Włodzimierz Gruszczyński
Instytut Języka Polskiego PAN
ORCID 0000-0001-9406-1354

KORPUSY JĘZYKOWE NARZĘDZIEM PRACY HISTORYKA JĘZYKA

WSTĘP

Korpusy językowe, czyli duże zbiory tekstów w formie cyfrowej opatrzonych dodatkowymi informacjami, zwykle udostępniane wraz z narzędziami ułatwiającymi ich przeszukiwanie, stały się w ostatnich latach jednym z podstawowych źródeł danych dla lingwistów. Od kilkunastu lat funkcjonuje pojęcie lingwistyki korpusowej, której założenia metodologiczne sprowadzają się do wyrafinowanego wykorzystywania korpusów i wyciągania wniosków na temat języka na ich podstawie (por. na przykład [Lewandowska-Tomaszczyk 2005], [Świdziński 2006]). Początkowo korpusy były zbiorami tekstów służących jako podstawy materiałowe do badań synchronicznych współczesnych języków etnicznych. W Polsce pierwsze korpusy tworzono na przełomie XX i XXI w. Prace nad nimi były rozproszone, a korpusy powstające wtedy nie mogły imponować wielkością. Pierwszym większym korpusem języka polskiego był korpus IPI PAN [Przepiórkowski 2004], liczący ponad 100 mln segmentów, czyli jednostek odpowiadających w bardzo dużym przybliżeniu słowom graficznym. Korpus IPI PAN stanowił pierwszy etap tworzenia Narodowego Korpusu Języka Polskiego (NKJP) dostępnego od 2012 roku pod adresem http://nkjp.pl/. NKJP powstał dzięki współpracy czterech instytucji: Instytutu Podstaw Informatyki PAN (koordynator), Instytutu Języka Polskiego PAN, Wydawnictwa Naukowego PWN oraz Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego w ramach projektu badawczego rozwojowego finansowanego przez MNiSW. Jest to największy do dziś korpus tekstów polskich udostępniany publicznie (por. [Przepiórkowski i in. 2014]). Niestety nie jest aktualizowany i powiększany.

Po etapie tworzenia jednojęzycznych korpusów tekstów współczesnych podjęto próby tworzenia innych, bardziej wyrafinowanych korpusów, na przykład tak zwanych korpusów równoległych istotnych dla translatologii i językoznawstwa porównawczego [Gruszczyńska, Leńko-Szymańska 2016]. Przyszedł też czas na zainteresowanie się historią języka. Powstawać zaczęły korpusy dawnych tekstów zaopatrzone w narzędzia podobne do tych, które wcześniej stosowano w korpusach współczesnych (a niekiedy dokładnie te same). Obecnie istnieją korpusy historyczne (bardzo różnej wielkości) wielu języków (a właściwie tekstów w tych językach).

Niektóre z tych korpusów stanowią tak zwane podkorpusy historyczne dużych korpusów narodowych. Tak jest na przykład w korpusie czeskim (https://www.korpus.cz/) czy rosyjskim (http://www.ruscorpora.ru)[1].

Nie inaczej jest w Polsce. W ostatnich latach zaczęły powstawać korpusy tekstów dawnych, a niektóre z nich są już dostępne. Zanim jednak je omówimy, spróbujmy dokonać niezbędnych ustaleń terminologicznych i wskazać pewne zasadnicze różnice pomiędzy korpusem współczesnym a korpusem historycznym.

USTALENIA TERMINOLOGICZNE

REPOZYTORIA TEKSTÓW I BIBLIOTEKI A KORPUSY JĘZYKOWE

Od dość dawna istnieją zbiory tekstów dawnych w postaci cyfrowej, które nazywane są repozytoriami lub bibliotekami. Takie repozytoria i biblioteki (zwłaszcza jeśli zawarte są w nich teksty przeszukiwalne, a nie w postaci skanów[2]) są oczywiście w jakimś sensie korpusami. Mają jednak podstawową wadę z punktu widzenia lingwisty – każdy z nich stanowi odrębną całość, co sprawia, że wyszukiwanie w nich interesujących językoznawcę obiektów językowych jest bardzo trudne, a przede wszystkim czasochłonne. Niemożliwe jest w krótkim czasie uzyskanie istotnych danych frekwencyjnych. Korzystanie z tych zbiorów w celu pozyskania danych lingwistycznych jest po prostu niewygodne. Korpusy tworzone specjalnie do badań języka takich wad nie mają lub mieć nie powinny. Dlatego repozytoriów, bibliotek cyfrowych i tym podobnych zbiorów tekstów nie będziemy tu nazywać korpusami, choć w dalszym ciągu o niektórych z nich poinformujemy.

KORPUS JĘZYKA CZY KORPUS TEKSTÓW Z OKREŚLONEGO OKRESU?

Oczywiście niemożliwe jest utworzenie „korpusu języka" w dosłownym znaczeniu. Istniejące korpusy są zbiorami tekstów, ale – zwłaszcza te współczesne – ze względu na ich wielkość i takie cechy, jak zrównoważenie i reprezentatywność, przyjęło się traktować jako reprezentację systemu językowego. Takie podejście jest bez wątpienia uproszczeniem, jednak w dalszym ciągu – zgodnie ze zwyczajem środowiskowym – będziemy używać określenia „korpus języka X" jako równoważnego z określeniem „korpus tekstów w języku X".

ZRÓWNOWAŻENIE I REPREZENTATYWNOŚĆ KORPUSU HISTORYCZNEGO

Wspomniane wyżej cechy – zrównoważenie i reprezentatywność korpusu – rozumiemy tak jak twórcy NKJP: „Reprezentatywność to odnoszenie się do jakiejś rzeczywistości istniejącej poza korpusem. Zrównoważenie zaś to dbałość o taką budowę korpusu, by żaden składnik na żadnym z poziomów nie dominował nad innymi" [Przepiórkowski i in. 2012, 26].

---

[1] Por. [Hebal-Jezierska 2014].

[2] Możliwe jest oczywiście tworzenie dokumentów zeskanowanych, które można przeszukiwać, por. [Bień 2012].

Nawet w wypadku korpusów współczesnych tekstów zachowanie reprezentatywności i zrównoważenia jest bardzo trudne. Można do nich co najwyżej dążyć i osiągnąć względnie zadowalający rezultat. Znacznie trudniej jest zachować zrównoważenie i reprezentatywność w wypadku korpusu historycznego. Wynika to z ograniczeń charakterystycznych dla badań historycznojęzykowych. Mamy dostęp tylko do części tekstów (czasem dość przypadkowych, jak na przykład w okresie staropolskim) i są to wyłącznie teksty pisane. Wiedza o piśmiennictwie poszczególnych okresów pozostaje niekompletna. Nie znamy w pełni struktury zbioru tekstów powstałych w badanej epoce. Zachowane teksty pochodzą często tylko, albo głównie, z określonych regionów, ich autorzy pochodzą w znacznej większości ze środowisk dawnej elity intelektualnej. Nie może więc być mowy o reprezentatywności i zrównoważeniu w takim stopniu, w jakim możliwe jest to dla korpusów współczesnych [Adamiec 2015].

KORPUSY DIACHRONICZNE CZY HISTORYCZNE?

Oczywiście możliwe jest zbudowanie zarówno korpusu diachronicznego, zawierającego teksty z różnych okresów tego samego języka etnicznego, jak i korpusu historycznego (synchronicznego). Obecnie jednak powstają w Polsce przede wszystkim korpusy tego drugiego typu, choć oczywiście nie są one w pełni synchroniczne (zawierają teksty z różnej długości okresów – por. niżej). Na ich podstawie możliwe będzie, jak się wydaje, stworzenie jednego korpusu diachronicznego, który obejmie całą dobę piśmienną w historii języka polskiego. Plany utworzenia takiego korpusu już istnieją (por. [Król i in. 2019]).

POLSKIE KORPUSY HISTORYCZNE

Poniżej przedstawione zostaną bardzo skrótowo wybrane polskie korpusy historyczne[3], w części znajdujące się jeszcze w fazie budowy lub rozbudowy[4].

KORPUS STAROPOLSKI

Po raz pierwszy określenie „korpus" w odniesieniu do historycznej polszczyzny pojawiło się chyba w projekcie „Elektroniczny korpus ciągłych tekstów staropolskich (do 1500 r.)" realizowanym w IJP PAN w Krakowie pod kierownictwem Wacława

[3] Nie uwzględniamy tu wielu korpusów i repozytoriów, mogących w pewnym stopniu zastępować korpusy, ograniczonych tematycznie lub gatunkowo. Ciekawymi zasobami tego typu są na przykład Elektroniczne Repozytorium Rot Wielkopolskich eROThA dostępne pod adresem https://rotha.ehum.psnc.pl/ (por.: [Włodarczyk, Kopaczyk, Kozak 2020]) czy Szesnastowieczne Przekłady Ewangelii dostępne pod adresem https://ewangelie.uw.edu.pl/. Nie uwzględniamy też korpusów, do których dostęp jest w jakimś stopniu ograniczony, na przykład Polish Diachronic Online Corpus PolDi (informacje na temat tego korpusu podają [Pałka, Kwaśnicka-Janowicz 2017, 164–165]).

[4] Korpusy te omówione zostały wcześniej w publikacjach [Pałka, Kwaśnicka-Janowicz 2017] oraz [Pastuch i in. 2018], jednak informacje tam zawarte dotyczą sytuacji sprzed (ponad) trzech lat i warto je tu uaktualnić.

Twardzika, którego rezultatem miał być korpus tekstów staropolskich. Korpus ten dostępny jest w postaci transkrybowanych tekstów udostępnianych w formacie pdf lub xml pod adresem https://ijp.pan.pl/publikacje-i-materialy/zasoby/korpus-tekstow-staropolskich/. Korpus nie jest znakowany morfosyntaktycznie, nie jest też zaopatrzony w jakąkolwiek wyszukiwarkę. Jest to więc raczej repozytorium tekstów staropolskich, które stanowi dobry punkt wyjściowy do stworzenia korpusu staropolskiego w rozumieniu dziś powszechnie przyjmowanym. Prace nad stworzeniem takiego korpusu trwają w zespole kierowanym przez Ewę Deptuchową (por. [Klapper, Kołodziej 2014], [Klapper, Kołodziej 2015], [Deptuchowa i in. 2020]).

KORPUS IMPACT

W latach 2009–2012 w ramach dużego międzynarodowego projektu IMPACT (Improving Access to Text), którego celem było udoskonalenie narzędzi do automatycznego rozpoznawania tekstów dawnych, powstał korpus polskich tekstów z XVI, XVII i XVIII wieku o łącznej długości ok. 1,5 mln segmentów w wersji transkrybowanej [Bień 2014]. Ze względu na cel, jaki przyświecał twórcom korpusu, zawiera on w swej podstawowej wersji teksty niezwykle pieczołowicie transliterowane [Bień 2015]. Korpus dostępny jest on-line w dwóch wersjach (tak zwanej jedno- i dwuwymiarowej) pod adresem: https://szukajwslownikach.uw.edu.pl/IMPACT_GT_1/. Ważną jego cechą jest możliwość przechodzenia od każdego wyszukanego cytatu do odpowiedniej strony skanu oryginału.

KORPUS POLSZCZYZNY XVI WIEKU

Korpus powstaje w Pracowni Słownika Polszczyzny XVI w. Instytutu Badań Literackich PAN [Opaliński, Potoniec 2020]. Obecnie udostępniany jest zalążek korpusu (9 tekstów), będący wynikiem pierwszego etapu prac (2012–2017). Teksty są transliterowane zgodnie z założeniami przyjętymi w *Słowniku polszczyzny XVI wieku*, a więc bardzo dokładnie oddają grafię oryginałów. Oprócz tego teksty dostępne są w postaci transkrybowanej. Korpus można przeszukiwać zarówno w wersji oryginalnej (transliteracja), jak i transkrybowanej. Teksty w zalążku korpusu są znakowane fleksyjnie i strukturalnie, dzięki czemu można w nim wyszukiwać nie tylko formy tekstowe, ale i formy podstawowe. Dostęp: http://spxvi.edu.pl/korpus/.

Pod tym samym adresem znajduje się też repozytorium wszystkich tekstów z kanonu *Słownika polszczyzny XVI w.*

ELEKTRONICZNY KORPUS TEKSTÓW POLSKICH XVII I XVIII W. (KORBA)

Najbardziej zaawansowanym i największym obecnie historycznym korpusem tekstów polskich jest korpus obejmujący teksty z XVII i XVIII wieku (na pierwszym etapie tylko do 1772 roku), a więc z epoki baroku (stąd akronim KorBa, czyli Korpus Barokowy) i dużej części oświecenia (do 1800 roku)[5]. Korpus powstaje w Instytucie Języka Polskiego PAN we współpracy z Zespołem Inżynierii Lingwistycznej

[5] Dokładniejsze dane na temat korpusu dostępne są na stronie https://korba.edu.pl/ w zakładce „Informacje" oraz w artykule [Gruszczyński i in. 2020].

z Instytutu Podstaw Informatyki PAN w ramach projektów finansowanych przez NPRH[6], kierowanych przez autora niniejszego tekstu. KorBa w obecnie udostępnianej wersji zawiera ok. 13,5 mln segmentów z ok. 700 tekstów z lat 1601–1772. Docelowa objętość będzie dwukrotnie większa, tzn. ok. 25 mln segmentów, z czego ok. 8 mln z tekstów z oświecenia (tzn. z okresu 1740–1800).

Wszystkie teksty w korpusie dostępne są w dwu wersjach – w transliteracji i w transkrypcji. Transliteracja dokonywana jest z dokładnością do grafemów, nie uwzględnia ich wariantów (na przykład różne warianty grafemów „r" czy „ż" są zawsze transliterowane odpowiednio jako *r* i *ż*, a ligatury tansliterowane są jako osobne litery). Transkrypcja dokonywana jest automatycznie, co sprawia – rzecz jasna – że pojawiają się w niej niekiedy błędy[7]. Teksty są bardzo dokładnie oznakowane zarówno pod względem morfosyntaktycznym (por. [Kieraś i in. 2017]), jak i strukturalnym (dokładny podział na jednostki tekstu). Oznakowane są też wszystkie fragmenty w językach obcych, z dokładnością do języka. Znakowanie strukturalne i językowe (języki obce) zostało wykonane ręcznie, natomiast morfosyntaktyczne – automatycznie za pomocą dwóch tagerów (programów do znakowania) wytrenowanych na ręcznie oznakowanym podkorpusie zawierającym pół miliona segmentów. Korpus udostępniany jest za pośrednictwem wyszukiwarki MTAS pod adresem https://korba.edu.pl [Bronikowska 2015].

KORPUS POLSZCZYZNY 1830–1918

Ten stosunkowo niewielki korpus utworzony został w latach 2013–2017 w ramach grantu *Automatyczna analiza fleksyjna tekstów polskich z lat 1830–1918 z uwzględnieniem zmian w odmianie i pisowni* przez zespół pod kierunkiem Magdaleny Derwojedowej w Instytucie Języka Polskiego UW (por.: [Bilińska i in. 2016], [Bilińska, Kwiecień, Derwojedowa 2018], [Derwojedowa 2020]). Korpus utworzony został na innej zasadzie niż większość dużych korpusów służących najczęściej do badań leksykalnych. Składa się z wylosowanych próbek stosunkowo dużej liczby tekstów, dzięki czemu mimo małej objętości jest wystarczająco zróżnicowany, by móc służyć jako podstawa materiałowa do badań historii fleksji XIX-wiecznej polszczyzny. Jego objętość wynosi ok. 1 mln słów (1000 próbek o długości ok. 1000 słów każda). W korpusie zachowane zostało zrównoważenie pod względem stylistycznym[8], jest on też, oczywiście, oznakowany pod względem morfosyntaktycznym [Kieraś, Woliński 2018]. Korpus udostępniony jest pod adresem https://szukajwslownikach.uw.edu.pl/f19/ (wyszukiwarka Poliqarp), a poza tym pod adresem http://korpus19.nlp.ipipan.waw.pl/query_corpus/ wraz z korpusem 100 powieści Kraszewskiego anotowanym automatycznie za pomocą tagera Concraft-2 (wyszukiwarka MTAS).

---

[6] Pierwszy projekt pt. „Elektroniczny korpus tekstów polskich XVII i XVIII w." (0036/NPRH2/H11/81/2012) zrealizowany został w latach 2013–2018, drugi – pt. „Rozbudowa Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. i jego integracja z *Elektronicznym słownikiem języka polskiego XVII i XVIII w.*" (0413/NPRH7/H11/86/2018) realizowany jest obecnie (od 1 stycznia 2019 do 31 grudnia 2023).

[7] W kolejnej wersji korpusu zastosowany zostanie nowy program dokonujący transkrypcji (tak zwany transkryber), co prawdopodobnie istotnie zmniejszy liczbę błędów.

[8] Odmiany stylistyczne wyodrębniono według zasad przyjętych w *Słowniku frekwencyjnym polszczyzny współczesnej* [Kurcz i in. 1990].

CHRONOPRESS, CZYLI PORTAL TEKSTÓW PRASOWYCH

Korpus i wyszukiwarka ChronoPress powstały w ramach prac konsorcjum naukowego „CLARIN – Polskie wspólne zasoby językowe i infrastruktura technologiczna" pod kierunkiem Adama Pawłowskiego. Korpus zawiera ok. 56 tys. fragmentów tekstów prasowych z lat 1945–1954, wylosowanych według zasad gwarantujących reprezentatywność zbioru. Teksty są oznakowane morfosyntaktycznie. Każda próbka ma długość ok. 300 wyrazów tekstowych, co daje łącznie ok. 17 mln wyrazów tekstowych w całym korpusie. ChronoPress ma pewne niestandardowe, ciekawe funkcjonalności dające bardzo duże możliwości badawcze[9]. Dostęp: http://chronopress.clarin-pl.eu/#!start.

KORPUS DYSKURSU PARLAMENTARNEGO

Korpus ten stworzony i opracowany został w IPI PAN w latach 2011–13 i 2016–18 przez zespół pod kierunkiem Macieja Ogrodniczuka [Ogrodniczuk 2018]. Jest to zbiór anotowanych lingwistycznie tekstów z posiedzeń plenarnych Sejmu i Senatu RP, interpelacji i zapytań poselskich oraz posiedzeń komisji od roku 1919 do chwili obecnej (uzupełniany na bieżąco).

Dostęp: http://sejm.nlp.ipipan.waw.pl/ za pośrednictwem wyszukiwarki MTAS. Korpus ma oczywiście znaczenie nie tylko dla lingwistów, ale też dla historyków i politologów.

MOŻLIWOŚCI INTEGRACJI ZASOBÓW HISTORYCZNOJĘZYKOWYCH

Jak już wyżej stwierdzono, korpusy historyczne mogą, a nawet powinny być ze sobą łączone tak, by powstał korpus diachroniczny, najlepiej taki, który objąłby cały okres istnienia języka polskiego. Próby takie już podejmowano. W ramach projektu *Przebiegi zmian gramatycznych i leksykalnych w historii języka polskiego – metody korpusowe i kwantytatywne w językoznawstwie diachronicznym* kierowanego przez Rafała L. Górskiego powstał korpus utworzony w znacznej mierze z tekstów zaczerpniętych z już istniejących lub właśnie powstających korpusów oraz z niektórych bibliotek cyfrowych (m.in. Wolne Lektury). Korpus ma charakter oportunistyczny, jest znakowany tagerem z analizatorem morfologicznym Morfeusz (przeznaczonym do współczesnej polszczyzny). Ma więc wiele wad, ale mimo to mógł posłużyć jako podstawa do interesujących badań historycznojęzykowych, których metody i wyniki zostały przedstawione w książce [Górski, Król, Eder 2019].

Ważną i ciekawą próbę wykorzystania połączonych korpusów historycznych jako źródła wiedzy na temat zmian we fleksji stanowi projekt ChronoFleks[10] zrealizowany w Instytucie Podstaw Informatyki PAN pod kierunkiem Marcina Wolińskiego [por. Woliński, Kieraś 2020]. W ramach projektu powstało wiele narzędzi ułatwiających korzystanie z korpusów (w szczególności korpusów historycznych) oraz prototyp przekrojowego słownika fleksyjnego języka polskiego dostępnego przez Internet

---

9 Por. tekst A. Pawłowskiego w niniejszym tomie.

10 Pełna nazwa projektu: „Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja". Okres realizacji 2015–2019.

umożliwiającego wizualizację zmian paradygmatów w czasie (system ChronoFleks). Wyniki projektu i sam system dostępne są pod adresem http://chronofleks.nlp.ipipan.waw.pl/. Obecnie system wykorzystuje trzy korpusy: XVII–XVIII wieku, XIX wieku i Narodowy Korpus Języka Polskiego.

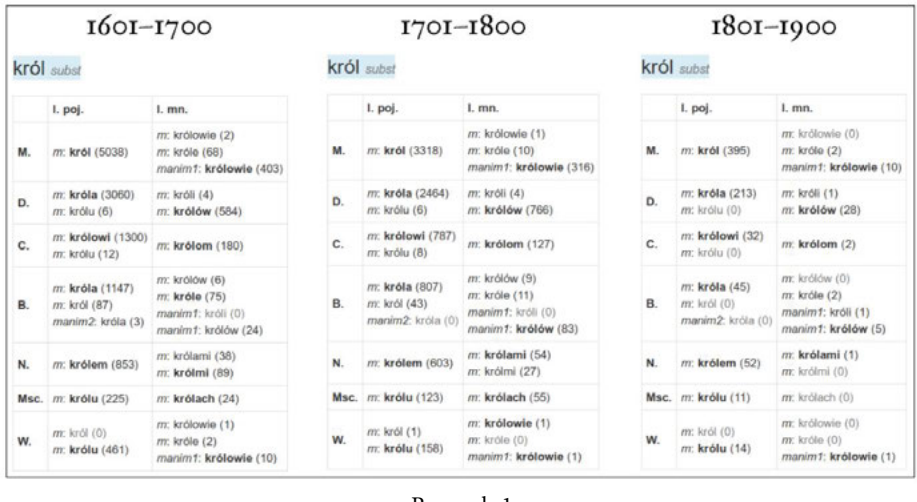
PRZYKŁADY WYKORZYSTANIA KORPUSÓW HISTORYCZNYCH

Poniżej podajemy wybrane przykłady wykorzystania niektórych spośród omówionych wcześniej korpusów w badaniach historycznojęzykowych. Interesujące jest to, że można je również wykorzystać do weryfikacji pewnych sądów funkcjonujących w językoznawstwie normatywnym, w szczególności takich, które odwołują się do kryterium historycznego oceny form językowych.

USTALANIE DYNAMIKI ZMIAN JĘZYKOWYCH

W okresie średniopolskim, jak wiadomo, zachodziło wiele istotnych zmian w systemie językowym, w szczególności we fleksji nominalnej. Utworzył się wówczas rodzaj męskoosobowy, zanikły pewne wykładniki form fleksyjnych. Na rysunku 1, który jest wydrukiem z systemu ChronoFleks, zaobserwować można wycofywanie się formy narzędnika *królmi* i zastępowanie jej formą *królami* oraz ekspansję formy biernika lm. równej dopełniaczowi, czyli tworzenie się rodzaju męskoosobowego.

Innym ważnym procesem, zachodzącym w okresie średniopolskim, było wycofywanie się form niezłożonych przymiotnika. Dzięki Korpusowi Języka Polskiego XVII i XVIII wieku i narzędziom ChronoFleksu można pokazać dynamikę tych zmian. Rysunek 2 to wizualizacja procesu wychodzenia z użycia form niezłożonych przymiotnika w okresie XVII–XVIII wieku, a rysunek 3 – wizualizacja tego samego procesu, ale w odniesieniu do leksemów (wykres pokazuje, od ilu leksemów przymiotnikowych tworzono te formy w danym przedziale czasu).



1601–1700

król *subst*

	l. poj.	l. mn.
M.	*m*: **król** (5038)	*m*: królowie (2) *m*: króle (68) *manim1*: **królowie** (403)
D.	*m*: **króla** (3060) *m*: królu (6)	*m*: króli (4) *m*: **królów** (584)
C.	*m*: **królowi** (1300) *m*: królu (12)	*m*: **królom** (180)
B.	*m*: **króla** (1147) *m*: król (87) *manim2*: króla (3)	*m*: królów (6) *m*: **króle** (75) *manim1*: króli (0) *manim1*: królów (24)
N.	*m*: **królem** (853)	*m*: królami (38) *m*: **królmi** (89)
Msc.	*m*: **królu** (225)	*m*: **królach** (24)
W.	*m*: król (0) *m*: **królu** (461)	*m*: królowie (1) *m*: króle (2) *manim1*: **królowie** (10)

1701–1800

król *subst*

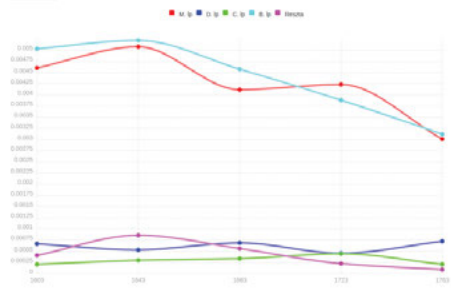
	l. poj.	l. mn.
M.	*m*: **król** (3318)	*m*: królowie (1) *m*: króle (10) *manim1*: **królowie** (316)
D.	*m*: **króla** (2464) *m*: królu (6)	*m*: króli (4) *m*: **królów** (766)
C.	*m*: **królowi** (787) *m*: królu (8)	*m*: **królom** (127)
B.	*m*: **króla** (807) *m*: król (43) *manim2*: króla (0)	*m*: królów (9) *m*: króle (11) *manim1*: króli (0) *manim1*: **królów** (83)
N.	*m*: **królem** (603)	*m*: **królami** (54) *m*: królmi (27)
Msc.	*m*: **królu** (123)	*m*: **królach** (55)
W.	*m*: król (1) *m*: **królu** (158)	*m*: **królowie** (1) *m*: króle (0) *manim1*: **królowie** (1)

1801–1900

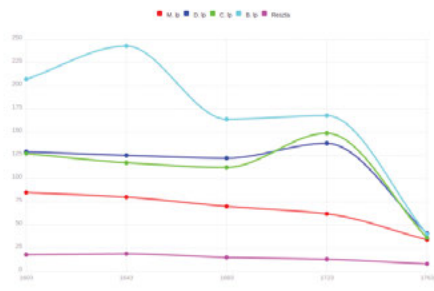
król *subst*

	l. poj.	l. mn.
M.	*m*: **król** (395)	*m*: królowie (0) *m*: króle (2) *manim1*: **królowie** (10)
D.	*m*: **króla** (213) *m*: królu (0)	*m*: króli (1) *m*: **królów** (28)
C.	*m*: **królowi** (32) *m*: królu (0)	*m*: **królom** (2)
B.	*m*: **króla** (45) *m*: król (0) *manim2*: króla (0)	*m*: królów (0) *m*: króle (2) *manim1*: króli (1) *manim1*: **królów** (5)
N.	*m*: **królem** (52)	*m*: **królami** (1) *m*: królmi (0)
Msc.	*m*: **królu** (11)	*m*: królach (0)
W.	*m*: król (0) *m*: **królu** (14)	*m*: królowie (0) *m*: króle (0) *manim1*: **królowie** (1)

Rysunek 1



Rysunek 2



Rysunek 3

WERYFIKACJA SĄDÓW NORMATYWNYCH ODWOŁUJĄCYCH SIĘ DO TRADYCJI

Jako przykłady zastosowania korpusów historycznych do weryfikacji sądów poprawnościowych przytoczymy dane dotyczące dwu konstrukcji językowych wzbudzających niekiedy kontrowersje normatywne.

W poradnikach językowych i słownikach normatywnych zaleca się, by w starannych wypowiedziach unikać łączenia rzeczowników rodzaju nijakiego zakończonych w mianowniku l. poj. na *-ę* z liczebnikami głównymi, ponieważ tradycyjnie łączyło się je z tak zwanymi rzeczownikami zbiorowymi. Jeśli tradycja ta sięgała XVII i XVIII wieku, to w korpusie tekstów z tego okresu nie powinno być przykładów takich połączeń albo, co najwyżej, powinny to być przykłady z tekstów mało starannych. Okazuje się jednak, że połączeń takich jest wiele, również w tekstach bardzo starannych, na przykład w *Biblii Gdańskiej*, w niektórych utworach Potockiego, w diariuszu Hieronima Radziwiłła czy przełożonych przez Otwinowskiego *Metamorfozach* Owidiusza, na przykład:

- Nie chciałeś z równą ciągnąć w małżeńskim chomącie, Bies po klaczy, **dwa źrebię** po starym drygancie. (PotFrasz1Kuk_II)
- Czyni tu Poetá wzmiánkę Zodyáku, koła źwierzęcego, ktore ná sobie nośi **dwoienaśćie źwierząt** z gwiazd wykształtowanych... (OvOtwWPrzem)
- Wielbłądźie odchowujących młode ze źrzbięty ich trzydźieśći/ krow czterdźieśći/ y wołow dźieśięć/ dwádźieśćiá oślic/ y **dźieśięć ośląt**. (BG_Rdz)
- *Eadem die* moi myśliwi też kontynuując polowanie, szczenną **siedmiu szczeniętami** uszczwali liszkę... (RadziwHDiar)
- Polowałem na łoszę z **dwoma cielętami**... (RadziwHDiar)
- **Trzech chłopiąt** ogień nie spalił Babyloński (ChmielAteny_IV)
- Czyli wystrászyć iaką dźiczyznę chcą z tey stáyni/ zwłaszcza lwá straszno ryczącego/ żeby **dwu bydląt** nie pożarł... (HinPlęsy)

W całym korpusie przykładów takich jest kilkadziesiąt, choć, przyznać trzeba, większość to połączenia z liczebnikami, które nie mają „odpowiedników" zbiorowych (*wiele, parę, trochę, sto*). Do myślenia daje jednak wynik wyszukiwania połączeń tego samego typu rzeczowników (na *-ę*) z liczebnikami zbiorowymi: w całym korpusie jest ich tylko 48. Czy tradycja była rzeczywiście mocno ugruntowana? A może to dopiero normatywne gramatyki XIX- i XX-wieczne wprowadziły zasadę, której przestrzeganie zaleca się do dziś?

Znacznie bardziej restrykcyjna jest zasada dotycząca tworzenia imiesłowów przysłówkowych od czasowników o odpowiednim aspekcie: uprzednich – od czasowników

dokonanych, a współczesnych – od niedokonanych. Reguła ta podawana jest nie tylko w wydawnictwach normatywnych, ale także w gramatykach opisowych i słownikach. Nie ulega jednak wątpliwości, że w historii języka polskiego powiązanie takie nie zawsze istniało[11]. W okresie średniopolskim stosowanie imiesłowów na *-(w)szy* od czasowników niedokonanych było na tyle częste (również w tekstach bardzo starannych), że nie można uznać tego za błąd. W Korpusie Barokowym występuje aż 1648 form takich imiesłowów, na przykład:

- Czego nie umiawszy inaczej wyrazić Historycy/ po staremu ją Scytia jeno że z przydatkiem Regia zowią (DembWyw-1633)
- Chrystus Pan/ jako był w żywocie Naświętszej Panny ogarniony/ bywszy Bogiem nieogarnionym (KorRoz-1645)
- Byłam tedy już w moim weselnym stroju, sama nie wiedziawszy. (GelPrzyp-1755)

Pojawiają się też, choć ich frekwencja jest znacznie mniejsza, imiesłowy na *-ąc* utworzone od czasowników dokonanych, na przykład:

- Przecię ja będę popływając sobie Kończył swój zamysł w zaczętym sposobie... (TwarKŁodz-1618)
- Jednak że na prawdę nie skąpo o Instygatory, i Medyka oblig jest ten, robić na dobro ludzkie, poniechając abo tlumiąc je kogoby mógł ratować, uczestnikiem się stawa, i męki i śmierci bliźniego, i winę z karami od Boga siebie ciągnie... (PetrJWod-1635)
- Dajac tę przyczynę/ że diabeł chytry/ wypędzisz go zjednego członka/ a on się w drugiem zatai/ i tak z egzorcyzmu nic. (WisCzar-1680)

Podobnie jak w wypadku konstrukcji z liczebnikami może pojawić się wątpliwość dotycząca tego, na ile reguła uznawana dziś za systemową wytworzyła się w języku „spontanicznie", a na ile jest wynikiem regulacji normatywnych z początku XIX wieku.

ZAKOŃCZENIE

Podane powyżej przykłady zastosowań korpusów historycznych w pracy historyka języka nie pozostawiają chyba wątpliwości, że korpusy historyczne mogą i powinny stanowić jedno z podstawowych narzędzi (a w zasadzie źródeł danych) do formułowania nowych hipotez badawczych i, co może jeszcze ważniejsze, weryfikowania tez zawartych w klasycznych monografiach dotyczących historii języka polskiego. Odnosi się to, rzecz jasna, również do leksykografii. Korpusy stały się już dość dawno podstawowym źródłem danych dla autorów słowników współczesnej polszczyzny[12], obecnie stają się takim źródłem także dla twórców słowników historycznych, zwłaszcza dla

---

[11] Do dziś formy takie występują w tekstach dość licznie, zwłaszcza w tekstach publikowanych w Internecie, a więc zazwyczaj nieopracowanych redakcyjnie. Więcej na ten temat: [Bojałkowska 2008] i [Gruszczyński 2008].

[12] Na podstawie danych korpusowych powstał najpierw *Inny słownik języka polskiego PWN* pod red. Mirosława Bańki (Warszawa 2000), a obecnie powstaje *Wielki słownik języka polskiego* pod red. Piotra Żmigrodzkiego (dostęp: https://wsjp.pl/).

Elektronicznego słownika języka polskiego XVII i XVIII w. Związek obu typów zasobów językowych (korpusów i słowników) jest niezwykle bliski i – co może być zaskakujące – dwustronny (por. [Żmigrodzki 2005] i [Bronikowska i in. 2016]).

Aby korpusy historyczne mogły spełniać swoją rolę w badaniach historycznojęzykowych, konieczne jest to, żeby powstał względnie jednorodny pod względem metodologicznym korpus diachroniczny obejmujący cały okres istnienia polszczyzny. Potrzeba taka była już wielokrotnie zgłaszana (por. m.in. [Pastuch i in. 2018]). Plany utworzenia takiego korpusu już istnieją (por. [Król i in. 2019]) i wiele wskazuje na to, że on powstanie.

LITERATURA

Adamiec D., 2015, Kryteria doboru tekstów do Elektronicznego Korpusu Tekstów Polskich z XVII I XVIII w. (do 1772 r.), „Prace Filologiczne", nr 67, s. 11–20.

Bień J. S., 2014, The IMPACT project Polish Ground-Truth texts as a DjVu corpus, „Cognitive Studies | Études Cognitives", Vol. 14, p. 75–84; https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.2014.008/174 (dostęp: 15.12.2019).

Bień J. S., 2012, Skanowane teksty jako korpusy, „Prace Filologiczne", nr 63, s. 25–36.

Bień, J., 2015, Problemy kodowania znaków w korpusach historycznych [w:] Semantyka a Konfrontacja Językowa, tom. 5, red. D. Roszko, J. Satoła-Staśkowiak, Warszawa, s. 67–79.

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

Bilińska J., Kwiecień M., Derwojedowa M., 2018, Microcorpus of Nineteenth-Century Polish [w:] Grammar and Corpora 2016, Heidelberg, s. 377–388.

Bojałkowska K., 2008, Użycie form typu „oglądawszy", „zjedząc" we współczesnym języku polskim, „Poradnik Językowy", z. 6, s. 31–44.

Bronikowska, R., 2015, Możliwości przeszukiwania korpusu barokowego – cele i założenia, „Prace Filologiczne", nr 67, s. 45–56.

Bronikowska R., Gruszczyński W., Ogrodniczuk M., Woliński M., 2016, The Use of Electronic Historical Dictionary Data in Corpus Design, „Studies in Polish Linguistics" 11 (2), s. 47–56.

Deptuchowa E., Jasińska K., Klapper M., Kołodziej D., 2020, O projekcie Korpusu Polszczyzny do 1500 roku, „Poradnik Językowy", z. 8, s. 7–16.

Derwojedowa M., 2020, Mikrokorpus Gronowy Polszczyzny 1830-1918, „Poradnik Językowy", z. 8, 52–65.

Górski R.L., Król M., Eder M., 2019, Zmiana w języku. Studia kwantytatywno-korpusowe, Kraków.

Gruszczyńska E., Leńko-Szymańska A. (red.), 2016, Polskojęzyczne korpusy równoległe, Warszawa.

Gruszczyński W., 2008, Czy we współczesnej polszczyźnie istnieje forma *widziawszy*? (rekonesans badawczy) [w:] Reverendissimae Halinae Satkiewicz cum magna aestimatione, red. G. Dąbkowski, Warszawa, s. 115–122.

Gruszczyński W., Adamiec D., Bronikowska R., Wieczorek A., 2020, Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe, „Poradnik Językowy", z. 8, s. 32–51.

Hebal-Jezierska M. (red.), 2014, Praktyczny przewodnik po korpusach języków słowiańskich, Warszawa.

Kieraś W., Komosińska D., Modrzejewski E. and Woliński M., 2017, Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish [w:] Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science, Vol. 10415, red. K. Ekštein, V. Matoušek, Springer, p. 308–316. https://doi.org/10.1007/978-3-319-64206-2_35 (dostęp: 15.10.2019).

Kieraś W., Woliński M., 2018, Manually annotated corpus of Polish texts published between 1830 and 1918 [w:] Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), red. N. Calzolari i in., Paryż, s. 3854–3859. http://www.lrec-conf.org/proceedings/lrec2018/pdf/675.pdf (dostęp: 15.10.2019).

Klapper M., Kołodziej D., 2014, Elektroniczny Korpus Tekstów Staropolskich do 1500 r. Perspektywy i problemy, „Prace Filologiczne", nr 65, s. 203–210.

Klapper M., Kołodziej D., 2015, Elektroniczny Tezaurus Rozproszonego Słownictwa Staropolskiego do 1500 roku. Perspektywy i problemy, „Polonica", nr 35, s. 87–101. https://polonica.ijp.pan.pl/index.php/polonica/article/view/76 (dostęp: 15.10.2019).

Król M., Derwojedowa M., Górski R.L., Gruszczyński W., Opaliński K., Potoniec P., Woliński M., Kieraś W., Eder M., 2019, Narodowy Korpus Diachroniczny Polszczyzny. Projekt, „Język Polski", nr 49, s. 92–101.

Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J., 1990, Słownik frekwencyjny polszczyzny współczesnej, Kraków.

Lewandowska-Tomaszczyk B. (red.), 2005, Podstawy językoznawstwa korpusowego, Łódź.

Ogrodniczuk M., 2018, Polish Parliamentary Corpus [w:] Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora, D. Fišer, M. Eskevich, F. de Jong, p. 15–19, Paryż.

Opaliński K., Potoniec P., 2020, Korpus Polszczyzny XVI wieku, „Poradnik Językowy", z. 8, s. 17–31.

Pałka P., Kwaśnicka-Janowicz A., 2017, Przewodnik po elektronicznych zasobach językowych dla polonistów (słowniki, kartoteki, korpusy, kompendia), Kraków. http://tmjp.pl/images/pdf/przewodnik_po_elektronicznych_zasobach.pdf

Pastuch M., Duda B., Lisczyk K., Mitrenga B., Przyklenk J., 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", nr 33/4, s. 857–873.

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.

Świdziński M., 2006, Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy, „LingVaria", nr 1, s. 23–34.

Włodarczyk M., Kopaczyk J., Kozak M., 2020, Multilingualism in Greater Poland court records (1386–1446): Tagging discourse boundaries and code-switching (Short corpus report), „Corpora", nr 15 (3). https://www.researchgate.net/publication/335259728_Multilingualism_in_Greater_Poland_court_records_1396-1446_Tagging_discourse_boundaries_and_-code-switching (dostęp: 20.03.2020).

Woliński M., Kieraś, 2020, Analiza fleksyjna tekstów historycznych i zmienność fleksji polskiej z perspektywy danych korpusowych, "Poradnik Językowy", z. 8, s. 66–80.

Żmigrodzki P., 2005, Słownik jako korpus tekstów – korpus tekstów jako słownik. Perspektywy polskiej leksykografii naukowej, „Poradnik Językowy", z. 6, s. 3–14.