Magdalena Majdak
(Polska Akademia Nauk)

SŁOWA KLUCZE W MATERIALE HISTORYCZNYM – WYZWANIA I OGRANICZENIA[1]

WPROWADZENIE

Powstające w ostatnim czasie narzędzia elektroniczne umożliwiają wyszukiwanie w ogromnych zbiorach tekstów słów najistotniejszych z punktu widzenia danej kultury, czasu, odbioru otaczającej rzeczywistości i sposobu jej przeżywania. Wyniki takich działań opierają się na analizie zróżnicowanego materiału współczesnego, mogącego zawierać dane pochodzące z różnych odmian, stylów i rejestrów języka, pisane i mówione rozmaitych gatunków. Jednym ze sposobów wykorzystania korpusu tekstów (ograniczonego np. do określonego podkorpusu tekstów publicystycznych) może być poszukiwanie słów kluczy. Na gruncie polskim zadania takie podejmowały zespoły Korpusu Języka Polskiego PWN, następnie Narodowego Korpusu Języka Polskiego, a obecnie Instytutu Języka Polskiego Uniwersytetu Warszawskiego. Należy tu wspomnieć o portalu *Słowa dnia*[2] i stronie internetowej projektu *Słowa klucze* realizowanego w Instytucie Języka Polskiego Uniwersytetu Warszawskiego[3]. Domeną takich przedsięwzięć jest rejestracja leksyki znajdującej najżywszy oddźwięk w życiu społecznym. Wyniki uzyskane dzięki analizie zawartości najświeższych tytułów prasowych i kanałów internetowych pozwalają śledzić dynamikę bieżących zainteresowań, a także obserwować pojawianie się nowych wyrazów i zmian znaczeniowych.

SŁOWA KLUCZE W MATERIALE HISTORYCZNYM

Cechą definicyjną słów kluczy w rozumieniu lingwistyki korpusowej jest ich znacząco wyższa częstotliwość w analizowanym tekście w stosunku do korpusu

[1] Tekst niniejszy powstał w ramach projektu badawczego „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)" realizowanego przez Pracownię Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego PAN we współpracy z Zespołem Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN. Grant finansowany jest ze środków Narodowego Programu Rozwoju Humanistyki na lata 2013–2018 (nr projektu: 0036/NPRH2/H11/81/2012).

[2] http://slowadnia.clarin-pl.eu/#/default/343, wcześniej nkjp.uni.lodz.pl/WordsOfDay.

[3] http://www.slowanaczasie.uw.edu.pl.

referencyjnego. Przykładowo: w badaniach zespołu Korpusu Języka Polskiego PWN prowadzonych we współpracy z redakcją „Rzeczpospolitej" w latach 2000–2005 „tygodniową frekwencję względną danego słowa porównywano z jego frekwencją względną w okresie 3 miesięcy i w ten sposób otrzymywano listę słów kluczowych danego tygodnia"[4]. Z punktu widzenia historyka języka interesująca byłaby możliwość zastosowania metody wydobywania słów kluczy w odniesieniu do tekstów dawnych. Badanie leksyki historycznie istotnej mogłoby pomóc w odkryciu ważnych wówczas tematów społecznych, politycznych, a nawet w próbie odtworzenia wizji świata – rzecz jasna przy wszystkich ograniczeniach, jakie wynikają z natury materiału i możliwości jego analizy. Wyzwania takie w odniesieniu do polszczyzny z powodzeniem podejmowali Władysław Kuraszkiewicz, Jerzy Woronczak, Jadwiga Sambor czy Edward Stachurski[5].

Poszukiwanie słów kluczy rozumianych jako centra tematyczne, wykładniki aktualnych tematów społecznych, pozostających w ścisłej relacji z tekstami tworzonymi na bieżąco, będącymi doraźnymi opisami i komentarzami rzeczywistości w odniesieniu do badań nad dawną polszczyzną, musi podlegać stosownym modyfikacjom. Dostępności tekstów, umiejętności ich tworzenia, powszechności udziału oraz dynamiki powstawania nie da się porównywać ze współczesnymi. Można jednak podejmować próbę zdania sprawy z tematów istotnych, która oparłaby się na konkretnym doborze tekstów, np. pochodzących z druków ulotnych lub prasy codziennej. Byłaby to siłą rzeczy próba bardziej statyczna niż dzisiejsze ze względu na mniejsze możliwości przeprowadzania systematycznych sond (np. z dokładnością nie co do dnia czy miesiąca, ale do dziesięcioleci). Jest to zatem raczej poszukiwanie nasilenia występowania pewnych słów w danym czasie niż bezwzględnej liczby wystąpień. Z uwagi na brak regularności powstawania tekstów, nakład i tempo pracy, kanał odbioru, recepcję, elitarność piśmiennictwa czy wreszcie wybór źródeł, jakie dotrwały do naszych czasów, liczba poświadczeń poszczególnych form wyrazowych nie przesądza o rzeczywistej częstotliwości użycia w tamtych czasach. Mimo że reprezentatywność nie może być mocną stroną w badaniach języka dawnego, a frekwencja nie jest kryterium przesądzającym o wadze słowa w kulturze (por. np. Wierzbicka[6]), wielomilionowe próby wyrazów i narzędzia do ich przeszukiwania pozwalają dziś na znacznie większe przybliżenia niż dotąd.

---

[4] Marek Łaziński, *Słowa klucze prasy polskiej. Słowa dnia i słowa roku UW*. Spośród wielu można tu jeszcze wspomnieć o projekcie zatytułowanym *A Needle in a Haystack* realizowanym przez Instytut Czeskiego Korpusu Narodowego Uniwersytetu Karola w Pradze oraz Wydział Języków Słowiańskich Uniwersytetu Browna w USA http://trost.korpus.cz/kwords.

[5] Wybrane pozycje znajdują się w bibliografii na końcu artykułu.

[6] „«Słowa kluczowe» danej kultury to słowa, które są dla niej w jakiś szczególny sposób ważne i które mogą o niej wiele powiedzieć. [...] Nie istnieje coś takiego jak skończony zbiór «słów kluczy» danej kultury, nie ma też żadnych obiektywnych procedur umożliwiających ich zidentyfikowanie. Żeby ustalić, czy dane słowo jest w danej kulturze szczególnie ważne, trzeba po prostu rozważyć różne «za» i «przeciw». [...] niektóre słowa [klucze] można analizować jako swoiste centra, wokół których organizują się poszczególne zjawiska kulturowe. Studiując takie centra, być może zdołamy dotrzeć do jakiejś ogólniejszej zasady, która nadaje strukturę i spójność określonej sferze kultury jako pewnej całości i która nierzadko zachowuje swą moc wyjaśniającą także i w innych sferach", Anna Wierzbicka, *Słowa klucze. Różne języki – różne kultury*, przeł. Izabela Duraj-Nowosielska, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2007, s. 42–43.

KORPUSY HISTORYCZNE

Podstawę materiałową do badań tego rodzaju dają będące *in statu nascendi* korpusy polszczyzny dawnej – nowość w polskiej lingwistyce korpusowej. Oto lista przedsięwzięć ułożona według chronologii opisywanego materiału:

- *Biblioteka zabytków polskiego piśmiennictwa średniowiecznego* przygotowana w Pracowni Języka Staropolskiego Instytutu Języka Polskiego PAN przez zespół pod kierunkiem Wacława Twardzika. Dostępna na płycie DVD (2006) i (z modyfikacjami) jako *Korpus tekstów staropolskich do roku 1500* na stronie https://www.ijp-pan.krakow.pl/publikacje-elektroniczne/korpus-tekstow-staropolskich. Jeszcze w 2012 roku pisano: „To jak dotąd jedyny korpus historyczny języka polskiego”[7]. Obecnie trwają prace nad przygotowaniem *Elektronicznego Korpusu Tekstów Staropolskich do 1500 r.* (szacowana objętość to milion segmentów) oraz *Elektronicznego Tezaurusa Rozproszonego Słownictwa Staropolskiego*. Będzie to „największa baza polskiego piśmiennictwa do 1500 r.”[8].
- *Korpus Tekstów i Korespondencji Jana Dantyszka* powstający w Pracowni Edytorstwa Źródeł Wydziału „Artes Liberales” Uniwersytetu Warszawskiego pod kierunkiem Anny Skolimowskiej, http://dantiscus.ibi.uw.edu.pl.
- *Korpus polszczyzny XVI wieku. Etap I: Digitalizacja źródeł oraz stworzenie narzędzi informatycznych i udostępnienie materiałów testowych korpusu* przygotowywany przez zespół toruńskiej Pracowni Słownika Polszczyzny XVI wieku Instytutu Badań Literackich PAN pod kierunkiem Patrycji Potoniec. Zakończenie pierwszego etapu prac obejmującego transliterację 272 źródeł planowane jest na 2017 rok, http://www.spxvi.edu.pl/korpus/.
- *Korpus IMPACT* obejmuje zestaw wybranych dokumentów historycznych z lat 1570–1756 udostępniony przez Zespół Bibliotek Cyfrowych Poznańskiego Centrum Superkomputerowo-Sieciowego wraz z wyszukiwarką opracowaną pod kierunkiem Janusza S. Bienia, http://poliqarp.wbl.klf.uw.edu.pl/pl/IMPACT_GT_2/.
- *Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.) „Korba”* powstaje w Pracowni Historii Języka Polskiego XVII–XVIII wieku Instytutu Języka Polskiego PAN przy udziale Zespołu Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN, projekt realizowany w latach 2013–2018 pod kierunkiem Włodzimierza Gruszczyńskiego, https://korba.ijp-pan.krakow.pl.
- Projekt *Automatyczna analiza fleksyjna tekstów polskich z lat 1830–1918 z uwzględnieniem zmian w odmianie i pisowni* zaplanowany na lata 2013–2016 powstaje w Instytucie Języka Polskiego Wydziału Polonistyki Uniwersytetu Warszawskiego pod kierunkiem Magdaleny Derwojedowej. Jednym z celów tego przedsięwzięcia jest stworzenie jednomilionowego korpusu języka polskiego lat 1830–1918 wraz z analizatorem morfologicznym, www.f19.uw.edu.pl.

---

[7] *Narodowy Korpus Języka Polskiego*, red. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Wydawnictwo Naukowe PWN, Warszawa, s. 6, http://www.nkjp.pl/settings/papers/NKJP_ksiazka.pdf [dostęp 15.03.2016].

[8] Magdalena Klapper, Dorota Kołodziej, *Elektroniczny Korpus Tekstów Staropolskich do 1500 r. Perspektywy i problemy*, „Prace Filologiczne” 2014, nr 65, s. 211.

Oprócz publikacji elektronicznych, w których tytule umieszczono słowo *korpus*, należy także wymienić następujące:

- *Piętnastowieczne przekłady Nowego Testamentu – elektroniczna konkordancja staropolska. Internetowa baza danych* przygotowana w Pracowni Języka Staropolskiego Instytutu Języka Polskiego PAN przez zespół pod kierunkiem Mariusza Leńczuka, http://www.stnt.ijp-pan.krakow.pl/teksty/index/1.
- *Przeglądarka wersetów równoległych szesnastowiecznych Ewangelii* powstała w Instytucie Języka Polskiego Wydziału Polonistyki Uniwersytetu Warszawskiego pod kierunkiem Izabeli Winiarskiej-Górskiej. Zawiera edycje krytyczne dziesięciu polskich translacji Ewangelii drukowanych w latach 1551–1599, http://www.ewangelie.uw.edu.pl/.

Jak widać z powyższego zestawienia, w ciągu najbliższych lat współpraca lingwistów z informatykami zaowocuje kilkoma korpusami obejmującymi materiał dawnej polszczyzny. Całość będzie wymagała jeszcze wielu działań, uzupełnień (np. o materiał z lat 1773–1829), główny zrąb jednak już powstał.

KORPUS „KORBA”

Podstawą prezentowanych tu analiz jest materiał ze wspomnianego korpusu tekstów średniopolskich „Korba”. Zawiera on utwory z przedziału czasowego 1601–1772, docelowo będzie liczył 12 milionów segmentów[9]. Autorzy dążą do zrównoważenia materiału, jednak mimo przemyślanego doboru tekstów trudno na tym etapie prac autorytatywnie stwierdzić, czy korpus będzie reprezentatywny dla słownictwa epoki. Zgodnie z przyjętymi założeniami za istotny uznaje się czas powstania dzieła (podział stuleci na części), region (dzięki temu odzwierciedla m.in. aktywność wydawniczą ośrodków – przewagę Małopolski) i charakter tekstu, który można opisać ze względu „typ mowy” (wierszowana, niewierszowana, mieszana), rodzaj (epika, liryka, dramat, utwory synkretyczne, wiadomości prasowe, literatura faktograficzna, listy), gatunek (np. fraszki, sielanki, opisy podróży, lamenty, przepisy kucharskie, romanse, tragedie), tematykę (alchemia, anatomia, architektura, filozofia, fizyka, gramatyka, górnictwo, hutnictwo, kulinaria, matematyka, medycyna, muzyka, obyczajowość, polityka, religia, retoryka, wojskowość, zielarstwo, żeglarstwo i inne)[10]. Możliwość sprecyzowania kryteriów doboru materiału pozwala na stworzenie zbioru tekstów określonego typu, zestawienie go z korpusem referencyjnym i wydobycie słów kluczy.

[9] Włodzimierz Gruszczyński, Dorota Adamiec, Maciej Ogrodniczuk, *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu badawczego*, „Polonica” 2014, nr 33, s. 311–318.

[10] Szerzej na ten temat w artykule Doroty Adamiec, *Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)”*, „Prace Filologiczne” 2015, nr 67, s. 11–20.

PRZYKŁAD

Dla przykładu wyłoniono korpus „instrukcji górniczych" (tekstów regulujących sposób pracy w kopalniach), a następnie za pomocą programu AntConc porównano go z zasobami „Korby" i wygenerowano słowa o podwyższonej w stosunku do nich frekwencji (*keywords*). Oto fragment utworu, z którego pochodzą: „Pp. warcabni wiarygodni mają być, aby wiernie robotnika popisywali; wieczór rozmówiwszy się z pp. stygarami wprzód kierat naznaczyć, trybarzów, potym walaczów, aby z niemi senior zjeżdżał [...] A potym zszedłszy na dół litanie zwyczajne wszyscy wespół przed zaczęciem roboty skarbowej śpiewać, po prześpiewaniu których ażeby ośmiogodzinnej szychty nie opóźnili, więc bez utraty czasu każdy zaraz do swojej udać się roboty powinni będą"[11].

Lista pierwszych dwustu pozycji (po usunięciu słów łacińskich i synsematycznych części mowy) wygląda następująco: *gr[oszy], miar, soli, dróżek, będą, będzie, sól, roboty, beczki, szychtę, szybu, łoju, beczek, powinni, robotnika, bałwanów, robocie, oprawy, dole, zł[otych], powinien, robotnik, robotę, fl[orenów], ma, żupach, kłapcia, miarki, skarbowej, szyb, tydzień, kopacz, mci, oficyjalistów, stygar, miarę, szybika, miary, wielickich, bałwany, par, mają, komory, pp[anowie], robotnikowi, robota, dróg, dróżka, kłapeć, rachować, bocheńskich, myta, górach, dół, kopaczowi, oficyjalistom, szerzyznę, płacy, dolni, należycie, powinna, den, płacić, oficyjalistowie, żupnej, każdą, solą, każdy, górę, żup, łój, żupy, komorze, szybik, wieliczce, beczkę, powinno, komorach, żupie, miarek, stygarowie, dwóch, robić, bałwana, kliny, cieśle, pieca, ognia, powinny, frochtarze, garce, piecowi, piecowych, bałwan, szychta, dolnych, każdego, należało, dołu, złp, sposobem, robią, ściany, kopackiej, końcy, kłapcie, solach, szycht, pisać, osobna, stygarów, szychty, wszerz, miejscach, komór, osady, skarbowi, bochni, frochtarz, krakowskich, podżupek, wyboja, wybój, każdym, karani, robót, rachując, wyżej, żupa, doglądać, kieratu, kopacza, kłapci, seraf, spód, starejsi, wielickiej, zjeżdżać, piszą, winą, statek, dawać, frochtarzów, imć, otworu, podobnymże, podżupka, szybem, szybikach, walaczów, walenia, warcabni, wynosków, odbierać, skarbowych, potrzeba, ławy, gór, instrukcja, górą, pisze, kwartał, kamieniu, dolnym, janina, mazowieckich, pakunku, szromy, walacze, wybojami, skarb, piec, statki, skarbu, zwierzchności, okoliczności, górami, mieć, kolekty, ściana, cetnarów, danielowiec, hutmani, kaszt, lampy, majster, otwór, robotnikiem, robotników, rumów, skarbowego, wyboje, wymierzyło, żupna, regestru, osobliwie, wzwyż, złotemu, buna, dolnego*.

Choć analiza treści słów kluczy nie jest przedmiotem tego artykułu, warto zauważyć, że na ich liście znalazły się m.in. rzeczowniki będące nazwami osób różnej specjalności i rangi, obiektów związanych z infrastrukturą kopalni, sprzętów, zasobów mineralnych, należności finansowych i powinności zawodowych. Tekst jest utrzymany w stylistyce instrukcji, stąd obecność charakterystycznych form czasownikowych, takich jak *powinni*, *mają*, *należy*, *musi* (*coś zrobić*)[12]. Ogląd słów kluczy potwierdza intuicję – odzwierciedlają one tematykę utworu i charakterystyczne elementy średniopolskiej rzeczywistości górniczej.

[11] *Instrukcje górnicze dla żup krakowskich z XVI–XVIII wieku*, oprac. Antonina Keckowa, Ossolineum, Wrocław 1963, s. 52 i 145.

[12] Przyjrzenie się samym tylko formom fleksyjnym może być cenne. Po przeprowadzeniu analizy morfologicznej i przyporządkowaniu ich właściwemu leksemowi część informacji, np. o braku form pierwszej osoby, takich jak *mam* czy *jestem powinna* (*coś zrobić*), znika.

CEL, MATERIAŁ I METODA

Skromnym, a zarazem głównym celem niniejszego artykułu jest przegląd form wyrazowych rzeczowników o najwyższej frekwencji, pochodzących z tekstów polskojęzycznych XVII i XVIII wieku. Praca ta, sytuująca się nieco z boku głównego nurtu refleksji nad słowami kluczami, nawiązuje tematycznie do badań słownictwa z wykorzystaniem nowych narzędzi.

Całość materiału będącego podstawą analiz liczy 9 572 149 segmentów (*word tokens*) i 797 876 typów słów (*word types*) pochodzących z 653 tekstów. Do badań wykorzystano program AntConc w rozszerzonej wersji 3.5.0 z 2015 roku[13]. Pozwala on zarówno na wygenerowanie listy słów (*word list*), jak i słów kluczy (*keywords*), a także na jej porządkowanie według wybranych kryteriów. Z uwagi na specyfikę materiału i założony cel zastosowano sortowanie według frekwencji wystąpienia formy wyrazu.

Kolejność postępowania badawczego była następująca:

1. Wygenerowano listę słów liczącą 798 013 pozycji.
2. Wybrano wszystkie rzeczowniki, których przynajmniej jedna z form odmiany przekroczyła tysiąc wystąpień. Okazało się, że lista zawiera sto pozycji.
3. Uzgodniono warianty morfologiczne, fonetyczne, pisowniowe, ustalono postać mianownika i przypisano formy deklinacyjne.
4. Odtworzono i zestawiono listę paradygmatów wymaganą przez program AntConc z uwzględnieniem możliwych postaci graficznych poszczególnych form. Oto przykładowy fragment:

czas → [czas] [czasu] [czasowi] [czasem] [czasie] [czaśie] [czasy] [czasow] [czasów] [czasom] [czasami] [czasámi] [czasach] [czasách] [czás] [czásu] [czásem] [czásie] [czáśie] [czásy] [czásow] [czásów] [czásom] [czásach] [czásami] [czásách]

panna → [panna] [panná] [panny] [pannie] [pannę] [panną] [panien] [pannom] [pannami] [pannámi] [pannach] [pannách] [pánna] [pánná] [pánny] [pánnie] [pánnę] [pánną] [pánien] [pánnom] [pánnami] [pánnámi] [pánnach] [pánnách]

książę → [książę] [księcia] [książęcia] [księćia] [książęćia] [księciá] [książęciá] [księćiá] [książęćiá] [księciu] [książęciu] [księćiu] [książęćiu] [księciem] [książęciem] [księćiem] [książęćiem] [książęta] [książętá] [książąt] [kśiążętom] [kśięciom] [kśięciom] [książętom] [książętámi] [książętami] [kśiążętámi] [kśiążętami] [książętach] [książętach] [kśiążętách] [kśiążętach] [xiążę] [xięcia] [xiążęcia] [xięćia] [xiążęćia] [xięciá] [xiążęciá] [księćiá] [xięćiá] [xiążęćiá] [xięciu] [xiążęciu] [xięćiu] [xiążęćiu] [xięciem] [xiążęciem] [xięćiem] [xiążęćiem] [xiążęta] [xiążętá] [xiążąt] [xięćiom] [xiążętom] [xiążętámi] [xiążętami] [xiążętach] [xiążętách]

5. Po zsumowaniu liczby wystąpień form deklinacyjnych dla poszczególnych stu rzeczowników otrzymano listę rzeczywistej frekwencji leksemów.
6. Aby zwiększyć stopień pewności, że w wykazie znajdą się rzeczowniki o frekwencji powyżej tysiąca, poszerzono go o kolejne, których suma frekwencji

[13] Serdecznie dziękuję Prof. Markowi Łazińskiemu za możliwość uczestnictwa w seminarium *Metody korpusowe w pracy filologa*, które odbywało się na Wydziale Polonistyki UW w roku akademickim 2014/15.

poszczególnych form odmiany przekracza tysiąc wystąpień. W ten sposób uzyskano 240 rzeczowników o najwyższej frekwencji w obrębie badanego materiału, z czego w przypadku pierwszych 140 (lista A) zliczono dokładnie sumę frekwencji wszystkich form, a w przypadku kolejnych stu przyjęto dane szacunkowe.

7. Ostateczna lista (B) zawiera alfabetyczny wykaz 240 rzeczowników o frekwencji wyższej niż tysiąc wystąpień. Forma podstawowa podawana jest w pisowni współczesnej.

UWAGI

Praca wymagała zmierzenia się z typowymi problemami, z jakimi borykają się osoby analizujące historyczny materiał językowy, takimi jak określenie wartości form gramatycznej i znaczeń, ujednoznacznienie nieprzewidzianych i nieprzewidywalnych wariantów morfologicznych, fonetycznych i graficznych (kreskowania liter oznaczających zarówno samogłoski, jak i spółgłoski, np. *wycinać*, oraz w zakresie pisowni łącznej lub rozdzielnej). Kwestie te mogą w znaczący sposób wpływać na wynik pomiarów, np. formy odmiany rzeczowników *koło* czy *siła* mogłyby być również formami innych części mowy, nieanalizowany w tym artykule czasownik *być* w 3. osobie liczby pojedynczej w pisowni *iest* liczy 38 845 wystąpień, *jest* 10 963, w połączeniu *toiest* (niezależnie od tego, czy jest to oryginał, czy błąd w przepisywaniu) 881 wystąpień, w postaci *tojest* – 13.

Aby określić formę podstawową i rozstrzygnąć o rzeczywistej obecności rzeczownika, przeprowadzano testy. Na przykład postać *miedzy*, licząca 4408 wystąpień, mogła być zarówno wykładnikiem odmiany rzeczownika *miedza*, jak i przyimkiem. Sprawdzono dla porównania frekwencję innych form odmiany i otrzymany wynik: *miedza* – 2, *miedzá* – 1, *miedzą* – 6 wystąpień przesądził o usunięciu tej formy z listy. Formy takie jak *myśli*, *woli*, których znaczna część kontekstów potwierdzała przynależność do paradygmatu rzeczownikowego, pozostawiono. W przypadku wysokiego stopnia synkretyczności form trudnych do określenia bez szczegółowej analizy przykładów sporadycznie zdecydowano się zamieścić paradygmat niepełny, pozbawiony form identycznych. Dotyczy to np. rzeczownika *rzeka*, którego forma biernika liczby pojedynczej *rzekę* może być homonimiczna z formą 1. osoby liczby pojedynczej czasownika, czy rzeczownika *droga*, którego dopełniacz liczby pojedynczej może być homonimiczny z formą przymiotnikową. W sytuacjach wątpliwych sumowano wystąpienia (suma niebudzących wątpliwości form odmiany musiała przekraczać tysiąc) i na tej podstawie podejmowano decyzję, czy rzeczownikiem podstawowym będzie *wino* czy *wina*, *dobro* czy *dobra*, *stop* czy *stopa*. Nie brano pod uwagę prawdopodobnych przymiotników substantywizowanych, takich jak *złoty*, *drogi*, *zły*, *święty*.

Dopiero szczegółowa analiza kontekstów pozwoli zdefiniować znaczenia i odpowiedzieć na pytanie, czy *miłość* to nazwa uczucia, czy element zwrotu grzecznościowego, jaka jest proporcja łacińskich form *nos*, *soli* w stosunku do polskich itd.[14]

[14] Dzięki odpowiednim znacznikom formy łacińskie w Korpusie „Korba" nie będą się ujawniały.

Na obecnym etapie prac celem było wyzyskanie blisko 10-milionowego materiału do wydobycia rzeczowników o frekwencji powyżej tysiąca.

WYNIKI

Lista A: rzeczowniki o frekwencji powyżej tysiąca. Gwiazdką oznaczono te, których żadna pojedyncza forma fleksyjna nie przekracza tysiąca wystąpień. Lista uporządkowana jest według frekwencji:

pan 27564, bóg 23675, król 19561, czas 18646, rok 18285, człowiek 16296, dzień 15301, miasto 15248, rzecz 13750, ziemia 11883, miejsce 11271, prawo 10971, syn 10040, serce 9930, świat 9773, woda 9630, ręka 9525, część 9451, raz 9345, wojsko 9090, kościół 8346, lud 8205, dom 8058, ojciec 7844, głowa 7769, strona 7681, słowo 7607, oko 7532, góra 7199, sposób 7050, śmierć 6855, droga 6788, niebo 6774, książę 6135, potrzeba 6131, sprawa 5980, morze 5748, krew 5617, dusza 5560, miłość 5553, pani* 5302, ciało 5276, ogień 5222, wiara 5214, rada 4982, koń 4837, noc 4801, biskup 4794, Polska 4790, Jan* 4757, wojna 4736, brat 4712, królestwo 4705, imię 4617, panna* 4523, pokój 4513, grzech 4434, wiek 4412, mila 4370, słońce 4321, pole 4309, poseł* 4304, drzewo 4291, cnota 4222, wola 4087, cesarz 4073, moc 4059, osoba 4055, siła 4041, koniec 4022, łaska 4022, wino* 4008, hetman 3984, żona 3950, przyczyna 3950, życie 3945, list 3889, mąż 3885, nauka 3814, koło 3797, ojczyzna* 3744, noga 3714, naród* 3704, żywot 3651, rozum 3634, Chrystus 3552, dwór 3551, rzeka 3493, złoto* 3475, korona 3456, linia 3411, duch* 3383, zdrowie 3365, dobro* 3337, powietrze 3203, krzyż* 2986, sława 2917, księga* 2872, szkoda* 2864, głos 2856, sąd 2815, myśl 2791, zamek 2774, wolność* 2665, marszałek* 2622, wojewoda 2622, choroba* 2595, Rzym 2578, Turek* 2535, drzwi 2484, twarz 2439, chwała* 2397, starosta* 2393, matka* 2380, obóz* 2375, robota* 2346, chleb* 2343, żal* 2244, sejm* 2211, bok 2166, Rzeczpospolita 2136, pomoc* 2072, obraz 2067, ostatek* 2042, zakon* 2014, wieś 1985, człek* 1926, dekret* 1924, chałupa* 1902, Piotr* 1875, rozdział* 1841, wierzch 1840, pieśń 1774, piec* 1723, prowincja* 1690, dziecko 1598, pieniądz 1541, kształt 1392, zło* 1382, stopa* 1333, Rzplita 1251.*

Lista B: 240 rzeczowników o najwyższej frekwencji w badanym materiale – kolejność alfabetyczna:

August, biskup, bok, bóg, brama, brat, brzeg, bydło, centrum, cesarz, chałupa, chęć, chleb, chłop, chorągiew, choroba, Chrystus, chwała, ciało, ciepło, cnota, cyrkuł, czas, część, człek, człowiek, dekret, dobra, dom, droga, drzewo, drzwi, duch, duchowny, dusza, dwór, działo, dziecko, dzień, figura, fortuna, głos, głowa, gniew, godzina, góra, grób, grunt, grzech, grzywna, gwałt, gwiazda, herb, hetman, honor, imię, izba, Jan, Jezus, język, kamień, kanclerz, koło, kometa, koniec, koń, korona, korzeń, kościół, kość, kraj, krew, król, królestwo, królewic, królowa, krzyż, ksiądz, książę, księga, kształt, kwadrat, liczba, linia, list, lud, łaska, łokieć, łut, łza, marszałek, matka, mąż, męka, miasto, miecz, miejsce, mila, miłość, moc, morze, Moskwa, mość, mowa, msza, mur, myśl, naród, natura, nauka, niebo, niewola, noc, noga, nos, obiad, obóz, obraz, obrona, ogień, ojciec, okno, oko, osoba, ostatek, pałac, pan, pani, panna, państwo, papież, para, piechota, pieniądz, piec, pieśń, Piotr, pismo, początek, pokój, Polak, pole, Polska, południe, pomoc, poseł, potrzeba, powietrze, pożytek, północ, praca, prawda, prawo, proch, prowincja, przyczyna, punkt, rada, rana, raz, ręka, robota, rok, rozdział, rozum, ryba, rząd, rzecz, Rzeczpospolita, rzeka, rzplta, Rzym, sąd, sejm, sejmik, sen, serce, siła, skutek, sława, słońce, słowo, sługa, sól, sposób, sprawa, stan, Stanisław, starosta, stół, stopa, strona, syn, szczęście, szkoda, sztuka, ściana, śmierć, świat, trzęsienie, Turek, twarz, tydzień, tytuł, Warszawa, wesele, wiadomość, wiara, wiatr, wiek, wierzch, wieś, wino, władza, własność, włosy, woda, wojewoda, województwo, wojna, wojsko, wola, wolność, wieś, wschód, wyspa, wzrok, zakon, zamek, zboże, zdrowie, ziele, ziemia, złość, złoto, znak, zwyczaj, żal, żona, życie, żywot.

ANALIZA I WNIOSKI

Do badań wybrano wszystkie formy odmiany rzeczownika o frekwencji wyższej niż tysiąc wystąpień. Mieszczą się one wśród pierwszych około 670 na 797 876 wszystkich typów słów. Wyodrębnienie kategorii znaczeniowych wymaga zdefiniowania znaczenia – na etapie pracy z inwentarzem słów nie jest to możliwe. Zarys ogólny mógłby jednak wyglądać następująco: części ciała (*głowa*, *ręka*, *noga*, *bok*, *stopa*, *twarz*, *nos*, *oko*, *włosy*, *serce*, *kość*, *krew*), nazwy miary czasu, wagi i odległości (*łut*, *stopa*, *łokieć*, *mila*, *dzień*, *tydzień*, *rok*), hierarchia państwowa, kościelna i społeczna – nazwy godności i urzędów (*król*, *królewicz*, *książę*, *kanclerz*, *marszałek*, *hetman*, *wojewoda*, *starosta*, *cesarz*, *papież*, *biskup*, *ksiądz*, *pan*, *pani*, *sługa*), nazwy stopnia pokrewieństwa (*matka*, *ojciec*, *brat*, *mąż*, *żona*, *dziecko*, *syn*), nazwy domów, pomieszczeń mieszkalnych i ich elementów konstrukcyjnych (*pałac*, *zamek*, *chałupa*, *dom*, *dwór*, *izba*, *mur*, *ściana*, *brama*), nazwy emocji (*złość*, *żal*, *gniew*), inne *abstracta* (*miłość*, *własność*, *wolność*, *niewola*, *prawda*, *wiara*, *honor*, *szczęście*, *myśl*, *chęć*, *moc*, *rozum*, *cnota*, *grzech*, *łaska*, *życie*), imiona (*Jan*, *Piotr*, *Stanisław*), nazwy miast i państw (*Rzym*, *Warszawa*, *Moskwa*, *Rzeczpospolita*), nazwy jednostek wojskowych (*chorągiew*, *piechota*), nazwy narodowości (*Polak*, *Turek*), słownictwo matematyczne (*bok*, *kwadrat*, *liczba*, *linia*), nazwy przyrodnicze (*ogień*, *ziemia*, *powietrze*, *woda*, *niebo*, *słońce*, *gwiazda*, *kometa*, *wiatr*, *morze*, *rzeka*, *wyspa*, *brzeg*, *świat*, *natura*, *ziele*, *zboże*), nazwy obszarów i jednostek administracyjnych (*województwo*, *państwo*, *kraj*, *miasto*, *wieś*, *prowincja*, *królestwo*), pory dnia i nocy, kierunki geograficzne (*dzień*, *noc*, *południe*, *północ*, *zachód*), nazwy zwierząt (*bydło*, *ryba*, *koń*).

Warto dostrzec także, że na szczycie listy – pierwszych trzech miejscach pod względem częstości wystąpień – znajdują się słowa *pan*, *Bóg* i *król*. W przyszłości można by porównać wyniki z podobnymi opracowaniami dotyczącymi epok sąsiadujących i stworzyć postulowaną wielokrotnie przez badaczy siatkę porównawczą leksyki historycznej i współczesnej.

ZAKOŃCZENIE

Możliwość pracy z materiałem pochodzącym z jednego z największych korpusów tekstów historycznych w Europie odkrywa nowe ścieżki badawcze. Daje nadzieje na poznanie słów najistotniejszych dla osób posługujących się polszczyzną XVII–XVIII wieku. Przedstawione tu wyniki należy traktować orientacyjnie, jako wyraz tendencji i zarys proporcji. Zgodnie z zastrzeżeniem, że kryterium frekwencji nie musi przesądzać o „kluczowości" słowa, warto dostrzec, że w odniesieniu do materiału historycznego pozwala ono domyślać się ukrytego za nim świata. Poszukiwanie słów najczęstszych, najważniejszych, sztandarowych, kluczowych wpisuje się w dociekania Anny Wierzbickiej (2007), Walerego Pisarka (2002) i współczesnych językoznawców korpusologów. Z jednej strony nawiązuje do klasycznych badań słownictwa (np. Kuraszkiewicz 1969), z drugiej otwiera się na takie operacje, jakie dla języka prasy PRL proponuje Adam Pawłowski w realizowanym

przez siebie projekcie *ChronoPress*[15]. Wydaje się, że warto także pomyśleć o przyjęciu wypadkowej metody statystycznej i kulturowej jako adekwatnej do badania wielkich zbiorów leksykalnych.

Wniknięcie w językowego ducha odległej epoki jest zadaniem trudnym. Wymaga mierzenia się z problemami typowymi dla twórców korpusów tekstów historycznych – różnorodną grafią, dawnymi formami odmiany, nieczytelnościami. Czynniki te i inne wpływają na ostateczny wynik badań, dlatego niniejszy artykuł należy traktować jako badawczy rekonesans.

BIBLIOGRAFIA

Adamiec Dorota, *Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)”*, „Prace Filologiczne” 2015, nr 67, s. 11–20.

Andrzejczuk Anna, Łaziński Marek, *Słowa dnia*, w: *Narodowy Korpus Języka Polskiego*, red. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Wydawnictwo Naukowe PWN, Warszawa, s. 275–279, http://www.nkjp.pl/settings/papers/NKJP_ksiazka.pdf [dostęp 15.03.2016].

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

Derwojedowa Magdalena, Kieraś Witold, Skowrońska Danuta, Wołosz Robert, *Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych*, „Polonica” 2014, nr 34, s. 21–27.

Gruszczyński Włodzimierz, Adamiec Dorota, Ogrodniczuk Maciej, *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu badawczego*, „Polonica” 2014, nr 33, s. 311–318.

Hajnicz Elżbieta, *Najbardziej znane korpusy tekstów*, Prace IPI PAN, Warszawa 2011, http://nlp.ipipan.waw.pl/Bib/hajn:11h.pdf [dostęp 15.03.2016].

Instrukcje górnicze dla żup krakowskich z XVI–XVIII wieku, oprac. Antonina Keckowa, Zakład Narodowy im. Ossolińskich, Wrocław 1963.

Klapper Magdalena, Kołodziej Dorota, *Elektroniczny Korpus Tekstów Staropolskich do 1500 r. Perspektywy i problemy*, „Prace Filologiczne” 2014, nr 65, s. 203–210.

Kuraszkiewicz Władysław, *Rzeczowniki w „Wizerunku” Mikołaja Reja*, „Pamiętnik Literacki” 1969, nr 60/4, s. 103–136.

Łaziński Marek, *Słowa klucze prasy polskiej. Słowa dnia i słowa roku UW*, http://www.slowanaczasie.uw.edu.pl/wp-content/uploads/klucze.pdf [dostęp 15.03.2016].

Narodowy Korpus Języka Polskiego, red. Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Wydawnictwo Naukowe PWN, Warszawa, http://www.nkjp.pl/settings/papers/NKJP_ksiazka.pdf [dostęp 15.03.2016].

Pawłowski Adam, Sambor Jadwiga, *Jerzy Woronczak – twórca polskiej lingwistyki kwantytywnej*, w: *Od starożytności do współczesności. Język – literatura – kultura. Księga poświęcona pamięci Profesora Jerzego Woronczaka*, red. Irena Kamińska-Szmaj, Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław 2004, s. 393–403.

Pisarek Walery, *Polskie słowa sztandarowe i ich publiczność*, Universitas, Kraków 2002.

Sambor Jadwiga, *Badania statystyczne nad słownictwem (na materiale „Pana Tadeusza”)*, Zakład Narodowy im. Ossolińskich, Wrocław 1969.

Słownik frekwencyjny polszczyzny współczesnej, Ida Kurcz, Andrzej Lewicki, Jadwiga Sambor [et al.], red. Z. Saloni, PAN, Kraków 1990.

[15] http://clarin.pelcra.pl/chronopress/.

Stachurski Edward, *Słowa-klucze polskiej epiki romantycznej*, Wydawnictwo Naukowe WSP, Kraków 1998.
Urbańczyk Stanisław, *Słownictwo staropolskie a wyższa kultura*, w: *Prace z dziejów języka polskiego*, Zakład Narodowy im. Ossolińskich, Wrocław 1979.
Wierzbicka Anna, *Słowa klucze. Różne języki – różne kultury*, przeł. Izabela Duraj-Nowosielska, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2007.

KEYWORDS IN HISTORICAL MATERIAL – CHALLENGES AND CONSTRAINTS

Summary

The article *Keywords in Historical Material – Challenges and Constraints* is an attempt to present the use of currently developed, multimillion lexical databases. The author reviews currently created corpora of Polish historical language and evaluates possibilities of extracting keywords in the old language material similar to undertakings relating to the contemporary language. Based on usage frequency, she selects 240 top ranking words in the material *Electronic Corpus of the 17th and 18th Century Polish Texts* (up to 1772) "Korba". It is an introduction to further analyses and comparisons.

Trans. Izabela Ślusarek