PROJEKTY I PROPOZYCJE BADAWCZE

MAGDALENA KRÓL[I] | INSTYTUT JĘZYKA POLSKIEGO POLSKIEJ AKADEMII NAUK, KRAKÓW
MAGDALENA DERWOJEDOWA[II] | UNIWERSYTET WARSZAWSKI
RAFAŁ L. GÓRSKI[III] | INSTYTUT JĘZYKA POLSKIEGO POLSKIEJ AKADEMII NAUK, KRAKÓW; UNIWERSYTET JAGIELLOŃSKI, KRAKÓW
WŁODZIMIERZ GRUSZCZYŃSKI[IV]
INSTYTUT JĘZYKA POLSKIEGO POLSKIEJ AKADEMII NAUK, KRAKÓW
KRZYSZTOF OPALIŃSKI[V], **PATRYCJA POTONIEC**[VI]
INSTYTUT BADAŃ LITERACKICH POLSKIEJ AKADEMII NAUK, WARSZAWA
MARCIN WOLIŃSKI[VII], **WITOLD KIERAŚ**[VIII]
INSTYTUT PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK, WARSZAWA
MACIEJ EDER[IX] | INSTYTUT JĘZYKA POLSKIEGO POLSKIEJ AKADEMII NAUK, KRAKÓW; UNIWERSYTET PEDAGOGICZNY W KRAKOWIE

Narodowy Korpus Diachroniczny Polszczyzny. Projekt[1]

Słowa kluczowe: korpus, historia języka polskiego, diachronia, językoznawstwo historyczne, językoznawstwo korpusowe.

doi: http://dx.doi.org/10.31286/JP.99.1.8

Wprowadzenie[2]

Językoznawstwo historyczne, jeśli tylko dotyczy czasów, które zostawiły po sobie świadectwo pisane, od zawsze jest językoznawstwem „korpusowym" w tym znaczeniu, że źródłem

[I]magdalena.krol@ijp.pan.pl, ORCID: 0000-0003-0392-0921, [II]derwojed@uw.edu.pl, ORCID: 0000-0002-6515-2940, [III]rafal.gorski@ijp.pan.pl, ORCID: 0000-0003-4727-2639, [IV]wlodzimierz.gruszczynski@ijp.pan.pl, ORCID: 0000-0001-9406-1354, [V]krzysztof.opalinski@ibl.waw.pl, ORCID: 0000-0001-8775-4953, [VI]patrycja.potoniec@ibl.waw.pl, ORCID: 0000-0002-5911-5422, [VII]marcin.wolinski@ipipan.waw.pl, ORCID: 0000-0002-7498-1484, [VIII]witold.kieras@ipipan.waw.pl, ORCID: 0000-0002-8062-5881, [IX]maciej.eder@ijp.pan.pl, ORCID: 0000-0002-1429-5036.

1 Magdalena Król, Magdalena Derwojedowa oraz Rafał L. Górski odpowiadają za ostateczny kształt tekstu; Włodzimierz Gruszczyński, Witold Kieraś, Patrycja Potoniec, Krzysztof Opaliński, Marcin Woliński oraz Maciej Eder brali udział w opisywaniu poszczególnych korpusów i narzędzi; koncepcję opracowali M. Eder i M. Król. Wszyscy autorzy brali udział w pisaniu artykułu.

2 W skład NKDP wejdą korpusy i rozwiązania powstałe w ramach projektów: „Automatyczna analiza fleksyjna tekstów polskich z lat 1830–1918 z uwzględnieniem zmian w odmianie i pisowni" (f19; NCN DEC-2012/07/B/HS2/00570), „Elektroniczny korpus tekstów polskich XVII i XVIII w. – do 1772 roku" (KorBa; NPRH 0036/NPRH2/H11/81/2012), „Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja" (Chronofleks, NCN 2014/15/B/HS2/03119), „Korpus polszczyzny XVI wieku" (KXVI; NPRH 0138/FNiTP/H11/80/2011), jak również „Korpus tekstów staropolskich do roku 1500" (Korpus Staropolski) oraz teksty pozyskane z innych źródeł, takich jak: Biblioteka Narodowa, edycje krytyczne, Biblioteka Literatury Polskiej w Internecie i Wolne Lektury.

danych językowych nie jest kompetencja badacza, ale są nim teksty. Mogłoby się więc wydawać, że o ile pojawienie się elektronicznego korpusu w językoznawstwie synchronicznym spowodowało znaczącą zmianę w sposobie uprawiania nauki, o tyle w językoznawstwie historycznym sprowadzało się ono do wdrożenia narzędzia, które tylko usprawnia pracę badacza. A jednak prześledzenie niemal trzydziestu lat[3] obecności korpusów elektronicznych w językoznawstwie historycznym prowadzić może do wniosku, że mamy tu do czynienia z dogłębną zmianą jakościową. Językoznawstwo historyczne – często czerpiąc z doświadczeń synchronistów – powoli, ale nieodwracalnie zmienia swoje oblicze: praca z narzędziem, jakim jest korpus elektroniczny, pozwala gromadzić znacznie więcej danych językowych dużo mniejszym nakładem pracy. Ta obfitość danych pozwala mniej koncentrować się na tym, co jednostkowe, a zarazem wymusza użycie – choćby najprostszych – metod statystycznych.

Przykładem badania, które wykorzystuje potencjał korpusu diachronicznego, jest śledzenie zmiany w języku. Językoznawstwo kwantytatywne, opierające się na doświadczeniach nauk ścisłych, przez ostatnich sto lat wykształciło sporą liczbę metod statystycznych dających wgląd w mierzalne zjawiska języka naturalnego. Podstawowym narzędziem stosowanym w badaniu chronologii zmian językowych jest metoda szukania trendu: badane cechy językowe mierzy się przez przedstawienie frekwencji omawianego zjawiska jako funkcji czasu (Ellegård 1953; Bajerowa 1968; Borecki 1974). Metodę tę – co prawda w bardzo ograniczonym zakresie – wykorzystywano również w analizie rozwoju języka polskiego w XVII (Ostaszewska (red.) 2002), XVIII (Bajerowa 1964; Ostaszewska (red.) 2002) i XIX wieku (Bajerowa 1986–2000) oraz w analizie zmian leksykalnych w kolejnych numerach „Trybuny Ludu" z 1953 roku (Pawłowski 2006). Najbardziej spektakularnym przykładem użycia omawianej metody jest korpus kilkunastu milionów dokumentów angielskich wraz z towarzyszącym mu programem Google Books Ngram Viewer (Michel i in. 2011). Podobne korpusy, oparte na tekstach nieoznakowanych pod względem cech fleksyjnych, dają wgląd przede wszystkim w zmiany leksykalne badanego języka, ponieważ pozwalają na wyszukiwanie wskazanych wyrazów tekstowych[4]. Ewolucję cech gramatycznych w ich wypadku można zobrazować tylko w niewielkim stopniu, tworząc skomplikowane zapytania na podstawie wyrazów tekstowych. Dociekania o naturze zjawisk gramatycznych umożliwia korpus znakowany fleksyjnie. Takim przykładem jest NKJP, który – obejmując polszczyznę ledwie kilku ostatnich dekad – daje na swój sposób namiastkę spojrzenia diachronicznego na język. Naszym celem jest stworzenie korpusu, który łączy możliwość prowadzenia na nim badań leksykalnych i gramatycznych, pragmalingwistycznych i stylistycznych, diachronicznych i synchronicznych, jakościowych i ilościowych.

Korpusy diachroniczne – przegląd

Przeszłość wielu języków jest dokumentowana przez większe i mniejsze korpusy diachroniczne. Przykłady to wspomniany Helsinki Corpus of English Texts (Rissanen i in. (red.) 1991),

3 Za początek uznajemy stworzenie Helsinki Corpus of English Texts, który powstawał w latach 1984–1991.

4 Praca na tekście niezawierającym rozpoznanych lematów, a co za tym idzie – i form odmiany wyrazów, w wypadku języków o rozbudowanej fleksji sprowadza się do konstruowania zapytań złożonych z wszystkich form fleksyjnych danego leksemu. Jeśli nałożymy na to problem zróżnicowania wariantów ortograficznych, zapytanie staje się rozbudowanym ciągiem kryteriów, uwzględniającym wszystkie potencjalnie możliwe konfiguracje.

Corpus of Historical American English – COHA (Davies 2010), DIAKORP – czeski korpus diachroniczny[5], diachroniczna część Narodowego Korpusu Języka Rosyjskiego – NKJR[6], obejmująca teksty od połowy XVIII do końca XIX wieku, Korpus Języka Hiszpańskiego (Corpus del Español)[7], szwedzkie korpusy z lat 1520–1850[8], chiński korpus języka mandaryńskiego[9] czy włoski korpus zawierający teksty historyczne[10], a także wiele innych (por. Davies 2002, 2012; Hajnicz 2011).

Na gruncie polskim podejmowano szereg działań służących tworzeniu zbioru zdygitalizowanych danych. Pierwszy korpus dawnych tekstów polskich, spełniający standardy obowiązujące dziś przy tworzeniu takich zasobów, powstał na potrzeby dużego międzynarodowego projektu IMPACT (*Improving Access to Texts*[11]). Korpus ten został stworzony w latach 2009–2012 w Katedrze Lingwistyki Formalnej Uniwersytetu Warszawskiego i zawiera teksty pochodzące z XVI, XVII i XVIII wieku. Cechą charakterystyczną tego korpusu jest wynikająca z celów całego projektu niezwykła wierność transliteracji – rozróżniane są w nim wszystkie kształty grafemów występujące w tekstach podstawowych (por. Bień 2014)[12].

Korpusy, które gromadzą dane o wybranych okresach rozwoju polszczyzny, to: „Słownik pojęciowy języka staropolskiego”[13], „Elektroniczny korpus łaciny średniowiecznej na ziemiach polskich”[14], Elektroniczny Tezaurus Rozproszonego Słownictwa Staropolskiego, XV- i XVI-wieczne przekłady Pisma Świętego Nowego Testamentu[15], Narodowy Fotokorpus Języka Polskiego[16] czy Polish Diachronic Online Corpus – PolDi[17]. Zbiory te różnią się podejściem do opracowywanego materiału – część stara się o jak najwierniejsze cyfrowe odtworzenie pierwotnego tekstu, priorytetem innych jest jak największa użyteczność, nawet kosztem pewnego braku wierności względem oryginału, niektóre zaś są zbiorami materiałów bardzo się od siebie różniących, a przy tym opisanych według różnych konwencji. Ważną bazę tekstów, przygotowaną nie tylko z myślą o naukowcach, ale stanowiącą materiał, z którym mogą pracować również badacze, oferują ponadto zasoby Biblioteki Narodowej[18], Wolnych Lektur[19], Federacji Bibliotek Cyfrowych[20] oraz Biblioteki Literatury Polskiej

5 http://ucnk.korpus.cz/english/diakorp.php; DIAKORP zdaje się korpusem rzadko aktualizowanym, niewątpliwie stanowi jednak cenne źródło informacji o historii języka czeskiego.

6 http://www.ruscorpora.ru/en/.

7 http://www.corpusdelespanol.org/.

8 https://spraakbanken.gu.se/eng/home.

9 http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/.

10 https://corpora.dipintra.it/.

11 Por. http://www.impact-project.eu.

12 Korpus dostępny jest obecnie pod adresami: https://szukajwslownikach.uw.edu.pl/IMPACT_GT_1/ oraz https://szukajwslownikach.uw.edu.pl/IMPACT_GT_2/.

13 http://spjs.ijp.pan.pl/.

14 http://scriptores.pl/elexicon/.

15 https://ewangelie.uw.edu.pl/.

16 http://nfjp.pl/.

17 http://rhssl1.uni-regensburg.de/SlavKo/korpus/poldi.

18 http://polona.pl.

19 https://wolnelektury.pl.

20 https://fbc.pionier.net.pl/.

w Internecie[21]. Zbiory te mają różne formaty danych, co utrudnia (a nawet uniemożliwia) prowadzenie badań polegających na śledzeniu zjawisk językowych wykraczających poza jeden okres historyczny.

W Polsce wciąż nie ma spójnego zbioru tekstów, który mógłby posłużyć do badań nad polszczyzną w jej rozwoju. W ciągu wielu lat rozproszonych badań nad dziejami naszego piśmiennictwa zdygitalizowano szereg zasobów opracowanych z zastosowaniem tradycyjnych metod (dzięki przedsięwzięciom takim jak Repozytorium Cyfrowe Instytutów Naukowych, POLONA, Federacja Bibliotek Cyfrowych). Inicjatywy dygitalizacyjne umożliwiły oglądanie niezliczonych tekstów dawnych – zarówno starodruków, notatek z podróży, pocztówek, dzienników, jaki i gazet z XVIII wieku oraz wielu, wielu innych dokumentów – częściej jako obrazów niż jako tekstów. Spektakularnym przykładem współczesnych możliwości cyfryzacyjnych jest rękopis *Rozmyślania przemyskiego* z początku XVI wieku, który można obejrzeć bardzo dokładnie dzięki niespotykanej dotąd szczegółowości odwzorowania rzędu 80 milionów pikseli[22]. Coraz lepsza jakość algorytmów optycznego rozpoznawania tekstu (*Optical Character Recognition* – OCR) sprawia, że spora część ucyfrowionych zbiorów jest dostępna – choć niestety często z licznymi błędami – w postaci umożliwiającej przeszukiwanie tekstu. Dla lingwisty to rzecz cenna, ale w żadnym razie nie może ona zastąpić korpusu językowego, ten bowiem powinien mieć trzy istotne cechy, których brak wirtualnym bibliotekom: łatwość przeszukiwania, reprezentatywność i znakowanie lingwistyczne. Jeśli chodzi o pierwszy parametr, powiedzmy, że często zadowalamy się wiernym obrazem tekstu, który można czytać z ekranu. Tymczasem lingwista nie zamierza czytać tekstu *in extenso*, a jedynie wyszukać te informacje, które są dla niego istotne. Patrząc zatem z perspektywy językoznawcy, kluczowe jest, by tekst był czytelny dla komputera. Druga z tych cech to po prostu taki dobór tekstów, by korpus jak najlepiej odgrywał rolę wiarygodnej próbki języka (por. np. Adamiec 2015); inna sprawa, że wśród językoznawców brak zgody co do tego, jak ten cel osiągnąć – zresztą im dawniejsze czasy, tym mniejszym jest to problemem, gdyż stan zachowania piśmiennictwa każe z pieczołowitością traktować każdy tekst. Trzecia z cech to w wersji minimum opatrzenie każdego słowa informacją dotyczącą formy fleksyjnej, w jakiej ono wystąpiło, a dla języków o bogatej morfologii – także sprowadzenie słowoformy do postaci słownikowej (czyli tzw. lematyzacja). W wersji idealnej to informacja o strukturze składniowej zdania[23]. Trzeba jednak podkreślić, że coraz większa liczba dostępnych zdygitalizowanych tekstów znacząco ułatwia językoznawcom korpusowym pierwszy etap pracy, jakim jest pozyskanie tekstu. Kluczowe z naszej perspektywy jest jednak takie opracowanie materiałów źródłowych, żeby stanowiły one wiarygodne i użyteczne zasoby danych językowych. Z takim rozwiązaniem wiąże się oczywiście szereg decyzji dotyczących samej postaci tekstów[24] oraz metadanych, zaopatrzenia słów w specjalne znaczniki, które umożliwiłyby identyfikację wyrazów bez względu na ich zapis

21 http://literat.ug.edu.pl.

22 Tekst dostępny w serwisie polona.pl.

23 Niewiele jest historycznych banków drzew, tj. korpusów z pełnym rozbiorem zdań, takich jak Penn Parsed Corpora of Historical English.

24 Tutaj kluczowe jest pytanie o zakres normalizacji tekstu. Kwestia ta będzie rozstrzygnięta w kolejnym etapie realizacji projektu.

ortograficzny[25], czy dostosowania już istniejących narzędzi (chociażby przygotowywanych do opracowywania Narodowego Korpusu Języka Polskiego, narzędzi powstałych przy realizacji korpusów KorBa, f19 i Chronofleks) do pracy z dawnymi tekstami. Wreszcie trzeba sobie odpowiedzieć na pytanie, w jakim stopniu powinien on być kompatybilny z NKJP, który zasadniczo rejestruje polszczyznę współczesną. Warto wszak zauważyć, że NKJP w obecnej wersji dla diachronisty stanowi ostatni z serii dużych korpusów dokumentujących poszczególne epoki dziejów języka polskiego.

Cel

W pierwszym etapie prac tworzymy więc korpus oportunistyczny[26], łączący zbiory, które już istnieją. Kolejnym etapem będzie jednolita, jeśli chodzi o założenia koncepcyjne, anotacja fleksyjna[27], dostosowana do każdej epoki. Naturalnie anotacja ta musi być uzależniona od epoki, do której ją stosujemy – nie można chociażby przypisywać tekstom XVI-wiecznym tych samych kategorii rodzaju co tekstom XIX-wiecznym. Istotne jest jednak to, by za etykietkami kategorii gramatycznych, dopasowanymi do poszczególnych epok, stała jednolita koncepcja opisu fleksji[28].

Należy w tym miejscu bardzo mocno podkreślić, że pod pojęciem korpusu oportunistycznego nie rozumiemy mechanicznego złączenia wszelkich możliwych zasobów dostępnych w wersji cyfrowej. Przeciwnie, celem naszym jest pozyskanie tekstów najwyższej jakości, w starannym opracowaniu filologicznym, takich wreszcie, których opracowanie będzie odporne na niedostatki technologii OCR. Nie interesuje nas zatem masowe tworzenie obrazów cyfrowych z książek znajdujących się w zasobach bibliotek, odrzucamy podejście określane jako *big and dirty*. Jeśli mówimy o korpusie oportunistycznym, mamy na myśli teksty, które zostały zdygitalizowane, są dostępne w postaci cyfrowej jako tekst, są edytowalne i mają dobrą jakość.

Naszym celem jest również przygotowanie takiego opracowania danych, które umożliwi prowadzenie badań nad rozwojem języka poszczególnych epok oraz rozwojem polszczyzny w dowolnie wybranym zakresie czasowym, tak pod względem leksykalnym, jak i ściśle gramatycznym (fleksyjnym i składniowym), ale także pragma- i socjolingwistycznym. Wśród zasobów, które planujemy włączyć do opracowywanego korpusu, znajdą się przede wszystkim

25 Chodzi o znaczniki umożliwiające wyszukiwanie słów, które będziemy traktować jako warianty ortograficzne (lub fonetyczne) tej samej jednostki leksykalnej (przy pewnym rozumieniu takiej jednostki), np. słów lematyzowanych jako *pirwej* i *pierwej* czy *barzo* i *bardzo*, będących odpowiednio wykładnikami tekstowymi tej samej (w rozumieniu historycznym) jednostki leksykalnej. Zasady łączenia ze sobą lematów za pomocą takich znaczników wymagają szczegółowej dyskusji. O obecnych możliwościach przeszukiwania pisze m.in. Renata Bronikowska (2015).

26 Terminem tym w językoznawstwie korpusowym określa się korpus, który składa się z dowolnych dostępnych tekstów, bez zwracania uwagi na reprezentatywność i zrównoważenie.

27 W językoznawstwie korpusowym powszechnie mówi się o znakowaniu morfosyntaktycznym, my wolimy jednak termin „znakowanie fleksyjne", a anotacja fleksyjna będzie dotyczyła opracowania tagsetu dla polszczyzny XV i XVI w. Systemy znakowania w korpusach KorBa i f19 wywodzą się w znacznym stopniu z systemu stworzonego na potrzeby NKJP (Kieraś, Woliński 2018; Kieraś i in. 2017).

28 Warto wspomnieć, że w projekcie „Chronofleks" opracowano prototypowe wersje słowników dla analizatora fleksyjnego Morfeusz pozwalające analizować teksty barokowe i XIX-wieczne oraz odpowiadające im modele dla tagera Concraft-PL. Powstało także narzędzie Anotatornia 2 służące do weryfikacji i uzupełniania anotacji fleksyjnej tekstów historycznych.

różnego rodzaju zbiory tekstów powstałe w Instytucie Języka Polskiego Polskiej Akademii Nauk, najczęściej jako baza materiałowa tworzonych słowników: „Korpusu staropolskiego" (Górski, Twardzik 2003), obejmującego teksty ciągłe do 1500 roku, „Elektronicznego korpusu tekstów polskich XVII i XVIII w. (do 1772 roku)", przygotowywanego w Instytucie Badań Literackich „Korpusu polszczyzny XVI wieku". Kolejnym zasobem będzie korpus obejmujący lata 1830–1918, powstały na Uniwersytecie Warszawskim w ramach realizacji grantu „Automatyczna analiza fleksyjna tekstów polskich z lat 1830–1918 z uwzględnieniem zmian w odmianie i pisowni". Narodowy Korpus Diachroniczny Polszczyzny obejmie zatem lata 1380–1918 i będzie się charakteryzował nierówną dystrybucją tekstów w odpowiednich okresach historycznych, zwłaszcza w latach 1772–1830. Realizacja w IJP drugiego etapu korpusu KorBa oraz planowany korpus polszczyzny w czasach rozbiorów i w okresie międzywojennym (projekt złożony do recenzji) wypełnią lukę powstałą w pierwszym etapie prac nad NKDP. Usunięcie zróżnicowanego dostępu do tekstów umożliwi śledzenie przebiegu interesujących badaczy zmian językowych oraz pozwoli patrzeć na źródła polszczyzny ułożone w sekwencji czasowej. Korpusy włączane do NKDP tworzone były na podstawie różnych założeń, a ich konstruowanie zostało już zakończone (jak w wypadku f19, SStp oraz I etapu projektu KorBa) lub zakończy się dopiero za kilka lat (KXVI). Częściowo zbiory są lub będą dostępne, co – przy opublikowaniu korpusów f19 i KorBa – będzie stanowiło silny impuls dla badań diachronicznych.

Problemy dygitalizacji

Podstawą badań prowadzonych na korpusie są zdygitalizowane teksty. To, czego możemy się z nich dowiedzieć, jest uwarunkowane sposobem ich opracowania (zachowanie standardów i dobrych praktyk tworzenia korpusu) oraz tym, jakich narzędzi można użyć, pracując z danym zasobem. Jakość korpusu pozyskiwanego z tekstów zdygitalizowanych zależy od jakości rozpoznawania pisma przez programy OCR, które z kolei jest uzależnione od wielu czynników po stronie źródła drukowanego (jakość odbitki, zanieczyszczenia, obecność marginaliów, żywej paginy, elementów graficznych, składu wielokolumnowego itp.) oraz od jakości sprzętu i oprogramowania użytego do skanowania. Dlatego pozyskany w ten sposób tekst trzeba zestawić z zeskanowaną wersją oryginału i wprowadzić korekty. Jest to pracochłonne i kosztowne. Z tego powodu w ramach prac polegających na powiększaniu korpusu o nowe teksty planujemy wykorzystać już istniejące metody maszynowe[29] (od narzędzi do optycznego rozpoznawania tekstów, tj. ABBYY FineReader czy TRANSKRIBUS, do zapisu w ustalonej konwencji TEI, por. Burnard, Bauman (red.) 2007).

Obecnie w naszych zasobach znajdują się mniej lub bardziej współczesne edycje tekstów. Dwa duże zbiory tego typu to Wolne Lektury i Biblioteka Uniwersytetu Gdańskiego. Chcielibyśmy jednak dotrzeć do pierwotnych wersji publikacji, by jak najpełniej zdać sprawę z pierwodrukowej postaci tekstu, a zarazem z faktycznych historycznych źródeł, korzystając – tam gdzie to możliwe – z automatycznego tworzenia transkrypcji na podstawie pierwotnych

29 Wykorzystamy wyniki grantu „Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych", kierowanego przez Janusza S. Bienia (por. https://bitbucket.org/jsbien/ndt) w ramach polskiej części europejskiego projektu IMPACT, czyli *IMProving ACces to Text* (por. http://www.man.poznan.pl/online/pl/projekty/117/IMPACT.html).

wersji tekstów, tak jak robiono to podczas tworzenia korpusu KorBa. Zamierzamy również udostępnić wersje transliterowane i transkrybowane tekstów dawnych, aby zaspokoić oczekiwania naukowe i użytkowe badaczy historii rozwoju piśmiennictwa i tekstów dawnych względem prezentowanego korpusu.

Nieodzownym elementem procesu przeszukiwania korpusów jest wyszukiwarka korpusowa, której stworzenie nie jest rzeczą łatwą – zwłaszcza gdy ma ona służyć do przeglądania zróżnicowanych zasobów i przeszukiwania różnych korpusów. Wyszukiwarka będzie musiała wykorzystywać wspomniane wcześniej znaczniki, które będą umożliwiały identyfikację wyrazów bez względu na ich zapis ortograficzny, a które przypisane będą wszystkim słowom, tak by użytkownik, znając jedynie współczesną formę wyrazu, mógł znaleźć jej warianty historyczne, poświadczone w tekstach[30]. Funkcją wyszukiwarki odpowiadającej NKDP będzie również wskazywanie lokalizacji szukanego słowa w tekście. W kolejnych etapach prac przewidujemy ulepszanie wyszukiwarki korpusowej przez stwarzanie możliwości budowania bardziej złożonych zapytań lub wyszukiwania w wybranej grupie tekstów.

Narodowy Korpus Diachroniczny Polszczyzny

W zamierzeniu Narodowy Korpus Diachroniczny Polszczyzny będzie rozwijany, poniższe uwagi dotyczą zatem stanu obecnego (wersji oznaczonej przez nas jako NKDP z dnia 20.04.2018 r.). Korpus obejmuje około 2400 tekstów sięgających od końca XIV wieku do roku 1918, zawierających 24 miliony słów.

I tak: „Korpus tekstów staropolskich do roku 1500" zawiera 17 tekstów o łącznej objętości około 400 tysięcy słów, „Korpus polszczyzny XVI wieku" to docelowo 271 tekstów zawierających około 8 milionów słów, „Korpus tekstów polskich z XVII i XVIII wieku" to 718 tekstów, co daje około 12 milionów słów (w II etapie rozwoju projektu KorBa planuje się dołączenie kolejnych 12 milionów słów), f19 to tysiąc tekstów składających się na milion słów. Teksty pozyskane z bibliotek wirtualnych to około dwóch milionów słów. Na potrzeby prowadzonych w IJP badań nad zmianą w diachronii pozyskano dodatkowo nieco ponad milion słów, pracując na tekstach skanowanych i rozpoznawanych optycznie przez Bibliotekę Narodową. Planowane prace korpusowe w IJP PAN przyczynią się do powiększenia korpusu o kolejne teksty. Korpus diachroniczny jako całość nie jest aktualnie dostępny w wersji elektronicznej, z jego poszczególnymi komponentami można się zapoznać na stronach internetowych poszczególnych projektów.

Z biegiem czasu i wraz z rozwojem języka w ogóle przyrost tekstów zdaje się postępować w ciągu geometrycznym, co sprawia, że równomierna reprezentacja kolejnych okresów w korpusie diachronicznym jest co najwyżej mrzonką. Tak samo wygląda to i w naszym korpusie: liczba tekstów oraz, przede wszystkim, sumaryczna liczba słów przyrastają w sposób lawinowy z każdym kolejnym okresem historycznym.

Opracowując korpusy składowe, odpowiedzialne za nie zespoły dążyły do tego, by zgromadzone teksty – na ile to tylko możliwe – odzwierciedlały wielość gatunkową reprezentowaną w danej epoce. Czyniły tym samym pierwszy krok w stronę zrównoważenia korpusu.

30 W projektach dotyczących korpusów powstały już wyszukiwarki, które będziemy chcieli udoskonalać.

Ze względu na specyfikę rozwoju piśmiennictwa trudno mówić o takim zrównoważeniu korpusu, jakim poszczycić się może NKJP. Teksty SStp i „Korpus polszczyzny XVI wieku" siłą rzeczy opierają się na jedynych dostępnych źródłach (por. Twardzik i in. (red.) 2005), KorBa z kolei jest przykładem korpusu zmierzającego w stronę jak największego zróżnicowania tekstów (Adamiec 2015), natomiast f19 to korpus gronowy. Tworząc Narodowy Korpus Diachroniczny Polszczyzny, chcemy połączyć możliwości, jakie oferuje nam dostępność tekstów, oraz założenia, które należy przyjąć, by powstał korpus zrównoważony (opierając się na uwagach zawartych w pracach Sinclaira 1991; Sambor 1972 i in.).

Jeśli chodzi o rozkład tekstów na linii czasu, uwagę zwraca zwłaszcza niedoreprezentowanie czasów saskich (początek XVIII w.) oraz, może nawet w większym stopniu, przełomu wieków XVIII i XIX, a także nieobecność dużej liczby tekstów z okresu międzywojennego. Stąd też, jak już zostało powiedziane, naszym zamiarem jest jak najpełniejsze uzupełnienie korpusu pod tym względem. W ramach tych samych prac planowane jest również stworzenie korpusu, który w połączeniu ze zrównoważonymi podkorpusami f19 i korpusu KorBa stworzy zrównoważoną, na ile to tylko możliwe w wypadku korpusu historycznego, diachroniczną próbkę polszczyzny.

Nasz korpus odzwierciedla stan zachowanego piśmiennictwa z dawnych wieków, jest więc determinowany wyborami gatunkowymi i tematycznymi tekstów z poszczególnych epok. Nic zatem dziwnego, że teksty sprzed roku 1500 (a nawet 1550) są zdominowane przez tematykę religijną, z kolei w renesansowej części korpusu widać wyraźny udział poezji, wieki XVIII i XIX zawierają wiele źródeł prasowych, w dwudziestoleciu międzywojennym dużą rolę odgrywają powieści etc. Wszystko to sprawia, że badania diachroniczne będą do pewnego stopnia zaburzone przez statystycznie istotny sygnał gatunku. Pamiętamy również o tym, że kilka największych objętościowo utworów prozaicznych z XVI i XVII wieku (np. Stanisława Herakliusza Lubomirskiego *Rozmowy Artaksesa i Ewandra*) to zarazem teksty wybitne, odciskające swoje piętno na całej literaturze tego czasu, co oczywiście może prowadzić do zaburzonych wyników w badaniu dynamiki zmiany językowej. Rozmiar tekstów istotnych dla danej epoki może zatem odkształcać obraz korpusu, uda się jednak tę słabość przezwyciężyć, włączając do zbioru obszerne fragmenty ważnych z punktu widzenia historii tekstów (Gruszczyński i in. 2013).

Zastosowania

Jak wspomniano na początku, elektroniczny korpus diachroniczny to łatwy dostęp do znaczących ilości danych. Nawet w odniesieniu do tak dobrze poznanego języka, jakim jest język angielski, korpusy historyczne pozwoliły ustalić wiele wcześniej nieznanych faktów. Staraliśmy się więc pokazać, że sam korpus umieszczony w Internecie, dostępny dla wszystkich, stanowić może źródło danych dla dociekań naukowych, popularnonaukowych czy ogólnojęzykowych (warto przyjrzeć się chociażby pracom powstałym na zrębach korpusu diachronicznego, takim jak: Bronikowska i in. 2016; Derwojedowa i in. 2016; Eder i in. 2015; Górski, Król 2018; Majdak 2016). Patrząc na osiągnięcia w dziedzinie korpusów historycznych w innych krajach,

wierzymy, że nasz korpus okaże się narzędziem przydatnym w wielu obszarach językoznawstwa diachronicznego, synchronicznego czy historycznego.

W celu pokazania możliwości przyszłych badań przywołamy przykład, którym posłużyliśmy się w poprzedniej części niniejszego artykułu – pracy Ireny Bajerowej. W efekcie opracowania korpusu po kilkudziesięciu latach od ukazania się tej przełomowej rozprawy (Bajerowa 1964), zogniskowanej na czterech zaledwie punktach węzłowych okresu średniopolskiego, będzie można nie tylko zweryfikować postawione wtedy hipotezy na znacznie większym materiale językowym, ale i zadać wiele następnych pytań: jak kształtował się rozwój polskiej składni? Jaki rozwój znaczenia poszczególnych słów proponują dane modele ilościowe w ramach semantyki dystrybucyjnej? Jak duże jest zróżnicowanie stylu autorów w wybranych okresach historycznych? Jaką tradycję mają w polskiej prasie ogłoszenia drobne? Jak kształtuje się rozwój grupy nominalnej w języku polskim w kontekście innych języków słowiańskich? Badania filologiczne można będzie wzbogacić o komponent ilościowy, a w ramach badań kwantytatywnych zaplanować analizę dużo większych danych, niż jest to teraz możliwe.

Bibliografia

Adamiec D. 2015: *Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)"*, „Prace Filologiczne" LXVII, s. 11–20.

Bajerowa I. 1964: *Kształtowanie się systemu polskiego języka literackiego w XVIII wieku*, Zakład Narodowy im. Ossolińskich, Wrocław.

Bajerowa I. 1968: *Frekwencja form i badanie procesów historycznojęzykowych*, „Biuletyn Polskiego Towarzystwa Językoznawczego" XLI, s. 69–81.

Bajerowa I. 1986–2000: *Polski język ogólny XIX wieku. Stan i ewolucja*, t. 1–3, Wydawnictwo Uniwersytetu Śląskiego, Katowice.

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

Bronikowska R. 2015: *Możliwości przeszukiwania korpusu barokowego – cele i założenia*, „Prace Filologiczne" LXVII, 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", nr 11(2), s. 47–56.

Borecki M. 1974: *Kształtowanie się normy językowej w drukach polskich XVI wieku (na przykładzie oboczności typu* pirwszy || pierwszy*)*, Zakład Narodowy im. Ossolińskich, Wydawnictwo Polskiej Akademii Nauk, Wrocław–Warszawa–Kraków–Gdańsk.

Burnard L., Bauman S. (red.) 2007: *Guidelines for Electronic Text Encoding and Interchange (TEI P5)*, The TEI Consortium (online: http://www.tei-c.org, dostęp: 20 kwietnia 2018).

Davies M. 2002: *Corpus del Español: 100 million words, 1200s–1900s* (online: http://www.corpusdelespanol.org/hist-gen/, dostęp: 20 kwietnia 2018).

Davies M. 2010: *The Corpus of Historical American English (COHA): 400 million words, 1810–2009* (online: https://corpus.byu.edu/coha/, dostęp: 20 kwietnia 2018).

Davies M. 2012: *Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English*, „Corpora", nr 7(2), s. 121–157.

Derwojedowa M., Kieraś W., Bilińska J., Kwiecień M. 2016: *Dynamika zmian fleksyjnych i ortograficznych między reformami 1830–1918*, „Język Polski" XCVI, s. 24–35.

Eder M., Klapper M., Kołodziej D. 2015: *Dawna polszczyzna i nowe technologie: testowanie metod przetwarzania języka naturalnego na materiale polskiego piśmiennictwa od średniowiecza po wiek XX*, „Biuletyn Polskiego Towarzystwa Językoznawczego", z. 71, s. 189–202.

Ellegård A. 1953: *The auxiliary do: The establishment and regulation of its use in English*, Almquist & Wiksell, Stockholm.
Górski R.L., Twardzik W. 2003: *Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie*, [w:] S. Gajda (red.), *Językoznawstwo w Polsce. Stan i perspektywy*, Wydawnictwo Uniwersytetu Opolskiego, Opole, s. 155–157.
Górski R.L., Król M. 2018: *Polish Adverbial Perfect Participle. A corpus-based study*, [w:] W. Guz, B. Szymanek (red.), *Canonical and non-canonical structures in Polish*, Wydawnictwo Katolickiego Uniwersytetu Lubelskiego, Lublin.
Gruszczyński W., Adamiec D., Ogrodniczuk M. 2013: *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)*, „Polonica" XXXIII, s. 311–318.
Hajnicz E. 2011: *Najbardziej znane korpusy tekstów. Opracowanie przeglądowe*, Wydawnictwo Instytutu Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.
Kieraś W., Woliński M. 2018: *Manually annotated corpus of Polish texts published between 1830 and 1918*, [w:] N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida i in. (red.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Paris, s. 3854–3859.
Kieraś W., Komosińska D., Modrzejewski E., Woliński M. 2017: *Morphosyntactic annotation of historical texts. The making of the Baroque corpus of Polish*, [w:] *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017, Proceedings*, „Lecture Notes in Computer Science", nr 10415, s. 308–316.
Majdak M. 2016: *Słowa klucze w materiale historycznym – wyzwania i ograniczenia*, „Przegląd Humanistyczny", nr 60(3), s. 45–55.
Michel J.-B., Shen Y.K., Aiden A.P., Veres A., Gray M.K., The Google Books Team i in. 2011: *Quantitative analysis of culture using millions of digitized books*, „Science", nr 331(6014), s. 176–182 (online: https://doi.org/10.1126/science.1199644).
Ostaszewska D. (red.) 2002: *Polszczyzna XVII wieku*, Wydawnictwo Naukowe „Śląsk", Katowice.
Twardzik W., Deptuchowa E., Szelachowska-Winiarzowa L. (red.) 2005: *Opis źródeł* Słownika staropolskiego, Wydawnictwo Instytutu Języka Polskiego Polskiej Akademii Nauk, Kraków.
Pawłowski A. 2006: *Chronological analysis of textual data from the „Wrocław Corpus of Polish"*, „Poznań Studies in Contemporary Linguistics", t. 41, s. 9–29.
Rissanen M., Kytö M., Kahlas-Tarkka L., Kilpiö M., Nevanlinna S., Taavitsainen I., Nevalainen T., Raumolin-Brunberg T. (red.) 1991: *The Helsinki Corpus of English Texts*, University of Helsinki, Helsinki.
Sambor J. 1972: *Słowa i liczby. Zagadnienia językoznawstwa statystycznego*, Zakład Narodowy im. Ossolińskich, Wrocław.
Sinclair J. 1991: *Corpus, concordance, collocation*, Oxford University Press, Oxford.
SStp: *Słownik staropolski*, red. S. Urbańczyk, t. 1–9, Zakład Narodowy im. Ossolińskich, Wydawnictwo PAN, Wrocław–Warszawa–Kraków 1953–1987, t. 10–11, Instytut Języka Polskiego PAN, Kraków 1988–2002 (online: http://rcin.org.pl/dlibra/publication?id=39990&from=pubindex&dirids=105&tab=1&lp=236).

Summary

The Diachronic Corpus of Polish (DCP). A project

Keywords: corpus, history of the Polish language, diachrony, historical linguistics, corpus linguistics.

The paper presents the project of the Diachronic Corpus of Polish (DCP), which is intended as a cohesive collection of smaller corpora, arising in various scientific centers and covering various time ranges. This corpus will include texts covering the years 1380–1939 and will complement the National Corpus of Polish. The aim of the project is creating a balanced corpus presenting the history of the development of the Polish language and constituting a data basis for language researchers and a point of comparison for the historical corpora of European and world languages. The different parts of the corpus consist of 24 million words in total, while the target NPDC will constitute a repository of 40 million words.