer 2016/23/G/HS2/00922) i Deutsche Forschungsgemeinschaft (DFG; numer 380115654). Wkład autorów w powstanie tekstu jest równy i wynosi po 33,3%.

Wczesne językoznawstwo korpusowe, z jego traktowaniem świadectwa tekstu jak rodzimego użytkownika języka, przypominało metodę filologiczną. Szybko jednak oko badacza zaczęło rejestrować przede wszystkim to, co w korpusie seryjne, ciężar dowodu zaś przesunął się na argumentację przede wszystkim ilościową. Nie pojedyncza osobliwa konstrukcja czy forma, ale właśnie to, co typowe, staje się przedmiotem zainteresowania badaczy. Tą drogą też coraz częściej podążają lingwiści interesujący się przeszłością języka. Szersze zastosowanie zaawansowanych technik statystycznych wymaga powiększenia skali korpusów. W językoznawstwie historycznym brak tekstów zawsze stanowił wąskie gardło, warto jednak zauważyć, że było to mniej dotkliwe, dopóki filolog pracował z fiszką i piórem, gdyż największym ograniczeniem było jego tempo pracy. Współcześnie, gdy przeszukiwanie zbiorów o objętości milionów czy nawet miliardów słów nie stanowi problemu, to właśnie niedostateczna liczba dawnych tekstów staje się największą przeszkodą dla badacza.

Nie znaczy to oczywiście, że w epoce poprzedzającej powstanie korpusów elektronicznych w językoznawstwie historycznym świadomość roli danych ilościowych nie istniała. Znakomitą pracą, w której przebieg zmian jakościowych jest śledzony poprzez precyzyjny opis ilościowy, jest opracowanie Ireny Bajerowej (1964), podobnie – tekst Anny Wierzbickiej (1966), by wymienić tylko dwie dawne prace.

Wróćmy więc do samych korpusów historycznych. Z niewielkim ryzykiem pomyłki można powiedzieć, że pierwsze takie korpusy dokumentowały język angielski od jego początków do XVIII w. (Rissanen 1992). Korpusem dawnej polszczyzny, który powstał jako pierwszy, a zarazem dokumentuje najstarszą warstwę języka, jest Korpus tekstów staropolskich (stworzony przez zespół Słownika staropolskiego IJP PAN, a opisany w pracy Twardzik, Górski 2003)[2]. Korpus ten obejmuje zasadniczo wszystkie znane polskie teksty ciągłe do roku 1500. Wiek XVI reprezentuje korpus tworzony przez Pracownię Słownika Polszczyzny XVI Wieku IBL PAN[3]. Oba te korpusy nie są lematyzowane ani opatrzone anotacją fleksyjną (morfosyntaktyczną). Okres 1600–1772 pokrywa Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772 r.) (KorBa, por. Gruszczyński, Adamiec, Ogrodniczuk 2013; Gruszczyński, Adamiec, Bronikowska, Kieraś, Modrzejewski, Wieczorek, Woliński 2022). Wiek XIX reprezentuje korpus polszczyzny XIX w. (Derwojedowa et al. 2014; Bilińska et al. 2016). Sami jego twórcy nazywają go mikrokorpusem, został on bowiem stworzony w określonym celu, który od początku ograniczał jego rozmiary, o czym szerzej piszemy niżej[4]. Wreszcie Narodowy Korpus Języka Polskiego

2 Korpus ten nie daje się przeglądać przez Internet, natomiast teksty w postaci plików XML można pobrać z adresu https://ijp.pan.pl/wp-content/uploads/2018/10/KorpusStp.zip. Można je przeglądać przy użyciu dowolnego programu konkordancyjnego pracującego z plikami tekstowymi.

3 Jest on dostępny publicznie pod adresem http://spxvi.edu.pl/korpus/.

4 Autorzy artykułu i jednocześnie twórcy korpusu składają w tym miejscu podziękowanie Magdalenie Derwojedowej oraz innym twórcom korpusów i historycznych oraz zbiorów tekstów literackich, którzy udostępnili teksty do naszego projektu.

(NKJP, por. Przepiórkowski et al. 2012) stanowi reprezentatywny obraz polszczyzny początków XXI w., jakkolwiek zawiera on również niewielkie ilości tekstów drugiej połowy XX w., a także trochę wcześniejszej prozy[5]. Widzimy więc, że wymienione korpusy nie pokrywają całej historii polszczyzny. Opisywany w niniejszym artykule korpus ma na celu wypełnienie tej luki.

W tym miejscu warto poczynić istotne rozróżnienie między korpusami historycznymi i diachronicznymi. Te pierwsze to korpusy synchroniczne dokumentujące pewną epokę historyczną. Te drugie to chronologicznie uporządkowane serie korpusów synchronicznych. Pierwsze pozwalają opisywać pewien stan języka w przeszłości, drugie – proces zmiany językowej.

Przykładem korpusu diachronicznego jest The Corpus of Late Modern English Texts (CLMET, por. De Smet 2005). Korpus ten, dokumentujący lata 1710–1920, składa się z trzech podkorpusów o bardzo zbliżonej budowie i podobnych rozmiarach, z których każdy obejmuje siedemdziesiąt lat. Pozwala to na dokonywanie bezpośrednich porównań pomiędzy tymi trzema podkorpusami.

Oczywiście granica pomiędzy korpusem historycznym i diachronicznym może być płynna. Przykładowo KorBa zasadniczo nie jest skonstruowana jako korpus diachroniczny. Pokrywa on jednak okres 172 lat (a więc niewiele mniej niż CLMET), okres, w którym zaszło wiele zmian, w tym zmian systemowych, doskonale więc może służyć do badania ich przebiegu. Użytkownik dzięki metadanym może tworzyć dowolne chronologicznie uporządkowane podkorpusy, jakkolwiek musi pamiętać o tym, że będą się one różniły zapewne zarówno wielkością, jak i budową.

Problem zróżnicowanej budowy podkorpusów korpusu diachronicznego jest zresztą nieusuwalny. Zauważmy, że w wypadku polszczyzny wiek XV reprezentują niemal wyłącznie teksty religijne i prawne, współcześnie dalece nie najważniejsze. Stopniowo pojawiają się nowe typy tekstów, a podstawowa dzisiaj prasa wyłania się na szerszą skalę dopiero w XIX w.

Piszemy o tym w kontekście planowanego przedsięwzięcia – stworzenia Narodowego Diachronicznego Korpusu Polszczyzny, który miałby scalić istniejące korpusy historyczne tak, by reprezentując wszystkie epoki, stanowiły korpus diachroniczny (Król et al. 2019), lecz także by bardzo wyraźnie podkreślić, że opisywany tutaj korpus jest korpusem historycznym, ale też i synchronicznym.

5 Ten krótki przegląd korpusów dokumentujących poszczególne epoki języka polskiego jest oczywiście niepełny. Istnieją mniejsze, bardziej wyspecjalizowane korpusy, np. tłumaczeń Biblii. Zainteresowanego czytelnika odsyłamy do artykułu: Pastuch et al. 2018.

2. Korpusy historyczne w polsko-niemieckim w projekcie BEETHOVEN 2. Geneza i założenia

Opisywany korpus został stworzony na potrzeby projektu naukowego w ramach grantu Narodowego Centrum Nauki oraz Deutsche Forschungsgemeinschaft BEETHOVEN 2: „Rozwój polskiego systemu aspektowego w ostatnich 250 latach na tle sąsiednich języków słowiańskich" (numer projektu w NCN: 2016/23/G/HS2/00922, numer projektu w DFG: 380115654).

Zestawienie historii aspektu czasownika, czyli opozycji czasowników dokonanych i niedokonanych, z lingwistyką korpusową może wydawać się zaskakujące, ale punktów zbieżnych było dosyć, by przedsięwzięcie się udało. Głównym zadaniem projektu było prześledzenie rozwoju systemu aspektowego polszczyzny od etapu trójek typu *czytać – przeczytać – przeczytywać*, *czytać – odczytać – odczytywać* do dzisiejszego stanu par aspektowych i rzadkich trójek, np. *tworzyć – stworzyć – stwarzać*. Ten schemat rozwoju zilustrowaliśmy w bazie 1816 dawnych i współczesnych trójek aspektowych od XVIII w. do dziś[6]. Strona projektu z bazą, korpusami i publikacjami to diaspol.uw.edu.pl.

Najważniejszym z punktu widzenia projektu był stworzony specjalnie korpus polsko-niemiecki, obejmujący ostatnie 270 lat, w którym łatwo sprawdzić tożsamość leksykalną czasowników polskich poprzez zestawienie ich z przekładem na język niemiecki. W korpusie tym (http://diaspol.uw.edu.pl/polniem/ i http://diaspol.uni-mainz.de/Beethoven/) otagowano potencjalne pary i trójki aspektowe (Łaziński, Meger, Woźniak 2022).

Kiedy tworzyliśmy bazę trójek i poszukiwaliśmy kontekstów dla członów opozycji w różnych okresach, stało się jasne, że luka w polskich korpusach, jaką stanowią późny wiek XVIII, wiek XIX w. i początki wieku XX, uniemożliwiałaby realizację takiego projektu. Dlatego zbudowaliśmy korpus polszczyzny 2. połowy XVIII i XIX w. wielkości 12 mln słów (diaspol.uw.edu.pl/korpus-tekstów-xix-w). Korpus, który powstaje niejako na marginesie większego projektu, musi być korpusem oportunistycznym[7]. Pod tym względem nasza koncepcja jest z gruntu odmienna niż koncepcja, która stała za stworzeniem Mikrokorpusu Gronowego Polszczyzny 1830–1918 (Derwojedowa et al. 2014; Bilińska et al. 2016). Ten ostatni miał być korpusem niezwykle starannie zrównoważonym na wzór korpusu *Słownika frekwencyjnego polszczyzny współczesnej* – co więcej, z uwzględnieniem zrównoważenia chronologicznego – a to dlatego, że powstał na użytek badań nad fleksją. Jego rozmiar miał więc znaczenie drugorzędne, dużo mniejsze niż zrównoważenie[8]. Korpus, który jest

6 Szczegółowy opis projektu zob.: Wiemer, Wrzesień-Kwiatkowska, Łaziński 2020.

7 Korpus oportunistyczny definiowany jest jako korpus, w którym nie dba się o zrównoważenie, ale do którego włącza się każdy dostępny tekst. W ten sposób poświęca się zrównoważenie dla zwiększenia rozmiaru korpusu (Halliday et al. 2004).

8 W tym miejscu warto dodać, że twórcy Mikrokorpusu, zdając sobie sprawę ze stale powiększającego się wolumenu zdygitalizwanych tekstów z epoki, skoncentrowali się na stworzeniu

przedmiotem niniejszego artykułu, powstał na potrzeby projektu, którego jądrem jest położona na styku gramatyki i semantyki leksykalnej kategoria aspektu, musi więc być możliwie duży. Współcześnie inicjatywy takie jak Wolne Lektury, Wikiźródła, Polona czy (dla polszczyzny w bardzo ograniczonym zakresie) Projekt Gutenberg dostarczają tekstów niechronionych prawem autorskim, które jest silną barierą w tworzeniu korpusów współczesnych. Trzeba jednak pamiętać, że proces dygitalizacji w dużej mierze nie jest motywowany potrzebami językoznawców, lecz badaczy z różnych dziedzin czy wręcz „konsumentów kultury", co powoduje, że dużo łatwiej o reprezentację literatury wysokiej niż np. tekstów prasowych, naukowych czy religijnych. Wreszcie nie należy zapominać, że jakość zapisu bywa różna; o ile w wypadku Wolnych Lektur teksty przechodzą pełną korektę wydawniczą, o tyle serwis Polona z założenia dostarcza wysokiej jakości reprodukcję książki, surowy OCR zaś jest tylko uzupełnieniem. W tym miejscu zastrzegamy, że z tego ostatniego serwisu wybieraliśmy jedynie te teksty, w których jakość optycznego rozpoznania obrazu nie budziła większych zastrzeżeń.

3. Struktura korpusu

Na korpus składa się 380 tekstów napisanych przez 148 różnych autorów (oraz jeden tekst anonimowy) o łącznej objętości 12 377 900 segmentów[9].

Najobszerniejszy tekst to *O przeprowadzeniu odosobnienia w zakładach więziennych* Aleksandra Moldenhawera (1840–1909) wydany w roku 1866. Najmniejsze teksty, pojedyncze wiersze lub opowiadania, liczą po kilkaset segmentów.

Zróżnicowanie chronologiczne korpusu ukazuje poniższa tabela:

Tabela 1. Zróżnicowanie chronologiczne korpusu

okres	liczba segmentów	procent całości
1800–1825	244548	~2%
1826–1850	397157	~3%
1851–1875	2078207	~17%
1876–1900	4510698	~37%
1901–1933	5060184	~41%

analizatora morfologicznego, który uwzględniał nie tylko ówczesną fleksję, ale także wciąż nie do końca znormalizowaną ortografię. Opisywany projekt miał przede wszystkim stworzyć narzędzie dla twórców korpusów.

9 Pojęcie segmentu, wprowadzone w pracy Adama Przepiórkowskiego (2004) i powszechnie używane w polskim językoznawstwie korpusowym, obejmuje tradycyjnie rozumiane wyrazy, ruchome końcówki i znaki interpunkcyjne, a także inicjały, symbole i tym podobne mniej istotne elementy tekstu.

Ponad połowa jednostek odnotowanych w korpusie pochodzi z tekstów, których miejscem wydania jest Warszawa. Lwów (13%) i Kraków (10%) pozostają daleko w tyle. Wśród miejsc wydania uwzględnionych w korpusie tekstów widnieją również Paryż, Mikołów, Kijów czy Podgórze (obecnie dzielnica Krakowa).

Nie unikaliśmy tłumaczeń – jest ich w korpusie 37.

4. Platforma techniczna

Korpus udostępniony jest on-line pod adresem http://www.diaspol.uw.edu.pl/korpus-tekstow-xix-w/. Program konkordancyjny jest oparty na CWB (The IMS Open Corpus Workbench, http://cwb.sourceforge.net), bardzo popularnym narzędziu (a właściwie zespole narzędzi) służącym do przeszukiwania korpusów. Z kolei po stronie użytkownika (tzw. front-end) został oparty na narzędziu ParaVoz 2, z którego wcześniejszej wersji korzysta także wielojęzyczny korpus Parasol (von Waldenfels 2015; https://bitbucket.org/rvwfels/paravoz2/src/master/).

Teksty korpusu są zapisane w formacie kolumnowym[10], tj. wyraz tekstowy oraz lemat i tag gramatyczny umieszczone zostały w jednym wierszu, oddzielone tabulatorem. Nagłówek każdego z plików (*header*), który zawiera informację natury bibliograficznej, ma format XML.

Nagłówek nie jest bardzo rozbudowany i zawiera następujące informacje: autor, tytuł tekstu, data wydania, miejsce wydania, wydawca, tekst oryginalny czy tłumaczenie, wreszcie źródło, z którego pozyskaliśmy tekst (Korpus XIX w., Wolne Lektury, Wikiźródła, Polona). Wszystkie te informacje mogą służyć do ograniczenia wyszukiwania, a więc w korpusie można np. wyszukać wyraz *pies* tylko w tekstach Józefa Ciechońskiego bądź tylko w powieści *Wodzirej*. Ponieważ korpus zawiera również tłumaczenia, można je łatwo wykluczyć z wyszukiwania.

Korpus został otagowany morfologicznie (morfosyntaktycznie) za pomocą tagera dla wieku XIX (Bilińska et al. 2016).

5. Wyszukiwanie w korpusie

Zapytania można formułować w prostych w obsłudze okienkach w sekcji „Basic Search". Okienko „Token" służy do wyszukiwania słów tekstowych (form wyrazowych

10 Format kolumnowy (VRT) może być uznany za krok wstecz w stosunku do formatu w pełni opartego na XML, jak to ma miejsce np. w Narodowym Korpusie Języka Polskiego (por. Przepiórkowski et al. 2012). Ten ostatni co prawda jest bardziej elastyczny i umożliwia zaopatrzenie korpusu w znacznie bogatszą anotację, z drugiej strony jednak pliki w formacie kolumnowym są dużo mniejsze i czytelniejsze.

w tekście), okienko „Lexeme" do wyszukiwania leksemów z uwzględnieniem odmiany. Wyszukiwać można wyrazy zaczynające się lub kończące daną literą albo ciągiem liter za pomocą pól wyboru „Begins with" lub „Ends with". Można też przy pomocy funkcji „Case sensitive" uwzględnić wielkie i małe litery, z założenia ignorowane w wyszukiwaniu.

Korzystanie z trzeciego okienka sekcji Basic Search: „Gram. tag", umożliwiającego wyszukiwanie za pomocą cech gramatycznych, wymaga znajomości tagów części mowy stosowanych w Narodowym Korpusie Języka Polskiego. Użytkownicy przyzwyczajeni do pracy z wyszukiwarką morfosyntaktyczną NKJP Poliqarp mogą w pierwszej chwili uznać sposób dodawania ograniczeń gramatycznych za nieco skomplikowany, tagset stosowany w polskich korpusach ma bowiem charakter pozycyjny, tj. stanowi sekwencję kodów oznaczających wartości poszczególnych kategorii gramatycznych. Przykładowo tag oznaczający przymiotnik liczby pojedynczej w narzędniku i rodzaju żeńskim niezaprzeczony ma postać: adj:sg:inst:f:pos. O ile w Poliqarpie można wyszukać przymiotniki w rodzaju żeńskim i liczbie pojedynczej za pomocą zapytania [pos=adj & gnd=f & nmb=sg], o tyle w opisywanym interfejsie tag jest traktowany jako pewna całość, trzeba więc użyć wyrażeń regularnych, by nieinteresujące nas elementy zmienić na dowolne ciągi liter. Odnośne zapytanie miałoby więc postać: adj:sg:.*:f:.*, pytanie zaś o dowolny segment rodzaju żeńskiego miałoby postać: .*f.*[11].

Warunki zapytania można ze sobą łączyć. Jeśli np. wpiszemy w okienku „Token" literę „c", zaznaczymy opcję „Ends with", a w okienku „Gramm. Tag" wpiszemy „inf", otrzymamy wszystkie bezokoliczniki zakończone na *-c*.

Każde zapytanie sformułowane w sekcji Basic Search pokazuje się w postaci poprawnego zapytania w języku zapytań CQP w oknie „CQP Search", z którego mogą korzystać bezpośrednio bardziej doświadczeni korpusowo użytkownicy. Dla bezokoliczników zakończonych na *-c* zapytanie CQP ma postać [word=".*c"%c & tag=".*inf.*"]. W tym oknie można również modyfikować zapytanie skonstruowane w sekcji podstawowej w celu uzyskania bardziej skomplikowanych zapytań (np. użycie operatora alternatywy „|" albo nierówności „!=").

Wyszukiwanie leksemów zilustrujemy pytaniem ze złożonymi warunkami. Wystąpienia leksemu PIES w dopełniaczu wyszuka zapytanie:

[lemma="pies" & tag=".*gen.*"].

To zapytanie możemy dodatkowo ograniczyć za pomocą metadanych, tak by program przeprowadził wyszukiwanie tylko w tekstach Gabrieli Zapolskiej wydanych przed rokiem 1896:

11 Kropka w wyrażeniach regularnych zastępuje dowolny znak, gwiazdka oznacza, że kropka może być powtórzona dowolną liczbę (w tym zero) razy – w praktyce oznacza to „dowolny ciąg znaków".

[lemma=”pies” & tag=”.*gen.*”]::match.meta_autor=”Gabriela Zapolska”& int(match.meta_data_wydania)<1896].

Wyszukiwanie słów tekstowych, leksemów oraz form określonych przez wartości kategorii gramatycznych może obejmować wiele wyrazów lub alternatywę. Służy do tego znak „+”, który otwiera dodatkowy wiersz z polami wyszukiwania oraz pozwala oznaczyć długość odstępu między wyszukiwanymi słowami.

Wyniki wyszukiwania prezentowane są w postaci konkordancji. Każde dopasowanie wraz z szerszym kontekstem przedstawione jest w pojedynczym wierszu. W panelu górnym dostępne są informacje o zapytaniu, liczbie wyników w korpusie i o języku głównym. Możliwa jest zmiana sposobu wyświetlania wyników na tryb KWIC (Keyword in Context), który może być wygodniejszy w niektórych przypadkach. Dostępne są także ikony pozwalające zobaczyć metadane dla każdego dopasowania, sortowanie wyników i zapisywanie ich do pliku w formacie TSV (Tab-Separated Values). Ten ostatni format zawiera także wszystkie metadane, co pozwala pracować nad danymi w arkuszu kalkulacyjnym lub programie statystycznym (R, Statistica itp.).

6. Zakończenie

Prezentowany korpus zapełnia lukę w narzędziach przeznaczonych do badań historycznojęzykowych polszczyzny, a mianowicie dostarcza materiału empirycznego do badań nad wiekiem XIX i pierwszą połową wieku XX. Korpus ten z pewnością nie może zastąpić korpusu starannie zrównoważonego. Bez wątpienia obarczony jest pewnymi słabościami: niedostateczną reprezentacją pierwszej połowy XIX w. oraz niektórych stylów funkcjonalnych, równoczesną przewagą tekstów fikcjonalnych. Nawet jednak jeśli weźmiemy te słabości pod uwagę, dochodzimy do wniosku, że prezentowany korpus oferuje badaczowi dostęp do dość obfitych danych. Ponadto z każdym rokiem przybywa zdygitalizowanych tekstów reprezentujących interesujący nas okres. Stwarza to możliwość rozwoju korpusu zarówno w kierunku jego powiększania, jak i zapewnienia większej reprezentatywności. Oczywiście podstawowym zastosowaniem korpusu są badania historycznojęzykowe, służy on jednak również jako narzędzie do badań filologicznych, historycznych czy prawniczych. Twórcy korpusu będą się cieszyli, jeśli będzie on wykorzystywany szeroko.

Literatura

Bajerowa I., 1964, *Kształtowanie się systemu polskiego języka literackiego w XVIII wieku*, Wrocław.

Bilińska J., Derwojedowa M., Kieraś W., Kwiecień M., 2016, *Mikrokorpus polszczyzny 1830–1918*, „Komunikacja Specjalistyczna" nr 11, s. 149–161.

Campbell L., 1999, *Historical Linguistics. An Introduction*, Cambridge, Mass.

CLMET: The Corpus of Late Modern English Texts, [on-line:] https://perswww.kuleuven.be/~u0044428/clmet.htm (dostęp: 20 III 2023).

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

De Smet, 2005, *A Corpus of Late Modern English Texts*, „Icame Journal" 29, s. 69–82.

Gruszczyński W., Adamiec D., Bronikowska R., Kieraś W., Modrzejewski E., Wieczorek A., Woliński M., 2022, *The Electronic Corpus of 17th- and 18th-century Polish Texts*, „Language Resources and Evaluation" 56, s. 309–332, https://doi.org/10.1007/s10579-021-09549-1.

Gruszczyński W., Adamiec D., Ogrodniczuk M., 2013, *Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 roku) – prezentacja projektu badawczego*, „Polonica" XXXIII, s. 309–316.

Halliday M.A.K., Teubert W., Yallop C., Čermáková A., 2004, *Lexicology and Corpus Linguistics: An Introduction*, London – New York.

KorBa: Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772 r.), [on-line:] https://korba.edu.pl.

Król M., Derwojedowa M., Górski R.L., Gruszczyński W., Opaliński K.W., Potoniec P., Woliński M., Kieraś W., Eder M., 2019, *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" IC, s. 92–101.

Łazinski M., Meger A., Woźniak M., 2022, *Korpus Polsko-Niemiecki Uniwersytetu Warszawskiego i Uniwersytetu Gutenberga (PolGerCorp)*, „Tekst i Dyskurs – Text und Dyskurs" 16, s. 379–390, https://doi.org/ 10.7311/tid.16.2022.18.

NKJP: Narodowy Korpus Języka Polskiego, [on-line:] http://nkjp.pl/.

Pastuch M., Duda B., Lisczyk K., Mitrenga B., Przyklenk K., Sujkowska-Sobisz K., 2018, *Digital Humanities in Poland from the Perspective of the Historical Linguist of the Polish Language: Achievements, Needs, Demands*, „Digital Scholarship in the Humanities" 33/4, s. 857–873, https://doi.org/10.1093/llc/fqy008.

Przepiórkowski A., 2004, *Korpus IPI PAN. Wersja wstępna*, Warszawa.

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.

Rissanen M., 1992, *The Diachronic Corpus as a Window to the History of English*, [w:] J. Svartvik (red.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991*, Berlin – New York, s. 185–210, https://doi.org/10.1515/9783110867275.185.

Twardzik W.B., Górski R.L., 2003, *Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie*, [w:] S. Gajda (red.), *Językoznawstwo w Polsce. Stan i perspektywy*, s. 155–157.

von Waldenfels R., 2015, *ParaViz: A Vizualization Tool for Crosslinguistic Functional Comparisons Based on a Parallel Corpus*. [w:] G. Grigonytė, S. Clematide, A. Utka, M. Volk

(red.), *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, May 11–13, 2015, Vilnius, Lithuania*, Linköping, s. 32–36.

Wiemer B., Wrzesień-Kwiatkowska J., Łaziński M., 2020, *Badania aspektu w językach polskim, czeskim i rosyjskim za pomocą korpusów i baz danych (pierwsze podsumowanie tematu)*, „Forum Lingwistyczne" nr 7, s. 45–58, https://doi.org/10.31261/FL.2020.07.04.

Wierzbicka A., 1966, *System składniowo-stylistyczny prozy polskiego renesansu*, Warszawa.

Corpus of the 19th Century of the Warsaw University and IJP PAN
Abstract

The article describes a historical corpus which documents the 19th and early 20th century. The corpus was created as part of a research grant whose objective was to investigate the development of the aspectual system of Polish in the last 250 years against the background of Czech and Russian. An important resource for this investigation was a database of aspectual triplets, which, in turn, was based on materials such as text corpora. Since there was no large corpus of the 19th and early 20th century available, there was a need to bridge this gap. In the course of the project, such corpus was made and it is now publicly accessible with no restrictions. This comprehensive corpus contains over 12 million contemporary words. Its texts originate from major Polish virtual libraries. It is POS-tagged with a tagger dedicated for 19th century texts. A web-based concordancer, an adjusted version of ParaVoz, allows for querying the corpus. The queries may be constrained by metadata.