PORADNIK JĘZYKOWY

TKJ

INDEKS 369616
ISSN 0551-5343
NAKŁAD 500 egz.

8

(777)

TOWARZYSTWO KULTURY JĘZYKA
DOM WYDAWNICZY ELIPSA
WARSZAWA 2020

PORADNIK JĘZYKOWY

MIESIĘCZNIK ZAŁOŻONY W R. 1901
PRZEZ ROMANA ZAWILIŃSKIEGO
ORGAN TOWARZYSTWA KULTURY JĘZYKA

8

Zarząd Główny
ul. Krakowskie Przedmieście 26/28, 00-927 Warszawa
Wydział Polonistyki – Uniwersytet Warszawski
http://www.tkj.uw.edu.pl

TOWARZYSTWO KULTURY JĘZYKA
DOM WYDAWNICZY ELIPSA
WARSZAWA 2020

KOLEGIUM REDAKCYJNE

prof. dr hab. Stanisław Dubisz (redaktor naczelny), dr hab. Jolanta Chojak,
dr hab. Wanda Decyk-Zięba (zastępca redaktora naczelnego),
dr Ewelina Kwapień, prof. dr hab. Radosław Pawelec,
dr Marta Piasecka

RADA REDAKCYJNA

Prof. dr hab. Stanisław Dubisz (przewodniczący, Warszawa),
doc. dr Mirosław Dawlewicz (Wilno – Litwa),
prof. dr hab. Andrzej Markowski (Warszawa),
prof. dr hab. Alicja Nagórko (Berlin – Niemcy),
prof. dr Marta Pančiková (Bratysława – Słowacja),
prof. dr hab. Józef Porayski-Pomsta (Warszawa),
prof. dr hab. Danuta Rytel-Schwarz (Lipsk – Niemcy),
prof. dr Olga Šapkina (Moskwa – Rosja),
prof. dr hab. Hèléne Włodarczyk (Paryż – Francja)

Sekretarz Redakcji
dr Milena Wojtyńska-Nowotka

Redaktor naukowy zeszytu
dr Marta Piasecka

Redaktor językowy
Urszula Dubisz

Tłumacz
Monika Czarnecka

Korektor
Halina Maczunder

Adres redakcji
00-189 Warszawa, ul. Inflancka 15/198

http://www.poradnikjezykowy.uw.edu.pl; e-mail: poradnikjezykowy@uw.edu.pl
Dział Handlowy DW ELIPSA: tel. (48) 22 635 03 01, e-mail: sklep@elipsa.pl
Księgarnia internetowa: http://www.elipsa.pl

Czasopismo jest rejestrowane na listach ERIH i IC Master List,
indeksowane – w bazach Scopus, CEEOL, Pol-index, Scholar Google

Zeszyt opublikowany w wersji pierwotnej

© Copyright by Towarzystwo Kultury Języka and Dom Wydawniczy ELIPSA,
Warszawa 2020

PL ISSN 0551-5343

Wydanie zeszytu sfinansowane ze środków Ministra Nauki i Szkolnictwa Wyższego.
Decyzja nr 652/P-DUN/2019

Ark. wyd. 7,4. Ark. druk. 6,63. Papier offsetowy 80 g/m^2

W ZESZYCIE

– Elektroniczny Korpus Polszczyzny do 1500 r. tworzony jest w ramach projektu „Baza leksykalna średniowiecznej polszczyzny (do 1500 roku). Fleksja". Podstawą materiałową są dotychczasowe zbiory cyfrowe zabytków staropolskich, a celem opis fleksyjny wszystkich wyrazów odmiennych oraz anotowany morfosyntaktycznie zbiór tekstów z tego okresu.

– Pierwotnym celem powstania Korpusu Polszczyzny XVI w. było ocalenie bazy materiałowej *Słownika polszczyzny XVI wieku*. Zaczątek tego korpusu zawiera 135 tysięcy segmentów. Docelowo planowana jest integracja Korpusu Polszczyzny XVI w. z Narodowym Korpusem Diachronicznym Polszczyzny.

– Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. [KorBa] jest dużym anotowanym dostępnym online korpusem historycznym. Obejmuje 13,5 miliona segmentów, prezentujących teksty do 1772 r., które są zróżnicowane pod względem chronologicznym, geograficznym, stylistycznym i tematycznym. W pracy nad korpusem zastosowano narzędzia do przetwarzania języka naturalnego.

– Mikrokorpus Gronowy Polszczyzny 1830–1918 jest małym korpusem historycznym. Składa się z tysiąca próbek o przeciętnej długości 1027 segmentów z podziałem na podkorpusy: stylu popularnonaukowego, drobnych wiadomości prasowych, publicystyki, prozy artystycznej, dramatu artystycznego. Jest przydatny do badań zagadnień ortograficznych, fleksyjnych i składniowych.

– System komputerowy Chronofleks moduluje fleksję polską na podstawie materiału korpusowego, wizualizuje zmiany w czasie paradygmatów fleksyjnych poszczególnych leksemów, umożliwia także badanie zmienności frekwencji form fleksyjnych wyróżnionych na podstawie różnych kryteriów.

***

Historia języka polskiego – elektroniczny korpus tekstów – korpus diachroniczny polszczyzny – przetwarzanie języka naturalnego – fleksja – składnia – ortografia – stylometria – leksykografia.

Red.

SPIS TREŚCI

ARTYKUŁY I ROZPRAWY

Ewa Deptuchowa, *Katarzyna Jasińska*, *Magdalena Klapper*, *Dorota Kołodziej*: O projekcie Korpusu Polszczyzny do 1500 roku 7
Krzysztof Opaliński, *Patrycja Potoniec*: Korpus Polszczyzny XVI wieku 17
Włodzimierz Gruszczyński, *Dorota Adamiec*, *Renata Bronikowska*, *Aleksandra Wieczorek*: Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe 32
Magdalena Derwojedowa: Mikrokorpus Gronowy Polszczyzny 1830–1918 52
Marcin Woliński, *Witold Kieraś*: Analiza fleksyjna tekstów historycznych i zmienność fleksji polskiej z perspektywy danych korpusowych 66

OBJAŚNIENIA WYRAZÓW I ZWROTÓW

Agnieszka Grażul-Luft: O ewolucji znaczeniowej wyrazu *zdalny* w dobie pandemii koronawirusa 81

SPRAWOZDANIA, UWAGI, POLEMIKI

Hanna Pułaczewska, *Anna Zielińska*: Sprawozdanie z konferencji naukowej ICOMP – International Conference on Multilingual Practice, Uniwersytet Szczeciński, Szczecin, 24–26.02.2020 93

RECENZJE

Stanisław Dubisz: Jacek Warchala, *Formy perswazji*, Wydawnictwo Uniwersytetu Śląskiego, Katowice 2019 99

BIOGRAMY I WSPOMNIENIA

Hanna Popowska-Taborska: Profesor dr hab. Anna Kowalska (15.04.1936–25.12.2019) 103

CONTENTS

PAPERS AND DISSERTATIONS

Ewa Deptuchowa, *Katarzyna Jasińska*, *Magdalena Klapper*, *Dorota Kołodziej*: On the Corpus of Polish until 1500 project 7
Krzysztof Opaliński, *Patrycja Potoniec*: Corpus of the 16th-century Polish language 17
Włodzimierz Gruszczyński, *Dorota Adamiec*, *Renata Bronikowska*, *Aleksandra Wieczorek*: Electronic Corpus of 17th- and 18th-century Polish Texts – theoretical and workshop problems 32
Magdalena Derwojedowa: Cluster Micro-Corpus of the Polish Language of 1830–1918 52
Marcin Woliński, *Witold Kieraś*: Inflectional analysis of historical texts and variability of Polish inflection from the perspective of corpus data 66

EXPLANATIONS OF WORDS AND EXPRESSIONS

Agnieszka Grażul-Luft: On the semantic evolution of the word *zdalny* (*remote*) during the coronavirus pandemic 81

REPORTS, COMMENTS, POLEMICS

Hanna Pułaczewska, *Anna Zielińska*: Report on the academic conference ICOMP – International Conference on Multilingual Practice, University of Szczecin, Szczecin, 24–26 February 2020 93

REVIEWS

Stanisław Dubisz: Jacek Warchala, *Formy perswazji* (*Forms of persuasion*), Wydawnictwo Uniwersytetu Śląskiego, Katowice 2019 99

BIOGRAMS AND MEMOIRES

Hanna Popowska-Taborska: Professor Anna Kowalska (15 April 1936 – 25 December 2019) 103

Ewa Deptuchowa
(e-mail: ewa.deptuchowa@ijp.pan.pl)
ORCID: 0000-0002-3461-070x
Katarzyna Jasińska
(e-mail: katarzyna.jasinska@ijp.pan.pl)
ORCID: 0000-0002-9982-0644
Magdalena Klapper
(e-mail: magdalena.klapper@ijp.pan.pl)
ORCID: 0000-0003-3085-339X
Dorota Kołodziej
(e-mail: dorota.kolodziej@ijp.pan.pl)
ORCID: 0000-0002-5621-4679
(Instytut Języka Polskiego Polskiej Akademii Nauk, Kraków)

DOI: 10.33896/PorJ.2020.8.1

O PROJEKCIE KORPUSU POLSZCZYZNY DO 1500 ROKU

Celem niniejszego artykułu jest opisanie koncepcji Korpusu Polszczyzny do 1500 r. Najpierw przybliżymy założenia projektu, w ramach którego powstaje to opracowanie, następnie omówimy dotychczasowe zbiory cyfrowe tekstów staropolskich i przejdziemy do przedstawienia zasad doboru tekstów i ich opracowania w naszym Korpusie.

1. BAZA LEKSYKALNA ŚREDNIOWIECZNEJ POLSZCZYZNY (DO 1500 ROKU)

W Instytucie Języka Polskiego PAN w Krakowie realizowany jest projekt „Baza leksykalna średniowiecznej polszczyzny (do 1500 roku). Fleksja", finansowany ze środków NPRH, przyznanych na lata 2018–2023 na podstawie decyzji nr 0201/NPRH6/H11/85/2018. Kierownikiem projektu jest Ewa Deptuchowa. Wniosek został przygotowany przez członków zespołu Pracowni Języka Staropolskiego Instytutu Języka Polskiego PAN w Krakowie w następującym składzie: Ewa Deptuchowa, Katarzyna Jasińska, Magdalena Klapper, Dorota Kołodziej, Mariusz Frodyma i Mariusz Leńczuk.

Podstawowe cele naszych badań to opracowanie fleksji wszystkich wyrazów (odmiennych) z czasu do 1500 r.[1] oraz zbudowanie korpusu tekstów z tego samego okresu. Projekt zakłada przygotowanie interne-

[1] Projekt nie przekroczy zasięgu chronologicznego *Słownika staropolskiego*, aby nie kolidował z ramami czasowymi *Słownika* i *Korpusu polszczyzny XVI wieku*.

towej bazy leksykalnej składającej się z dwóch powiązanych aplikacji – Słownika i Korpusu.

Słownik będzie zawierał rozbudowaną informację gramatyczną, pogłębioną i poszerzoną w stosunku do *Słownika staropolskiego*,[2] zamieszczoną przy każdym haśle (w module fleksyjnym). W tym elemencie mikrostruktury hasła zostanie podana informacja na temat przynależności leksemu do danej części mowy oraz katalog jego form fleksyjnych. Będą w nim zawarte wyłącznie formy poświadczone w materiale źródłowym – nie zakładamy uzupełniania paradygmatów formami rekonstruowanymi. Katalog form uwzględni ich wariantywność fleksyjną i fonetyczną. Zapisy, których nie można jednoznacznie lub wcale zinterpretować (np. wyrazy z uszkodzoną końcówką), będą opatrzone komentarzami. Opis fleksyjny haseł Słownika dostarczy danych do interpretacji materiału leksykalnego w Korpusie.

Druga aplikacja, czyli Korpus, będzie zawierać obszerną kolekcję tekstów najstarszego okresu w dziejach polskiego języka piśmiennego. Zabytki w transliteracji i transkrypcji zostaną opatrzone metadanymi. Korpus umożliwi wyszukiwanie i porównywanie materiału leksykalnego w różnych tekstach oraz dostarczy dodatkowych poświadczeń wyrazów i ich form gramatycznych opisanych w Słowniku.

Badania nad najstarszymi zabytkami polszczyzny, prowadzone w ramach przedstawianego projektu, wpisują się w koncepcję opracowania Narodowego Korpusu Diachronicznego Polszczyzny.[3] Złożą się nań również kolekcje tekstów XVI, XVII i XVIII oraz XIX w. W połączeniu z Narodowym Korpusem Języka Polskiego[4] ułatwią studia nad piśmienną polszczyzną na przestrzeni całych jej dziejów, głównie w zakresie słownictwa i gramatyki. Zamieszczony w Internecie i dostępny dla wszystkich NKDP będzie narzędziem przydatnym w prowadzonych badaniach z zakresu językoznawstwa zarówno diachronicznego, jak i synchronicznego [por. Król i in. 2018].

2. DOTYCHCZASOWE ZBIORY CYFROWE TEKSTÓW STAROPOLSKICH

Znane są w zasadzie dwie elektroniczne edycje zbiorów tekstów średniowiecznych. Jedna to opublikowana w 2006 r. na płycie DVD *Biblioteka zabytków polskiego piśmiennictwa średniowiecznego*.[5] Zawiera ona nową edycję krytyczną 49 najważniejszych zabytków piśmiennic-

[2] Dalej skrót: Sstp.
[3] Dalej skrót: NKDP.
[4] Dalej skrót: NKJP.
[5] Została opracowana w ramach projektu grantowego KBN pod kierownictwem Wacława Twardzika w latach 2002–2006.

twa w języku polskim do 1500 roku. *Biblioteka* nie jest reedycją wcześniejszych opracowań. Znajdują się w niej zweryfikowane i poprawione transliteracje tekstów, ich nowe transkrypcje, komentarze do obu wersji oraz cyfrowe podobizny rękopisów. W te prace zaangażowany był przede wszystkim zespół Pracowni Języka Staropolskiego Instytutu Języka Polskiego PAN w Krakowie. Przykładowo Elżbieta Belcarzowa opracowała transliterację, transkrypcję i komentarze do *Kodeksu Suleda*, Ewa Deptuchowa i Zofia Wanicowa opracowały zweryfikowaną transliterację *Biblii królowej Zofii* oraz transkrypcję i komentarze do całości, Ludwika Szelachowska-Winiarzowa w taki sam sposób opracowała *Modlitwy Wacława*, Felicja Wysocka – *Ortyle magdeburskie*, Zofia Wójcikowa – *List tatarski* i *Ewangeliarz Zamoyskich*, a Wacław Twardzik część wierszy i pieśni. W opracowaniu *Biblioteki* brali też udział specjaliści z innych ośrodków naukowych, m.in. z Krakowa, Poznania i Warszawy. Kolekcja zawiera cenny materiał w postaci pełnotekstowych wersji średniowiecznych zabytków językowych wraz z aparatem krytycznym. Zamieszczone na płycie DVD narzędzia informatyczne zostały zaprojektowane do prezentacji transliteracji, transkrypcji i fotografii danego tekstu, ale nie zapewniały możliwości przeszukiwania ani porównywania zawartości całej kolekcji.

Druga edycja elektroniczna to opublikowany na stronie Instytutu Języka Polskiego PAN *Korpus tekstów staropolskich do roku 1500* w postaci plików .xml. Składa się nań ponad 130 tekstów ciągłych (w tym 114 tekstów pod wspólnym tytułem *Polskie zabytki wierszowane do końca XV wieku*, obejmujących legendy, wiersze i pieśni w różnych odpisach). *Korpus tekstów staropolskich* zawiera więcej zabytków językowych niż *Biblioteka*,[6] ale opublikowane są one wyłącznie w transkrypcji. Ten zbiór tekstów może być i jest wykorzystywany w badaniach korpusowych, ale ponieważ nie opracowano dotąd narzędzi wyspecjalizowanych w analizie morfosyntaktycznej polszczyzny średniowiecznej, posługiwanie się nim jest nieefektywne.

Obie kolekcje nie spełniają wymogów stawianych dzisiaj elektronicznym korpusom tekstów. Trzeba jednak podkreślić wkład pracy wykonawców i rzetelność opracowań od strony językoznawczej i edytorskiej oraz to, że stanowią one dużą pomoc w interpretacji zabytków staropolskich.

Wobec potrzeby zbudowania nowego korpusu tekstów średniowiecznej polszczyzny, łączącego wysoką jakość opracowania z łatwością użytkowania, rozpoczęto w Pracowni Języka Staropolskiego IJP PAN prace nad jego przygotowaniem [por. Klapper, Kołodziej 2014].

---

[6] M.in. *Modlitewnik Nawojki* (oprac. Mariusz Frodyma) czy *Rozmyślanie przemyskie* (oprac. Wacław Twardzik).

3. CHARAKTERYSTYKA KORPUSU POLSZCZYZNY DO 1500 R.

Korpus powstający w ramach projektu „Baza leksykalna średniowiecznej polszczyzny (do 1500 r.). Fleksja" będzie zawierał o wiele więcej danych niż dotychczasowe elektroniczne edycje. Nacisk zostanie położony na zwiększenie reprezentacji zabytków językowych sprzed 1500 r., wszechstronne opracowanie metadanych tekstów oraz odpowiednią anotację morfosyntaktyczną.

3.1. Zawartość Korpusu

Opracowywany Korpus obejmie średniowieczne dokumenty zapisane w języku polskim – teksty o charakterze ciągłym,[7] począwszy od najstarszych aż do tych datowanych na przełom XV i XVI w. Znajdą się w nim wszystkie teksty ciągłe z kanonu źródeł Sstp [por. Twardzik, Deptuchowa, Szelachowska-Winiarzowa 2005], sukcesywnie będzie uzupełniany o kolejne, nowo odkrywane zabytki.

Z uwagi na strukturę polskiego piśmiennictwa w średniowieczu i jego stan zachowania Korpus nie będzie zrównoważony. Wśród zachowanych zabytków sprzed 1500 r. przeważa tematyka religijna i prawna, a zdecydowana większość utworów to przekłady. Brakuje niektórych gatunków, w tym literatury naukowej, zwłaszcza dziejopisarstwa, słabo reprezentowane są m.in. epistolografia i proza artystyczna o tematyce świeckiej. W stosunku do dotychczasowych kolekcji tekstów staropolskich, w których dominowały utwory religijne, pewne zwiększenie liczby i różnorodności zabytków najstarszej polszczyzny jest jednak możliwe.

W Korpusie Polszczyzny do 1500 r. pojawi się duża grupa zabytków urzędowych i sądowych, w tym statuty cechowe, teksty skarg, przysiąg, umów, wojskowy rejestr popisowy oraz roty z różnych regionów Polski. Uwzględnione zostaną ponadto teksty kilku recept lekarskich. Zwiększy się też reprezentacja prozy religijno-dydaktycznej i epistolarnej, jak również poezji i drobnych zabytków języka polskiego.

Z konieczności włączamy do Korpusu zabytki niekompletne i zdefektowane (np. *Rozmyślanie przemyskie*), zachowane w wielu przekazach (np. modlitwy codzienne), zawierające wtręty obcojęzyczne (granicznym przykładem są *Kazania świętokrzyskie*). Ponadto uwzględniamy teksty bardzo drobne (np. zagadki, wierszyki mnemotechniczne, westchnienia modlitewne itp. oraz pojedyncze zdania w języku polskim cytowane w łacińskich kronikach). Nie będziemy jednak umieszczać w Korpusie łacińskich tekstów glosowanych. Przewidujemy prezentację tego typu źródeł w odrębnych kolekcjach.

---

[7] Objaśnienie tego określenia w dalszej części artykułu.

Ponieważ oprócz obszernych zabytków łatwych do zidentyfikowania jako teksty ciągłe w Korpusie uwzględniamy drobne zabytki języka polskiego towarzyszące tekstom łacińskim, konieczne jest wyznaczenie kryterium umożliwiającego odróżnienie takich krótkich tekstów od glos. Dlatego została sformułowana robocza definicja tekstu ciągłego.

Do tekstów ciągłych zaliczamy:

1. Przynajmniej dwuelementowe ciągi polskich wyrazów pospolitych, z których przynajmniej jeden jest czasownikiem w formie osobowej i tworzy sensowne zdanie (np. *Gorze się nam stało*).
2. Sekwencje wyrazów polskich lub polskich i łacińskich, które mają walory artystyczne (np. *Dworak szkoda*, *Dum bibo piwo*).

Zapisy zawierające wyłącznie polskie nazwy własne oraz większość glos będących polskimi odpowiednikami pojedynczych wyrazów łacińskich nie mieszczą się w tej definicji. Najkrótsze zabytki w naszym korpusie będą zatem jednozdaniowe.

3.2. Forma prezentacji zabytków w Korpusie

Przewidujemy zamieszczenie poszczególnych zabytków językowych w transliteracji i transkrypcji. Teksty zostaną przygotowane na podstawie istniejących wydań i będą sukcesywnie weryfikowane w miarę dostępności fotografii oryginału. Dotychczasowe opracowania, w tym te, z których korzystano w Sstp, różnią się pod względem kompletności, sposobu odwzorowania grafii zabytku i przyjętych zasad transkrypcji. Mimo to postanowiliśmy posłużyć się nimi, aby zebrać jak największy zasób słownictwa staropolskiego. Edycje te traktujemy jako podstawę udoskonalania odczytań materiału źródłowego.

Korpus uzupełnią teksty dotąd niepublikowane. Są to w znacznej mierze drobne zabytki identyfikowane podczas ekscerpcji źródeł łacińskich na potrzeby opracowania suplementu Sstp. Zgromadzenie obszernego materiału leksykalnego w Korpusie będzie pomocne przy obserwacji zjawisk językowych na maksymalnej liczbie poświadczeń.

3.3. Metadane w Korpusie

Teksty w Korpusie Polszczyzny do 1500 r. zostaną opatrzone metadanymi wskazującymi m.in. na ich czas powstania, miejsce pochodzenia, charakterystykę gatunkową i tematyczną. Sposób opisu zastosowany w metryczkach źródeł będzie w znacznej mierze zbieżny z tym stosowanym w innych korpusach historycznej polszczyzny. Pojawią się jednak pewne odstępstwa. Wiąże się to ze specyfiką średniowiecznych tekstów źródłowych.

Najstarsze zabytki polszczyzny, będące w zdecydowanej większości rękopisami, mają często niejednorodną, złożoną budowę i bogatą trady-

cję, co wymaga szczegółowego opisu; np. rękopis Biblioteki Czartoryskich w Krakowie nr 1418 zawiera statuty królewskie w polskim tłumaczeniu Świętosława z Wojcieszyna oraz statuty książąt mazowieckich w przekładzie Macieja z Rożana. Oba teksty przepisał Mikołaj Suled [Twardzik, Deptuchowa, Szelachowska-Winiarzowa 2005, 188]. Dlatego w korpusowej metryczce *Kodeksu Suleda* konieczne jest zaznaczenie, iż jest to przekład, rozróżnienie anonimowego autora (czy raczej autorów) łacińskiego pierwowzoru, tłumaczy i pisarza zabytku, a także wskazanie dat powstania poszczególnych statutów oraz kopii pióra M. Suleda (będącej właściwym źródłem korpusu).

Specjalnie opisane będą takie utwory jak: modlitwy codzienne, wiersze katechetyczne, niektóre pieśni zachowane w więcej niż jednym przekazie. Wskazanie jednego reprezentatywnego poświadczenia tekstu, np. *Bogurodzicy*, uznaliśmy za niecelowe i zdecydowaliśmy się na uwzględnienie wszystkich znanych przekazów zróżnicowanych pod względem fonetycznym, fleksyjnym i leksykalnym. W metryczce tego typu zabytków zostanie umieszczona informacja o istnieniu paralelnych wersji danego utworu. Pozwoli to na weryfikację wyników wyszukiwania w Korpusie, które mogłyby być zaburzone obecnością powielonych poświadczeń z różnych kopii tego samego tekstu.

3.4. Anotacja morfosyntaktyczna tekstów w Korpusie

W naszym zamierzeniu wszystkie teksty składające się na Korpus Polszczyzny do 1500 r. zostaną poddane automatycznej anotacji morfosyntaktycznej[8] weryfikowanej ręcznie. Segmentom zostaną przyporządkowane znaczniki określające ich formę podstawową, klasę gramatyczną, a klasom odpowiednie wartości. Opracowany przez zespół grantowy tagset, tj. zestaw klas i kategorii gramatycznych oraz ich wartości, oparty jest w znacznej mierze na tagsecie używanym przez Elektroniczny Korpus Tekstów Polskich z XVII i XVIII wieku (do 1772 r.),[9] zawiera jednak pewne zmiany. Wprowadzone doń modyfikacje wynikają z potrzeby opisania zjawisk typowych dla polszczyzny średniowiecznej, lecz nieobecnych bądź gasnących w późniejszych dobach języka polskiego.

Przykładowo wyróżniliśmy nową kategorię gramatyczną – typ odmiany z wartościami: prosta, złożona.[10] Stanowi ona kategorię fleksyjną

---

[8] Możliwości automatycznej anotacji tekstów średniowiecznej polszczyzny badali m.in. M. Eder, M. Klapper i D. Kołodziej [2015].

[9] Dalej skrót: KorBa; szczegółowe informacje na temat tagsetu można odnaleźć w opracowanej przez W. Gruszczyńskiego i R. Bronikowską instrukcji dostępnej na stronie: https://korba.edu.pl/manual [dostęp: 18.02.2020 r.].

[10] W wypadku form identycznych dla obu typów odmiany umownie przyjmujemy wartość: odmiana złożona. Znacznik odmiana prosta rezerwujemy tylko dla form niewątpliwych.

następujących fleksemów:[11] przymiotnik, liczebnik przymiotnikowy, imiesłów przymiotnikowy czynny, imiesłów czasu przeszłego czynny II,[12] imiesłów przymiotnikowy bierny. Tym samym znane z KorBy fleksemy z wariantem „odmiana niezłożona" zostały w naszym Korpusie połączone z odpowiednimi fleksemami podstawowymi.

Decyzja o utworzeniu kategorii gramatycznej „typ odmiany" była podyktowana wyraźnie zaznaczoną obecnością zarówno form odmiany złożonej, jak i prostej w polszczyźnie średniowiecznej. Zróżnicowanie to przejawia się w formach gramatycznych kilku części mowy (przede wszystkim przymiotnika, ale również liczebnika i niektórych imiesłowowych form czasownika). Przykładowe użycia tekstowe form wyrazowych występujących w odmianie prostej to:

a) przymiotnik: *Dobitczø, gesz to moze obyatowano bycz panv, ... <u>svyøto</u> bødze* (*sanctum erit*) *BZ* Lev 27,9 [Sstp IX, 75], forma mianownika liczby pojedynczej rodzaju nijakiego przymiotnika *święty*;
b) imiesłów czasu przeszłego czynny II: *Jaco kedy ma Jaschek obeslal panyczem, hyzbych mi* (pro: *mu*) *vroczil... czloweka..., w ty czasy v mpnye nye był <u>oszadl</u>* ('osiadły') *any moy byl* 1436 *Pozn* nr 1479 [Sstp V, 641], forma mianownika liczby pojedynczej rodzaju męskiego imiesłowu *osiadły* (czasownik *osiąść*);
c) imiesłów przymiotnikowy bierny: *Ktory bi vkradl czlowyeka a przedal, <u>doszwyatczon</u> wyny (convictus noxae) szmyerczyøø vmrzecz ma BZ* Ex 21,16 [Sstp II, 162], forma mianownika liczby pojedynczej rodzaju męskiego imiesłowu *doświadczony* (czasownik *doświadczyć*).

Dodanie tej kategorii umożliwiło precyzyjne opisanie elementów wyrażeń przyimkowych typu:

a) *z nagła: A z <u>nagla</u> stala szyą yest z angyolem vyelykoszcz ryczerstva nyebyeskyego EwZam* 292 [Sstp V, 43], forma dopełniacza liczby pojedynczej przymiotnika *nagły*,
b) *po gotowu*: *Joachymye y Anno... bosczie wi szamy zasluzili miecz czorką Matką Bozą..., a po <u>gotowu</u> iesztesczie... Jesu Crista ieden dziadem, druga babø MW* 45a [Sstp II, 475], forma celownika liczby pojedynczej przymiotnika *gotowy*.

Zamiast wyodrębnienia fleksemów: przymiotnik odmiana niezłożona (jak w KorBie) czy przymiotnik poprzyimkowy (jak w NKJP) człony przymiotnikowe wyrażeń przyimkowych będą opisywane jako przymiotniki w odpowiednim przypadku z wartością „prosta" przypisaną kategorii „typ odmiany".

Specyfika fleksji języka polskiego doby średniowiecza wymusiła również konieczność wprowadzenia nowych fleksemów, głównie czasowniko-

---

[11] Aby zachować ciągłość terminologiczną i zgodność z zasadami innych korpusów języka polskiego (NKJP, KorBa), w naszym projekcie również opieramy się na pojęciu fleksemu, por. Bień 1991.

[12] W KorBie i NKJP pseudoimiesłów.

wych. Są to m.in. forma aorystu i imperfektu. Określają one archaiczne z dzisiejszego punktu widzenia formy czasu przeszłego, których ślady można jeszcze odnaleźć w staropolszczyźnie. Przykładowe użycia fleksemów w tekstach to:

a) aoryst: *Gdy molwych gym Fl* 119, 6 [Sstp IV, 353], forma pierwszej osoby liczby pojedynczej czasownika *mołwić* (tj. *mówić*),
b) imperfectum: *Wzywali sø gospodna a on wisluchawa ie a we slupe obloka molwasze k nim Fl* 98,7 [Sstp IV, 353], forma trzeciej osoby liczby pojedynczej czasownika *mołwić* (tj. *mówić*).

Niejednoznaczność niektórych form wyrazowych poskutkowała wyróżnieniem grupy fleksemów odwołujących się do dwóch możliwości identyfikacyjnych, np.:

a) czasu teraźniejszego albo aorystu: *Seszla do domv y zrzucy s syebye cylycyum (abstulit a se cilicium) BZ* Judith 10, 2 [Sstp XI, 482], forma trzeciej osoby liczby pojedynczej czasownika *zrzucić*,
b) imperfektu albo aorystu: *Iaco ne winidzechø sz Ganowa domu y ne wcradzechø Woitcovi coni* 1401 *Kal* nr 17 [Sstp IX, 366], forma trzeciej osoby liczby mnogiej czasownika *ukraść*.

W tagsecie Korpusu Polszczyzny do 1500 r. fleksemy tego typu zostały określone odpowiednio: „niejednoznaczna forma czasownikowa aoryst albo praesens"; „niejednoznaczna forma czasownikowa aoryst albo imperfectum".

Ponadto poszerzył się repertuar form czasownika *być* jako elementu składowego konstrukcji czasownikowych. W opracowanym przez nas tagsecie leksem *być* w funkcji słowa posiłkowego może występować w następujących formach: przyszłej, teraźniejszej, przeszłej, rozkaźnikowej oraz w postaci aglutynantu i aglutynantu aorystycznego.[13] Taki podział odpowiada potrzebom opisu wieloskładnikowych form czasownikowych wyrażających w staropolszczyźnie m.in. czas zaprzeszły czy tryb przypuszczający. Przykładowo we fragmencie: *Allecz nasz Xt mily szafszecz gest gim on dobrego cloueka dal byl, genszecz gest ge o gich sloscz karal byl Gn* 11a [Sstp III, 242]; forma trzeciej osoby liczby pojedynczej rodzaju męskiego czasu zaprzeszłego: *gest karal byl* (transkr. *jest karał był*) składa się z formy teraźniejszej czasownika *być* (w funkcji składnika konstrukcji czasownikowej), imiesłowu czasu przeszłego czynnego II czasownika *karać* oraz formy przeszłej czasownika *być* (w funkcji składnika konstrukcji czasownikowej).

Na liście części mowy tagsetu Korpusu Polszczyzny do 1500 r. znalazły się również nowe pozycje, m.in. tzw. hybryda. Do tego leksemu

[13] W KorBie wyróżniono fleksemy: forma *być* – wykładnik czasu przyszłego, forma *być* – wykładnik czasu zaprzeszłego, aglutynant *być*, aglutynant aorystyczny *być*.

przyporządkowujemy te polskie wyrazy, które wykazują w tekście fleksję łacińską, np.:

a) *Liripepium nigrum ex czyndalino* 1495 *RocznKrak* XVI 63 [Sstp I, 340], forma rzeczownika *cyndalin* z końcówką łacińskiego ablatywu liczby pojedynczej,
b) *Item ducant poduodas in ordine de uilla ad uillam proximam* 1230 *KodMazK* nr 278 [Sstp VI, 282], forma rzeczownika *podwoda* z końcówką łacińskiego biernika liczby mnogiej.

Trzeba podkreślić, że w Korpusie hybryda nie ma przypisanych żadnych kategorii fleksyjnych i selektywnych.

Pomimo że hybrydy pojawiają się przede wszystkim w kontekstach łacińskich (teksty glosowane, makaronizowane) i przypuszczalnie niewiele z nich zostanie zanotowanych w Korpusie, to wyodrębnienie takiego leksemu wydaje się konieczne. Dzięki temu w aplikacji Słownik zbierającej pod danym hasłem wszystkie formy fleksyjne będzie możliwe uwzględnienie poświadczeń wyrazu z końcówką łacińską niezależnie od rodzaju tekstu, w jakim zostały zanotowane. Ponadto hybrydy będą przydatne w przyszłości przy anotacji tekstów nieciągłych.

Dostosowanie tagsetu KorBy do potrzeb staropolszczyzny objęło nie tylko dodanie nowych, ale i modyfikację wybranych oraz usunięcie zbędnych elementów. Odnosi się to do listy leksemów i fleksemów, zestawu kategorii gramatycznych i ich wartości oraz przyporządkowania kategorii gramatycznych do poszczególnych fleksemów. Zakładamy, że tagset Korpusu Polszczyzny do 1500 r. może jeszcze ulec zmianom po wstępnej anotacji wybranych tekstów.[14]

ZAKOŃCZENIE

Zbudowanie Korpusu Polszczyzny do 1500 r. oznakowanego fleksyjnie usprawni prowadzenie badań nad leksyką i systemem gramatycznym języka tego okresu. Dzięki temu zasobowi możliwe będzie porównywanie słownictwa wielu średniowiecznych zabytków, analizowanie zróżnicowania w zakresie leksyki i gramatyki, a także obserwowanie zmian w polszczyźnie najstarszej doby. W połączeniu ze zbiorami tekstów z późniejszych okresów Korpus ułatwi badanie przebiegu ewolucji zjawisk językowych od średniowiecza po współczesność.

[14] W artykule przedstawiono jedynie wybrane modyfikacje, mające uwypuklić konieczność dokonania pewnych zmian w tagsecie polszczyzny barokowej. Pełen zestaw klas i kategorii gramatycznych oraz ich wartości zostanie opublikowany po udostępnieniu Korpusu.

Bibliografia

J.S. Bień, 1991, *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, Warszawa.

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

M. Klapper, D. Kołodziej, 2014, *Elektroniczny korpus tekstów staropolskich do 1500 r. Perspektywy i problemy*, „Prace Filologiczne" LXV, s. 203–210.

Korpus tekstów staropolskich do roku 1500, https://ijp.pan.pl/publikacje-i materialy/zasoby/korpus-tekstow-staropolskich/ [dostęp: 18.02.2020 r.].

M. Król, M. Derwojedowa, R.L. Górski, W. Gruszczyński, K. Opaliński, M. Woliński, W. Kieraś, M. Eder, 2018, *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" XCIX, s. 92–101.

Sstp: S. Urbańczyk (red.), 1953–2002, *Słownik staropolski*, t. I–XI, Warszawa–Wrocław–Kraków.

W. Twardzik (red.), 2006, *Biblioteka zabytków polskiego piśmiennictwa średniowiecznego*, Kraków (publikacja multimedialna – płyta DVD).

W. Twardzik, E. Deptuchowa, L. Szelachowska-Winiarzowa, 2005, *Opis źródeł* Słownika staropolskiego, Kraków.

W. Wydra, 1998, *Historyczny i kodykologiczny opis rękopisu* [w:] *Rozmyślanie przemyskie*, t. 1–3, s. XXXVIII–LV.

On the Corpus of Polish until 1500 project

Summary

This paper presents the assumptions of the Corpus of Polish until 1500, which is being developed as part of the project titled *Baza leksykalna* średniowiecznej *polszczyzny (do 1500 roku). Fleksja* (*Lexical Database of Medieval Polish (until 1500). Inflection*). It introduces the fundamental objectives of the project, namely preparing an inflectional description of all (inflected) words from the time until 1500 and building a morphosyntactically annotated collection of texts from the same period. Afterwards, the authors discuss the present digital collections of Old Polish texts. In the main part of the paper, they present the criteria for selecting sources for the Corpus under creation and their elaboration methods, which refer to the solutions developed in the Electronic Corpus of Polish Texts from the 17th and 18th centuries (until 1772). Their major modifications aimed to adapt the structural and morphosyntactic annotations for the purpose of describing Mediaeval Polish are discussed on selected examples.

Keywords: electronic text corpus – Old Polish – Mediaeval Polish – inflectional description

Trans. Monika Czarnecka

DOI: 10.33896/PorJ.2020.8.2

Krzysztof Opaliński
(e-mail: krzysztof.opalinski@ibl.waw.pl)
ORCID: 0000-0001-8775-4953
Patrycja Potoniec
(e-mail: patrycja.potoniec@ibl.waw.pl)
ORCID: 0000-0002-5911-5422
(Instytut Badań Literackich Polskiej Akademii Nauk, Warszawa)

KORPUS POLSZCZYZNY XVI WIEKU

1. ZAŁOŻENIA

Pierwotną przyczyną podjęcia prac nad korpusem szesnastowiecznej polszczyzny są problemy, przed którymi stanął na początku XXI wieku zespół Pracowni Słownika Polszczyzny XVI wieku IBL PAN. Baza materiałowa tego słownika, przygotowana w połowie XX wieku (z wykorzystaniem dostępnych wówczas środków) na nośnikach o małej trwałości, po 50 latach użytkowania wykazywała znaczne ślady zniszczenia.

Baza ta składa się z kopii powielaczowych transliteracji tekstów szesnastowiecznych, stanowiących podstawę źródłową *Słownika polszczyzny XVI wieku* (dalej: SPXVI), oraz ze słynnej (uznawanej za najliczniejszą w dziejach polskiej leksykografii [por. np. Piotrowski 2001, 102]) kartoteki kartkowej liczącej 8 milionów fiszek. Dziś ogrom pracy włożonej w jej przygotowanie, wykonywanej w zasadzie ręcznie (z użyciem maszyny do pisania i powielacza), wydaje się nie do wyobrażenia.[1] Dzięki niej każdy z tekstów wybranych jako źródło dla haseł SPXVI otrzymał swój odpowiednik w postaci transliterowanych matryc tekstowych.

Oprawione egzemplarze tych transliteracji (zwane egzemplarzami archiwalnymi) przez pół wieku stale były używane w Pracowni Słownika Polszczyzny XVI wieku, w obliczu jednak postępującej degradacji niezbędne było opracowanie planu ich ocalenia przed całkowitą destrukcją.

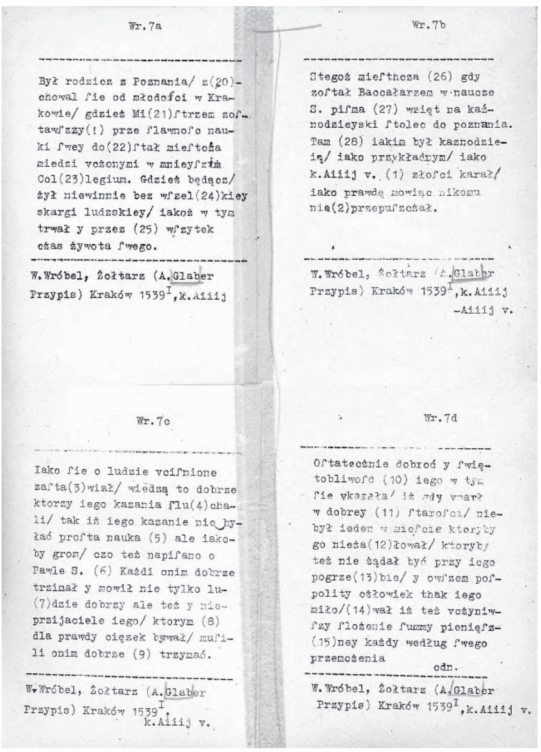
Najoczywistszym rozwiązaniem wydawało się jakiekolwiek utrwalenie materiałów w postaci cyfrowej. Dzięki możliwościom, które daje komputerowe przetwarzanie i zapisywanie danych, droga rozwoju badań językoznawczych (zwłaszcza na dużych zasobach danych) jest od dawna wytyczona: „nowoczesne nośniki informacji stwarzają (...) szeroki dostęp (także przez sieci komputerowe) i do wielkich zbiorów tekstów (zwanych

[1] Wystarczy powiedzieć, że miesięczna norma pracy – przepisanie, skontrolowanie (trzykrotne!), opatrzenie hasłami i uporządkowanie alfabetyczne fiszek – dla zespołu pięcioosobowego wynosiła 10 000 fiszek [SPXVI, zeszyt próbny, s. VII].

korpusami), które można stosunkowo swobodnie przeszukiwać" – pisał Zygmunt Saloni w 1996 roku [s. 15].[2]

Mając na uwadze unikatowość bazy źródłowej SPXVI (zbiór ponad 250 tekstów transliterowanych według jednolitych zasad), celowe wydawało się nie tylko stworzenie kopii cyfrowych (np. skanów) istniejących materiałów, ale i takie ich opracowanie, by mogły być przydatne nie tylko leksykografom przygotowującym SPXVI.

Ilustracja 1. Strona z egzemplarza archiwalnego (transliterowanej matrycy tekstowej).



Wr.7a

Był rodzicz z Poznania/ z(20)-
chował ſie od młodoſci w Kra-
kowie/ gdzieſ Mi(21)ſtrzem zoſ-
tawſzy(!) prze ſławnoſc nauki
ſwey do(22)ſtał mieſtcza
miedzi vczonymi w mnieyſzym
Col(23)legium. Gdzieſ będącz/
żył niewinnie bez wſzel(24)kiey
skargi ludzkiey/ iakoś w tym
trwał y przez (25) wſzytek
czas żywota ſwego.

W.Wróbel, Żołtarz (A.Glaber
Przypis) Kraków 1539[I],k.Aiiij

Wr.7b

Stegoż mieſtnicza (26) gdy
zoſtał Baccałarzem w nauce
S. piſma (27) wzięt na kaź-
nodzieyski ſtolec do poznania.
Tam (28) iakim był kaznodzie-
ią/ iako przykładnym/ iako
k.Aiiij v. (1) złoſci karał/
iako prawdę mowiąc nikomu
nie(2)przepuſzczał.

W.Wróbel, Żołtarz (A.Glaber
Przypis) Kraków 1539[I],k.Aiiij
-Aiiij v.

Wr.7c

Iako ſie o ludzie vciſnione
zaſta(3)wiał/ wiedzą to dobrze
ktorzy iego kazania ſłu(4)cha-
li/ tak iż iego kazanie nie by-
łaś proſta nauka (5) ale iako-
by grom/ czo też napiſano o
Pawle S. (6) Każdi onim dobrze
trzinał y mowił nie tylko lu-
(7)dzie dobrzy ale też y nie-
przijaciele iego/ ktorym (8)
dla prawdy ciężek bywał/ mu-
li onim dobrze (9) trzymać.

W.Wróbel, Żołtarz (A.Glaber
Przypis) Kraków 1539[I],
k.Aiiij v.

Wr.7d

Oſtatecznie dobroć y ſwię-
tobliwoſc (10) iego w tym
ſie vkazała/ iż gdy vmarł
w dobrey (11) ſtaroſci/ nie-
był ieden w mieſcie ktoryby
go nieża(12)łował/ ktoryby
też nie żądał być przy iego
pogrze(13)bie/ y owſzem poſ-
polity człowiek tak iego
miło/(14)wał iż też vczyniw-
ſzy ſłożenie ſummy pieniężz-
(15)ney każdy według ſwego
przenożenia
odn.

W.Wróbel, Żołtarz (A.Glaber
Przypis) Kraków 1539[I],k.Aiiij v.

Pojawienie się na początku lat dwutysięcznych radykalnych „postulatów, dotyczących naukowego słownika przyszłości" [Żmigrodzki 2005, 9], jak zgłoszony w 2005 roku przez Piotra Żmigrodzkiego: „słowniki dokumentacyjne mogłyby bez większych strat dla nauki zostać zastąpione

[2] Szczegółowy opis metod gromadzenia danych językowych (tworzenia korpusu tekstów dla słowników) za pomocą komputerów przedstawił na przykładzie Collin Birmingham University International Language Database (COBUILD) T. Piotrowski w 1993 roku [s. 210–218; o historii tego projektu zob. też: https://www.collinsdictionary.com/cobuild/].

elektronicznymi korpusami tekstów" [s. 9], wskazywało, że korpus tekstów szesnastowiecznych, i to korpus dostępny online, jest niezbędny i przez użytkowników pożądany [zob. uwagi w: Żmigrodzki 2005, 7–8], niezależnie od tego, czy uznać można słuszność postulatu zastąpienia korpusem tak zaawansowanego przedsięwzięcia leksykograficznego jak SPXVI.[3]

Trzeba jednak brać pod uwagę, że „należyte opracowanie korpusu wymaga dużej pracy, wyspecjalizowanego sprzętu komputerowego i oprogramowania oraz znacznych nakładów finansowych" [Saloni 1996, 15]. Dla Pracowni Słownika Polszczyzny XVI wieku IBL PAN podjęcie się realizacji takiego zadania oznaczałoby zawieszenie prac nad SPXVI, a było to (i jest) wykluczone, ponieważ jedyną właściwą drogą dla tak daleko posuniętych i tak długo trwających prac jak opracowanie SPXVI jest ich jak najszybsze zakończenie.

Dopiero więc uruchomiony w lutym roku 2011 Narodowy Program Rozwoju Humanistyki [https://www.gov.pl/web/nauka/narodowy-program-rozwoju-humanistyki, dostęp: 20 lutego 2020 r.] dawał szansę zdobycia odpowiednich środków finansowych na realizację takiego przedsięwzięcia i zatrudnienie osób spoza pracowni. 16 maja 2011 roku do Ministerstwa Nauki i Szkolnictwa Wyższego został więc złożony wniosek pt. *Korpus Polszczyzny XVI wieku. Etap I: Dygitalizacja źródeł oraz stworzenie narzędzi informatycznych i udostępnienie materiałów testowych korpusu*, którego realizację przewidziano na maksymalny możliwy dla projektów finansowanych przez NPRH okres 60 miesięcy. Przy objętości materiałów, które miano poddać dygitalizacji i anotacji, oczywiste było już w momencie składania wniosku, że niemożliwe będzie przygotowanie w ciągu 5 lat w pełni funkcjonalnego korpusu językowego. Tym bardziej, że duża część prac miała być wykonana siłami ludzkimi, a nie za pomocą narzędzi automatycznych. Na etapie tworzenia wniosku wydawało się to najlepszą drogą do zadowalającego ograniczenia liczby błędów w powstającym zasobie. Przemawiał za tym stan zachowania materiałów – część kopii powielaczowych jest tak uszkodzona, że ich skanowanie byłoby zbyt skomplikowane, a inne tak nieczytelne, że stanowiłyby dla programu do automatycznego rozpoznawania tekstu (w roku 2011) wyzwanie zbyt duże. Planowane zaś wstępne (tylko dla niektórych tekstów) anotowanie znacznikami formy podstawowej i form morfosyntaktycznych (wprowadzenie takiej anotacji uznaliśmy za niezbędne minimum, by można było mówić o korpusie,[4] a nie jedynie o bazie tekstów) mu-

---

[3] Nawet w roku 2005, gdy ukazał się w „Poradniku Językowym" artykuł zawierający postulat P. Żmigrodzkiego, prace nad SPXVI były doprowadzone niemalże do końca haseł na literę *P*, tom 33. kończący publikację haseł na tę literę jest datowany na rok 2009. Obecnie, w roku 2020, trwa redakcja haseł z zakresu *srożyć się – święty*.

[4] To dość oczywiste założenie, powtarzane często przy definiowaniu *korpusu* – por. np. Szałkiewicz 2013, 133.

siało być przeprowadzane ręcznie ze względu na znaczną wariantywność graficzną i morfologiczną szesnastowiecznej polszczyzny [por. Opaliński 2007, 113], co powodowało, że niemożliwe byłoby zastosowanie do niej narzędzi automatycznych używanych do polszczyzny współczesnej.

Warto podkreślić, że w roku 2011 nie istniał w zasadzie żaden diachroniczny korpus polszczyzny, na którym moglibyśmy się wzorować. Jak w roku 2019 wskazali autorzy artykułu poświęconego zintegrowanemu korpusowi diachronicznemu,

> pierwszy korpus dawnych tekstów polskich, spełniający standardy obowiązujące dziś przy tworzeniu takich zasobów, powstał na potrzeby dużego międzynarodowego projektu IMPACT [Król i in. 2019, 94]

i został udostępniony w styczniu 2012 [zob. Bień 2014, 76]. Korpus polszczyzny XVI wieku był pomyślany w gruncie rzeczy jako dodatek do SPXVI. Biorąc pod uwagę postulaty językoznawców, miał udostępniać materiały, którymi dysponowała Pracownia, a jednocześnie służyć leksykografom przygotowującym SPXVI. To zdeterminowało decyzje merytoryczne, które zostały podjęte przy opracowaniu założeń projektu. Jego celem, wskazanym we wniosku było:

> **Ocalenie, renowacja, poprawa i udostępnienie bazy transliterowanych tekstów szesnastowiecznych** wykorzystywanych do tworzenia *Słownika polszczyzny XVI wieku*,

a spodziewany wynik stanowić miało:

- **powstanie komputerowej bazy transliterowanych tekstów szesnastowiecznych** o objętości około 28 000 stron standardowego maszynopisu udostępnionej na witrynie internetowej Pracowni do ogólnego użytku,
- **ocalenie bazy materiałowej *Słownika polszczyzny XVI wieku***, co zapewni jego ukończenie,
- utworzenie zasobu, który zostanie przetworzony w korpus polszczyzny XVI wieku,
- przygotowanie „małego korpusu" pozwalającego na przetestowanie możliwości planowanego w przyszłości korpusu pełnego („mały korpus" zostanie udostępniony na witrynie internetowej Pracowni w celu zbadania potrzeb użytkowników).[5]

Dzięki skoncentrowaniu planowanego korpusu wokół SPXVI do prac nad realizacją projektu mogliśmy przystąpić z:

- wyłonioną bazą źródłową,
- jednolitą bazą materiałową (teksty transliterowane),
- ustalonym zestawem wyróżnianych klas gramatycznych i odpowiednich dla nich form gramatycznych,
- przeprowadzonym hasłowaniem.

---

[5] Cel i opis wyników to dosłowne (łącznie z wyróżnieniami) cytaty z „opisu projektu", będącego częścią wniosku o finansowanie złożonego w MNiSW 16 maja 2011 roku.

1.1. Baza źródłowa

Baza źródłowa została opisana we wstępie do pierwszego tomu SPXVI.[6] Teksty do niej dobrano zgodnie z następującymi założeniami (nieodbiegającymi od formułowanych w odniesieniu do współczesnych korpusów językowych [por. np. Górski, Łaziński 2012]:

> Materiał staraliśmy się wybrać tak, by znalazła się w nim reprezentacja możliwie wszystkich form piśmienniczych (...). Chcieliśmy także w naszym materiale widzieć wypowiedzi o możliwie różnym stopniu literackości i możliwie różnej genezie społecznej (...). Nie chcieliśmy zaniedbać możliwie pełnej reprezentacji wszystkich dzielnic Polski. Chcieliśmy też w miarę możności dobrać materiał tak, by reprezentował pełny chronologiczny rozwój w obrębie opracowywanego stulecia [Mayenowa 1966, VIII; wyróżnienia PP].

Założenia te obwarowano od razu pewnymi zastrzeżeniami, powtarzalnymi w odniesieniu do słowników historycznej polszczyzny [por. np. Adamiec 2015, 13]:

– „ułamkowość materiału jest losem każdego historycznego słownika",
– „jednostronność Słownika naszego wynika stąd, że jego podstawowym materiałem są druki",
– „materiał rękopiśmienny został wprowadzony głównie jako swoiste tło dla języka druków" [Mayenowa 1966, IX].

I ostatecznym wnioskiem: „*Słownik* zatem jest przede wszystkim *thesaurusem* języka literackiego w najrozmaitszych jego odmianach" [Mayenowa 1966, IX], który znajduje swoje rozwinięcie w uwadze poniekąd usprawiedliwiającej zawężenie materiału słownikowego:

> nie jest także pozbawione argumentacji stanowisko, które właśnie wybitnym indywidualnościom przypisuje odzwierciedlanie głównego toru rozwojowych tendencji języka [Mayenowa 1966, IX].

Baza źródłowa opiera się głównie na pierwszych wydaniach szesnastowiecznych utworów piśmienniczych (nieliczne teksty ekscerpowano z wydań późniejszych, na ogół w wypadku, gdy dostęp do pierwodruków był niemożliwy lub bardzo utrudniony) oraz niewielu rękopisach. Większe objętościowo teksty zostały włączone tylko częściowo, zgodnie z założeniem, że część taka przy dużej objętości tekstu jest w pełni reprezentatywna:

---

[6] Składające się na nią teksty (tzw. kanon podstawowy) wykazywane są w każdym tomie SPXVI w *Skorowidzu tekstów źródłowych I.*

objętość tekstu w literach	**procent ekscerpcji**
powyżej 300 000	20
150 000–300 000	50
poniżej 150 000	100

Założenie to zostało poddane weryfikacji, porównanie zasobu słownictwa w tekście literackim – *Psałterzu Dawidowym* Kochanowskiego [z 1579 roku] i w przekładzie psałterza wchodzącym w skład *Biblii Leopolity* [z 1561 roku], wykazało znaczną przewagę tego pierwszego. To oraz zewnętrzne przyczyny (powiązanie SPXVI z serią wydawniczą *Biblioteka Pisarzów Polskich*) wpłynęło na decyzję o modyfikacji wypracowanych założeń statystycznych i włączeniu wielu tekstów literackich w 100%, niezależnie od ich objętości.

Szczegółową analizę zasobu słownictwa w różnych tekstach kanonu źródłowego SPXVI przedstawił w swojej części wstępu do tomu 1. Władysław Kuraszkiewicz [1966, XIV–XXV].

1.2. Baza materiałowa

Szczególną wartość tej bazy określa fakt, że transliteracja tekstów szesnastowiecznych została w większości wykonana na potrzeby SPXVI,[7] a teksty te opracowano zgodnie z *Zasadami wydawania tekstów staropolskich. Projekt*.[8] Jednolitość tego zasobu stanowi o jego wyjątkowości, szczególnie gdy weźmie się pod uwagę znaczne zróżnicowanie tekstów drukowanych w XVI wieku, zarówno pod względem graficznym, jak i ortograficznym.

Opracowane według tych zasad teksty transliterowane powielono wielokrotnie, tak by dysponować odpowiednią liczbą kopii całości (egzemplarzy archiwalnych) oraz by każda jednostka danego tekstu otrzymała zawierającą ją fiszkę wraz z kilkuzdaniowym kontekstem.[9] Udostępnienie

---

[7] Z nielicznymi wyjątkami, gdy wykorzystano materiały już istniejące lub powstające niemalże jednocześnie z kartoteką SPXVI, jak np. *Zapiski i roty polskie XV–XVI wieku z ksiąg sądowych ziemi warszawskiej* wydane w 1950 r.

[8] Zresztą w opracowaniu tegoż (dziś już nieco przestarzałego i poddawanego licznym modyfikacjom przez redakcje i edytorów, choć nadal obowiązującego) zbioru zasad przetwarzania tekstu dawnego (transkrypcji i transliteracji o różnych poziomach szczegółowości) brali udział członkowie redakcji SPXVI i współpracownicy zespołu leksykograficznego, a przede wszystkim: M.R. Mayenowa, F. Pepłowski i J. Woronczak oraz W. Kuraszkiewicz i W. Taszycki [por. Hrabec 1966, V–VI].

[9] Zajęło to zresztą wiele lat (gdy ukazywał się zeszyt próbny SPXVI, w 7 lat od powstania Pracowni, przygotowanie materiałów nie było jeszcze zakończone), podczas których opracowano też zasady redakcyjne SPXVI. Por. też opis tradycyjnych metod ekscerpcji [Żmigrodzki 2009, 30–31].

takiej bazy jako zasobu internetowego umożliwiłoby badaczom dawnej polszczyzny korzystanie z najobszerniejszego zbioru tekstów renesansowych transliterowanych według jednolitych zasad.

1.3. Zestaw klas i form gramatycznych

W hasłach SPXVI obowiązuje tzw. tradycyjny podział na części mowy i przysługujące im formy gramatyczne. Zważywszy na okres, w którym powstawały założenia metodologiczne SPXVI jest to oczywiste. Podział ten został przez Krystynę Wilczewską scharakteryzowany ogólnie we wstępie do tomu 1. SPXVI [Wilczewska 1966, XXVI–XLII], a bardziej szczegółowo w *Instrukcji redakcyjnej „Słownika"* [patrz: http://spxvi.edu.pl/instrukcja, rozdz. IV i V]. Równie oczywiste w połowie XX wieku wydawało się przyjęcie w słowniku naukowym[10] konwencji opisu gramatycznego za pomocą skrótów nazw łacińskich. Kierując się wymogami, jakie narzuca redagowanie haseł SPXVI, zakładaliśmy używanie obowiązujących w nim konwencji klasyfikacji i opisu gramatycznego również w planowanym korpusie.

1.4. Hasłowanie

Kartoteka fiszkowa w Pracowni Słownika Polszczyzny XVI wieku jest w zasadzie uporządkowana. Po zebraniu i transliteracji źródeł przeprowadzono hasłowanie i wszystkie wyekscerpowane na potrzeby SPXVI jednostkowe reprezentacje tekstowe danego leksemu zakwalifikowano do formy podstawowej, a zawierające je fiszki umieszczono w odpowiednim dla danego hasła zbiorze materiałów.

Między innymi dzięki temu mógł powstać indeks zbiorczy wszystkich haseł SPXVI wraz z orientacyjną liczbą użyć tekstowych dla każdego z nich.[11]

W klasyfikacji materiałów zdarzają się, oczywiście, błędy,[12] wariantywność morfologiczna i ortograficzna dawnej polszczyzny nastrę-

---

[10] Na uznanie SPXVI za słownik naukowy wskazują już zawarte we wstępie do niego uwagi mówiące o nim jako o „podstawie materiałowej" dla językoznawstwa i stylistyki historycznej oraz historii literatury [Mayenowa 1966, VIII]. Spełnia on także wszystkie kryteria przypisywane słownikowi naukowemu przez P. Żmigrodzkiego [2005, 5].

[11] Indeks ten jest wciąż uzupełniany o hasła nowe, wyekscerpowane ze źródeł spoza podstawowego kanonu tekstów – jego aktualna wersja znajduje się na portalu internetowym SPXVI pod adresem: http://spxvi.edu.pl/indeks/

[12] Zwodnicze bywają, jak wiadomo, formy lub leksemy homonimiczne czy homografy – np. w materiale hasła STRONA do tej pory znajdują się zarówno fiszki o znaczeniu 'część instrumentu', jak i należące do znaczeń związanych z opisami stosunków przestrzennych, choć w indeksie wyróżnione są hasła 1.STRONA i 2 STRONA (jednak ich liczebność nie odpowiada stanowi faktycznemu).

cza trudności nawet doświadczonym badaczom [por. np. Gruszczyński 2010, 28–30]. Błędy te korygowane są na ogół w momencie opracowywania kolejnych haseł, więc w części materiałów (rozpoczynających się na litery *S–Ż*) korekty takiej jeszcze nie przeprowadzono. Nawet jednak niedoskonała siatka haseł może stanowić dobry punkt odniesienia dla planowanego korpusu tekstów szesnastowiecznych – zbiór leksemów, do których przyporządkowane muszą być wszystkie jednostki tekstu w całym zasobie. Ograniczenie zbioru możliwych form podstawowych to pomoc dla osób zajmujących się przypisywaniem odpowiednich znaczników jednostkom tekstu w korpusie oraz sposób na uniknięcie tworzenia przez wykonawców form nieistniejących.[13]

Ilustracja 2. Kartoteka fiszkowa *Słownika polszczyzny XVI wieku*.

2. REALIZACJA

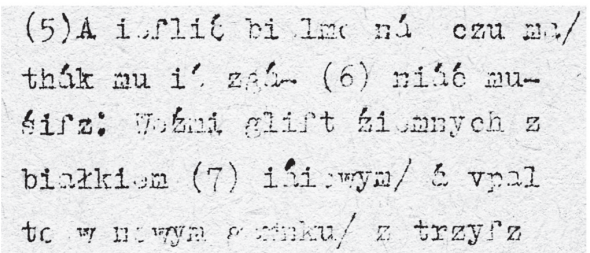
Przy opracowaniu koncepcji korpusu tekstów XVI-wiecznych założenia przyjęte kilka dekad temu dla materiału SPXVI należało zweryfikować na podstawie doświadczeń w pracach redakcyjnych oraz dostosować je do możliwości nowych technologii. Z racji ogromu przedsięwzięcia prace należało podzielić na kilka etapów, z których pierwszy miał polegać na uzyskaniu spójnej elektronicznej bazy tekstów.

[13] Prawdopodobne jest, że np. nieistniejąca dziś forma pierwszej osoby liczby mnogiej czasu teraźniejszego *porzem* większości anotatorów nie skojarzyłaby się z formą podstawową PROĆ (dziś: PRUĆ) [zob. SPXVI, t. 30, s. 325], a raczej sprowokowałaby wytworzenie *PORAĆ.

2.1. Baza tekstów

Początkowo rozważano pozyskanie tekstów w sposób możliwie zautomatyzowany. Wykonano w tym celu próby skanowania egzemplarzy archiwalnych i cyfrowego rozpoznawania tekstu. Niestety efekt tych prób nie był zadowalający, głównie ze względu na licznie występujące na kopiach powielaczowych i rozmywające tekst plamy tuszu czy niedotłoczone i nieodbite znaki oraz skazy słabej jakości papieru.[14]

Ilustracja 3. Fragment matrycy tekstowej egzemplarza archiwalnego.


(5)A i.rlić bi lmc ná ozu mo/
thák mu i' zgá- (6) niáć mu-
śirz: Wźmi glift źiemnych z
białkiem (7) iáiowym/ á vpal
to w nowym rondku/ z trzyfz

Dodatkową trudnością była organizacja tekstu na stronach egzemplarzy archiwalnych, która polega na podziale tekstu na tzw. ćwiartki, odpowiadające wielkością fiszce w formacie A6, oraz tekstem dodatkowym umieszczonym na każdej ćwiartce, zawierającym informacje o lokalizacji w tekście źródłowym i numerze ćwiartki (zob. ilustracja 1.). Nadto niektóre teksty na ćwiartkach miały dopisywane nierzadko uwspółcześnione i skrócone rozszerzenia kontekstu z poprzednich ćwiartek, by łatwiej było zrozumieć dany fragment. Dopiski te oraz informacje lokalizacyjne po cyfrowym rozpoznaniu tekstu należałoby usunąć, a to niepotrzebnie wydłużyłoby pracę nad tekstem. Przeciwko przetwarzaniu kopii powielaczowych przemawiały również błędy w przepisaniu niektórych tekstów.[15] Błędy te skrupulatnie i wielokrotnie poprawiane w pracach redakcyjnych i tak implikowałyby konieczność przeprowadzenia korekt, które trudno byłoby zautomatyzować.

W rezultacie czynności przygotowawczych wyłoniły się zarówno założenia co do zakresu prac, jak i koncepcja merytorycznej zawartości bazy tekstów i w dalszej perspektywie zalążka korpusu, który jako propozycja miał być udostępniony do dalszych dyskusji.

U podstaw części bazowej korpusu legła myśl, by teksty, tak samo jak SPXVI, służyły nie tylko lingwistom, ale by stanowiły także ma-

[14] Niestety nie zachowały się tzw. matryce będące wzorem dla kopii powielaczowych.

[15] Niektóre błędy dostrzeżono już na etapie przygotowywania matryc powielaczowych i ręcznie naniesiono poprawki, które dodatkowo utrudniają OCR.

teriał źródłowy dla specjalistów z innych dziedzin, np. historyków czy badaczy dawnego prawa i społeczeństwa, dlatego istotne było, by uzyskały formę elektroniczną o możliwie uniwersalnej postaci [Prinke 2000, 120 i nast.]. Przyjęto zatem kodowanie tekstów w języku XML w powszechnie stosowanym formalizmie TEI (Text Encoding Initiative) w wersji P5,[16] który jest w stanie przechować metadane o niemal wszystkich cechach tekstu oryginalnego. Zalecenia TEI są bardzo obszerne, należało zatem dokonać selekcji elementów i atrybutów, które będą miały zastosowanie w tekstach XVI-wiecznych. Dla wyboru języka XML istotne znaczenie miała też możliwość zakodowania informacji o niestandardowych znakach pojawiających się w dawnych tekstach, zatem do zestawu znaczników dodano listę encji znakowych, które miały reprezentować glify spoza standardowego zestawu znaków. Implikowało to z kolei konieczność wyboru odpowiedniego zestawu czcionek, który byłby w stanie wyświetlić zakodowane encje znakowe. W momencie opracowywania założeń najbardziej adekwatnym, a przy tym dostępnym na licencji *open source*, zestawem znaków był font *Palemonas MUFI*, opracowany w ramach projektu The Medieval Unicode Font Initiative.[17]

Jednym z istotnych zadań projektu było zbudowanie narzędzia do wygodnego przepisywania tekstów XVI-wiecznych, które byłoby maksymalnie przyjazne dla użytkownika i jednocześnie pozwalało na zapis plików w formacie XML. Posłużył do tego edytor opracowany przez Bartosza Białego na potrzeby edycji haseł do elektronicznej wersji SPXVI. Edytor został wyposażony w dodatkowy moduł oparty na specjalnym schemacie (wybór z TEI), dzięki któremu możliwe stało się zakodowanie pełnej informacji lokalizacyjnej (strona, kolumna, wiersz), a dla tekstów biblijnych dodatkowo również podział na księgi, rozdziały i wersety. Nadto by oddać układ strony pierwowzoru, można było także wyróżnić takie elementy typograficzne tekstu jak marginesy, nagłówki, ozdobne inicjały, tabele, elementy graficzne, żywe paginy, kustosze. Możliwe również było oznaczenie zmiany kroju pisma, co jest istotne, nierzadko bowiem w XVI-wiecznych drukach pociągała ona za sobą zmianę oznaczania niektórych diakrytów, zwłaszcza pochyleń samogłosek.

By ułatwić późniejsze automatyczne przetwarzanie korpusu, zastosowano także wyróżnianie obcojęzycznych wstawek. Dla przyszłego zastosowania tekstów w pracach redakcyjnych nad SPXVI ważne było również zachowanie numeracji tzw. matryc, czyli poszczególnych fiszek składających się na egzemplarz archiwalny. Numery matryc zapisane w specjal-

---

[16] https://tei-c.org/guidelines/P5/

[17] https://skaldic.abdn.ac.uk/m.php?p=mufi lub alias http://www.mufi.info/ Niestety wybrany font został zoptymalizowany tylko dla środowiska Linux, a w programach działających pod kontrolą systemu MS Windows pojawiają się niedokładności w kształtach znaków, co będzie w przyszłości skutkowało koniecznością wyboru innego fontu lub opracowania specjalnego zestawu.

nych tagach dodatkowo identyfikują miejsce w tekście i pozwalają łatwo odnieść się do szerszego kontekstu podczas analizowania danej fiszki, ponieważ nie zawsze na fiszkach oznaczano numery wierszy.

Przepisane teksty zostały poddane dwóm korektom. Pierwsza z nich sprawdzała zgodność z egzemplarzem archiwalnym, druga zaś polegała na kontroli względem oryginalnego tekstu w postaci fotokopii lub kopii cyfrowej. Sprawdzone pliki XML z pomocą arkuszy XSLT oraz XSL-FO poddane zostały przekształceniom, by uzyskać ich odpowiedniki w formacie HTML do publikacji w serwisie http://spxvi.edu.pl oraz PDF.

Każdy tekst z bazy w nagłówku pliku w formie metadanych zawiera informacje o autorze, tytule tekstu i roku wydania. Poza tym w nagłówku umieszczono główne informacje dotyczące projektu oraz nazwiska osób pracujących przy każdym tekście. Oczywiście istnieje możliwość, by w następnych etapach projektu umieścić w pliku szczegółowe metadane lub wyposażyć sam tekst w kolejne elementy opisujące go na różnych poziomach.

Obecnie zdeponowane w bazie teksty można przeszukiwać jedynie za pomocą wbudowanego w każdą przeglądarkę narzędzia (Ctrl+F lub F3) pod warunkiem, że znana jest przynajmniej przybliżona postać szukanego ciągu i to tylko w obrębie jednego tekstu (lub jego części w wypadku długich tekstów) załadowanego do przeglądarki. Jest to oczywiście rozwiązanie dalece niedoskonałe i potrzebne jest specjalne narzędzie pozwalające na dokładniejsze wyszukiwanie w ramach całej bazy, przy możliwości stosowania różnego rodzaju filtrów i kryteriów. Opracowanie takiego narzędzia zaplanowano w drugim etapie prac.

2.2. Zalążek korpusu

Część korpusową projektu stanowi zalążek korpusu, którego podstawowym założeniem była pełna integracja z elektroniczną wersją SPXVI. W myśl tego założenia korpus miał się posługiwać stosowaną w SPXVI nomenklaturą i sposobem opisu gramatycznego, by ułatwić wzajemne odniesienia między słownikiem a korpusem. Podstawę utworzenia zalążka korpusu stanowiła baza tekstów XVI-wiecznych opracowana w pierwszym etapie projektu. Przyjęto również, że fragmenty tekstów użyte w zalążku zostaną wyposażone w warstwę transkrypcji, a każdy segment będzie skorelowany z indeksem haseł słownika. Oprócz standardowego elementu `base` każdy segment otrzymał element `base_id` zawierający identyfikator hasła w indeksie. Za podstawę tagsetu korpusu posłużyła lista kategorii gramatycznych i form fleksyjnych opracowana przez A. Luto-Kamińską na podstawie *Instrukcji redakcyjnej „Słownika”* [Opaliński 2016]. Ponieważ tagset zalążka nie jest zgodny ze standardem tagsetu NKJP, elementy stosujące nietypowe wartości zostały nazwane inaczej jako `xvi_ctag` dla wartości kategorii gramatycznej oraz `xvi_msd`

dla wartości formy fleksyjnej. Dzięki takiemu rozwiązaniu poszczególnym segmentom będzie można dodać standardowe elementy `ctag` i `msd` stosujące wartości ujednolicone dla korpusów diachronicznych. Każdy wreszcie segment korpusu zawiera atrybut, którego wartość jednoznacznie identyfikuje zarówno tekst, jak i miejsce segmentu w tekście.

```
<seg xml:id="GornDworz_page_T_line_13_pos_2">
     <fs type="words">
           <f name="orth">
                <string>poſzypławſzy</string>
           </f>
           <f name="transcript">
                <string>poszypławszy</string>
           </f>
           <f name="base">
                <string>poszypłać</string>
           </f>
           <f name="base_id">
                <symbol value="90026"/>
           </f>
           <f name="xvi_ctag">
                <symbol value="vb:pf"/>
           </f>
           <f name="xvi_msd">
                <symbol value="part:praet:act"/>
           </f>
     </fs>
</seg>
```

Tagowanie tekstów zalążka korpusu było ręcznie wykonywane przez anotatorów. Do tego procesu został ponownie wykorzystany edytor służący przepisywaniu tekstów z dodatkowym prostym modułem pozwalającym na konwersję tekstu do formatu korpusu, a następnie na ręczną segmentację i tagowanie. Program jedynie podpowiadał możliwe wartości, ich ustalenie należało do anotatorów.

W wyniku ręcznego tagowania powstał zbiór 135 tysięcy segmentów, które zostały zdeponowane w bazie danych zarządzanej przez aplikację internetową zintegrowaną z serwisem SPXVI.[18] Prosta wyszukiwarka tymczasem pozwala na znalezienie żądanego ciągu na podstawie formy podstawowej (element base) lub transkrypcji (element transcript).

[18] Wyszukiwarka korpusowa znajduje się pod adresem: https://spxvi.edu.pl/korpus/probka/szukaj

kontekst	źródło	lokalizacja (strona/wiersz/pozycja)	artykuł hasłowy
... Bog tákową ráną náwiedził Bárwierz **pofzypławfzy** 11 trochę wpuśćił knoth między zálawthy ...	GórnDworz	1 / 13 / 2	poszypłać

Wynik wyszukiwania oprócz znalezionego segmentu zawiera kontekst rozszerzony o pięć poprzedzających i pięć kolejnych segmentów. Postać tego kontekstu można przełączać między transliteracją a transkrypcją. Na podstawie atrybutu identyfikującego segment `xml:id` formułowany jest link do odpowiedniego miejsca w tekście źródłowym, dzięki czemu można poszerzyć kontekst, wyświetlając cały tekst z zaznaczoną lokalizacją segmentu, element `base_id` zaś pozwala na utworzenie linku do odpowiedniego hasła w elektronicznej wersji słownika.

3. PLANY

Obecnie trwają prace przygotowujące do połączenia zarówno istniejących, jak i planowanych korpusów diachronicznych w jeden Narodowy Korpus Diachroniczny Polszczyzny [zob. Król i in. 2019]. W ramach Korpusu Polszczyzny XVI wieku zamierzone prace obejmują konwersję tagsetu do ujednoliconego systemu znakowania w przyszłym korpusie diachronicznym, rozszerzenie ręcznie znakowanego korpusu o kolejne 100 tys. segmentów, by wytworzyć bazę dla automatycznego tagera, a także wspomaganą programowo transkrypcję tekstów transliterowanych. Do tych prac będą adaptowane narzędzia sprawdzone w projekcie KorBa, który obecnie wyznacza standardy dla planowanego Narodowego Korpusu Diachronicznego Polszczyzny.

Poza wspólnymi działaniami akcesorów NKDP w ramach samego Korpusu Polszczyzny XVI wieku będą podejmowane działania mające na celu jego dalszy rozwój. Zasadnicze prace będą polegały na włączeniu do bazy jak największej liczby tekstów spoza tzw. kanonu SPXVI. Obecnie jest to lista licząca ponad pół tysiąca tekstów różnej objętości i o różnej tematyce. Uwzględnienie tego zbioru pozwoli na znacznie większe zrównoważenie bazy przede wszystkim pod względem językowym, ale również jeśli chodzi o różnorodność reprezentowanych dziedzin. Ambicją twórców bazy jest zawarcie jak największej liczby tekstów polskich powstałych w XVI wieku. Wszystkie teksty znajdujące się w bazie mają być zaopatrzone w metadane możliwie najdokładniej opisujące zarówno językowe, jak i pozajęzykowe cechy tekstu, co pozwoli na wielopłaszczyznowe obserwacje. Planuje się również dodanie do transliterowanych tekstów warstwy transkrypcji, co będzie miało niebagatelne znaczenie dla późniejszego wykorzystania ich w korpusie, ale także znacznie ułatwi użytkownikom kontakt z tekstem. Ponadto w zamierzeniach jest też wykorzystanie powstałego w ramach NKDP rozszerzonego i otagowanego

korpusu do wspomagania prac redakcyjnych nad słownikiem. Pozwoli to z jednej strony na przyspieszenie i do pewnego stopnia zautomatyzowanie procesu redakcji (zwłaszcza w części statystycznej), z drugiej zaś będzie okazją do weryfikacji automatycznej anotacji segmentów korpusu.

Bibliografia

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

J. Bień, 2014, *The IMPACT Project Polish Ground-Truth Texts as a DjVu Corpus*, „Cognitive Studies | Études cognitives" nr 14, s. 75–84 (DOI: 10.11649/cs.2014.008).

K. Górski i in. (red.), 1955, *Zasady wydawania tekstów staropolskich. Projekt*, Wrocław.

R. Górski, M. Łaziński, 2012, *Reprezentatywność i zrównoważenie korpusu* [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy korpus języka polskiego*, Warszawa, s. 25–36.

W. Gruszczyński, 2010, *Jednostka opisu leksykograficznego w słowniku historycznym na przykładzie „Słownika języka polskiego XVII i 1. połowy XVIII wieku"*, „Poradnik Językowy" z. 4, s. 26–40.

S. Hrabec, 1966, *Historia „Słownika"* [w:] *Słownik polszczyzny XVI wieku*, t. 1, Wrocław–Warszawa–Kraków, s. V–VII.

M. Król i in., 2019, *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" nr 1, s. 92–101.

W. Kuraszkiewicz, 1966, *Uwagi o statystyce w „Słowniku"* [w:] *Słownik polszczyzny XVI wieku*, t. 1, Wrocław–Warszawa–Kraków, s. XIV–XXV.

M.R. Mayenowa, 1966, *Charakterystyka „Słownika"* [w:] *Słownik polszczyzny XVI wieku*, t. 1, Wrocław–Warszawa–Kraków, s. VIII–XIII.

K. Opaliński, 2007, *Problemy kodowania korpusów historycznych* (*na przykładzie tekstów XVI-wiecznych*) [w:] J. Kamper-Warejko, I. Kaproń-Charzyńska (red.), *Z zagadnień leksykologii i leksykografii języków słowiańskich*, Toruń, s. 107–114.

K. Opaliński, 2016, *Korpus polszczyzny XVI wieku. Tagset (na podstawie opracowania A. Luto-Kamińskiej) – do użytku wewnętrznego*, Toruń.

T. Piotrowski, 2001, *Zrozumieć leksykografię*, Warszawa.

T. Piotrowski, 1993, *Z zagadnień leksykografii*, Wrocław.

R. Prinke, 2000, *Fontes ex machina. Komputerowa analiza źródeł historycznych*, Poznań.

A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), 2012, *Narodowy korpus języka polskiego*, Warszawa.

Z. Saloni, 1996, *Perspektywy polskiej leksykografii jednojęzycznej*, „Poradnik Językowy" z. 7, s. 1–18.

Słownik polszczyzny XVI wieku. Zeszyt próbny, 1956, Wrocław.

Ł. Szałkiewicz, 2013, *Lematyzacja w ręcznej anotacji milionowego podkorpusu Narodowego Korpusu Języka Polskiego – ciekawe przypadki*, „Polonica" r. 33, s. 133–156.

[K. Wilczewska], 1976, *Instrukcja redakcyjna „Słownika polszczyzny XVI wieku”*, Toruń – druk IBL PAN do użytku wewnętrznego, udostępniony za zgodą autorki i IBL PAN pod adresem: http://spxvi.edu.pl/instrukcja

K. Wilczewska, 1966, *Zasady redakcyjne „Słownika”* [w:] *Słownik polszczyzny XVI wieku*, t. 1, Wrocław–Warszawa–Kraków, s. XXVI–XLII.

Zapiski i roty polskie XV–XVI wieku z ksiąg sądowych ziemi warszawskiej, 1950, wyd. W. Kuraszkiewicz i A. Wolff, Kraków.

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

P. Żmigrodzki, 2009, *Wprowadzenie do leksykografii polskiej*, Katowice.

Corpus of the 16th-century Polish language

Summary

The original purpose of creating the corpus of the 16th Polish language was to preserve the material basis of *Słownik polszczyzny XVI wieku* (*Dictionary of the 16th-Century Polish Language*) (SPXVI) comprising 272 texts transliterated in accordance with standardised principles, which is of great value. The project described here consists in creating an online base of the resources and using a part of it as a germ of a language corpus with texts designated with morphosyntactic markers.

The works adopted XML encoding in the TEI (Text Encoding Initiative) formalism, version P5, adjusted to a 16th-century text. Typographical elements as well as grammatical categories and forms of words were designated in the texts. The germ of the corpus of the 16th-century Polish language comprises 135 thousand segments and it will be expanded by another 100 thousand in the future to provide material for an automated form designation tool. Ultimately, integration with the Diachronic Corpus of Polish is planned.

Keywords: lexicography – history of Polish – diachronic corpus of Polish

Trans. Monika Czarnecka

Włodzimierz Gruszczyński
(e-mail: wlodzimierz.gruszczynski@ijp.pan.pl)
ORCID: 0000-0001-9406-1354
Dorota Adamiec
(e-mail: dorota.adamiec@ijp.pan.pl)
ORCID: 0000-0001-8179-038X
Renata Bronikowska
(e-mail: renata.bronikowska@ijp.pan.pl)
ORCID: 0000-0001-9000-5355
Aleksandra Wieczorek
(e-mail: aleksandra.wieczorek@ijp.pan.pl)
ORCID: 0000-0002-8829-559X
(Instytut Języka Polskiego Polskiej Akademii Nauk, Warszawa)

DOI: 10.33896/PorJ.2020.8.3

ELEKTRONICZNY KORPUS TEKSTÓW POLSKICH Z XVII I XVIII W. – PROBLEMY TEORETYCZNE I WARSZTATOWE

1. WSTĘP

Tworzenie korpusów tekstów dawnych stawia przed ich autorami zupełnie nowe problemy, inne niż te, które pojawiają się podczas prac nad korpusami współczesnymi. Dotyczą one m.in. wyboru reprezentatywnych dla danej epoki tekstów, sposobu oddania graficznej postaci tekstu oraz opracowania takiego zestawu znaczników morfosyntaktycznych, który wiernie oddawałby strukturę gramatyczną języka na ówczesnym etapie rozwoju. Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w., który przedstawiamy w niniejszym artykule, jest dobrym przykładem tego, z czym muszą mierzyć się twórcy korpusów tekstów dawnych. Zastosowane rozwiązania umożliwiają w dużej mierze przezwyciężenie tych trudności, jednak w niektórych obszarach okazują się niedoskonałe. Mamy nadzieję, że zdanie sprawy zarówno z mocnych, jak i słabszych stron korpusu pozwoli jego użytkownikom bardziej świadomie z niego korzystać.

Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772 r.) jest najważniejszym rezultatem projektu realizowanego w latach 2013–2018 przez Pracownię Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego PAN we współpracy z Zespołem Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN.[1] Nieformalną nazwę korpusu stanowi określenie KorBa, będące akronimem wyrażenia *kor-*

[1] Projekt był finansowany w ramach Narodowego Programu Rozwoju Humanistyki na lata 2013–2018 (NPRH nr 0036/NPRH2/H11/81/2012).

pus barokowy.[2] W 2019 r. rozpoczęliśmy nowy projekt mający na celu m.in. rozbudowę korpusu (projekt KorBa 2).[3] W dalszej części artykułu, przy podawaniu informacji dotyczących tylko jednej z wersji korpusu, będziemy stosować oznaczenie odpowiedniej wersji (KorBa 1.2 – korpus obecnie dostępny na stronie, KorBa 2.0 – korpus, który będzie wynikiem końcowym obu projektów), kiedy zaś będziemy opisywać rozwiązania przyjęte w obu projektach, będziemy używać po prostu określenia KorBa.

Prezentowany korpus to pierwszy stosunkowo duży elektroniczny korpus dawnych tekstów polskich. Jest on też jedynym w świecie słowiańskim tak obszernym, anotowanym i dostępnym online korpusem prezentującym teksty sprzed XIX stulecia. KorBa 1.2 liczy prawie 13,5 miliona segmentów (w rozumieniu przyjętym przez twórców Narodowego Korpusu Języka Polskiego, dalej: NKJP). Korpus jest zamieszczony w Internecie pod adresem: https://korba.edu.pl. Można go przeszukiwać za pomocą wyszukiwarki MTAS [Brouwer i in. 2017] wykorzystującej powszechnie używany Corpus Query Language (CQL).[4]

2. REPREZENTATYWNOŚĆ I ZRÓWNOWAŻENIE?

Przy budowie korpusu KorBa uwzględniliśmy cechy uznawane za fundamentalne dla korpusów językowych, czyli reprezentatywność i zrównoważenie. Właściwości te rozumieliśmy 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].

Realizacja tych założeń w odniesieniu do materiału historycznego okazała się znacznie trudniejsza niż w korpusie języka współczesnego. Wynika to z ograniczeń charakterystycznych dla badań historycznojęzykowych. Mamy fragmentaryczny dostęp do materiału i są to wyłącznie teksty pisane. Wiedza o piśmiennictwie epoki pozostaje niekompletna. Nie znamy w pełni struktury zbioru tekstów powstałych w badanej epoce. Poza tym w trakcie gromadzenia tekstów do budowy korpusu uwidoczniły się cechy charakterystyczne dla zasobu piśmiennictwa z tego okresu. Wyrazisty przykład stanowią istotne różnice ilościowe pomię-

---

[2] Korpus obejmuje okres, którego większa część przypada na panowanie w literaturze polskiej stylu barokowego [por. Hernas 2002, 20].

[3] Prace nad projektem „Rozbudowa Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. i jego integracja z *Elektronicznym słownikiem języka polskiego XVII i XVIII w.*" są zaplanowane na lata 2019–2023 i finansowane w ramach NPRH (0413/NPRH7/H11/86/2018).

[4] Szczegółowa instrukcja przeszukiwania korpusu [Gruszczyński, Bronikowska 2018] jest dostępna na stronie KorBy.

dzy regionami historycznymi ze znaczącą przewagą tekstów powstałych w Małopolsce. Ważne jest również zróżnicowanie ilościowe i jakościowe na przestrzeni czasu – piśmiennictwo pierwszej połowy XVII w. charakteryzuje się różnorodnością tematyczną i formalną, podczas gdy na początku XVIII w. dochodzi do istotnego ograniczenia liczby wydawanych tekstów, a ich poziom zwykle odzwierciedla załamanie kulturowe, które wynikało z ówczesnej sytuacji politycznej Rzeczypospolitej.

Wyrazem dążenia do zrównoważenia KorBy była decyzja o włączaniu do korpusu wielu obszernych tekstów jedynie we fragmencie obejmującym najczęściej około 100 stron starodruku. Takie rozstrzygnięcie było konieczne, gdyż szacunki dotyczące objętości ważnych tekstów[5] potwierdzały, iż wykorzystanie tych skądinąd bardzo ciekawych pozycji w większym zakresie całkowicie zburzy zrównoważenie korpusu.

Historyczny materiał językowy ogranicza także realizację założenia reprezentatywności opartej na kryterium gatunkowym. Zazwyczaj najmniej dotyczy to twórczości literackiej, często wydawanej ponownie w odróżnieniu od tekstów użytkowych. Trudno więc było zrealizować założenie tylko kilkunastoprocentowego udziału literatury pięknej w całym korpusie [Przepiórkowski i in. 2012, 34]. Istotny postulat przy tworzeniu korpusów stanowi także dążenie do odzwierciedlenia struktury czytelnictwa. Wobec okresu historycznego dysponujemy znikomą wiedzą na ten temat i opiera się ona wyłącznie na wnioskowaniu pośrednim, na przykład na podstawie liczby wydań.

2.1. Kryteria doboru tekstów

W korpusie uwzględniliśmy następujące typy źródeł: rękopisy, starodruki i wydania tekstów XVII- i XVIII-wiecznych, które ukazały się po 1800 r. Oryginalne teksty z epoki (rękopisy i starodruki) stanowią 64% korpusu.[6] Ponieważ nie zachowały się liczne ważne dzieła, uznaliśmy, że lepszym rozwiązaniem niż pominięcie jest umieszczenie w korpusie tekstu w edycji późniejszej, nawet tak niedoskonałej jak wydania dziewiętnastowieczne, w których teksty często podlegały zmianom pod względem ortograficznym, a nawet gramatycznym.

---

[5] Np. cztery tomy *Kazań* T. Młodzianowskiego obejmujące ponad 1000 stron *in folio* zawierają szacunkowo ponad pół miliona segmentów. *Zielnik* S. Syreńskiego (1540 stron *in folio*) szacujemy na 1,1 mln segmentów. Z drugiej strony w całości zostały włączone do korpusu cztery tomy *Nowych Aten* B. Chmielowskiego (1,1 mln segmentów). Tekst w postaci elektronicznej udostępnił J.S. Bień. Uznaliśmy, że różnorodność tematyczna, encyklopedyczny charakter dzieła przemawiają za jego pełnym wykorzystaniem w KorBie.

[6] Wszystkie podane w tym podrozdziale liczby odnoszą się do KorBy 1.2. W wersji 2.0 proporcje pomiędzy poszczególnymi rodzajami tekstów mogą z różnych powodów ulec pewnym zmianom.

Rozstrzygnięcia wymagało również pytanie o miejsce rękopisów w korpusie. Stulecia siedemnaste i osiemnaste to nadal czasy tekstów w znacznej mierze rękopiśmiennych. Trudno jednak uwzględnić ten fakt w korpusie, szczególnie mając świadomość ich bardzo ograniczonego kręgu odbiorców i tym samym znacznie mniejszego udziału w kształtowaniu języka ogólnego w porównaniu z drukami. Istotną przeszkodę we włączaniu rękopisów do korpusu stanowią również trudności z pozyskaniem tego typu tekstów. Ich odczytanie zazwyczaj wymaga niezwykle wysokich kompetencji i jest czasochłonne, a więc generuje koszty znacznie przewyższające przepisywanie nawet trudnych i słabo czytelnych starodruków. Jednocześnie rękopisy to nieprzeceniony materiał w badaniach nad idiolektami czy regionalizmami. Warto więc w przyszłości podjąć prace nad rozbudową korpusu także o podkorpus rękopisów.[7]

Dobór tekstów do korpusu opierał się na następujących kryteriach: chronologicznym, geograficznym i genologicznym oraz kryterium różnorodności tematycznej. Wewnętrzne cezury czasowe wprowadzone w okresie 172 lat, które obejmuje korpus, mają charakter wyłącznie umowny. Udział w korpusie tekstów z wyróżnionych podokresów przedstawia się następująco: 1601–1650: 38,4% segmentów; 1651–1700: 29,2%; 1701–1750: 16,3%; 1751–1772: 16,1%.

Dominacja pierwszej połowy XVII w. pod względem liczby segmentów w korpusie jest uzasadniona (jak już wspomniano wyżej) faktem, że w tym okresie powstało wiele obszernych tekstów ważnych i popularnych w całej epoce polskiego baroku (np. zbiory kazań F. Birkowskiego, Sz. Starowolskiego, *Zielnik* Sz. Syreńskiego, *Biblia gdańska*, poezje K. Miaskowskiego i J. Jurkowskiego, utwory S. Twardowskiego, przekłady popularnych w ówczesnej kulturze europejskiej tekstów: *Orlanda szalonego* L. Ariosta, *Relacji powszechnych* G. Botera). Konsekwencją (której staraliśmy się uniknąć) ścisłego zrównoważenia chronologicznego byłoby okrojenie tego różnorodnego materiału na rzecz tekstów z pierwszej połowy XVIII w. Te teksty z kolei charakteryzują się najczęściej nieurozmaiconą tematyką (religijną, panegiryczną) i w korpusie doszłoby do nadreprezentacji tekstów o podobnej tematyce. Konstruowanie korpusu było – jak widać – sztuką balansowania i kompromisu pomiędzy różnymi jego walorami.

Kwalifikując teksty do korpusu, braliśmy pod uwagę także ich zróżnicowanie pod względem pochodzenia geograficznego. Teksty zgromadzone w korpusie zostały przyporządkowane regionom na podstawie miejsca ich wydania, a w wypadku braku miejsca wydania z epoki uwzględnialiśmy miejsce powstania (jeśli jest znane). Trzeba zauważyć, że nierównomierny udział tekstów z poszczególnych regionów to przede wszystkim odzwier-

[7] Ze względu na przedstawione ograniczenia rękopisy do KorBy 1.2 zostały włączone w niewielkim zakresie. Przykłady takich tekstów to: *Pamiętniki* J.Ch. Paska, *Kopie listów do (...) Krzysztofa Paca* M. Czartoryskiego, *Księga grodzka owrucka*.

ciedlenie aktywności głównych ośrodków wydawniczych. Największa część materiału językowego pochodzi z Małopolski (30,4%). Udział pozostałych regionów przedstawia się następująco: Ziemie Ruskie 12,4%, Wielkie Księstwo Litewskie 7,8%, Wielkopolska 7,6%, Mazowsze 7,4%, Pomorze i Prusy 5,2%, Śląsk[8] 1,4%, Inflanty 0,04%. W KorBie 1.2 nie ma tekstów pochodzących z Podlasia i Kurlandii.[9] Teksty polskie, które zostały wydane ówcześnie za granicą (np. w Lipsku), stanowią osobną klasę, obejmującą 1,1% segmentów. Niestety nie udało się ustalić, z jakiego regionu pochodzą teksty stanowiące 26,7% materiału. Dążyliśmy do tego, aby zawartość korpusu odzwierciedlała udział poszczególnych regionów w produkcji piśmienniczej, co oczywiście musiało się odbyć kosztem zrównoważenia pod względem udziału regionów w korpusie.

Na korpus złożyły się teksty sklasyfikowane przez nas w jedenastu rodzajach (w tym cztery literackie, sześć nieliterackich oraz *Biblia*[10]). Teksty literackie stanowią 23,3% całego korpusu, teksty nieliterackie to 74,3%, pozostałe zaś 2,4% korpusu to teksty biblijne. Klasyfikacja na rodzaje koresponduje – na tyle, na ile było to możliwe – z klasyfikacją zastosowaną do tekstów współczesnych w NKJP. Wyodrębnione w KorBie rodzaje to: epika (8,7%), liryka (8,6%), dramat (1,8%), utwory synkretyczne (4,2%), wiadomości prasowe i druki ulotne (1,5%), teksty naukowo-dydaktyczne lub informacyjno-poradnikowe (24,5%), teksty perswazyjne (17,8%), literatura faktograficzna (21,3%), teksty urzędowo-kancelaryjne (7,4%), listy (1,8%). W obrębie poszczególnych rodzajów wyróżniamy gatunki (np. dla epiki – bajki, żywoty świętych i in., dla liryki – epitafia, ody i in., dla tekstów naukowo-dydaktycznych lub informacyjno-poradnikowych – encyklopedie, kompendia, podręczniki, książki kucharskie i in., dla tekstów urzędowo-kancelaryjnych – inwentarze, testamenty i in.). Ze względu na wielką różnorodność wyodrębniliśmy ok. 60 gatunków[11] i nie traktujemy tej listy jako zamkniętej.

2.2. Metadane

Korpus jest opatrzony licznymi metadanymi, które pozwalają na filtrowanie wyników wyszukiwania. Są to dane bibliograficzne oraz informacje opisane w punkcie 2.1. Dzięki nim użytkownik może ograniczyć wyszukiwanie np. do tekstów z wybranego przedziału czasowego, tek-

---

[8] Śląsk nie leżał wprawdzie w granicach Rzeczypospolitej, jednak język polski był tam używany i publikowano po polsku.

[9] Staramy się pozyskać teksty z tych regionów do rozbudowywanego obecnie korpusu.

[10] Poszczególne fragmenty *Biblii* należą do różnych gatunków literackich i nieliterackich, stąd najlepszym rozwiązaniem wydało nam się wyodrębnienie tego szczególnego tekstu.

[11] Spis gatunków przypisanych poszczególnym rodzajom znajduje się na stronie internetowej korpusu w zakładce *O korpusie*.

stów jednego autora czy pochodzących z jednego regionu. Wyodrębnienie regionów geograficznych może posłużyć do śledzenia zróżnicowania dialektalnego w tekstach, a możliwość zawężania wyszukiwania do przedziałów czasowych – do śledzenia zmian chronologicznych.

Każdy tekst ma unikatowy identyfikator, którym jest skrót utworzony od nazwiska autora, tłumacza i tytułu. Poza tym użytkownik korpusu może uwzględnić przy wyszukiwaniu następujące dane bibliograficzne: tytuł, autor, tłumacz (w wypadku tekstów przetłumaczonych), miejsce wydania, drukarnia, data wydania. Rzecz jasna, nie dla każdego tekstu wszystkie te informacje są dostępne, część utworów jest zatem oznaczona jako anonimowe, z nieznanym miejscem wydania czy też z nieznaną bądź tylko przybliżoną datą wydania. Wydania XIX-wieczne i późniejsze opatrzone są odpowiednią informacją i danymi bibliograficznymi wydania spoza epoki.

Każdy tekst został też opatrzony licznymi informacjami dotyczącymi stylistyki, genologii i tematyki. Określamy zatem typ mowy (wierszowana, niewierszowana, mieszana), rodzaj, gatunek i zakres tematyczny. Zaznaczamy także, czy tekst powstał w konwencji poetyki żartu. Kategoria ta dotyczy różnorakich utworów satyrycznych i ma umożliwić badanie nacechowanych, także idiolektalnych środków językowych. Rozróżnienie na mowę wierszowaną i niewierszowaną również może być pomocne w badaniach, gdyż w poezji użycie środków językowych bywa podporządkowane rymom i rytmowi.

Określenie gatunku tekstu bywało kłopotliwe. Podobnie było z tematyką – tylko w wypadku niektórych typów utworów mogliśmy wskazać ją jednoznacznie (np. w wypadku tekstów naukowych mogła to być astronomia, biologia, fizyka, matematyka itd., w wypadku akt sejmikowych – polityka i prawo, a kazań – religia). Jednemu dziełu przypisujemy więc niekiedy więcej niż jeden gatunek i zakres tematyczny. W szczególnych wypadkach dopuszczaliśmy nieprzypisywanie tekstowi żadnego gatunku bądź zakresu tematycznego. Taką decyzję podjęliśmy np. w odniesieniu do kalendarzy, którym nie przypisaliśmy żadnej tematyki.

3. OZNAKOWANIE STRUKTURY DOKUMENTU I FRAGMENTÓW OBCOJĘZYCZNYCH

Anotacja przepisanych tekstów obejmuje odwzorowanie struktury dokumentu źródłowego i oznakowanie fragmentów obcojęzycznych (a także oznakowanie morfosyntaktyczne każdego segmentu, co omówimy niżej w punkcie 5.).

3.1. Znakowanie strukturalne

Dzięki znakowaniu strukturalnemu użytkownik zyskuje m.in. precyzyjną informację o lokalizacji poszukiwanego fragmentu tekstu z dokładnością do strony. Dokładne określenie lokalizacji szukanego wyrażenia

w źródle ułatwia wykorzystanie cytatów z korpusu w pracach naukowych i leksykograficznych oraz odszukanie odpowiedniego fragmentu w podstawie przepisywania.[12]

Wbrew pozorom ustalenie zasad wprowadzania identyfikatora strony nie było proste. Jeśli cały tekst poddawany transliteracji miał w oryginale ciągłą paginację, to jedynym problemem były błędy (występujące nierzadko) i ciągi stron nieliczbowanych. W pierwszym wypadku wprowadzane były tylko numery stron zgodne z tymi, które figurują w starodruku.[13] W drugim początkowe strony oznaczano jako *nlb.* (nieliczbowane) z odpowiednim numerem, natomiast stronom nienumerowanym następującym po stronach numerowanych nadawano kolejne numery „odtworzone" i zapisywano je w nawiasach ostrych, np. <154>, <155> itd. Starodruki, w obrębie których występował więcej niż jeden ciąg paginacyjny, podzielono na części i każdą z nich traktowano jako odrębną pozycję bibliograficzną [por. np. G. Boter, *Relacje powszechne*].

Znakowane są także inne elementy struktury dokumentu, które dają użytkownikowi pełniejszą wiedzę o kontekście, w jakim występuje szukane przez niego wyrażenie. Służą temu m.in. następujące typy znaczników:

- znaczniki strony tytułowej i jej poszczególnych elementów (np. drukarnia), znaczniki fragmentów występujących przed tekstem głównym (np. dedykacje), znaczniki fragmentów stanowiących dodatek do tekstu głównego (np. notki marginesowe, żywa pagina, kustosze) itp.;
- oznaczenia pominiętych podczas przepisywania elementów tekstu, takich jak: dłuższe fragmenty obcojęzyczne, wzory matematyczne itp.;
- oznaczenia fragmentów nienależących do tekstu oryginału, czyli wstawek odredakcyjnych pochodzących z wydań późniejszych (od XIX w.) oraz komentarzy wprowadzonych przez osoby przepisujące teksty, takich jak: korekty literówek, oznaczenia wątpliwości co do postaci wyrazu itp.

3.2. Znakowanie językowe

Każdy niepolski segment tekstu został oznakowany znacznikiem informującym o jego przynależności do konkretnego obcego języka.[14] Oznaczenie fragmentów obcojęzycznych było konieczne przede wszystkim ze

---

[12] Większość starodruków stanowiących podstawę tekstów włączonych do korpusu jest dostępna w bibliotekach cyfrowych. Planowane jest dodanie linków do tych materiałów na stronie KorBy.

[13] W korpusie KorBa 2.0 wprowadzimy modyfikację tej zasady: będzie podany zarówno numer oryginalny, jak i skorygowany, co ułatwi odnalezienie cytatu w starodruku.

[14] Fragmenty obcojęzyczne wprowadzaliśmy do korpusu tylko wtedy, kiedy znajdują się w strukturze zdania polskiego lub łączą się z nim w logiczną całość. Dłuższe fragmenty obcojęzyczne były przy przepisywaniu pomijane.

względu na obecność w polskich tekstach XVII- i XVIII-wiecznych wielu wtrętów łacińskich. Oprócz łaciny w korpusie reprezentowane są również następujące języki: arabski, czeski, francuski, grecki, hebrajski, hiszpański, litewski, niemiecki, węgierski i włoski. W paru wypadkach za pomocą jednego znacznika zostały oznakowane całe grupy językowe, takie jak: skandynawska, turecko-tatarska, południowosłowiańska i wschodniosłowiańska.[15]

4. WIERNA TRANSLITERACJA CZY POSTAĆ UWSPÓŁCZEŚNIONA?

Grafia i ortografia tekstów historycznych różnią się od współczesnej pisowni nie tylko sposobem zapisu niektórych głosek czy ich połączeń, lecz także – co jest dla nas szczególnie ważne – znacznie mniejszą konsekwencją i standaryzacją. Ów niższy stopień standaryzacji i będąca jego rezultatem duża liczba wariantywnych zapisów tej samej słowoformy nawet w jednym tekście stanowią ogromną przeszkodę dla automatycznego przetwarzania takich tekstów oraz dla ich szybkiego i skutecznego przeszukiwania. Utrudnieniem są także rozbieżności pomiędzy językiem epok dawniejszych a współczesnym na innych poziomach języka, przede wszystkim na poziomie fleksji oraz leksyki, które powodują, że znaczna część form jest nierozpoznawalna dla narzędzi do automatycznego przetwarzania tekstu. Z tych powodów w korpusach historycznych teksty podlegają zwykle jakiejś formie normalizacji, a zakres ingerencji bywa bardzo różny.

Decyzja o zakresie normalizacji zależy od specyficznych uwarunkowań danego języka i celów, jakie stawiają sobie twórcy korpusu. W większych korpusach pewien poziom standaryzacji jest konieczny, aby możliwe było zastosowanie narzędzi do ich automatycznego przetwarzania (nie jest możliwe ręczne znakowanie dużych korpusów). Jednak trwała ingerencja w tekst powoduje poważne straty danych lingwistycznych. Z tych powodów w opisywanym korpusie postanowiliśmy zastosować dwie warstwy zapisu, powiązane ze sobą: warstwę transliteracji i warstwę do pewnego stopnia normalizowaną. Istnienie tej drugiej warstwy umożliwia stosowanie narzędzi do anotacji i lematyzacji. Jednocześnie zachowujemy i udostępniamy odbiorcom nie tylko wersję uwspółcześnioną, lecz także zapis jak najbliższy oryginałowi.[16]

---

[15] To rozwiązanie zostało zastosowane do języków, które w tekstach z tej epoki trudno jednoznacznie uznać za należące do któregoś ze spokrewnionych ze sobą języków (np. czasem nie da się stwierdzić, czy mamy do czynienia ze słowem duńskim, norweskim czy szwedzkim).

[16] W świecie słowiańskim KorBa jest pierwszym korpusem historycznym, w którym zastosowano takie rozwiązanie. Wśród korpusów słowiańskich oryginalną grafię oddają korpusy historyczne rosyjskie [RNC], natomiast w korpusach historycznych czeskich [CNC] i słowackich [SNC] zastosowano transkrypcję.

Przyjęliśmy zasadę, że normalizacji (nazywanej dalej transkrypcją) podlegają wyłącznie zjawiska ortograficzne, natomiast niezmieniona pozostaje postać brzmieniowa wyrazu – na tyle, na ile możemy ją rekonstruować na podstawie zapisu[17] – a także dawne końcówki fleksyjne[18] oraz historyczne słownictwo. Decyzja ta ma zasadnicze konsekwencje dla dalszego procesu automatycznego przetwarzania tekstów: konieczne było dostosowanie dostępnego analizatora morfologicznego do stanu fleksji polskiej w XVII i XVIII w. (więcej na ten temat zob. w p. 5.1.).

4.1. Transliteracja

Instrukcja transliterowania tekstów do korpusu oparta została na regułach stosowanych do wydań typu A w *Zasadach wydawania tekstów staropolskich* [Górski i in. 1955, 52–63]. Reguły te przystosowała do potrzeb projektu M.B. Majewska [Majewska 2014]. W wypadku wydań i rękopisów z XVII i XVIII w. ortografia została zachowana. Zachowane zostały np. takie cechy oryginalnej ortografii, niezgodne z ortografią współczesną, jak stosowanie liter *ś*, *ź*, *ć* przed literą *i*, zapis *cż* zamiast *cz*, oryginalne użycie liter *y* oraz *i*, a także *o*, *ó* oraz *u*, kreskowane *á* i *é*. Zachowujemy również oryginalną pisownię łączną lub rozdzielną[19] oraz wielkość liter. Skróty pozostawiamy nierozwiązane zgodnie z oryginałem. Ujednolicona została natomiast pisownia liter oraz znaków diakrytycznych mających tę samą funkcję, np. nie zachowujemy różnych wariantów liter *z*, *s*, *r* (por. rys. 1.); litera *ż* zapisywana jest zawsze z kropką, choć w oryginałach bywa zapisywana również jako *ž* lub *z*; jedynie we fragmentach obcojęzycznych dokładnie oddajemy postać znaków diakrytycznych. Ligatury zapisujemy za pomocą dwóch liter (np. *ß* jako *sz*[20]). W tekstach pozyskanych z wydań XIX-, XX- i XXI-wiecznych zachowujemy w całości oryginalną pisownię tych wydań.

W korpusie słoweńskim nie tylko uspójniono pisownię tekstów, lecz także zostały one uwspółcześnione pod względem fleksyjnym, a nawet leksykalnym [por. Erjavec 2015, 13–14].

[17] Pewne odstępstwa od tej reguły wprowadziliśmy dla niektórych zjawisk z pogranicza ortografii i fonetyki. Np. geminaty w wyrazach pochodzenia łacińskiego były upraszczane zgodnie ze współczesną ortografią (np. *massa* → *masa*), pomimo że zapis ten mógł oddawać ich ówczesną wymowę.

[18] Niekiedy ortografia średniopolska uniemożliwia jednoznaczną interpretację fleksyjną. Dotyczy to form dopełniacza liczby mnogiej rzeczowników rodzaju żeńskiego typu *produkcja*, *galanteria*, gdyż nie wiemy dziś, czy zapis *produkcyi*, *galanteryi* należy odczytywać jako [produkcii̯], [galanteri̯i] czy też [produkcyi̯], [galanteryi̯].

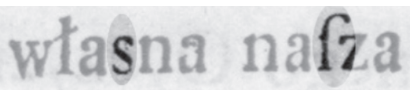
[19] Nie zachowujemy jedynie przypadkowych spacji wewnątrz wyrazu, np. *N áuká* → *Náuká* oraz braku spacji wynikającego z braku miejsca w danej linii.

[20] Dotyczy to tekstu w języku polskim, we fragmentach obcojęzycznych pozostawiamy *ß*.

Rysunek 1. Przykłady transliteracji różnych grafemów jako *r*, *s* i *z*.

dobr

Regestru


własna nasza

názwány

trzy

4.2. Transkrypcja

Transliterowane teksty zostały poddane automatycznej transkrypcji.[21] Głównym celem transkrypcji było sprowadzenie różnych zapisów jednej słowoformy do wspólnej postaci (np. *źiemiá*, *ziemiá*, *zięmiá*, *ziemia* → *ziemia*). Dzięki temu proces automatycznego znakowania morfosyntaktycznego był prostszy i bardziej spójny, a obecnie użytkownicy korpusu mają możliwość wyszukania interesującej ich formy niezależnie od różnych wariantów jej zapisu.

Zgodnie z przyjętymi założeniami transkrybowany tekst ma w warstwie ortograficznej jak najbardziej przypominać tekst pisany współczesnym językiem polskim. Teksty transkrybowane zapisujemy dzisiejszym alfabetem. Użycie znaków diakrytycznych zostało zmodyfikowane zgodnie z obecną ortografią (np. *ktore* → *które*, *śiebie* → *siebie*, *dobrże* → *dobrze*). Występujące w dawnej grafii litery *á*, *é* zastąpiono przez *a* i *e* (np. *bárzo* → *barzo*, *pierwéj* → *pierwej*). Niewystępujące w polskim alfabecie litery *q*, *x*, *v* zostały zamienione na odpowiadające im fonetycznie *k*, *ks*, *u* lub *w* (np. *quadrans* → *kwadrans*, *xięgi* → *księgi*, *vpadnie* → *upadnie*, *conversácye* → *konwersacje*). Użyte niezgodnie ze współczesną normą litery *y*, *i* zostały zmienione na litery *i* lub *j*, np. *iest* → *jest*, *naszey* → *naszej*. Wyrazy zapisane zgodnie z wymową (a wbrew współczesnej normie ortograficznej) doprowadzono do postaci współczesnej, np. *prętko* → *prędko*. Pozostawiono pisownię odbiegającą od współczesnej jedynie

[21] Transkrypcja ręczna nie była przeprowadzana na żadnym etapie prac ze względu na zbyt dużą czasochłonność. Jedynie podczas anotacji ręcznej tzw. korpusu treningowego mogły być dokonywane poprawki w transkrypcji (więcej o korpusie treningowym w p. 5.1.).

w pewnych wypadkach szczególnych (np. pozostawiono literę *q* w nieistniejącym współcześnie wyrazie *Tlaquaciow* z uwagi na niemożność określenia, jak wyglądałaby ta forma zgodnie z obecną normą ortograficzną). Transkrypcji nie podlegają fragmenty w językach obcych.

W projekcie KorBa 1 do transkrypcji automatycznej użyliśmy narzędzia opracowanego pod kierownictwem J.S. Bienia [Bień 2014]. Transkryber ten zamieniał pewne ciągi znaków (odpowiadające fragmentom wyrazów lub całym wyrazom) na inne zgodnie z podanymi regułami. Do projektu KorBa 1 zostały napisane dwa zestawy reguł (dla tekstów oryginalnych i wydań późniejszych), każdy zawierający blisko 4 tys. reguł.[22]

Wielka objętość przetwarzanych tekstów spowodowała, że pomimo tak znacznej liczby reguł nie udało się uwzględnić wszystkich miejsc w tekstach wymagających transkrypcji lub też w niektórych sytuacjach działanie reguł powoduje błędną transkrypcję. I tak np. pewna grupa reguł obejmuje zamianę litery *o* na *ó*, gdyż często w tekstach średniopolskich w miejscu dzisiejszego *ó* wpisywano właśnie *o*. Oczywiście zamianę tę należało przeprowadzić tylko w niektórych kontekstach, więc reguły musiały być w tym wypadku szczegółowe, czasem dotyczące tylko jednej formy, a czasem opatrzone licznymi wyjątkami. W przeważającej większości sytuacji działają one poprawnie, tzn. dobrze rekonstruują literę *ó*, a jednocześnie pozostawiają *o* we właściwych miejscach, np.: *listkow* → *listków*, *zioł* → *ziół*, *krolestwa* → *królestwa*, *zdolność* → *zdolność*, *pozdrowił* → *pozdrowił*, *Spektator* → *Spektator*. Jednak pomimo uwzględnienia wielu kombinacji litery *ó* z innymi literami nie udało się objąć regułami wszystkich sytuacji, w których należało ją zrekonstruować, np. nie została zrekonstruowana litera *ó* w wyrazie *wieczor*. Z drugiej strony w niektórych wyrazach zamiana ta została dokonana niepotrzebnie, np.: *Faktorowie* → *Faktórowie*, *Gorycz* → *Górycz*, *zdroznościom* → *zdróżnościom*.

W ramach projektu KorBa 2 opracowywane jest nowe narzędzie do transkrypcji automatycznej, uczące się reguł zamiany na wzorcowych tekstach, w których dokonano ręcznej korekty transkrypcji.[23]

W korpusie można wyszukiwać segmenty według zadanej transliteracji i transkrypcji jednocześnie, np. zapytanie [`orth="masa" & translit ="massa"`] pozwala odnaleźć wszystkie formy mianownika wyrazu *masa* z podwojonym *s* w pisowni oryginalnej. Wyniki wyszukiwania również dostępne są zarówno w transliteracji, jak i w transkrypcji – widok można zmieniać w każdej chwili za pomocą odpowiedniego przycisku.

---

[22] Ich autorami są Monika Kresa i Emanuel Modrzejewski.

[23] W chwili, gdy ten tekst ukaże się drukiem, na stronie internetowej korpusu będzie już zapewne dostępna wersja w nowej transkrypcji, wykonanej tym narzędziem.

4.3. Segmentacja

W dobie średniopolskiej obowiązywały nieco inne niż dzisiaj zasady pisowni łącznej i rozdzielnej i nie zawsze były stosowane konsekwentnie. Częstokroć spotykamy łączną pisownię przyimków z rzeczownikami, partykuły *nie* z osobowymi formami czasowników itp. Z drugiej strony możliwa jest rozdzielna pisownia przedrostków czasownikowych. Filolog czytający teksty z tej epoki może mimo tych różnic bez trudu wyodrębnić formy gramatyczne wyrazów w postaci zgodnej z przyjętymi obecnie konwencjami opisu językoznawczego, jednak dla narzędzi przeznaczonych do automatycznego przetwarzania języka inna pisownia oznacza zupełnie inną jednostkę.

Jednym z etapów automatycznej analizy tekstów jest ich podział na niepodzielne dalej jednostki, zwane segmentami. Segmenty są później poddawane znakowaniu morfosyntaktycznemu. Podział tekstu na segmenty został w KorBie przeprowadzony zgodnie z zasadami NKJP. Zwykle pojedynczy segment odpowiada wyrazowi pisanemu od spacji do spacji (lub znaku interpunkcyjnego), jednak w niektórych wypadkach ciąg taki dzielony jest na dwa lub więcej segmentów, które następnie podlegają anotacji morfosyntaktycznej i lematyzacji jako odrębne jednostki (np. *robili|by|śmy*). W tekstach średniopolskich ze względu na stosowanie pisowni łącznej (zresztą niekonsekwentne) tam, gdzie dziś stosujemy pisownię rozdzielną, więcej jest wypadków, w których ciąg liter pomiędzy spacjami lub innymi separatorami należy zinterpretować jako dwa lub więcej segmentów np. *nagorze*, *nárok*, *zawiele*, *Nieodiął*. Istnieją też sytuacje odwrotne – pisownia rozdzielna tam, gdzie spodziewalibyśmy się łącznej, np. *przy szyć*, *w chodzi*.

Uwspółcześnienie niezgodnej z dzisiejszą normą pisowni łącznej lub rozdzielnej nie jest możliwe w procesie transkrypcji automatycznej. Podczas automatycznej anotacji morfosyntaktycznej możliwe było podzielenie segmentów typu *nagorze*, *zawiele* (podobnie jak to przebiegało w NKJP i w KorBie w odniesieniu do form czasownikowych typu *robili|by|śmy*). Umożliwiały to reguły segmentacyjne dodane do analizatora morfologicznego na użytek KorBy. Niestety automatyczne odróżnienie przyimka od przedrostka słowotwórczego jest bardzo skomplikowane, co spowodowało niepotrzebne podzielenie wielu form, np. *przy|dáie*, *po|koj*. Zdarzają się też błędy polegające na braku podziału tam, gdzie byłby on potrzebny, np. pozostawienie w korpusie segmentu *będęć* (który powinien zostać podzielony na formę czasownika *być* oraz partykułę *ć*).

Problem automatycznego łączenia segmentów, które powinny stanowić całość (np. *przy sposobią*, *od prawiać*, *w padamy*, *z náyduię*, *od tąd*, *á petyt*, *w nątrze* 'wnętrze'), pozostał na razie nierozwiązany.[24]

---

[24] Zarówno podzielenie jakiegoś ciągu znaków na dwa lub więcej segmentów, jak i połączenie odrębnych segmentów w jeden było natomiast możliwe podczas anotacji ręcznej (na temat podkorpusu anotowanego ręcznie zob. p. 5.1.).

5. ANOTACJA MORFOSYNTAKTYCZNA

5.1. KorBa ręczna i KorBa automatyczna

Cechą wyróżniającą nowoczesne korpusy jest dokładne oznakowanie morfosyntaktyczne poszczególnych segmentów, składające się z informacji o ich formie podstawowej (lemacie) oraz o ich właściwościach gramatycznych (część mowy, wartości kategorii gramatycznych). To właśnie warstwa morfosyntaktyczna tekstu umożliwia przeprowadzanie precyzyjnych analiz lingwistycznych dotyczących np. częstości występowania form poszczególnych leksemów lub konstrukcji składniowych.[25]

Ręczne oznakowanie każdego segmentu korpusu byłoby zadaniem ogromnie czasochłonnym, dlatego jest ono dokonywane automatycznie przez narzędzia zwane tagerami. Aby jednak tager prawidłowo oznakował teksty zgromadzone w korpusie, musi zostać wytrenowany na materiale oznakowanym przez człowieka. W tym celu w ramach projektu KorBa 1 został zestawiony półmilionowy podkorpus, składający się z próbek tekstów włączonych do pełnego korpusu. Każda próbka była następnie znakowana w systemie Anotatornia 2 opracowanym w ramach projektów Chronofleks [Woliński i in. 2017] i KorBa 1. Do systemu były wprowadzane teksty poddane automatycznej transkrypcji, posegmentowane i przeanalizowane za pomocą analizatora morfologicznego Morfeusz 2 [Woliński 2014] wykorzystującego słownik fleksyjny polszczyzny XVII–XVIII-wiecznej o nazwie Korbeusz [Kieraś i in. 2017]. Działanie analizatora polegało na przypisaniu każdemu segmentowi wszystkich możliwych interpretacji fleksyjnych (bez uwzględnienia kontekstu). Podstawowym zadaniem anotatorów było ujednoznacznienie analizy, czyli wybór interpretacji odpowiedniej w danym kontekście. Anotatorzy mieli też możliwość zaproponowania własnej interpretacji w wypadku, gdy żadna z interpretacji podpowiedzianych nie była prawidłowa. Mogli też wprowadzać poprawki w warstwie transkrypcyjnej tekstu, w tym łączyć lub dzielić segmenty, oraz zmieniać podział na zdania. Nie mogli natomiast ingerować w warstwę transliteracyjną tekstu. Oznakowany ręcznie podkorpus został wykorzystany do wytrenowania dwóch tagerów: Concraft [Waszczuk i in. 2018] i Toygger [Krasnowska-Kieraś 2017], za pomocą których oznakowano pełny, 13,5-milionowy korpus.

Użytkownicy KorBy mają dostęp zarówno do podkorpusu oznakowanego ręcznie, jak i do pełnego korpusu oznakowanego automatycznie (w interfejsie wyszukiwarki są one określone odpowiednio jako „KorBa ręczna" i „KorBa automatyczna"). W KorBie automatycznej zawarte są co prawda wszystkie fragmenty znajdujące się w KorBie ręcznej, jednak

[25] Możliwości dokonywania takich analiz z użyciem narzędzia do komputerowego modelowania polskiej fleksji historycznej (Chronofleksu) opisane są w artykule M. Wolińskiego i W. Kierasia w niniejszym tomie [Woliński, Kieraś 2020].

interpretacje tych samych segmentów mogą być różne – w większości wypadków znakowanie dokonywane przez anotatorów jest poprawniejsze, choć tu też zdarzają się błędy. Pełny korpus dostępny jest w dwóch wersjach (jedna oznakowana przez Concraft, druga – przez Toygger), różniących się między sobą pod względem lematyzacji i przypisanych poszczególnym formom znaczników morfosyntaktycznych. Różnice pomiędzy poszczególnymi wersjami korpusu wynikłe z odmiennych sposobów znakowania są rzeczą nieuniknioną. Dostęp do podkorpusu znakowanego ręcznie oraz do dwóch wersji korpusu znakowanego automatycznie rozszerza możliwości korzystania z korpusu. Użytkownicy mogą porównywać trafność interpretacji obu tagerów, a także zestawiać interpretacje obu narzędzi z decyzjami podejmowanymi przez anotatorów.

5.2. Jaki tagset?

Podstawowym problemem lingwistycznym w zakresie anotacji morfosyntaktycznej jest zestawienie odpowiedniego tagsetu, czyli zbioru klas i kategorii gramatycznych dla danego języka w danym momencie rozwoju. Naturalnym punktem odniesienia dla tagsetów w korpusach historycznych jest tagset przyjęty w korpusie języka współczesnego, a zatem w wypadku polszczyzny – w Narodowym Korpusie Języka Polskiego.[26] Dostosowanie tagsetu współczesnego do systemu gramatycznego polszczyzny dawnej stanowi jedno z większych wyzwań stojących przed twórcami korpusu historycznego. Z jednej strony powinni oni starać się zachować możliwie największą zbieżność z tagsetem współczesnym, ułatwia to bowiem proces anotacji (zarówno jeśli chodzi o możliwość zastosowania narzędzi opracowanych z myślą o obsłudze tekstów współczesnych, jak i o usprawnienie pracy anotatorów) oraz późniejsze korzystanie z korpusu. Z drugiej strony – muszą dążyć do precyzyjnego oddania stanu klas i kategorii gramatycznych istniejących w danej epoce. Szczególnych trudności nastręczają tutaj korpusy obejmujące teksty powstałe w długim okresie, w których trzeba uwzględnić zmiany zachodzące w obrębie klas i kategorii gramatycznych na przestrzeni wielu lat (tak jest w wypadku KorBy obejmującej teksty powstałe w ciągu dwóch stuleci).

Stosunkowo łatwą do rozwiązania kwestię stanowi sytuacja, kiedy w opisywanym okresie rozwoju języka istniało zjawisko gramatyczne, które w języku współczesnym zanikło całkowicie. W takim wypadku wystarczy dopisać do tagsetu odpowiednią klasę lub kategorię gramatyczną. Przykładem takiego zjawiska w polszczyźnie XVII–XVIII w. są dawne formy aglutynacyjne czasownika posiłkowego *być*: *-(e)ch*, *-(e)chmy*. Konieczność ich uwzględnienia sprawiła, że tagset KorBy został roz-

[26] Opis tagsetu NKJP, którego założenia stały się również podstawą tagsetu KorBy, znajduje się w: Przepiórkowski i in. 2012.

szerzony o nowy fleksem[27] o nazwie *aglutynant aorystyczny* (agltaor) – analogicznie do klasy *aglutynantów* (aglt), obejmującej współcześnie istniejące formy typu *-(e)m*, *-(e)śmy*.

Większe zmiany w tagsecie są niezbędne w sytuacji, kiedy istniejąca dawniej kategoria gramatyczna obecnie zanikła, ale pozostawiła po sobie pewne relikty. Sposób traktowania takich reliktów w tagsecie współczesnym wynika z ich obecnego statusu gramatycznego, a nie z ich funkcji w dawnym systemie gramatycznym. Natomiast w tagsecie historycznym należy uwzględnić ich dawną funkcję, przypisując im wartości kategorii gramatycznych charakterystycznych dla nich w danej epoce. Dobrą ilustracją tego problemu może być potraktowanie przez twórców korpusów form dawnej niezłożonej odmiany przymiotnika. W tagsecie NKJP dwa reliktowe typy form przymiotnikowych potraktowano jako dwa różne fleksemy: pierwszy to przymiotnik predykatywny (adjc), obejmujący dawne formy niezłożone M. lp. r.m. niektórych przymiotników (np. *zdrów, gotów*); drugi to przymiotnik poprzyimkowy (adjp), obejmujący dawne niezłożone formy D. lp. r.m. i n. (np. *bliska*) oraz C. lp. r.m. i n. (np. *polsku*) zachowane współcześnie tylko w połączeniach z przyimkami (*z bliska*, *po polsku*). Ponieważ w XVII–XVIII w. istniały jeszcze inne formy dawnej niezłożonej odmiany przymiotnika (m.in. B. lp. r.ż., np. *gwałtownę*), w tagsecie KorBy utworzono odrębny fleksem o nazwie *przymiotnik w odmianie niezłożonej* (adjb), do której oczywiście włączono również niezłożone formy istniejące współcześnie.[28] Tym samym w tagsecie KorBy przestały być potrzebne fleksemy adjc i adjp wykorzystywane w NKJP.

Nieco inne problemy rodzi obecność we współczesnej polszczyźnie pozostałości liczby podwójnej, zachowanych w niektórych formach leksemów rzeczownikowych *oko*, *ucho* i *ręka*. Ponieważ współcześnie pełnią one funkcje form liczby pojedynczej lub mnogiej, zazwyczaj wariantywnych w stosunku do form prymarnie używanych w tej funkcji (np. *(w) ręku*, *rękoma*, *(te) ręce*), nie było potrzeby tworzenia dla nich osobnej kategorii w tagsecie NKJP. Natomiast w języku średniopolskim występowały formy liczby podwójnej rzeczowników i czasowników, choć stosowano je już rzadko i niekonsekwentnie (wymiennie z formami liczby mnogiej), zatem w tagsecie KorBy dodano do kategorii liczby wartość dualis (du). Podjęto również decyzję, że wszystkim formom liczby podwójnej bez względu na ich funkcję i składnię należy przypisać wartość du.

---

[27] Pojęcie fleksemu wprowadzone przez J.S. Bienia [Bień 1991] określa zbiór form, które można scharakteryzować za pomocą tych samych kategorii gramatycznych. Na pojęciu tym oparta jest koncepcja klas gramatycznych zastosowana zarówno w NKJP, jak i KorBie.

[28] Analogicznie powiększono zbiór imiesłowów przymiotnikowych o imiesłów przymiotnikowy w odmianie niezłożonej czynny (pactb), np. *będęcy*, i bierny (ppasb), np. *umęczon*.

Prawdopodobnie największym wyzwaniem dla twórcy tagsetu jest konieczność opisania zjawiska gramatycznego, które w dawnej epoce było jeszcze nieustabilizowane, co pociągało za sobą dużą wariantywność form, w których się przejawiało. Kategorie przyjęte do opisu ustabilizowanych form istniejących w języku współczesnym mogą w takim wypadku nie oddawać w pełni prawdziwie struktury gramatycznej języka dawnego. Nieustabilizowanie kategorii zwiększa też stopień niepewności współczesnego użytkownika języka co do postaci pewnych form występujących w języku dawnym. Wszystko to sprawia, że w tagsecie historycznym muszą zostać przyjęte takie rozwiązania, które pozwolą zdać sprawę z istnienia danego zjawiska tylko z takim stopniem precyzji, jaki jest możliwy dla współczesnego badacza języka.

W tagsecie KorBy największe problemy stwarzał opis kategorii rodzaju. Ze względu na zmiany, które zachodziły w rodzaju męskim w okresie średniopolskim, nie można było do jej opisu zastosować wartości przyjętych w tagsecie NKJP. Przypomnijmy, że przyjęto tam podział na trzy podrodzaje męskie (m1 – żywotny osobowy, np. *król*, m2 – żywotny nieosobowy, np. *lew*, m3 – nieżywotny, np. *słup*), wyróżnione na podstawie charakterystycznych układów form synkretycznych w paradygmacie. Tymczasem w XVII i XVIII w. odrębne podrodzaje m1 i m2 dopiero się kształtowały. Przejawiało się to w chwiejnej postaci B. lm. rzeczowników żywotnych (*król*, *lew*), która raz była synkretyczna z M. lm. (*króle*, *lwy*), a innym razem z D. lm. (*królów*, *lwów*). Również w M. lm. rzeczowniki żywotne przybierały czasem końcówki charakterystyczne dla podrodzaju m1, a czasem dla m2 (*królowie* : *króle*,[29] *lwowie* : *lwy*, *wilcy* : *wilki*). Możliwe było także zaburzenie zgody rodzajowej w związkach składniowych, np. rzeczowniki o końcówkach charakterystycznych dla dzisiejszego m2 mogły się łączyć z takimi formami czasowników i przymiotników, które współcześnie są wymagane przez rzeczowniki podrodzaju m1, np. *chłopy* (m2) *wołali* (m1); *oni* (m1) *męże* (m2).

Nieustabilizowanie podrodzajów męskich w języku średniopolskim uniemożliwia jednoznaczne przypisanie określonej wartości rodzaju całemu leksemowi rzeczownikowemu. W KorBie została więc wprowadzona zasada przypisywania odpowiednich podrodzajów męskich poszczególnym formom rzeczownika. W tych przypadkach gramatycznych, które pozwalają szczegółowo określić podrodzaj męski, rzeczownik otrzymuje wartość manim1 (żywotny 1), manim2 (żywotny 2)[30] lub po prostu m (męski), np. M. lm: *lwowie* (manim1): *lwy* (m); B. lm: *królów* (manim1) : *króle* (m). Natomiast w tych przypadkach, w których

[29] Również współcześnie w M. lm. podrodzaju m1 mogą występować dwie formy tego typu. Zjawisko to jest opisane w NKJP jako opozycja między formami niedeprecjatywnymi i deprecjatywnymi.

[30] Celowo zmieniliśmy oznaczenia przyjęte w tagsecie NKJP, tak żeby nie wprowadzać opozycji: osobowy – nieosobowy.

następuje neutralizacja podrodzajów męskich, stosowany jest tzw. rodzaj męski uogólniony (m), np. N. lm.: *królami* (m). Opis dwu form M. lm. typu *królowie* : *króle* w kategoriach różnic między dwoma podrodzajami męskimi pociąga za sobą rezygnację z wyróżniania form deprecjatywnych rzeczownika i wyodrębniania ich w osobne fleksemy (jak to zostało uczynione w NKJP).

Poza wyżej opisanymi modyfikacjami w tagsecie KorBy, wynikającymi z różnic między systemami gramatycznymi polszczyzny współczesnej i XVII–XVIII-wiecznej, w tagsecie tym wprowadzono także pomniejsze zmiany, uszczegóławiające opis zastosowany w NKJP.[31]

W stosunku do zestawu klas gramatycznych wyróżnionych w NKJP zostały zastosowane następujące modyfikacje:

- wyróżnienie dwóch dodatkowych fleksemów liczebnikowych: liczebnika przymiotnikowego (adjnum), np. *dwojaki*, pełniącego funkcje przymiotnika, i liczebnika przysłówkowego (advnum), np. *dwojako*, pełniącego funkcje przysłówka;[32]
- wyróżnienie użyć czasownika *być* w funkcji słowa posiłkowego czasów złożonych: przyszłego (fut), np. *będę* (*jechał*), i zaprzeszłego (plusq), np. *był* (*jechał*);[33]
- powiększenie zbioru imiesłowów przymiotnikowych o imiesłów przeszły (ppraet), np. *osiwiały*.

Repertuar kategorii gramatycznych oraz ich wartości wykazuje następujące zmiany w porównaniu z NKJP:

- rezygnacja z kategorii akomodacyjności form liczebników;
- dodatkowa wartość kategorii aspektu – aspekt podwójny (biasp), przypisywany czasownikom dwuaspektowym (np. *abdykować*) oraz tym, których aspekt jest niemożliwy do ustalenia ze względu na niewystępowanie w korpusie kontekstów diagnostycznych;[34]

---

[31] Pełny zestaw klas gramatycznych i kategorii gramatycznych wraz z przypisanymi im wartościami znajduje się w instrukcji korzystania z korpusu [Gruszczyński, Bronikowska 2018].

[32] Klasy adjnum – adj oraz advnum – adv nie różnią się pod względem formalnogramatycznym. Zostały one wyodrębnione na podstawie semantyczno-słowotwórczej ze względu na tradycję opisów historycznojęzykowych.

[33] Taki sposób znakowania ułatwia wyszukanie w korpusie form czasu przyszłego złożonego czasowników oraz form czasu zaprzeszłego, co może mieć zastosowanie np. w pracach leksykograficznych.

[34] To kolejny przykład konieczności modyfikacji tagsetu ze względu na niewiedzę współczesnych użytkowników języka dotyczącą dawnego systemu gramatycznego. Jeśli w materiale historycznym brak form wskazujących na aspekt czasownika (czyli form czasu przyszłego prostego lub złożonego oraz form imiesłowu przysłówkowego współczesnego lub uprzedniego), innym formom czasownika nie można przypisać wartości aspektu.

- dwie dodatkowe wartości kategorii rodzaju: rodzaj przymnogi osobowy (p1), przypisywany osobowym rzeczownikom *plurale tantum* (np. *królestwo* 'król i królowa', i przymnogi nieosobowy (p2), przypisywany nieosobowym rzeczownikom *plurale tantum* (np. *arcaby*).[35]

6. PERSPEKTYWY

Jak było powiedziane we wstępie, Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. w postaci opisanej w niniejszym artykule został udostępniony w 2018 r., a od 2019 r. prace nad korpusem są kontynuowane w ramach projektu KorBa 2. Rozbudowa korpusu będzie polegała zarówno na powiększeniu jego objętości w granicach chronologicznych zakreślonych dotychczas (1601–1772), jak i na poszerzeniu jego zakresu chronologicznego o lata 1773–1800. Łącznie wielkość korpusu planowana jest na 25 milionów segmentów. W projekcie przewidziane jest także zintegrowanie różnych zasobów językowych polszczyzny obejmujących okres XVII–XVIII w. [Ogrodniczuk, Gruszczyński 2019].

Liczymy na to, że doświadczenie zdobyte w pierwszym etapie prac nad korpusem pomoże udoskonalić postać KorBy 2.0. Będziemy dążyć do większego zrównoważenia korpusu pod względem chronologicznym, geograficznym, gatunkowym i tematycznym. Zastosowanie nowego transkrybera, wykorzystującego sieci neuronowe, pozwoli prawdopodobnie ulepszyć wygląd znormalizowanej warstwy tekstu. Planujemy także wprowadzenie pewnych zmian w tagsecie i w zasadach znakowania morfosyntaktycznego, które umożliwią wierniejsze oddanie struktury gramatycznej tekstu. Część z tych zmian będzie konieczna również ze względu na to, że do KorBy 2.0 włączone będą teksty z początku kolejnego okresu w dziejach języka – epoki nowopolskiej.

Kolejne zmiany w KorBie i ewentualne jej powiększenie będą zapewne powiązane z tworzeniem Narodowego Korpusu Diachronicznego Polszczyzny [Król i in. 2019].

Bibliografia

J.S. Bień, 1991, *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, Warszawa [http://bc.klf.uw.edu.pl/12/; dostęp: 21.05.2020].

J.S. Bień, 2014, *The IMPACT project Polish Ground-Truth texts as a DjVu corpus*, "Cognitive Studies | Études Cognitives" 14, s. 75–84 [https://ispan.

[35] W NKJP osobowym pluraliom tantum został przypisany rodzaj m1, a nieosobowym – rodzaj n.

waw.pl/journals/index.php/cs-ec/article/view/cs.2014.008/174; dostęp: 21.05.2020].

M. Brouwer, H. Brugman, M. Kemps-Snijders, 2017, *MTAS: A Solr/Lucene based multi tier annotation search solution* [w:] L. Borin (red.), *Selected papers from the CLARIN Annual Conference 2016 (Aix-en-Provence, 26–28 October 2016), Linköping Electronic Conference Proceedings* 136, s. 19–37 [http://www.ep.liu.se/ecp/136/002/ecp17136002.pdf; dostęp: 21.05.2020].

T. Erjavec, 2015, *The IMP Historical Slovene Language Resources*, "Language Resources and Evaluation" 49, s. 753–775 [https://doi.org/10.1007/s10579-015-9294-7; dostęp: 21.05.2020].

K. Górski, W. Kuraszkiewicz, F. Pepłowski, S. Saski, W. Taszycki, S. Urbańczyk, S. Wierczyński, J. Woronczak, 1955, *Zasady wydawania tekstów staropolskich. Projekt*, Wrocław.

W. Gruszczyński, R. Bronikowska, 2018, *Instrukcja korzystania z wyszukiwarki do Elektronicznego Korpusu Tekstów Polskich z XVII i XVIII wieku (do 1772 r.)* [https://www.korba.edu.pl/manual; dostęp: 21.05.2020].

C. Hernas, 2002, *Barok*, Warszawa.

W. Kieraś, D. Komosińska, E. Modrzejewski, M. Woliński, 2017, *Morphosyntactic annotation of historical texts. The making of the baroque corpus of Polish* [w:] K. Ekštein, V. Matoušek (red.), *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31*, *Lecture Notes in Computer Science 10415*, s. 308–316.

K. Krasnowska-Kieraś, 2017, *Morphosyntactic disambiguation for Polish with bi-LSTM neural networks* [w:] Z. Vetulani, P. Paroubek (red.), *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, s. 367–371.

M. Król, M. Derwojedowa, R.L. Górski, W. Gruszczyński, K.W. Opaliński, P. Potoniec, M. Woliński, W. Kieraś, M. Eder, 2019, *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" XCIX, z. 1, s. 92–101.

M.B. Majewska, 2014, *Zasady transliteracji źródeł w Korpusie polszczyzny barokowej*, niepublikowana instrukcja wewnętrzna.

M. Ogrodniczuk, W. Gruszczyński, 2019, *Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus* [w:] *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019*, Kuala Lumpur, Malaysia, proceedings, editors: A. Jatowt, A. Maeda, Sue Yeon Syn, LNISA volume 11853, s. 125–138.

A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa [http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf; dostęp: 21.05.2020].

J. Waszczuk, W. Kieraś, M. Woliński, 2018, *Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields* [w:] P. Sojka, A. Horák, I. Kopeček, K. Pala, (red.), *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, Lecture Notes in Artificial Intelligence 11107*, s. 188–196.

M. Woliński, 2014, *Morfeusz reloaded* [w:] N. Calzolari i in. (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, s. 1106–1111.

M. Woliński, W. Kieraś, D. Komosińska, 2017, *Anotatornia 2 – An Annotation Tool Geared towards Historical Corpora* [w:] Z. Vetulani, P. Paroubek (red.), *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, s. 158–162.

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

Cytowane elektroniczne zasoby językowe:

CNC – Czech National Corpus (Český národní korpus) [https://www.korpus.cz/; dostęp: 21.05.2020].

KorBa – Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772 r.) [https://www.korba.edu.pl; dostęp: 21.05.2020].

NKJP – Narodowy Korpus Języka Polskiego [http://nkjp.pl/; dostęp: 21.05.2020].

RNC – Russian National Corpus (Национальный корпус русского языка) [http://www.ruscorpora.ru; dostęp: 21.05.2020].

SNC – Slovak National Corpus (Slovenský národný korpus) [https://korpus.sk; dostęp: 21.05.2020].

Electronic Corpus of 17th- and 18th-century Polish Texts – theoretical and workshop problems

Summary

This paper presents the Electronic Corpus of 17th- and 18th-century Polish Texts (KorBa) – a large (13.5-million), annotated historical corpus available online. Its creation was modelled on the assumptions of the National Corpus of Polish (NKJP), yet the specific nature of the historical material enforced certain modifications of the solutions applied in NKJP, e.g. two forms of text representation (transliteration and transcription) were introduced, the principle of designating foreign-language fragments was adopted, and the tagset was adapted to the description of the grammatical structure of the Middle Polish language. The texts collected in KorBa are diversified in chronological, geographical, stylistic, and thematic terms although, due to e.g. limited access to the material, the postulate of representativeness and sustainability of the corpus was not fully implemented. The work on the corpus was to a large extent automated as a result of using natural language processing tools.

Keywords: electronic text corpus – historical corpus – 17th-18th-century Polish – natural language processing

Trans. Monika Czarnecka

DOI: 10.33896/PorJ.2020.8.4

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

MIKROKORPUS GRONOWY POLSZCZYZNY 1830–1918

1. OGÓLNA CHARAKTERYSTYKA KORPUSU

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

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

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

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

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

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

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

2. ŹRÓDŁA

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

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

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

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

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

3. MIKROSTRUKTURA KORPUSU

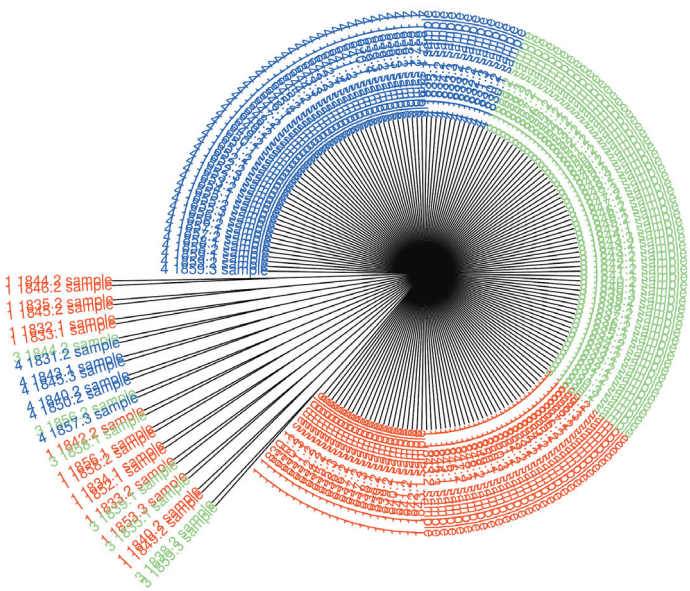
Każda próbka korpusu przechowywana jest w czterech plikach: 1) wybrany fragment tekstu (przykład (4)); 2) metadane, czyli informacja o autorze, tytule, miejscu i czasie wydania itd. (przykład (5)) oraz przynależności do podkorpusu; 3) plik źródłowy, czyli pełny skan dzieła (o ile był dostępny) lub –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

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

4. MAKROSTRUKTURA

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

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

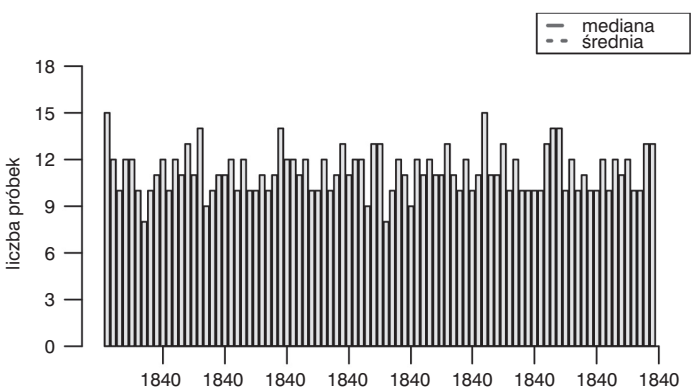


Ze względu na tematykę urywki PN z lat 30. dotyczą pieniądza (jako numizmatu i elementu operacji giełdowych), ponadto dwa teksty – anonimowy i Franciszka Grzymały – dotyczą kwestii politycznych. Próbki z lat 40. pochodzą z dzieł naukowych, w większości poświęconych naukom ścisłym (geometrii, biologii)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

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

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

Rysunek 3. Liczba próbek w poszczególnych latach.



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

5. ZRÓŻNICOWANIE

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

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

---

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

Tabela 1. Liczba próbek według miejsca wydania źródła.

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

6. UDOSTĘPNIENIE I ZASTOSOWANIE

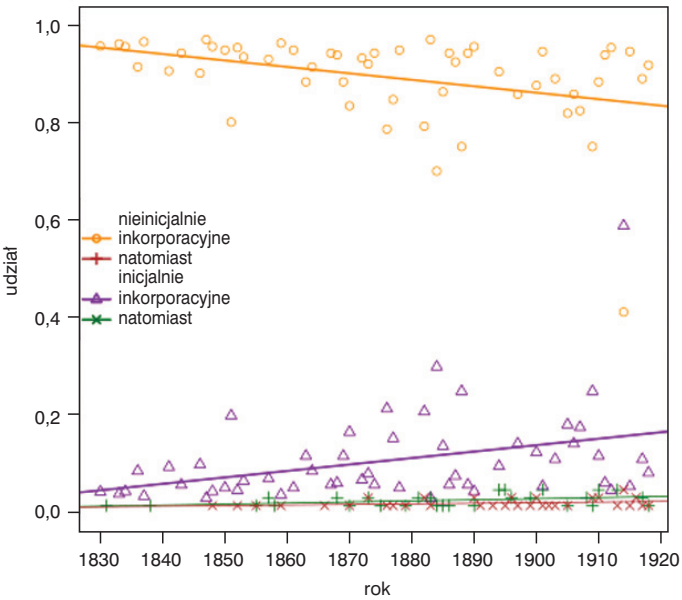
Korpus został udostępniony w postaci tekstowej, jako archiwa źródeł oraz skompilowany do przeglądarki Poliqarp-Marasca [por. Bień 2011; 2012] do użytku z programem off-line oraz jako serwis www pod adresem https://szukajwslownikach.uw.edu.pl/f19/ na licencji umożliwiającej całkowicie swobodne zastosowanie.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

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

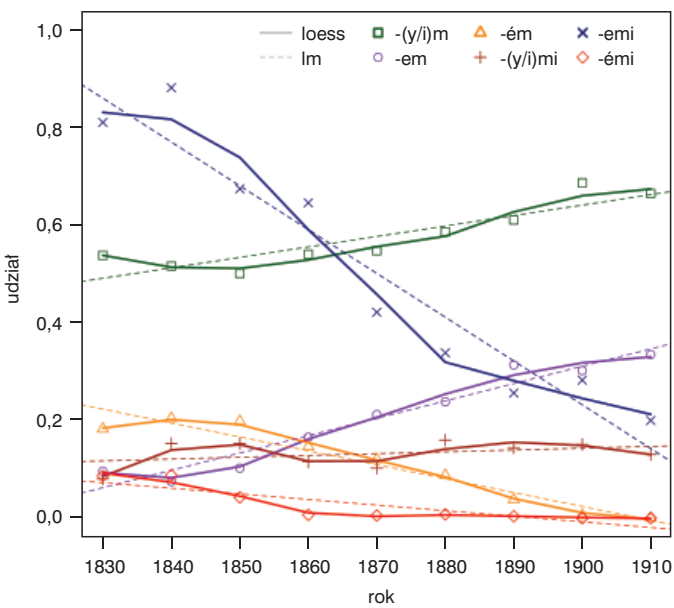
W porównawczym współczesnym korpusie milionowym NKJP [Przepiórkowski i in. 2012, 51, 58) z 508 wystąpień spójników inkorporacyjnych 128, czyli jedna czwarta, zajmuje pozycję inicjalną. Pozycje, które zajmuje NATOMIAST, pozostają w takich samych proporcjach, tzn. z lekką

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

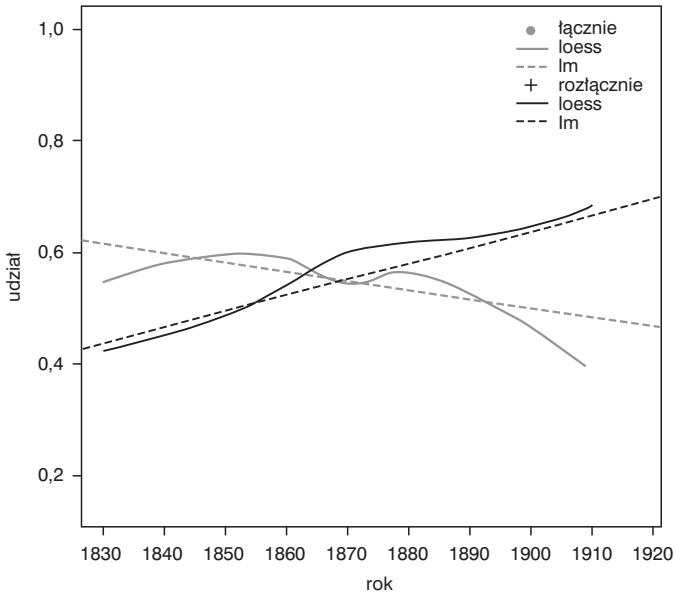
Rysunek 4. Pozycja linearna spójników inkorporacyjnych i NATOMIAST.



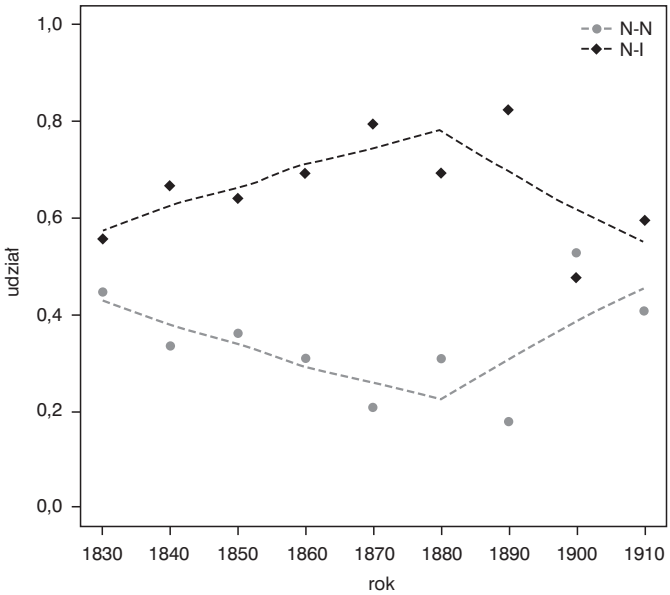
Rysunek 5. Końcówki N. i Msc. rodzaju męskiego i nijakiego przymiotników.



Rysunek 6. Pisownia łączna i rozdzielna Z DALA.



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



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

7. ZMIANY I KIERUNKI ROZWOJU

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

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

Przykłady

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

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

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

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

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

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

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

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

Bibliografia

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

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

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

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

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

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

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

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

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

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

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

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

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

Summary

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

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

Trans. Monika Czarnecka

Marcin Woliński
(e-mail: wolinski@ipipan.waw.pl)
ORCID: 0000-0002-7498-1484
Witold Kieraś
(e-mail: w.kieras@ipipan.waw.pl)
ORCID: 0000-0002-8062-5881
(Instytut Podstaw Informatyki
Polskiej Akademii Nauk, Warszawa)

DOI: 10.33896/PorJ.2020.8.5

ANALIZA FLEKSYJNA TEKSTÓW HISTORYCZNYCH I ZMIENNOŚĆ FLEKSJI POLSKIEJ Z PERSPEKTYWY DANYCH KORPUSOWYCH

Język polski, jak każdy język naturalny, podlega ciągłym zmianom. Historycy języka do niedawna dysponowali wyłącznie wyrywkowymi danymi, na podstawie których formułowali swoje hipotezy badawcze. Obecnie dostępne są korpusy tekstów dawnych, które można badać metodami automatycznymi na dużo większą skalę. Do prowadzenia takich badań niezbędne są wyspecjalizowane narzędzia informatyczne. Opracowanie narzędzi komputerowego modelowania polskiej fleksji historycznej było przedmiotem projektu „Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja" (projekt Narodowego Centrum Nauki nr 2014/15/B/HS2/03119) prowadzonego w IPI PAN w latach 2015–2019.

Głównym wynikiem projektu jest system komputerowy Chronofleks [http://chronofleks.nlp.ipipan.waw.pl/], który oferuje nowe spojrzenie na korpus diachroniczny. System składa się z dwóch części. Pierwsza z nich może być widziana jako prototyp słownika fleksyjnego prezentującego zmienność paradygmatów fleksyjnych poszczególnych leksemów w czasie. Druga część pozwala badać zmienność w czasie frekwencji wybranych form fleksyjnych konkretnych leksemów lub grup leksemów. Jest to pierwsze narzędzie dające łatwy dostęp do informacji ilościowej – możliwe jest badanie nie tylko tego, jakie formy danych leksemów są potwierdzone, ale też ile jest tych potwierdzeń i jak są one rozłożone w czasie. W dalszej części artykułu przedstawimy możliwości badawcze, jakie się dzięki temu otwierają.

Praca z tekstem historycznym różni się istotnie od pracy z tekstem współczesnym, ponieważ dotyczy ona w istocie języka martwego. Nie ma jego żyjących użytkowników, a badacz, zwłaszcza w wypadku bardziej odległych epok, nie może w pełni polegać na swoim wyczuciu językowym. Rekonstrukcja form niepoświadczonych w materiale może prowadzić do powstawania form fikcyjnych, nigdy niewystępujących w realnych tek-

stach. Dlatego w projekcie przyjęto zasadę uwzględniania w modelowaniu wyłącznie poświadczonych form fleksyjnych. Przedstawiane przez system Chronofleks paradygmaty mają potwierdzenie w postaci twardych danych ekstrahowanych ze znakowanych korpusów. Oczywiście w związku z tym prezentowane paradygmaty są kompletne tylko dla najczęstszych wyrazów.

Aby możliwa była taka analiza, konieczne jest odpowiednie opracowanie korpusu: wszystkie słowa muszą zostać poddane analizie morfologicznej, a więc muszą one zostać przypisane do leksemów (zlematyzowane) i scharakteryzowane co do pełnionej funkcji gramatycznej. W pierwszej części artykułu przedstawiono przyjęte zasady takiego opracowania oraz służące do tego narzędzia.

PRZETWARZANIE FLEKSYJNE TEKSTÓW ŚREDNIO- I NOWOPOLSKICH

Analiza morfologiczna i tagowanie stanowią jeden z podstawowych etapów przetwarzania tekstów zarówno w wypadku budowy korpusu, jak i w innych zadaniach NLP. Dla współczesnych języków o bogatej fleksji zagadnienie to jest zwykle rozwiązywane przez zestawienie obszernego słownika wykorzystywanego przez analizator i budowę modelu ujednoznaczniającego [Kobyliński, Kieraś 2016; Przepiórkowski i in. 2012]. Przy zastosowaniu tej metodologii do tekstów historycznych napotyka się problemy powodujące, że jakość znakowania jest dużo gorsza.

Problemy te obejmują: rozchwianie ortograficzne; dawne formy fleksyjne nienotowane przez współczesne słowniki; problem leksykalny – słownictwo nieobecne we współczesnym języku. Przy przetwarzaniu tekstów dawnych mamy problem języka obcego, dla którego brak odpowiednich zasobów. Rozwiązanie problemu jest analogiczne: próbujemy przekształcić zasoby dostępne dla podobnego języka – w tym wypadku współczesnej polszczyzny. Taką właśnie procedurę zastosowano dla historycznych tekstów polskich.

Brak ustabilizowanej ortografii w tekstach dawnych sprawia, że przy ich przetwarzaniu zwykle jako pierwszy krok stosuje się normalizację zapisu (transkrypcję tekstów). Może ona mieć różny zakres. Niektórzy badacze stosują transkrypcję w pełni uwspółcześniającą tekst (również pod względem fleksyjnym, a nawet leksykalnym) – tak postąpili m.in. twórcy korpusu historycznego języka słoweńskiego [Erjavec 2015]. Dzięki temu automatyczne przetwarzanie może się odbyć wyłącznie za pomocą narzędzi dla języka współczesnego, bez konieczności ich modyfikowania. Odwrotne podejście zastosowano w projekcie f19, w którym założono, że przetwarzaniu podlega tekst niepoddany jakiejkolwiek normalizacji [Bilińska i in. 2016]. W prezentowanej pracy przyjęliśmy podejście pośrednie – normalizacji podlega wyłącznie pisownia, nie ingeruje się zaś w zjawiska fleksyjne oryginalnego tekstu. Pozwala to na wykorzystanie

istniejących rozwiązań technicznych po uprzednim dostosowaniu ich do specyfiki tekstów dawnych, a jednocześnie nie pozbawia możliwości badania zmian fleksyjnych.

Analizę fleksyjną tekstów wykonano za pomocą analizatora Morfeusz [Woliński 2014] z odpowiednio zmienionym słownikiem – przystosowanym do danej epoki. Podstawowym źródłem danych do analizy fleksyjnej współczesnej polszczyzny jest *Słownik gramatyczny języka polskiego* [Saloni i in. 2015]. W wypadku polszczyzny dawniejszej dane SGJP są bardzo pomocne, ale niezbędna jest dodatkowa lista form niewystępujących w tekstach współczesnych. Źródłem takich danych może być inny słownik lub odpowiednia procedura modyfikująca („postarzająca") dane SGJP. Do stworzenia Korbeusza, czyli słownika fleksyjnego Morfeusza dla tekstów z XVII i XVIII wieku, wykorzystano obie metody [Kieraś i in. 2017], choć ogromna większość danych została uzyskana za pomocą drugiej z nich. Ważnym powodem tego jest fakt, że trzon danych SGJP stanowi siatka haseł SJPDor., którego zasób leksykalny sięga ostatniej ćwierci XVIII wieku, dzięki czemu słownik ten zawiera dużo słownictwa dawnego i przestarzałego, które w tekstach z XVII, XVIII czy XIX w. jest jeszcze w ogólnym użyciu.

Modyfikacja danych SGJP polegała przede wszystkim na dostosowaniu tych danych do tagsetu przyjętego w korpusie KorBa. Na tym etapie z form bazowych SGJP tworzone są także automatycznie niektóre dawne regularne formy fleksyjne, nieobecne w paradygmatach prezentowanych w SGJP, np. dla długiej serii rzeczowników żeńskich typu KONSTYTUCJA utworzono dawną formę dopełniacza, celownika i miejscownika zakończoną na *-ej*: *konstytucjej*.

Drugim źródłem danych fleksyjnych dla Korbeusza jest informacja fleksyjna pochodząca z *Elektronicznego słownika języka polskiego XVII i XVIII wieku* (e-SXVII), w którym notowano w poszczególnych hasłach formy zaświadczone w kanonie tekstów słownika. Znaczy to, że paradygmaty fleksyjne w e-SXVII są właściwie zawsze niekompletne, a bardzo często hasła są w postaci zalążkowej i zawierają jedynie formy hasłowe. W efekcie na dane fleksyjne e-SXVII składa się zaledwie ok. 76 tys. form fleksyjnych, choć słownik zawiera prawie 39 tys. haseł.

Dla korpusu tekstów z lat 1830–1918 powstał osobny analizator fleksyjny [Kieraś, Woliński 2018]. Procedura jego tworzenia jest właściwie uproszczoną wersją procedury tworzenia Korbeusza. Dla XIX wieku nie istnieje żaden dodatkowy zasób danych fleksyjnych, zatem analizator XIX-wieczny korzysta wyłącznie ze zmodyfikowanych danych SGJP. Zasób leksykalny SGJP (czyli w praktyce SJPDor.) okazał się do tego celu wystarczający, ponieważ analizator uzyskuje niemal równie dobre pokrycie tekstowe jak współczesny Morfeusz (odsetek form nierozpoznanych przez analizator XIX-wieczny dla tekstów z okresu 1830–1918 jest niemal równie mały jak Morfeusza współczesnego dla tekstów współczesnych).

Do automatycznego ujednoznacznienia fleksyjnego tekstów z XVII, XVIII i XIX w. użyto dwóch tagerów statystycznych: Concraft [Waszczuk

i in. 2018] oraz Toygger [Krasnowska-Kieraś 2017]. Z uwagi na większe zróżnicowanie tekstów, mniejszy stopień ustandaryzowania dawnej polszczyzny, a także mniejszą ilość ręcznie znakowanych wzorcowych danych zadanie automatycznego znakowania tekstów dawnych jest istotnie trudniejsze niż analogiczne zadanie dla tekstów współczesnych. Nie dziwi więc, że jakość takiego znakowania jest zauważalnie gorsza. Warto też jednak zauważyć – nie wdając się w techniczne szczegóły ewaluacji tego zadania – że wykwalifikowani anotatorzy z przygotowaniem filologicznym również słabiej sobie radzą ze znakowaniem fleksyjnym tekstów dawnych niż tekstów współczesnych.

USPÓJNIONY SPOSÓB OPISU (TAGSET)

Jak dotąd, procesowi tak rozumianego przetwarzania – czyli transliteracji, transkrypcji, analizy fleksyjnej i ujednoznaczniania – podlegały teksty z trzech okresów: współczesnej polszczyzny z przełomu XX i XXI w. (NKJP), polszczyzny XIX i początku XX w. (1830–1918) oraz polszczyzny XVII i XVIII w. (do 1772 r.). W dużym uproszczeniu można powiedzieć, że mamy tu do czynienia ze zbiorami tekstów reprezentującymi trzy różne epoki rozwoju polszczyzny: średniopolską, nowopolską i współczesną. Jak łatwo się domyślić, wszystkie trzy projekty różniły się nieco zbiorem stosowanych znaczników, zasadami anotacji i różnie podchodziły do szczegółowych rozstrzygnięć napotykanych w procesie przetwarzania tekstów. Aby móc wykorzystać wszystkie zasoby i narzędzia stworzone dla poszczególnych okresów polszczyzny, konieczne jest zniwelowanie przynajmniej najważniejszych różnic, czyli sprowadzenie anotacji fleksyjnej do wspólnego zbioru znaczników (tagsetu). Istotne jest też to, by proces uspójnienia mógł przebiegać automatycznie, ponieważ ręczna praca anotatorów jest bardzo czaso- i kosztochłonna. Z konieczności zatem uspójnienie prowadzi niekiedy do utraty pewnych szczegółowych informacji uwzględnionych w tagsetach dla konkretnych epok.

Tagsety poszczególnych projektów różnią się przede wszystkim systemem opisu rodzaju gramatycznego, zwłaszcza w obrębie podrodzajów męskich. W NKJP przyjęto klasyczny opis W. Mańczaka, w którym rozróżnia się pięć rodzajów, w tym trzy męskie: osobowy (m1), żywotny (m2) i nieżywotny (m3). Dodatkowo za osobne formy w paradygmacie rzeczownika męskiego osobowego uznano [za Salonim 1988 i SGJP] tzw. formy deprecjatywne, czyli formy mianownika i wołacza liczby mnogiej łączące się z przymiotnikami w rodzaju męskim żywotnym. Na mocy konwencji formom deprecjatywnym w tagsecie NKJP przypisano rodzaj m2. W znakowaniu korpusu XIX-wiecznego zasada jest podobna – system rodzajowy polszczyzny w tym okresie był już ustabilizowany i nie wymagał innego potraktowania rodzajów męskich. Zupełnie inaczej jest jednak w wypadku tekstów z XVII i XVIII w., w którym to okresie

kształtowała się kategoria męskoosobowości i co przejawia się przede wszystkim dużym rozchwianiem form rzeczowników męskich w obrębie paradygmatu. W tagsecie tego projektu przyjęto zatem, że notowany jest po prostu jeden rodzaj męski, a dodatkowo oznacza się jego użycia osobowe, żywotne i nieżywotne w formach diagnostycznych dla tych podrodzajów [Gruszczyński, Bronikowska 2018]. A zatem rzeczowniki męskie w formach biernika liczby pojedynczej mogą mieć oznaczenie manim2 lub mnanim (czyli odpowiednio: rodzaj męski w użyciu żywotnym i nieżywotnym), w formach zaś mianownika, biernika i wołacza liczby mnogiej może się pojawić dodatkowo oznaczenie manim1, czyli rodzaj męski w użyciu osobowym. Nie wyróżnia się oczywiście form deprecjatywnych, które – gdyby istniały w tej funkcji – nie różniłyby się od użyć żywotnych.

We wspólnym tagsecie zasadniczo przyjęto rozwiązanie z korpusu tekstów z XVII i XVIII w. – z jednym uproszczeniem. Za domyślny rodzaj męski uznaje się rodzaj nieżywotny. W diagnostycznych dla żywotności i osobowości formach pojawia się zatem wartość m, jeśli użycia są nieżywotne, lub manim1 i manim2 – w wypadku użyć osobowych i żywotnych. Znika zatem wartość mnanim. W danych korpusów epok późniejszych rodzaje m1, m2 i m3 w znacznikach fleksyjnych zostały skonwertowane na wartość m lub w wypadku biernika liczby pojedynczej oraz mianownika, biernika i wołacza liczby mnogiej – na wartości manim1 lub manim2.

Inną przykładową różnicą pomiędzy tagsetami poszczególnych korpusów jest opis form przymiotnikowych tzw. odmiany niezłożonej. W polszczyźnie współczesnej formy te są obecne już tylko szczątkowo, przede wszystkim w mianowniku męskim (*zdrów*, *godzien*), a także w wyrażeniach przyimkowych z celownikiem (*po angielsku*, *po cichu*) i dopełniaczem (*z wolna*, *od dawna*) o funkcji przysłówkowej. W NKJP formy poprzyimkowe zgromadzono wraz z innymi skostniałymi poprzyimkowymi formami przymiotnikowymi pod wspólną etykietą adjp. Z kolei formy mianownikowe (zakończone na spółgłoskę) rozdzielono między zwykłe formy przymiotnikowe i formy predykatywne o etykiecie adjc. W korpusie XIX-wiecznym etykietę adjp ograniczono już tylko do poprzyimkowych form odmiany niezłożonej wraz z określeniem ich przypadka (dopełniacza lub celownika). Nie wyróżniono też oddzielnej klasy adjc, formy mianownikowe odmiany niezłożonej opisano zaś identycznie jak ich odpowiedniki o odmianie złożonej. Zupełnie inny opis zastosowano w korpusie XVII i XVIII w., ponieważ w tym okresie formy odmiany niezłożonej występowały istotnie częściej niż w późniejszych epokach i obejmowały więcej form paradygmatu przymiotnikowego (w tym przede wszystkim biernikowe formy żeńskie, np. *piękną*). W tym tagsecie odmiana złożona oznaczana była oddzielną etykietą adjb uwzględniającą wszystkie kategorie przymiotnika. Analogicznie postąpiono z imiesłowami przymiotnikowymi, których formy odmiany niezłożonej również w tym okresie sporadycznie się pojawiają. Uspójniając tagsety, postanowiliśmy odtworzyć opis dawnych form odmiany złożonej ze szczątkowych

informacji, które można odzyskać z korpusów późniejszych okresów. I tak, formy predykatywne z NKJP uznano za formy mianownika męskiego liczby pojedynczej. Pozostałe formy przymiotnikowe (mianownikowe i biernikowe) zakończone na spółgłoskę uznano odpowiednio za formy męskie mianownika lub biernika liczby pojedynczej. Formy poprzyimkowe z NKJP i korpusu XIX w. uznano odpowiednio za formy dopełniacza lub celownika liczby pojedynczej – historycznie mogły być to formy rodzaju męskiego i nijakiego, ponieważ jednak zwykle nie występują w uzgodnieniu z żadnym rzeczownikiem, na mocy konwencji przypisuje się im rodzaj nijaki (podobną zasadę stosowano w korpusie XVII i XVIII w.). Wreszcie, formy przymiotnikowe żeńskie w bierniku zakończone na *-ę*, występujące niekiedy jeszcze w tekstach XIX-wiecznych, również uznano za formy odmiany niezłożonej.

Nie da się jednak zniwelować skutków niektórych zasadniczych decyzji, które podejmowane są w poszczególnych projektach związanych z korpusami historycznymi konkretnych okresów. I tak, przykładowo, korpus XVII i XVIII w. w swoim schemacie anotacji uwzględnia liczbę podwójną, która co prawda w średniopolszczyźnie była już zdecydowanie rzadka, ale jej obecność w tekstach była wciąż większa niż w wiekach późniejszych, gdy widać ją jedynie w skostniałych formach kilku leksemów. Dodatkowo zdecydowano, że za formy liczby podwójnej uznaje się również te, które zaczęły być używane w funkcji liczby mnogiej i są w użyciu do dziś, czyli przede wszystkim formy rzeczowników OKO, UCHO i RĘKA. W znakowaniu korpusów tekstów późniejszych tej zasady oczywiście nie stosowano i te same formy opisywane były po prostu jako mnogie. W efekcie zasięg liczby podwójnej w wykorzystanych danych sięga roku 1772 i urywa się nagle. Można by oczywiście ograniczyć się do uznania za formy liczby podwójnej tych trzech leksemów tylko tych, które faktycznie są użyte w takim znaczeniu (np. z liczebnikiem), ale nie da się takiej postaci danych uzyskać całkowicie automatycznie. Dlatego trzeba mieć świadomość, że prezentowane w Chronofleksie wyniki są obarczone wieloma szczegółowymi decyzjami, na które autorzy samej aplikacji nie mieli wpływu. Stosuje się to jednak do wszystkich zaawansowanych narzędzi i zasobów lingwistycznych.

Innym przykładem takich szczegółowych decyzji, które mają wpływ na możliwość śledzenia obecności pewnych form fleksyjnych w czasie mogą być zasady transkrypcji. W wypadku tekstów XIX-wiecznych wyrażenia dawniej pisane niekiedy rozdzielnie, a dziś tylko łącznie, np. *po krótce*, *na prędce*, w których obecne są formy miejscownikowe przymiotników KRÓTKI i PRĘDKI odmiany niezłożonej, uwspółcześniano, zapisując je łącznie i oznaczając jako przysłówki (zresztą miały już one w XIX w. wyraźnie przysłówkową funkcję i charakter). W korpusie XVII i XVIII w. tymczasem pozostawiano je w zapisie rozdzielnym i opisywano oba człony oddzielnie: jako przyimek i formę przymiotnikową odmiany niezłożonej. W tekstach współczesnych te wyrażenia są oczywiście zapisywane tylko łącznie i uznawane za przysłówki, a ich związek z przymiotnikami, od których pochodzą, jest już trudny do uchwycenia automatycznie.

SYSTEM CHRONOFLEKS

System Chronofleks został udostępniony w postaci aplikacji webowej pod adresem http://chronofleks.nlp.ipipan.waw.pl/. Ta instalacja systemu została zasilona danymi omawianych wyżej korpusów historycznych, a więc korpusu barokowego, oznaczonego jako KorBa [https://korba.edu.pl/] i korpusu tekstów XIX-wiecznych [http://korpus19.nlp.ipipan.waw.pl/]. Uwzględniono zarówno teksty znakowane ręcznie, jak i teksty znakowane za pomocą tagera automatycznego. Stanowią one osobne podkorpusy, które można uwzględnić lub pominąć w wyświetlanych wynikach (służy do tego przycisk wyboru korpusów pojawiający się w obu częściach systemu Chronofleks, por. rys. 1. i 4.). Dzięki temu użytkownik systemu może użyć w badaniach albo danych mniejszych, opracowanych dokładniej, albo danych większych, ale wtedy musi się liczyć z wpływem na wyniki mniej dokładnych metod automatycznych. Część współczesną używanych danych stanowią teksty jednomilionowego ręcznie znakowanego podkorpusu NKJP [Przepiórkowski i in. 2012]. W miarę pojawiania się innych korpusów zamierzamy dodawać dane do systemu, wzbogacając tym samym uzyskiwane wyniki.

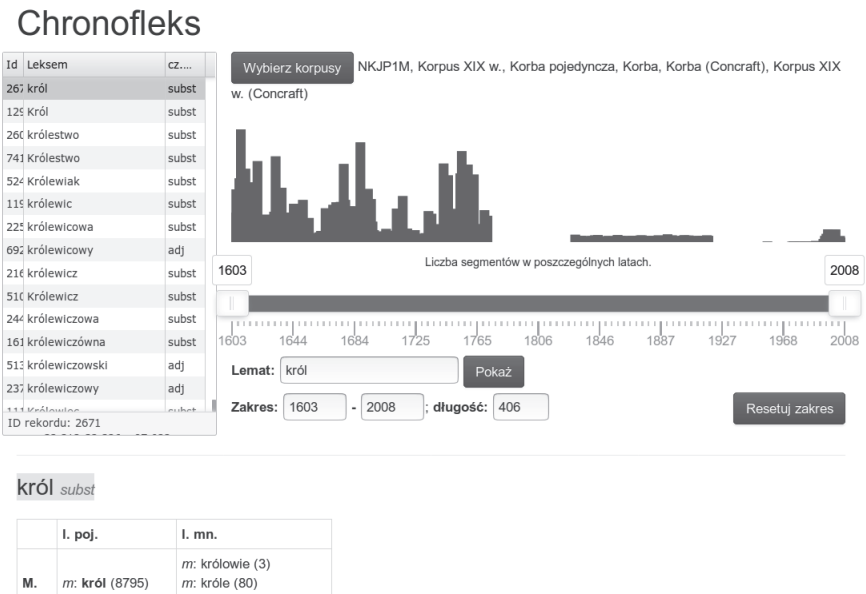
BADANIE ZMIENNOŚCI PARADYGMATÓW FLEKSYJNYCH Z WYKORZYSTANIEM SYSTEMU CHRONOFLEKS

Pierwsza część systemu (dostępna ze strony głównej za pomocą przycisku Paradygmaty) stanowi prototyp przekrojowego słownika fleksyjnego języka polskiego dostępny przez Internet i umożliwiający wizualizację zmian paradygmatów w czasie. Jak wspomniano wcześniej, słownik ten prezentuje wyłącznie formy odmiany poświadczone w zgromadzonych korpusach, podając ich zliczenia, przy czym użytkownik ma możliwość wskazania interesującego go okresu.

Rysunek 1. przedstawia interfejs systemu służący do tego celu. Przewijana lista po lewej stronie rysunku zestawia leksemy poświadczone w przeanalizowanych tekstach. Można z niej wybrać leksem stanowiący przedmiot zainteresowania. Po prawej stronie podano informację o użytych (pod)korpusach. Wykres przedstawia ilość tekstu (liczoną w segmentach) dostępną do analizy w poszczególnych latach (nie jest to wykres dla wybranego leksemu, tylko dla wybranych podkorpusów). Na rys. 1. przedstawiono wykres dla wszystkich dostępnych w systemie korpusów. Widać, że teksty barokowe mają dużo bogatszą reprezentację niż późniejsze. Widać też dwie luki w materiale korpusowym: w zgromadzonych dotychczas korpusach brak tekstów z okresu między rokiem 1773 a 1830 oraz między 1919 a 1956.

Umieszczony poniżej wykresu suwak pozwala wskazać podokres, z którego mają pochodzić wyświetlone niżej formy. Na rysunku 2. przedstawiono trzy migawki z systemu Chronofleks ukazujące formy odmiany

Rysunek 1. Interfejs części systemu Chronofleks prezentującej paradygmaty fleksyjne



leksemu KRÓL z tekstów z XVII, XVIII i XIX wieku. Paradygmat rzeczownika jest prezentowany w postaci tabeli, której kolumny odpowiadają wartości kategorii gramatycznej liczby (jeśli wykryto formy liczby podwójnej, pojawią się one w osobnej kolumnie), a wiersze – kategorii przypadka. W każdej z klatek pokazano różne możliwe wykładniki formy o danej charakterystyce. Przy każdym podano charakterystykę rodzajową (z uwzględnieniem form żywotnych i osobowych rodzaju męskiego) oraz liczbę wystąpień w rozpatrywanym okresie. Pogrubiono wykładnik najczęstszy.

Jeżeli na przykład skupić się na klatce paradygmatu reprezentującej formę narzędnika liczby mnogiej, można zaobserwować, jak zmniejsza się częstość formy *królmi*, która w XVII wieku dominuje nad *królami* (89 wystąpień *królmi* i 38 *królami*), później jednak proporcje się odwracają (w XVIII wieku odnotowano 54 wystąpienia *królami* i 27 *królmi*). Można uznać, że jeszcze w XVIII wieku forma *królmi* jest żywa. Zgromadzone dane nie pozwalają ustalić momentu jej zaniku, bo w tekstach z XIX w. (po roku 1830) jest tylko jedno poświadczenie narzędnika i jest to forma współczesna.

W obliczeniu uwzględniono formy z korpusów znakowanych automatycznie, co nie grozi przekłamaniem dzięki temu, że forma *królmi* nie jest homonimiczna z żadną inną formą fleksyjną.

Rysunek 2. Paradygmat leksemu *król* na podstawie form zanotowanych w tekstach z lat 1601–1700, 1701–1800 oraz 1801–1900

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)

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)

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)

Jeden z problemów, jakie napotykali znakujący korpusy historyczne, ilustruje leksem KOMETA (zob. rys. 3.). W tekstach dawnych leksem ten jest rozchwiany między rodzajem męskim a żeńskim (czasami na formy obu rodzajów natrafia się nawet u tego samego autora). Formy te były, na ile to było możliwe, przypisywane przez znakujących do właściwego rodzaju na podstawie kontekstu, a więc przede wszystkim na podstawie uzgodnień z formami przymiotnikowymi i czasownikowymi. Analogicznych przypisań uczyły się też narzędzia automatyczne. Wyniki obu metod można porównać na rysunku 3.

Rzucającym się w oczy elementem jest obecność kolumny liczby podwójnej w paradygmacie odpowiadającym znakowaniu ręcznemu. Występująca w niej jedna w całym korpusie forma pochodzi z tekstu Mateusza Bembusa *Kometa to jest pogróżka z nieba na postrach, przestrogę i upomnienie ludzkie* z roku 1619:

> *Dwie zasie* ***Komecie****/ które przyniósł rok Pański 1469. okrutnym krwie rozlaniem osobliwie w Niemczech/ Europie groziłu/ jako masz u Kromera.*

Przy znakowaniu automatycznym zjawiska tego nie udało się wychwycić: było ono zbyt rzadkie, żeby odcisnęło się w modelach statystycznych.

Narzędziom automatycznym udało się wyłuskać i poprawnie zinterpretować 8 wystąpień męskoosobowej formy *kometowie* nieobecnej w mniejszych danych znakowanych ręcznie. W wypadku form, które są homonimiczne dla rzeczownika męskiego i żeńskiego (np. *komecie*, *komety*) narzędzia automatyczne przypisywały im rodzaj na podstawie kontekstu. Proporcje form zakwalifikowanych do poszczególnych rodzajów są podobne w danych znakowanych ręcznie i automatycznie, więc można

domniemywać, że również te automatyczne zostały ustalone w większości wypadków poprawnie. Ciekawostkę stanowi forma *komety* interpretowana jako narzędnik. Niestety, chociaż taka jej interpretacja wydaje się możliwa, sprawdzenie cytatów w korpusie ujawnia, że we wszystkich wypadkach narzędzia automatyczne zakwalifikowały ją błędnie. Pokazuje to, że przy formułowaniu wniosków na podstawie znakowania automatycznego należy zachować ostrożność.

Rysunek 3. Paradygmat dwurodzajowego leksemu *kometa* na podstawie danych korpusowych znakowanych ręcznie (po lewej) i automatycznie (po prawej)

kometa *subst*

	l. poj.	l. podw.	l. mn.
M.	*m*: **kometa** (29) *f*: kometa (7)	*f*: **komecie** (1)	*m*: komety (3) *f*: **komety** (6)
D.	*m*: **komety** (15) *f*: komety (7)	-	*m*: komet (4) *f*: **komet** (8)
C.	*m*: **komecie** (1) *f*: **komecie** (1)	-	*m*: **kometom** (1)
B.	*m*: **kometę** (2) *manim2*: kometę (1) *f*: kometę (1)	-	*m*: komety (2) *f*: **komety** (3)
N.	*m*: **kometą** (2)	-	*m*: **kometami** (1)
Msc.	*f*: **komecie** (2)	-	-
W.	-	-	-

kometa *subst*

	l. poj.	l. mn.
M.	*m*: **kometa** (345) *f*: kometa (194)	*m*: komety (50) *manim1*: kometowie (8) *f*: **komety** (117)
D.	*m*: **komety** (325) *f*: komety (62)	*m*: komet (17) *m*: kometów (14) *f*: **komet** (193)
C.	*m*: **komecie** (7) *f*: **komecie** (7)	*m*: **kometom** (5) *f*: kometom (1)
B.	*m*: **kometę** (54) *f*: kometę (19) *f*: kometą (2)	*m*: komety (28) *f*: **komety** (37)
N.	*m*: **kometą** (13) *f*: kometą (10)	*m*: **komety** (18) *m*: kometami (4) *f*: kometami (3)
Msc.	*m*: komecie (9) *f*: **komecie** (27)	*m*: komeciech (1) *f*: **kometach** (37) *f*: komeciech (1)
W.	-	-

BADANIE ZMIAN FREKWENCJI GRUP FORM FLEKSYJNYCH W CZASIE Z WYKORZYSTANIEM SYSTEMU CHRONOFLEKS

Druga część systemu Chronofleks (do której można przejść ze strony głównej http://chronofleks.nlp.ipipan.waw.pl/ za pomocą przycisku Wykresy) przedstawia inne spojrzenie na te same dane. Pozwala ona mianowicie generować wykresy przedstawiające zmienność w czasie frekwencji wybranych grup form fleksyjnych.

Aby przeprowadzić badanie zmian częstości, trzeba zadać trzy elementy odpowiadające ramkom po lewej stronie rysunku 4. Po pierwsze, konieczne jest wskazanie badanego materiału (ramka Ustawienia), a więc wybór używanych podkorpusów i interesującego nas zakresu czasowego. Można także zadać sposób zliczania i grupowania wyników, co omówimy dalej.

Po drugie, należy zadać, jakie formy fleksyjne będą podlegać badaniu (ramka Segmenty). Sposób ich określenia przypomina formułowanie zapytania w wyszukiwarce korpusowej dotyczącego pojedynczego segmentu. W przykładzie pokazanym na rysunku 4. przedmiotem badania są formy o lemacie *król*, którym przypisano w korpusie cechy narzędnika liczby mnogiej. Warunki dookreślające wybierane formy dodaje się za pomocą przycisku opatrzonego symbolem Plus.

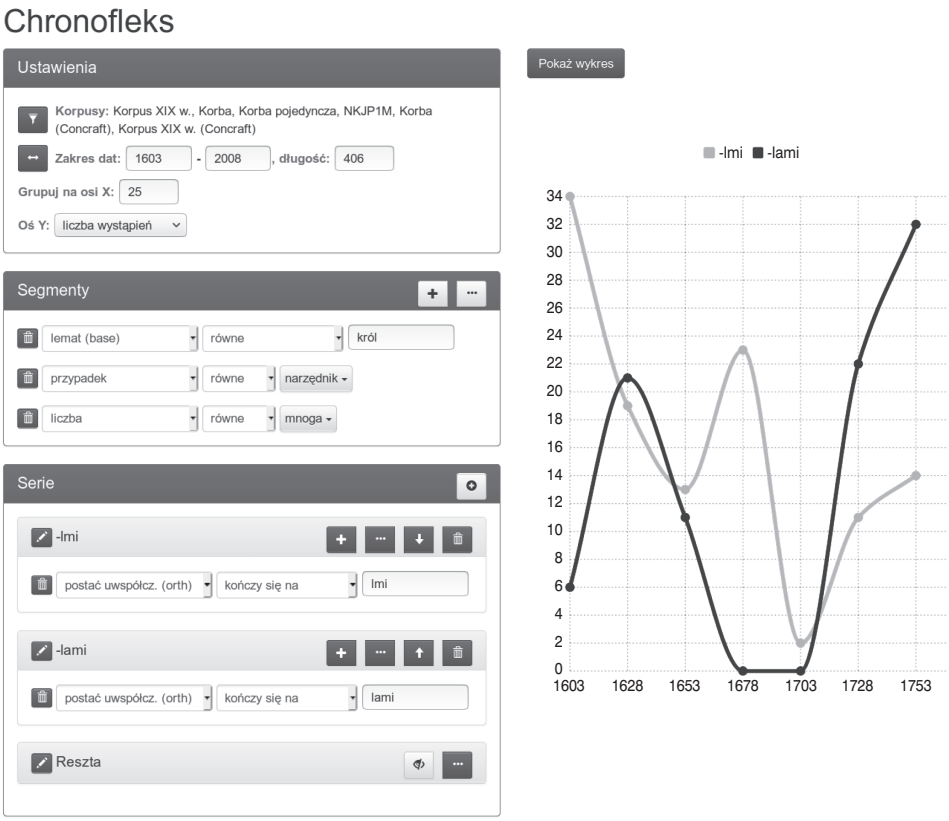
Trzecim elementem konstrukcji badania jest podział wyszukanych form na serie danych (ramka Serie). Każda seria zostanie wyświetlona jako oddzielna linia na wykresie. Określenie serii ma postać warunków podobnych do używanych w części drugiej, ale tym razem wyszukiwanie odbywa się wśród wskazanych uprzednio form (a więc w przykładzie form narzędnika liczby mnogiej leksemu KRÓL). Kolejność określania serii jest istotna, dane wystąpienie formy zostanie zakwalifikowane do pierwszej serii, w której spełnia zadane warunki. Formy niespełniające żadnego z warunków trafiają do predefiniowanej serii Reszta. W szczególności jeżeli nie zada się żadnych warunków w tej części, wszystkie formy znalezione w części drugiej trafiają do serii Reszta i ich frekwencja jest rysowana w postaci pojedynczej linii (por. rys. 5., przedstawiający zmiany frekwencji wszystkich form leksemu KRÓL).

W przykładzie na rysunku 4. zdefiniowano dwie serie danych, rozdzielając formy narzędnikowe na podstawie końcówki. Do serii pierwszej trafiły formy zakończone ciągiem liter *lmi*, a drugiej – *lami*. Serie te obejmują wszystkie wybrane formy, więc domyślna seria Reszta pozostaje pusta – nie pokazano jej na wykresie. Przy prowadzeniu tego rodzaju badań użyteczny bywa przycisk z wielokropkiem znajdujący się przy definicji serii. Pozwala on obejrzeć przykłady form, które trafiły do danej serii. Dzięki temu można sprawdzić, czy zadane warunki działają zgodnie z oczekiwaniami.

Uzyskany wykres przedstawia liczby form o danych zakończeniach w poszczególnych latach. Istotnym parametrem badania jest sposób grupowania danych. W przykładzie przyjęto zbieranie tekstów w grupy po 25 lat. Drobniejsze grupy wprowadzają więcej losowości, związanej z tym, że zgromadzone teksty nie są równo rozłożone w czasie. Grupowanie w dłuższych okresach wygładza wykres, ale zmniejsza jego szczegółowość. W praktyce dobranie najlepszego sposobu grupowania wymaga eksperymentów.

Na osi pionowej wykresu na rysunku 4. zobrazowano zliczenia form. Można więc na przykład odczytać, że w zgromadzonych tekstach z lat 1603–1627 wystąpiły 34 formy *królmi* i 8 form *królami*. Zliczenia te zależą oczywiście od ilości tekstu z danego okresu. Od tego czynnika można się uniezależnić, badając częstości względne form (liczbę wystąpień szukanych form dzieli się przez sumaryczną długość tekstów z danego okresu). W ramce Ustawienia przewidziano także trzeci sposób zliczania. Polega on na podawaniu liczby wystąpień *różnych* form w tekstach. Pozwala to odróżnić zjawiska reprezentowane przez małą liczbę częstych form od zjawisk reprezentowanych przez wiele różnych form.

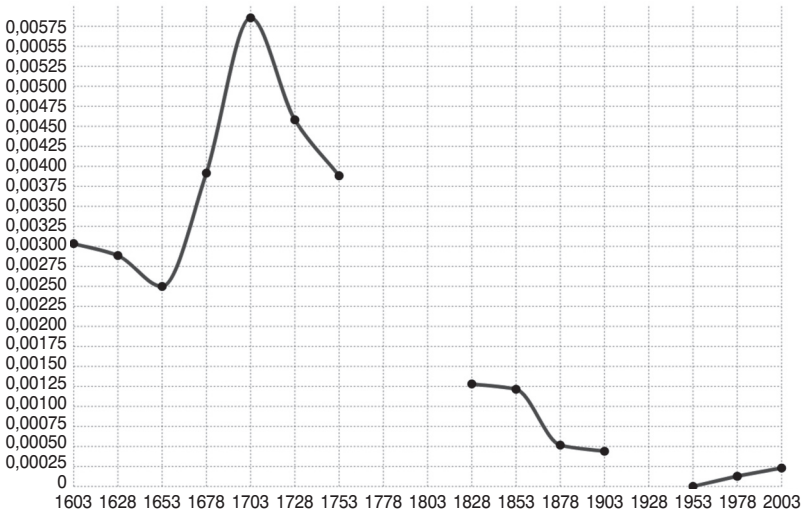
Rysunek 4. Interfejs tworzenia wykresów w systemie Chronofleks



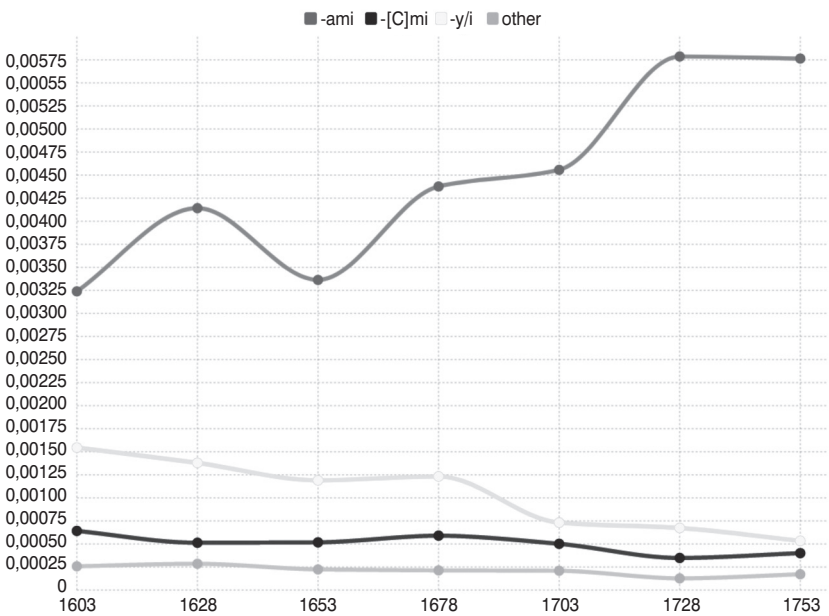
Niezależnie od wyboru omawianych ustawień z wykresu przedstawiającego rozkład w czasie form narzędnikowych leksemu KRÓL nie daje się odczytać użytecznej generalizacji. Form tych jest po prostu za mało. Ciekawszy jest, przedstawiony na rys. 5., wykres częstości względnej dowolnych form leksemu *król* – widoczna jest tendencja zniżkowa frekwencji tego leksemu (co jest zdroworozsądkowe wobec przemian w odzwierciedlanej językowo rzeczywistości).

Prawdziwa siła systemu Chronofleks polega jednak na możliwości zebrania na jednym wykresie form całej grupy leksemów. Na rysunku 6. przedstawiono frekwencje względne form narzędnika liczby mnogiej dla wszystkich rzeczowników męskich i nijakich odnotowanych w korpusie barokowym. Formy te podzielono na cztery serie: o zakończeniu *-ami*, o zakończeniu *-mi* po spółgłosce, o zakończeniu *-y* lub *-i* oraz pozostałe. W badanym okresie widoczna jest bardzo wyraźna tendencja rosnąca końcówki *-ami*, która dominuje współcześnie.

Rysunek 5. Zmienność w czasie frekwencji form leksemu *król*



Rysunek 6. Zmiany frekwencji w tekstach korpusu barokowego form rzeczowników męskonijakich w narzędniku według zakończeń



PODSUMOWANIE

Przedstawiony system umożliwia prowadzenie ciekawych badań fleksji historycznej z perspektywy frekwencyjnej. Jego zaletą jest swobodny dostęp przez Internet, może być on wykorzystywany zarówno w praktyce badawczej, jak i edukacyjnej przez profesjonalistów i amatorów.

Trzeba jednak pamiętać, że wprowadzone do systemu zasoby korpusowe są ograniczone, w związku z czym wyraziste wyniki uzyskuje się jedynie dla badań obejmujących zjawiska odpowiednio częste – dotyczące najczęstszych leksemów lub odpowiednio dużych grup form.

Planowane jest uzupełnianie zasobów korpusowych w miarę ich pojawiania się. Uzupełnienia będą dotyczyły powiększania bazy tekstów z okresów już uwzględnionych, jak i uzupełnienia wspominanych luk w reprezentowanych okresach historycznych [por. Król i in. 2016]. Dostęp do obszerniejszych korpusów znakowanych ręcznie pozwoli też ulepszyć narzędzia znakowania automatycznego, co przełoży się na jakość wyników uzyskiwanych z korpusów znakowanych automatycznie.

Bibliografia

J. Bilińska, M. Derwojedowa, W. Kieraś, M. Kwiecień, 2016, *Mikrokorpus polszczyzny 1830–1918* [w:] Ł. Karpiński, P. Michałowski, *Komunikacja Specjalistyczna* 11, s. 149–161.

W. Gruszczyński, R. Bronikowska, 2018, *Instrukcja korzystania z wyszukiwarki do Elektronicznego Korpusu Tekstów Polskich z XVII i XVIII wieku (do 1772 r.)*, https://www.korba.edu.pl/manual

T. Erjavec, 2015, *The IMP Historical Slovene Language Resources*, „Language Resources and Evaluation" 49(3), s. 753–75; https://doi.org/10.1007/s10579-015-9294-7

W. Kieraś, D. Komosińska, E. Modrzejewski, M. Woliński, 2017, *Morphosyntactic annotation of historical texts. The making of the baroque corpus of Polish* [w:] K. Ekštein, V. Matoušek (red.), *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31*, „Lecture Notes in Computer Science" 10415, s. 308–316.

W. Kieraś, M. Woliński, 2018, *Manually annotated corpus of Polish texts published between 1830 and 1918* [w:] N. Calzolari i in. (red.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, s. 3854–3859.

Ł. Kobyliński, W. Kieraś, 2016, *Part of speech tagging for Polish: State of the art and future perspectives* [w:] *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, Konya.

K. Krasnowska-Kieraś, 2017, *Morphosyntactic disambiguation for Polish with bi-LSTM neural networks* [w:] Z. Vetulani, P. Paroubek (red.), *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, s. 367–371.

M. Król, M. Derwojedowa, R.L. Górski, W. Gruszczyński, K.W. Opaliński, P. Potoniec, M. Woliński, W. Kieraś, M. Eder, 2019, *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" XCIX(1), s. 92–101.
A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.
Z. Saloni, 1988, *O tzw. formach nieosobowych rzeczowników męskoosobowych we współczesnej polszczyźnie*, „Biuletyn Polskiego Towarzystwa Językoznawczego" XLI, Kraków, s. 155–166.
Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, 2015, *Słownik gramatyczny języka polskiego*, wyd. III on-line, Warszawa; http://sgjp.pl
J. Waszczuk, W. Kieraś, M. Woliński, 2018, *Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields* [w:] P. Sojka, A. Horák, I. Kopeček, K. Pala (red.), *Text, Speech, and Dialogue: 21st International Conference, TSD 2018*, Brno, Czech Republic, s. 188–196.
M. Woliński, 2014, *Morfeusz reloaded* [w:] N. Calzolari i in. (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland, s. 1106–1111.

Inflectional analysis of historical texts and variability of Polish inflection from the perspective of corpus data

Summary

The subject matter of this paper is Chronofleks, a computer system (http://chronofleks.nlp.ipipan.waw.pl/) modelling Polish inflection based on a corpus material. The system visualises changes of inflectional paradigms of individual lexemes over time and enables examination of the variability of the frequency of inflected form groups distinguished based on various criteria.

Feeding Chronofleks with corpus data required development of IT tools to ensure an inflectional processing sequence of texts analogous to the ones used for modern language; they comprise a transcriber, a morphological analyser, and a tagger.

The work was performed on data from three historical periods (1601–1772, 1830–1918, and modern ones) elaborated in independent projects. Therefore, finding a common manner of describing data from the individual periods was a significant element of the work.

Keywords: electronic text corpus – natural language processing – inflection of Polish – history of language

Trans. Monika Czarnecka

DOI: 10.33896/PorJ.2020.8.6

Agnieszka Grażul-Luft
(Mazowiecka Uczelnia Publiczna, Płock,
e-mail: a_grazul@wp.pl)
ORCID: 0000-0001-9129-3002

O EWOLUCJI ZNACZENIOWEJ WYRAZU *ZDALNY* W DOBIE PANDEMII KORONAWIRUSA

Konieczność ograniczenia kontaktów bezpośrednich w związku z rozprzestrzenianiem się koronawirusa SARS-CoV-2 pociągnęła za sobą zmiany w sposobie funkcjonowania zarówno w przestrzeni publicznej, jak i prywatnej. I choć do tej pory komunikacja międzyludzka i tak w dużej mierze odbywała się z wykorzystaniem urządzeń elektronicznych, to jednak pewne obszary działania pozostawały poza jej istotnym wpływem. Sytuacja, w której właściwie wszystkie możliwe aktywności trzeba podejmować, unikając bezpośredniego kontaktu z drugim człowiekiem, znalazła swoje przełożenie na popularność przymiotnika *zdalny*. Wymuszona izolacja sięgnęła tak wielu sfer ludzkiego działania, że nową specyfikę zaczęto ujmować wyrażeniem *zdalne życie*[1] [www 1]. Przymiotnik *zdalny* pod koniec marca 2020 roku osiągnął na tyle dużą frekwencję, że przez trzy dni znajdował się wśród słów dnia [www 2].

WYRAZ *ZDALNY* W DEFINICJACH SŁOWNIKOWYCH

W *Słowniku języka polskiego* pod red. W. Doroszewskiego wyraz *zdalny* definiowany jest w następujący sposób:

> *zdalny* – techn., zdalne sterowanie, kierowanie itp. ‘sterowanie, kierowanie różnego typu mechanizmami, urządzeniami, pojazdami z pewnej odległości za pośrednictwem aparatury mechanicznej, elektrycznej, radiowej itp.’ [Doroszewski 1968, t. 10, 956].[2]

---

[1] Wyrażenie pojawiło się między innymi w głównym tytule na okładce tygodnika „Polityka” [2020, nr 13].

[2] W *Słowniku języka polskiego* pod red. M. Szymczaka [Szymczak 1981, t. 3, 988] leksem został zdefiniowany właściwie w ten sam sposób – z dwoma drobnymi różnicami: wyraz został oznaczony jako termin, a ponadto nie zachowano skrótu „itp.”, który w definicji w *Słowniku* pod red. W. Doroszewskiego pojawił się w części definicji: ‘zdalne sterowanie, kierowanie itp.’.

W *Praktycznym słowniku współczesnej polszczyzny* pod red. H. Zgółkowej definicja została rozwinięta poprzez wprowadzenie nowego elementu:

> *zdalny* – techniczny 'kierowany, sterowany z pewnej odległości za pomocą odpowiedniej aparatury elektronicznej, mechanicznej, radiowej itp.; zasilany energią na znaczną odległość za pośrednictwem odpowiedniej aparatury' [Zgółkowa 2004, t. 49, 113].

Z powyższego opisu znaczenia wynika, że wynikiem procesu związanego z działaniem zdalnym niekoniecznie musi być sterowanie, kierowanie czymś lub kimś (brak wskazania, że obiektem sterowania jest urządzenie czy mechanizm), czyli w konsekwencji doprowadzenie kogoś lub czegoś do jakiegoś rodzaju aktywności. Jak wynika z drugiej części definicji, z działaniem zdalnym mielibyśmy do czynienia również wtedy, gdy końcowym skutkiem procesu byłoby jedynie zasilenie (na odległość) obiektu jakimś rodzajem energii, która potencjalnie być może dawałaby mu zdolność wykonania działania, ale aktywność ta wcale nie musiałaby zostać podjęta. Energia właściwie mogłaby np. być tylko zmagazynowana.

W *Słowniku współczesnego języka polskiego* pod red. B. Dunaja *zdalny* to 'taki, który jest kierowany, sterowany z pewnej odległości, za pośrednictwem aparatury elektrycznej, radiowej itp.' [Dunaj 1996, 1350]. Podobnie jak w *Praktycznym słowniku współczesnej polszczyzny*, z definicji wynika, że zdalnie sterowany za pośrednictwem aparatury może być nie tylko mechanizm, ale prawdopodobnie również organizm. Wydaje się, że w obecnym stopniu rozwoju technologicznego jest to możliwe, np. w obszarze badań biologicznych czy medycyny.

W *Innym słowniku języka polskiego* pod red. M. Bańki znaczenie przymiotnika opisano w następujący sposób: 'Zdalne sterowanie urządzeniem odbywa się z daleka, za pomocą wysyłanych z jakiegoś miejsca sygnałów, np. radiowych' [Bańko 2000, t. 2, 1311]. Tym samym w definicji podkreśla się nie tyle fakt istnienia aparatury, dzięki której możliwe jest zdalne sterowanie, ile formy niematerialnej, która daje możliwość dokonania tego rodzaju operacji.

Pewnego rodzaju przełom definicyjny można dostrzec w opisie znaczenia w *Uniwersalnym słowniku języka polskiego* pod red. S. Dubisza: *zdalny* – techn. 'taki, który jest przeprowadzany z pewnej odległości za pomocą aparatury' [Dubisz 2003, t. 4, 940].[3]

Zastąpienie działania związanego ze „sterowaniem" czy „kierowaniem" aktywnością „przeprowadzania" pozwala włączyć w zakres opisu czynności zdalnej zdecydowanie szerszy obszar nowych działań, pojawiających się wraz z rozwojem technologicznym. Tego rodzaju ujęcie trafniej

---

[3] Ten sam opis znajduje się w *Wielkim słowniku języka polskiego* pod red. S. Dubisza [Dubisz 2018, t. 5, 813].

oddaje wszechstronność zdalnych aktywności, obejmując np. płatności przez Internet czy lekcje odbywane w kontakcie synchronicznym online. Definicję można zatem uznać za znacznie bardziej uniwersalną.

W *Wielkim słowniku języka polskiego* pod red. P. Żmigrodzkiego wyodrębnia się dwa znaczenia przymiotnika:
zdalny:

1. dostęp 'odbywający się z użyciem aparatury, która umożliwia wykonywanie określonych czynności w sytuacji, w której wykonawca jest oddalony od miejsca i obiektu tych czynności';
2. komputer, *inform.*, 'taki, który łączy się z jakimś urządzeniem albo systemem informatycznym za pośrednictwem sieci komputerowej' [www 3].

W większości wymienionych słowników języka polskiego w osobnych hasłach umieszczono też definicje przysłówka *zdalnie*. U W. Doroszewskiego wyraz *zdalnie*, opatrzony kwalifikatorem *techn.*, definiowany jest: 'nie bezpośrednio, na odległość (za pomocą odpowiednich urządzeń)' [Doroszewski 1968, t. 10, 956].[4] Definicja w *Praktycznym słowniku współczesnej polszczyzny* właściwie ma podobną zawartość treściową, a omówiona wyżej druga część definicji wyrazu *zdalny* (odnosząca do zasilania energią), nie znalazła analogicznego odpowiednika przy definiowaniu przysłówka: *zdalnie* – techniczny 'kierując, sterując urządzeniami, pojazdami, mechanizmami na odległość, nie bezpośrednio, lecz za pomocą odpowiedniej aparatury' [Zgółkowa 2004, t. 49, 113].

W *Słowniku współczesnego języka polskiego* brak natomiast kwalifikatora „techniczny", mimo iż omówienie znaczenia wydaje się nawet bardziej specjalistyczne: 'o kierowaniu urządzeniami: nie bezpośrednio, lecz z odległości za pomocą fal elektromagnetycznych' [Dunaj 1996, 1350].

W *Innym słowniku języka polskiego* przysłówek nie ma odrębnej definicji, traktowany jest jako podhasło derywowane od *zdalny* i opatrzony jest tylko przykładem użycia (*zdalnie sterowane modele latające*) [Bańko 2000, t. 2, 1311].

W *Wielkim słowniku języka polskiego* pod red. Żmigrodzkiego pojawia się natomiast nowy aspekt technologiczny, związany z wykorzystaniem sieci informatycznej: *zdalnie* – 'dzięki sterowaniu czymś przy użyciu określonych urządzeń i łączności radiowej lub informatycznej' [www 4].

Analiza definicji słownikowych prowadzi do wniosku, że wraz z postępem technicznym (szczególnie w obszarze informatycznym) i zmianami cywilizacyjnymi nastąpiły pewne przesunięcia znaczeniowe, które w większości słowników nie zostały jeszcze ujęte. W opisach często podkreślano aspekt związany ze sterowaniem na odległość przy użyciu pew-

---

[4] Taka sama definicja znajduje się w *Słowniku...* pod red. M. Szymczaka (brak kwalifikatora *techn.*) [Szymczak 1981, t. 3, 988], w *Uniwersalnym słowniku języka polskiego* [Dubisz 2003, t. 4, 940] oraz *Wielkim słowniku języka polskiego* [Dubisz 2018, t. 5, 813] pod red. S. Dubisza.

nych urządzeń czy narzędzi. W tym odniesieniu omawiany przymiotnik i przysłówek najczęściej występował w wyrażeniu *zdalnie sterowany* i *zdalne sterowanie*. Warto podkreślić, że choć to człowiek jest wykonawcą czynności, jaką jest zdalne sterowanie, wyrazy *zdalny*, *zdalnie* definiuje się przede wszystkim w odniesieniu do przedmiotu, obiektu (np. pocisku, samochodu czy komputera). To znaczenie jest nadal aktualne, a z przykładów użycia wynika, że najczęściej odnosi się do tematyki dziecięcych zabawek lub obszaru informatycznego, np.:

> Zabawki zdalnie sterowane to zdalnie sterowane samochody i pojazdy dla dzieci, które znakomicie sprawdzą się jako prezent dla chłopca. Same jeżdżą, biegają lub się toczą – zdalnie sterowane zabawki to prawdziwa frajda dla dzieci w każdym wieku, zarówno niemowlęcym, jak i przedszkolnym czy szkolnym [www 5].

> Windows 10 umożliwia zdalne sterowanie innymi komputerami z tą samą wersją systemu [www 6].

Analiza najnowszych przykładów użyć leksemu *zdalny* prowadzi do wniosku, że odpowiadają one najczęściej definicjom ze słowników pod red. S. Dubisza oraz pierwszemu z zaprezentowanych wyżej opisów znaczeń pochodzących z *Wielkiego słownika języka polskiego* pod red. P. Żmigrodzkiego. Również ta definicja obejmuje szeroki obszar ludzkiej aktywności: nacisk kładzie się nie na samo urządzenie czy przedmiot, tylko kluczowym czyni się to, co niematerialne, samą możliwość dostępu do kogoś lub czegoś mimo dzielącej odległości. Opis znaczenia, w którym najistotniejszym elementem czyni się sam dostęp, staje się bardziej uniwersalny.

Bardziej wszechstronne, w pewnym sensie rozszerzające ujęcia definicyjne, skutecznie pozwalają wychwycić zmianę realiów wynikającą z postępu technologicznego, dzięki któremu człowiek zyskał szereg narzędzi, które może wykorzystać do realizowania różnego rodzaju potrzeb i aktywności na odległość.

Zdalne działania można by podzielić na dwie podstawowe kategorie:

1. z wykorzystaniem narzędzi[5] do wykonania czynności i realizacji określonego celu na odległość bez konieczności bezpośredniego angażowania innych ludzi, np. wykonywanie płatności internetowych, zakup biletów;
2. z wykorzystaniem narzędzi do wykonania czynności i realizacji określonego celu na odległość, pociągające za sobą konieczność zaangażowania innych ludzi, np. zakupy w sklepie internetowym, złożenie wniosku w związku ze staraniem się o wydanie paszportu.

---

[5] Za narzędzia uznaję zarówno to, co materialne (np. komputer, telefon), jak i niematerialne (np. systemy informatyczne). Ludzkie działanie „na odległość" najczęściej odbywa się z wykorzystaniem narzędzi zarówno materialnych, jak i niematerialnych.

Warto jednak zwrócić uwagę, że choć z analizy użyć analizowanych leksemów wynika, że wyrazy *zdalnie* i *zdalny* użytkownicy języka obecnie zwykli sytuować najczęściej w kontekście najnowszych technologii informatycznych, to zdalny dostęp nadal odnosi się również do działań z wykorzystaniem mniej nowatorskich rozwiązań, np. zamówienia pizzy przez telefon czy regulacji głośności przy użyciu pilota do telewizora.

PRZYMIOTNIK *ZDALNY* W „POGOTOWIU MYŚLOWYM"

Z danych Narodowego Korpusu Języka Polskiego wynika, że wyraz *zdalny* nie osiągał przed 2020 rokiem szczególnie dużej frekwencji.[6] Pandemia koronawirusa nie tylko wpłynęła na częstotliwość leksemu, ale również na rozszerzenie zakresu jego użycia.

Punktem wyjścia, wynikającym z sytuacji epidemicznej, była konieczność podejmowania aktywności i wykonywania możliwie największej liczby czynności bez bezpośredniego kontaktu z drugim człowiekiem. Ten wymóg dotknął wielu obszarów, w tym m.in. pracy zawodowej, kształcenia, zakupów, kontaktów z urzędami i instytucjami, a nawet przychodniami zdrowia. Kolejnym krokiem było znalezienie i wykorzystanie wszelkich narzędzi, które umożliwiłyby realizację określonych zadań i kontakt z ludźmi na odległość. Wobec tego nastąpiło pewne przesunięcie, odwrócenie. Wydaje się, że dawniej najpierw pojawiało się jakieś narzędzie, zdolność wykonywania czynności bez bezpośredniego dostępu (np. zdalny dostęp do pulpitu komputera, zdalne sterowanie urządzeniami w przemyśle, e-learning) i w miarę potrzeb było ono wykorzystywane, często alternatywnie wobec formy działania w kontakcie bezpośrednim (np. możliwość załatwienia wybranych spraw w urzędzie). W obliczu pandemii priorytetem stał się brak kontaktu bezpośredniego i do tej potrzeby zaczęto dopasowywać najdogodniejsze i najłatwiej dostępne narzędzia. W tym kontekście wyraz *zdalny* pozwolił trafnie określić tę kluczową konieczność.[7] Rodzaj użytego narzędzia, aparatury pozostał sprawą wtórną i jego dobór niekoniecznie musiał wiązać się z wykorzystaniem nowoczesnych technologii informatycznych, np. *zdalna przychodnia* to taka, w której można otrzymać telefonicznie poradę bądź receptę.

Wobec potrzeby nazwania nowego sposobu funkcjonowania zaczęły powstawać liczne kolokacje z użyciem przymiotnika *zdalny*, które odnosiły zdalność, czyli przede wszystkim konieczność działania bez kontaktu bezpośredniego, do rozmaitych sfer życia. Obok wcześniej funkcjonujących, choć zdecydowanie nie tak popularnych jak obecnie kolokacji:

---

[6] W wyszukiwarce korpusowej PELCRA dla ogólnego zapytania zdaln* pojawia się 1616 wskazań [dostęp 20.04.2020].

[7] Wyraz *zdalny* pochodzi od dawnego (XVI w.) wyrażenia przyimkowego *z dala* 'z daleka' [www 11].

zdalna edukacja, *zdalna praca*, *zdalne zakupy* pojawiły się kolejne, np.: *zdalne przedszkole*, *zdalne lekcje*, *zdalna przychodnia*, *zdalna msza*, *zdalne spotkanie*, *zdalne obrady*, *zdalne wybory*, *zdalne głosowanie*, *zdalna spowiedź*, *zdalny ślub*, *zdalne randki*. Oto wybrane przykłady z tekstów medialnych:

> Stan zagrożenia epidemicznego. Zdalna Msza, zdalna praca, nauka i posłuszeństwo [www 7].

> Zdalne lekcje: czy będą w przedszkolu? Czy przedszkolaki będą miały zdalne nauczanie? [www 8].

> Wideokonferencja ślubna, w której udział biorą przyszli małżonkowie, urzędnik stanu cywilnego i świadkowie, to jednak zbyt mało, by usankcjonować taki zdalny ślub. Dlatego władze stanu Nowy Jork wprowadziły jeszcze jeden wymóg – konieczność przesłania faksem lub w innej, elektronicznej formie, podpisanych dokumentów o zawarciu małżeństwa [www 9].

Pojawiły się też kolokacje, świadczące o rozszerzeniu łączliwości wyrazu *zdalny*, w których wskazuje się uczestnika bądź sprawcę zdalnego działania, np. *zdalny pracownik*, *zdalny uczeń*, *zdalny ninja*:

> Co jest największą obawą przedsiębiorców? Że pracownik zdalny de facto nie pracuje. Obawa nie jest tak do końca nieuzasadniona [www 10].

Tego rodzaju kolokacje to wynik pewnego uproszczenia: np. zdalny uczeń to taki, którego edukacja przebiega w sposób zdalny, który uczy się, nie kontaktując się z nauczycielem bezpośrednio, tylko na odległość, przy użyciu narzędzi. Tego rodzaju kolokacje, osadzone w kontekście, są w pełni zrozumiałe dla odbiorcy.

W opisie funkcjonowania w oddaleniu pomocny okazał się też przysłówek *zdalnie*, np.:

> Wraz z rozprzestrzenianiem się epidemii koronawirusa SARS-CoV-2 firmy zachęcają pracowników do tego, aby wykonywali swoje obowiązki zdalnie [www 12].

> Planowana na koniec kwietnia sesja Rady Warszawy odbędzie się zdalnie [www 13].

> W końcu randkować można również zdalnie, czyli online. Istnieje szereg komunikatorów, które umożliwiają rozmowę lub rozmowę wideo [www 14].

Wzrost popularności przymiotnika *zdalny* tłumaczy funkcjonalność i szeroki zakres znaczeniowy tego określenia. Z analizy tekstów medialnych z marca i kwietnia 2020 roku wynika, że kolokacje z przymiotnikiem *zdalny* zastępuje się najczęściej wyrażeniami z *online* (np. *lekcje online*), wszelkiego rodzaju zestawieniami z *e-*, np. *e-learning*, *e-szkoła*, *e-praca*, *e-nauka*. Zdalne działania opisuje się też przy użyciu wyrażenia *na odległość*. W kontekście pracy stosowane jest również zapożyczenie *home*

office. Właściwie jednak żadne ze wskazanych nie jest tak znaczeniowo pojemne i wszechstronne jak wyraz *zdalny*, który mieści w sobie kluczowe aspekty: działanie na odległość, bez kontaktu bezpośredniego z innymi ludźmi, z użyciem narzędzi, ale bez precyzowania ich rodzaju. Ponadto, w odróżnieniu od wyrazów zapożyczonych, jak również wyrażenia *na odległość*, przymiotnik *zdalny* umożliwia w miarę potrzeb tworzenie z dużą łatwością nowych kolokacji czy złożonych konstrukcji składniowych.

Leksem *zdalny* stał się wyrazem, który, jak pisze Andrzej Markowski o wyrazach modnych, „narzuca się sam", „jest w pogotowiu myślowym" [Markowski 2012, 188]. Mamy tu do czynienia ze zjawiskiem, opisywanym przez Danutę Buttler w kontekście mody językowej, związanym z pewnego rodzaju automatyzmem mówiących, dawaniem przez nich pierwszeństwa elementom szczególnie często słyszanym.[8]

Nowością związaną z użyciem wyrazu *zdalny* jest stosowanie go w kontekstach, w których nie ma powiązania z wykorzystaniem jakiegoś technologicznego urządzenia, aparatury. Istotny staje się tylko komponent znaczeniowy związany z oddaleniem. Tak jest np. w kontekście wyborów korespondencyjnych, np.:

> Dziś Sejm ma rozstrzygnąć, czy w wyborach prezydenckich jedyną metodą oddania głosu będzie przesłanie go w sposób zdalny przez Pocztę Polską [www 15].

> Oprócz tego, aktywni listonosze najzwyczajniej w świecie boją się o swoje zdrowie. Już teraz brakuje im środków ochronnych, a dostarczanie kart wyborczych za potwierdzeniem odbioru najzwyczajniej w świecie mogłoby przyczynić się do roznoszenia wirusa i naraziłoby wielu listonoszy na agresywne zachowania wystraszonych ludzi. Do tego istnieje możliwość, że część aktywnych pracowników zwyczajnie odmówi pracy przy zdalnych wyborach ze względu na poglądy polityczne [www 16].

OD BIEGUNA POZYTYWNEGO DO NEGATYWNEGO

Jak pisze Jadwiga Puzynina,

> wśród licznych cech osobowości człowieka bardzo istotną jest niezbywalna skłonność do wartościowania najróżniejszych elementów otaczającego go „jego świata" [Puzynina 2017, 149].

Wiele wskazuje na to, iż sytuacja związana z pandemią koronawirusa nie pozostanie bez wpływu na nacechowanie emocjonalne analizowanego wyrazu.

Według Jerzego Bartmińskiego

> mówiąc o wartościach ma na myśli to, co w świetle języka i kultury ludzie przyjmują za cenne (...). Chodzi więc o cechy rzeczy, ale też o same rzeczy, w tym też byty poję-

[8] Cyt. za Andrzejem Markowskim [Markowski 2012, 188].

ciowe, postawy, stany i sytuacje, zachowania, funkcjonujące jako „przewodnie idee", będące przedmiotem dążeń, chęci posiadania, motywujące postawy i zachowania ludzi, także inspirujące ich do działania [Bartmiński 2012, 133].

Wydaje się, że wyraz *zdalny*, mimo iż przez lata zaliczany często do słownictwa specjalistycznego, nabrał cech nazw wartości, do których Jadwiga Puzynina zalicza

nazwy cech, stanów, procesów, czynności, postaw, odczuć, które nadawane jak i odbierane są jako mające znaczenia kodowe pozytywne albo negatywne [Puzynina 2017, 154].

Zgodnie z klasyfikacją badaczki wyraz *zdalny* można zaliczyć do nazw opisowo-wartościujących, takich jak np. *szybki*, *wysoki*, *nowoczesny*, *powolny*, którym „towarzyszą wartościowania tylko w jakimś stopniu upowszechnione" (konotacje skonwencjonalizowane) [Puzynina 2017, 154–155]. Leksem *zdalny*, niezależnie od tego, czy stosowany w kontekście np. komputerów, urządzeń przemysłowych, zabawek, pracy czy nauki, kojarzony był z nowoczesnością, rozwojem cywilizacji, postępem, czyli kierował w stronę przewodnich idei, które powszechnie uznawane są za wartościowe. Wobec tego sam analizowany leksem zyskał nacechowanie pozytywne, o czym świadczą przykłady użycia z minionych kilkunastu lat:

Coraz więcej programistów ceni sobie, a wręcz oczekuje dużej niezależności i możliwości pracy zdalnej, zarówno z domu, jak i z dowolnego miejsca na Ziemi. Wiele osób, zastanawia się jak zostać programistą, właśnie dlatego, że w tym zawodzie znaleźć można tak wiele możliwości dobrze płatnej pracy zdalnej (data publikacji: listopad 2018) [www 17].

Zdalny monitoring jest coraz bardziej integralną częścią codziennej pracy w wielu zakładach produkcyjnych, dzięki niższym kosztom i prostemu wdrażaniu, związanym z najnowszymi technologiami. Jednym z podstawowych czynników sprzyjających rozprzestrzenianiu się zdalnego monitorowania jest niższy koszt początkowy inwestycji, kolejna zaleta to skuteczność monitorowania przez mniejszą liczbę operatorów. Zdalny monitoring umożliwia operatorom nie tylko pracę z niemal dowolnego miejsca, ale także bardziej skuteczne monitorowanie dzięki mniejszej liczbie pracowników, co jest istotne w dzisiejszych trudnych warunkach gospodarczych (data publikacji: maj 2013) [www 18].

Możemy nabyć na przykład urządzenie do zdalnej opieki nad naszymi ulubieńcami. Jeśli mamy komputer z odpowiednim programem, z każdego miejsca na ziemi możemy wydzielać kotu karmę, spowodować samooczyszczenie się kuwet i podglądać na ekranie, co akurat robi nasz kot (data publikacji: maj 2008) [www 19].

Zmiana w sposobie funkcjonowania związana z rozprzestrzenianiem się koronawirusa i wynikające z tego faktu zwiększenie frekwencji leksemu *zdalny* nie pozostało bez wpływu na jego aksjologizację. Ze względu na fakt, iż konieczność odosobnienia i podejmowania aktywności bez bez-

pośredniego kontaktu z drugim człowiekiem pojawiła się właściwie z dnia na dzień, zabrakło odpowiedniego przygotowania i wystarczających narzędzi. Poza tym proces działania zdalnego z konieczności wprowadzono również do obszarów w sposób naturalny nie bardzo do tego przeznaczonych, takich jak np. lekcje pływania czy udzielanie porad lekarskich. Wprowadzony obowiązek pracy zdalnej (bez wychodzenia z domu) objął wszystkich pracowników firm i instytucji, łącznie z m.in. kierowcami i konserwatorami obiektów. Dlatego początkowy entuzjazm szybko zaczęło zastępować niezadowolenie i intensywna krytyka, rosnąca nie tylko w związku z nie najlepszymi nastrojami społecznymi w obliczu negatywnych konsekwencji samej pandemii, ale i z działaniami podejmowanymi zdalnie. Przymiotnik *zdalny* zaczął kontekstowo nabierać cech negatywnych. Zjawisko to ilustrują następujące przykłady:

> Praca zdalna może być koszmarem dla niektórych [www 20].
>
> Praca zdalna to raj dla cyberprzestępców [www 21].
>
> Nauka zdalna z problemami. Będzie pomoc: ponad 100 tabletów trafi do uczniów. Ale to wciąż kropla w morzu potrzeb [www 22].
>
> Szanse, aby firma, która nigdy wcześniej nie pracowała zdalnie, teraz zrobiła to dobrze są bardzo małe; zwłaszcza kiedy przechodzi na taki model pracy w trybie pilnym [www 23].
>
> 2000 zł – nawet o tyle może zmniejszyć się pensja nauczyciela, który naucza teraz zdalnie. Dyrektorzy łączą zajęcia i nie wypłacają nadgodzin. Okazuje się, że przepisy dotyczące wynagrodzeń pedagogów za okres pracy zdalnej nie są jasne i każdy interpretuje je inaczej [www 24].

ZDALNY* VS *NORMALNY

Wydaje się, że wyraz *zdalny* ze względu na poziom nacechowania emocjonalnego można zaliczyć do słów sztandarowych, definiowanych przez Walerego Pisarka jako wyrazy, które

> z racji swojej wartości denotacyjnej i konotacyjnej, a zwłaszcza emotywnej, nadają się na sztandary i transparenty, a więc do roli x lub y w strukturach typu *Niech żyje x! Precz z y!*". Są nacechowane dodatnio lub ujemnie i wyrażają lub też wywołują w świadomości partnera dyskursu pojęcia pozytywne bądź negatywne [Pisarek 2017, 539].

Ładunek emocjonalny przymiotnika *zdalny* jest ściśle powiązany z kontekstem: przed pandemią – pozytywny, po wybuchu pandemii – coraz częściej negatywny. Skali konsekwencji tak szeroko stosowanych aktywności zdalnych nie da się jeszcze ocenić, wobec tego nie wiadomo, czy to negatywne wartościowanie utrwali się i pozostanie stałą cechą. Niewykluczone, że jeśli ograniczenie kontaktów bezpośrednich potrwa

długo, to skutki działalności zdalnej mogą przełożyć się na utrwalenie komponentu negatywnie wartościującego. Z analizy użyć kontekstowych wynika, że właściwie już po kilku tygodniach od ogłoszenia pandemii przymiotnik *zdalny* przestał kojarzyć się z nowoczesnością, postępem cywilizacyjnym, a zaczął z czymś gorszej jakości, nieprofesjonalnym, nieprzygotowanym, nienaturalnym. Ponadto wyraz *zdalny* stał się przeciwieństwem wyrazu *normalny*, np.:

> W Polsce nie ma jeszcze kwarantanny jak we Włoszech, nie ma więc przeciwwskazań, by wszyscy normalnie pracowali. Ale ci, którzy mogą, powinni pracować zdalnie – radzi sanepid [www 25].

W dobie pandemii przymiotnik *normalny* przywołuje tęsknotę za spokojem, „lepszymi czasami". Nie wiadomo, czy wyraz *zdalny* utrwali się jako antonim pragmatyczny[9] wyrazu *normalny*, prawdopodobnie jednak na długo będzie budził skojarzenie z okresem nietypowym, trudnym i sytuacją zagrożenia.

Nowa sytuacja społeczna znalazła swoje przełożenie na wzrost kreatywności w wielu obszarach. Sięgnęła też języka, również w zakresie użycia leksemu *zdalny*. Jednym z jej interesujących przejawów jest parafraza „uczeń zdalny, ale leniwy", która pojawiła się w okresie, gdy przedmiotem niesłabnącej dyskusji było zdalne kształcenie.

Słowniki

W. Doroszewski (red.), *Słownik języka polskiego*, t. I–XI, Warszawa 1958–1969.
M. Szymczak (red.), *Słownik języka polskiego*, t. I–III, Warszawa 1979–1981.
H. Zgółkowa (red.), *Praktyczny słownik współczesnej polszczyzny*, t. I–C, Poznań 1994–2005.
B. Dunaj (red.), *Słownik współczesnego języka polskiego*, Warszawa 1996.
M. Bańko (red.), *Inny słownik języka polskiego*, t. I–II, Warszawa 2000.
S. Dubisz (red. nauk.), *Uniwersalny słownik języka polskiego*, t. I–IV (I–VI), Warszawa 2003.
P. Żmigrodzki (red.), *Wielki słownik języka polskiego*, od 2007, Kraków.
S. Dubisz (red. nauk.), *Wielki słownik języka polskiego*, t. I–V, Warszawa 2018.

Bibliografia

J. Bartmiński, 2012, *Językowe podstawy obrazu świata*, Lublin.
A. Markowski, 2012, *Kultura języka polskiego. Teoria. Zagadnienia leksykalne*, Warszawa.

---

[9] Rozumiany jako para wyrazów przeciwstawnych pragmatycznie, używanych dla nazwania zjawisk i obiektów uznawanych tradycyjnie (przede wszystkim kulturowo) za przeciwstawne [Markowski 2012, 120–121].

A. Markowski, 2012, *Wykłady z leksykologii*, Warszawa.
W. Pisarek, 2017, *Pojęcie słów sztandarowych* [w:] J. Bartmiński i in. (red.), *Etyka słowa. Wybór opracowań*, Lublin, s. 539–544.
J. Puzynina, 2017 [w:] G. Grochowski i in. (red.), *Metafora. Tekst. Dyskurs*, Warszawa, s. 149–159.

Źródła internetowe

[www 1] https://www.pressreader.com/poland/polityka/20200325/281479278505165 [dostęp 3.04.20].
[www 2] http://slowadnia.clarin-pl.eu/#/default/2278/true [dostęp 6.04.20].
[www 3] https://www.wsjp.pl/do_druku.php?id_hasla=99572&id_znaczenia=5241457 [dostęp 3.04.20].
[www 4] https://www.wsjp.pl/do_druku.php?id_hasla=99583&id_znaczenia=5241470 [dostęp 3.04.20].
[www 5] https://brykacze.pl/zabawki-zdalnie-sterowane-32 [dostęp 3.04.20].
[www 6] https://www.download.net.pl/jak-zdalnie-sterowac-komputerem-za-pomoca-windows-10/n/8957/ [dostęp 3.04.20].
[www 7] https://www.pch24.pl/stan-zagrozenia-epidemicznego--zdalna-msza--zdalna-praca--nauka-i-posluszenstwo,74591,i.html#ixzz6J2LM2rmq [dostęp 6.04.20].
[www 8] https://www.radiozet.pl/Co-gdzie-kiedy-jak/zdalne-lekcje-czy-beda-w-przedszkolu-czy-przedszkolaki-beda-mialy-zdalne-nauczanie [dostęp 6.04.20].
[www 9] https://cyfrowa.rp.pl/technologie/46147-slub-online-teraz-to-mozliwe [dostęp 22.04.20].
[www 10] https://alebank.pl/praca-zdalna-w-czasie-pandemii-koronawirusa-cztery-najwieksze-bledy-i-jak-ich-uniknac/ [dostęp 22.04.20].
[www 11] https://www.wsjp.pl/do_druku.php?id_hasla=99572&id_znaczenia=0 [dostęp 6.04.20].
[www 12] https://www.computerworld.pl/news/Przestrzegajac-tych-zasad-bedziemy-pewni-ze-zdalna-praca-jest-bezpieczna,419652.html [dostęp 22.04.20].
[www 13] https://tvn24.pl/tvnwarszawa/srodmiescie/koronawirus-w-warszawie-kolejna-sesja-rady-warszawy-odbedzie-sie-zdalnie-4545060 [dostęp 22.04.20].
[www 14] https://nto.pl/randki-w-dobie-koronawirusa-jak-sie-zakochac-podczas-epidemii/ar/c1-14903427 [dostęp 22.04.20].
[www 15] https://www.dziennik.pl/artykul-prasowy/6477759,sondaz-wybory-prezydenckie-koronawirus-covid-19-epidemia-polityka.html [dostęp 22.04.20].
[www 16] https://innpoland.pl/158957,wybory-zdalne-moga-zmusic-poczte-polska-do-zawieszenia-innych-uslug [dostęp 22.04.20].
[www 17] https://www.cyfrowinomadzi.pl/jak-zostac-programista-i-pracowac--zdalnie/ [dostęp 23.04.20].
[www 18] https://www.utrzymanieruchu.pl/zdalne-monitorowanie/ [dostęp 23.04.20].

[www 19] https://www.wsjp.pl/index.php?id_hasla=99572&id_znaczenia=5241457&l=29&ind=0 [dostęp 23.04.20].
[www 20] http://www.eskarock.pl/eska_rock_news/praca_zdalna_moze_byc_koszmarem_dla_niektorych_zabawne_wideo_pokazuje_jak_ludzie_nie_radza_sobie_podczas_kwarantanny/159762 [dostęp 23.04.20].
[www 21] https://zdrowie.wprost.pl/koronawirus/10315758/praca-zdalna-to-raj-dla-cyberprzestepcow-uwaga-na-dane.html [dostęp 23.04.20].
[www 22]
https://gazetakrakowska.pl/krakow-nauka-zdalna-z-problemami-bedzie-pomoc-ponad-100-tabletow-trafi-do-uczniow-ale-to-wciaz-kropla-w-morzu-potrzeb/ar/c1-14921044 [dostęp 23.04.20].
[www 23]
https://alebank.pl/praca-zdalna-w-czasie-pandemii-koronawirusa-cztery-najwieksze-bledy-i-jak-ich-uniknac/ [dostęp 23.04.20].
[www 24]
https://www.asta24.pl/2020/04/17/za-nauke-zdalna-nauczyciele-zarobia-nawet-o-2-tys-zl-mniej [dostęp 23.04.20].
[www 25] https://www.rp.pl/Kadry/303099905-Koronawirus-Pracodawcy--powinni-byc-szczegolnie-czujni-pracownicy-najlepiej-w-pracy-zdalnej.html [dostęp 1.05.2020].

DOI: 10.33896/PorJ.2020.8.7

SPRAWOZDANIE Z KONFERENCJI NAUKOWEJ ICOMP – INTERNATIONAL CONFERENCE ON MULTILINGUAL PRACTICE, UNIWERSYTET SZCZECIŃSKI, SZCZECIN 24–26.02.2020

W dniach 24–26 lutego 2020 roku w Uniwersytecie Szczecińskim odbyła się konferencja naukowa ICOMP – International Conference on Multilingual Practice (w tłumaczeniu polskim: Międzynarodowa Konferencja nt. Praktyk Wielojęzyczności). Konferencję zorganizował zespół językoznawców reprezentujący Instytut Językoznawstwa Uniwersytetu Szczecińskiego pod kierownictwem dr hab. Hanny Pułaczewskiej, prof. Uniwersytetu Szczecińskiego. Poza Hanną Pułaczewską w skład komitetu organizacyjnego konferencji wchodzili dr hab. Agnieszka Stępkowska, prof. US, dr Małgorzata Sokół i dr Witosław Awedyk. W organizacji obrad pomagali studenci studiów norweskich Uniwersytetu Szczecińskiego. Językami konferencyjnymi były angielski, polski oraz niemiecki.

Konferencja zgromadziła 49 czynnych uczestników z 16 krajów, w tym 17 uczestników z dziewięciu instytucji polskich: Uniwersytetu Łódzkiego, Uniwersytetu Śląskiego, Uniwersytetu Kazimierza Wielkiego w Bydgoszczy, Uniwersytetu Mikołaja Kopernika, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Warszawskiego, Uniwersytetu Wrocławskiego, Instytutu Slawistyki Polskiej Akademii Nauk i wreszcie z występującego w roli gospodarza Uniwersytetu Szczecińskiego. Pozostali uczestnicy pochodzili z 12 krajów europejskich, takich jak Austria, Grecja, Hiszpania, Irlandia, Islandia, Luksemburg, Niemcy, Norwegia, Szwajcaria, Węgry, Turcja i Wielka Brytania oraz krajów bardziej odległych, takich jak Kazachstan, Singapur oraz Mongolia – dla trzech uczestniczek z tego kraju, reprezentujących dwie uczelnie wyższe, udział w konferencji oznaczał dwukrotne pokonanie trudów dwudobowej podróży. W konferencji uczestniczyła także w charakterze słuchaczy dziewięcioosobowa grupa studentów germanistyki interkulturowej z Hochschule Fulda pod przewodnictwem profesora Volkera Hinnenkampa.

Kilkunastu uczestników, którzy zapowiedzieli swój udział w konferencji, zrezygnowało z uczestnictwa na krótko przed jej rozpoczęciem ze względu na wzrastające dla osób podróżujących na trasach międzynarodowych zagrożenie pandemią. Był to ostatni moment, kiedy tak liczne i międzynarodowe grono mogło zebrać się na obrady przed ogłoszeniem pandemii przez Światową Organizację Zdrowia w marcu 2020.

Rozpoczynając konferencję, główna organizatorka podkreśliła szczególną rolę wielojęzyczności dla rejonu przygranicznego, w którym znajduje się Szczecin, i potrzebę prowadzenia badań w tej dziedzinie właśnie na Uniwersytecie Szczecińskim, położonym zaledwie kilkanaście kilometrów od granicy z Niemcami. Po czasach nie-

sprzyjających wielokulturowości i wielojęzyczności w dobie PRL akcesja Polski do Unii Europejskiej w roku 2004 i swoboda wyboru miejsca zamieszkania owocowała powstaniem licznych skupisk ludności polskiej tuż za granicą kraju, a w efekcie – specyficznymi dla tej sytuacji praktykami rodzinnej i instytucjonalnej wielojęzyczności. Hanna Pułaczewska stwierdziła, że rozwój nowych form życia społecznego i komunikacji w wyniku przygranicznej migracji został dostrzeżony przez badaczy stosunkowo późno. Zarazem podkreśliła, że intencją organizatorów konferencji było między innymi wzmocnienie i podniesienie rangi badań nad wielojęzycznością na Uniwersytecie Szczecińskim, pozyskanie partnerów krajowych i międzynarodowych do współpracy i wymiany myśli naukowej w tej dziedzinie oraz wzmocnienie rangi uczelni na mapie studiów nad wielojęzycznością w skali kraju, Europy i świata.

Prelegentami plenarnymi byli: koryfeusz badań nad wielojęzycznością François Grosjean – światowej sławy badacz z Uniwersytetu w Neuchatel, który uczestniczył w obradach za pośrednictwem profesjonalnej technologii wideokonferencyjnej, Bernhard Brehmer z Uniwersytetu w Greifswaldzie, znany z projektów i publikacji z zakresu dwujęzyczności polsko-niemieckiej i rosyjsko-niemieckiej w diasporach słowiańskich w Niemczech, oraz Volker Hinnenkamp z Hochschule Fulda, którego dorobek obejmuje liczne badania i publikacje dotyczące wielojęzyczności w kontekście integracji europejskiej, emigracji, tożsamości, etnografii mowy, hybrydyzacji kulturowej i językowej oraz funkcjonowania uczelni wyższej. Niestety ze względów rodzinnych i zdrowotnych w konferencji nie wziął udziału Władysław Miodunka, najwybitniejszy polski prekursor i znawca badań nad wielojęzycznością polsko-obcą, przewodniczący Zespołu Języka Polskiego poza Granicami Kraju przy Radzie Języka Polskiego PAN. Jego wykład na temat dwujęzyczności polsko-obcej w Ameryce Łacińskiej, zaplanowany na ostatni dzień konferencji jako wprowadzenie do obrad sesji polskojęzycznej, zastąpiło odczytanie listu profesora do uczestników i organizatorów konferencji przez Agnieszkę Stępkowską; stanowi on załącznik do niniejszego sprawozdania. W tym czasie odbył się ponadplanowy wykład Hanny Pułaczewskiej w języku angielskim na temat (nie)przekazywania języka odziedziczonego na emigracji jako aspektu rodzicielstwa i wychowania w rodzinie.

Pierwsza część wykładu plenarnego Françoisa Grosjeana koncentrowała się wokół zagadnienia złożoności biografii językowych i związanej z tym zmienności kompetencji językowej osób wielojęzycznych. Druga część wykładu poświęcona była hipotezie istnienia zmiennych stanów neuronalnych osób wielojęzycznych: monolingwalnego, pośredniego i bilingwalnego, odpowiadających zwiększonej bądź zmniejszonej gotowości do używania wielu języków. Volker Hinnenkamp nawiązał do najnowszych tendencji w socjolingwistycznym nurcie badań, ujmującym kompetencję wielojęzyczną jako repertuar wykraczających poza monolingwizm środków wyrazu o wielorakim zastosowaniu, a nie jako sumę kompetencji formalnych. W swoim wykładzie podkreślił subiektywny i płynny charakter praktyk wielojęzycznych w kontekście społecznym i sytuacyjnym. Pojęcie *code switching* jest już niewystarczające dla opisu różnorodnych wielojęzycznych praktyk, dlatego coraz częściej wykorzystywane są pojęcia takie jak *language crossing*, *metrolingualism*, *polylanguaging*, *translanguaging*. Volker Hinnenkamp przedstawił przykłady możliwości opisu wielojęzycznych praktyk na podstawie własnych badań. W ostatnim wykładzie plenarnym Bernhard Brehmer zapre-

zentował najnowsze badania dotyczące ukształtowanego historycznie pogranicza polsko-niemieckiego. Poruszył kwestię asymetrii dostępności nauczania języka niemieckiego w Polsce i języka polskiego w Niemczech w strefie pogranicznej. Przedstawił także projekt pedagogiczny związany z dwujęzycznym kształceniem dzieci. Zwrócił uwagę na specyficzne dla pogranicza zapotrzebowanie na dwujęzyczne usługi publiczne, np. w zakresie ratownictwa medycznego.

Poza wykładami plenarnymi uczestnicy dyskutowali w dwóch równoległych sekcjach. Spotkanie badaczy z wielu krajów, reprezentujących kilka różnych subdyscyplin i nurtów badawczych, stworzyło okazję do wieloaspektowego spojrzenia na wielojęzyczność. Prezentowano badania nad wielojęzycznością z perspektywy neurolingwistyki, psycholingwistyki, socjolingwistyki i pedagogiki. Uczestnicy konferencji mieli okazję wysłuchać 43 referatów w podziale na sześć kręgów tematycznych, takich jak: podtrzymywanie i rewitalizacja języków miejscowych (Maintenance and revival of indigenous languages), języki i osoby przemieszczone (Translocated languages and speakers), neurolingwistyka i psycholingwistyka wielojęzyczności (Neuro- and psycholinguistics of multilingualism), pedagogika języka (Language pedagogy), wielojęzyczność w obszarach specjalistycznych (Multilingualism in specialized domains) oraz wielojęzyczność i gramatyka (Multilingualism and grammar). Największą liczbę referatów zgromadziła sekcja poświęcona językom i osobom przemieszczonym. Zaprezentowane referaty dotyczyły problematyki integracji w środowisku szkolnym oraz relacji między językiem kraju przyjmującego a językiem kraju pochodzenia wśród uchodźców, emigrantów i dzieci z rodzin emigrantów, między innymi w Niemczech (Simone Ploger, Uniwersytet w Hamburgu; Hristo Kyuchukov, Uniwersytet Śląski), w Niemczech i w Izraelu (Galina Putjata, Universitat Münster), w Turcji (niemiecki, Işıl Erduyan, Boğaziçi University) oraz z perspektywy ogólnej (Sabine Ehrhardt z Uniwersytetu w Luxemburgu, wykład przygotowany wspólnie z nieobecną Isabelle Barth z Multilingual Cafe, Erquinghem-Lys we Francji). Język polski poza granicami kraju był tematem referatu Marii Obojskiej z Uniwersytetu w Luxemburgu, poświęconego polskim nastolatkom w Norwegii i ich rodzinom, oraz wystąpienia Krystyny Bojałkowskiej z Uniwersytetu Mikołaja Kopernika w Toruniu, badającej rozwój kompetencji gramatycznej u nastolatków polskiego pochodzenia uczących się polskiego w Niemczech. Joanna Kwiatkowska z Uniwersytetu Johannesa Gutenberga w Mainz przedstawiła analizę oddolnych inicjatyw rodzicielskich w Niemczech w zakresie podtrzymywania języka polskiego u dzieci. Zagadnienie to wiązało się tematycznie z wystąpieniem Hanny Pułaczewskiej, która wskazała na rolę przyjaźni rówieśniczych w Polsce jako kraju pochodzenia w rozwoju kompetencji kulturowej i językowej młodzieży polonijnej w Niemczech. Zagadnienia związane z nauczaniem i przekazywaniem języka polskiego za granicą pojawiły się również w innych wystąpieniach, m.in. Agnieszka Marciszewska z Uniwersytetu w Greenich omawiała kwestie związane z uczeniem się języka trzeciego z wykorzystaniem danych zgromadzonych wśród dzieci z polskich rodzin w Wielkiej Brytanii. Ponadto Urszula Chwesiuk z Uniwersytetu Szczecińskiego przedstawiła wyniki dotychczasowych badań (pod kierownictwem Hanny Pułaczewskiej) nad interferencją z języka angielskiego w polszczyźnie młodych szczecinian, tudzież plany rozszerzenia tychże badań na inne grupy wiekowe. Tym samym zagadnieniem z udziałem dodatkowych języków obcych zajmowała się Zofia Chłopek z Politechniki Śląskiej

na podstawie studium przypadku. Jacek Mianowski z Uniwersytetu im. Adama Mickiewicza w Poznaniu przedstawił wyniki badań na temat naprzemiennego używania języka polskiego i angielskiego przez członków społeczności graczy w rozgrywkach ligowych online. Nader rzadki temat dotyczący dwujęzyczności polsko-obcej w Polsce, z udziałem języka polskiego jako drugiego, podjęła współorganizatorka konferencji Agnieszka Stępkowska w wystąpieniu zatytułowanym *Bilingual couples in Poland: Patterns of communication*, poświęconym empirycznemu badaniu praktyk i ideologii językowych dwujęzycznych par polsko-obcych. Zmianie języka w kontekście historycznym poświęcony był referat Anny Zielińskiej z Instytutu Slawistyki Polskiej Akademii Nauk, która przedstawiła wyniki szeroko zakrojonego projektu na temat międzypokoleniowej utraty języka niemieckiego i polskiego w wyniku drugiej wojny światowej, powojennej homogenizacji kulturowo-językowej Ziem Zachodnich i emigracji z tych ziem do Niemiec. Zagadnienia gramatyczne zostały poruszone w referacie Jolanty Mazurkiewicz-Sokołowskiej z Uniwersytetu Szczecińskiego, która przedstawiła wpływ rodzajów gramatycznych rzeczownika na postrzeganie przedmiotu nazywanego w zależności od używanego języka. Nowatorski temat zaprezentowała Wioleta Piegzik z Uniwersytetu Szczecińskiego, która przedstawiła założenia i metodologię planowanych badań na temat intuicji językowej w języku francuskim Polaków uczących się francuskiego. Tematyka konferencji obejmowała różne języki i ich warianty, np. Steven Robert Adam z Nanyang Technological University przedstawił badania nad cechami gramatycznymi w pisanym akademickim wariancie języka angielskiego używanym w Singapurze. Magdalena Al-Sayadi z Uniwersytetu Kazimierza Wielkiego w Bydgoszczy omówiła zróżnicowanie na język wsi i język miasta na obszarze mówionego języka arabskiego. Aida Amirova z Nazarbayev University przedstawiła problematykę używania języka kazachskiego w sferze akademickiej. Kyriakos Antoniou i George Spanoudis z University of Cyprus zajęli się bidialektalnością w Grecji. Ulziikhishig Tsedendamba i Nyamsuren Buriad z Otgontenger University przedstawiły odległą dla nas tematyką pejzażu lingwistycznego w Ułan Bator.

Nie sposób jest chociażby krótko omówić wszystkie wygłoszone referaty. Podczas konferencji przedstawiono najnowsze kierunki badań nad wielojęzycznością prowadzonych w różnych ośrodkach naukowych na świecie. Konferencja pokazała, że badania nad wielojęzycznością rozwijają się dynamicznie, szczególnie w kontekście badań nad nauczaniem języków odziedziczonych i przekazem języka w rodzinie.

Pełna lista i streszczenia referatów w języku angielskim dostępne są na stronie internetowej konferencji pod adresem http://icomp.usz.edu.pl/programme/

Hanna Pułaczewska
(Uniwersytet Szczeciński,
e-mail: hanna.pulaczewska@usz.edu.pl)
ORCID: 0000-0002-7155-8306

Anna Zielińska
(Instytut Slawistyki Polskiej Akademii Nauk, Warszawa,
e-mail: azielinska@ispan.waw.pl)
ORCID: 0000-0002-0497-4916

List od Profesora Władysława Miodunki
Kraków, 24 lutego 2020 r.

Szan. Pani
dr hab. Hanna Pułaczewska, prof. Uniwersytetu Szczecińskiego
Instytut Filologii Angielskiej
Uniwersytetu Szczecińskiego

Wielce Szanowna Pani Profesor,

trwająca w Uniwersytecie Szczecińskim konferencja na temat praktyki wielojęzyczności jest bardzo ważnym wydarzeniem z kilku powodów, spośród których wymienię tylko kilka. Przede wszystkim jest to pierwsza konferencja w Polsce na temat wielojęzyczności. Polskie językoznawstwo, pochłonięte troską o zachowanie polszczyzny przez ludzi emigrujących z naszego kraju, długo nie dostrzegało faktu, że emigranci ci stają się w swych krajach osiedlenia dwujęzyczni i że duża ich część po kilku latach posługuje się lepiej językiem kraju osiedlenia niż polskim. To powodowało, że lingwiści polscy długo, bo aż do roku 2000, nie podejmowali problematyki dwu- i wielojęzyczności.

Późniejsze lata przyniosły wiele interesujących monografii na temat dwujęzyczności polsko-obcej: w Brazylii (Miodunka 2003), w Argentynie (Guillermo-Sajdak 2015), w Australii (Lipińska 2003; Dębski 2009), we Francji (Wróblewska-Pawlak 2004), w Austrii (Cieszyńska 2006; Kainacher 2007), w Szwecji (Laskowski 2009), w Wielkiej Brytanii (Błasiak 2011; 2019), na Białorusi (Golachowska 2012); na Ukrainie (Dzierżawin 2009; Levchuk 2019); w Niemczech (Besters-Dilger, Dąbrowska, Krajewski, Żurek (red.) 2016; Pułaczewska 2017; Żurek 2018), we Włoszech (Kowalcze-Franiuk 2018), a także w Polsce (Czykwin, Misiejuk 2002; Głuszkowski 2011; Stępkowska 2019). Jak wynika z tego wyliczenia, badania dwu- i wielojęzyczności stają się coraz bardziej popularne. Natomiast trzeba zauważyć, że brakowało nam dotychczas konferencji na ten temat. Brakowało konferencji, które by włączały dorobek polskich badaczy do dorobku światowego, znanego w naszym kraju dość dobrze. Stąd szczególna rola i znaczenie konferencji odbywającej się w Uniwersytecie Szczecińskim: konferencji, która zgromadziła wielu referentów z różnych krajów świata, podejmujących w swych badaniach ciekawe tematy; konferencji, na której referaty plenarne wygłaszają tak znani specjaliści jak François Grosjean, Volker Hinnenkamp i Bernhard Brehmer. **Nie ulega wątpliwości, że jest to konferencja, jakiej dotąd w Polsce nie było.**

Przy Radzie Języka Polskiego, która jest wyspecjalizowanym organem w zakresie polskiej polityki językowej, działa od 2012 r. Zespół Języka Polskiego poza Granicami Kraju, który zajmuje się zachowaniem tam polszczyzny i przemia-

nami, jakim ona podlega, w tym nauczaniem naszego języka w warunkach dwu- i wielojęzyczności. Od roku 2019 aktywnie działa w tym Zespole prof. Hanna Pułaczewska. Jak widać, działa bardzo energicznie i skutecznie, o czym świadczy trwająca w Uniwersytecie Szczecińskim konferencja. Działa tak dobrze, że jestem z Niej dumny.

Bardzo żałuję, że ze względu na problemy zdrowotne w rodzinie nie mogę uczestniczyć w tej konferencji, zapewniam jednak, że prześlę organizatorom swój referat. Kończąc, życzę uczestnikom konferencji wielu interesujących dyskusji, a jej organizatorom satysfakcji z nawiązanych kontaktów oraz tego, żeby Uniwersytet Szczeciński stał się nowym centrum badań wielojęzyczności w Polsce.

Proszę przyjąć, Szanowna Pani Profesor, wyrazy prawdziwego poważania

Przewodniczący Zespołu Języka Polskiego
poza Granicami Kraju przy Radzie Języka Polskiego

prof. dr hab. *Władysław T. Miodunka*

R E C E N Z J E

DOI: 10.33896/PorJ.2020.8.8

JACEK WARCHALA, *FORMY PERSWAZJI*, Wydawnictwo Uniwersytetu Śląskiego, Katowice 2019, ss. 428

Jest to monografia podejmująca problematykę częstokroć prezentowaną w różnych publikacjach, co tłumaczy obszerność bibliografii, zamieszczonej na 22 stronach druku (s. 405–426). Jest to książka, której temat fascynował humanistów przez całe dzieje naszej kultury intelektualnej – poczynając od Platona i Arystotelesa, a na współczesnych badaczach kończąc. Jest to wreszcie rozprawa naukowa, która ma ambicję ukazania zjawiska perswazji *sensu latiore*, w szerokiej perspektywie czasowej (od starożytności przez epoki dawne po teraźniejszość), w ujęciu interdyscyplinarnym i w różnych kontekstach metodologicznych. Tę złożoność (kompleksowość, kompletność?) obrazu sygnalizuje, a może podsumowuje, tytuł pracy ***Formy*** *perswazji*.

Wyraz *perswazja* (< łac. *persuasio* 'przekon(yw)anie') w polszczyźnie ogólnej należy do słownictwa książkowego, erudycyjnego, i w tej odmianie języka ma następującą definicję słownikową – 'tłumaczenie komuś czegoś, namawianie lub odradzanie z przytoczeniem argumentów popierających słuszność zdania, przekonywanie kogoś o czymś; perswadowanie' [USJP III, 99]. W swej pracy Jacek Warchala dwukrotnie formułuje specjalistyczną definicję perswazji – wyjściową (opartą na wstępnym przeglądzie literatury przedmiotu) i finalną (będącą wynikiem przeprowadzonych spekulacji i analiz):

> (...) perswazję chciałbym na wstępie ująć jako proces organizujący działanie komunikacyjne, a jego charakter i sposób postrzegać w kategoriach szerszych niż tylko językowe. Perswazja byłaby zatem działaniem (procesem) zarazem językowym, jak i pozajęzykowym, wykorzystującym systemy kodowe wielorodne; perswadować możemy słowem, ale i gestem konwencjonalnym czy dźwiękiem niewerbalnym oraz obrazem. Perswazja jako proces obejmuje także psychologiczne wyposażenie uczestników komunikacji, ich postawy, nastawienia na zrozumienie swoich racji lub konfrontację poglądów, ale też na podjęcie próby przekonywania i otwarcie na sytuacje bycia przekonywanym, zaakceptowanie odrębności innego partnera, a nie przedmiot oddziaływania – nawet wówczas, gdy mamy do czynienia ze stosunkiem władzy i podległości [Warchala 2019, 13].

> Perswazja jest działaniem komunikacyjnym podjętym świadomie i zorientowanym na osiągnięcie celu, którym jest wywarcie wpływu na odbiorcę (odbiorców), co powinno skutkować zmianą jego stanowiska w jakiejś kwestii lub wytrąceniem go z dotychczasowego stanu przekonań. Najszersza formuła działań perswazyjnych, wynikająca

> z idealistycznych koncepcji komunikacji Jürgena Habermasa, zakłada taki typ strategicznych działań komunikacyjnych, który jest obliczony na wpływ, a nie obliczonej [sic! obliczony? – S.D.] na porozumienie. (...) Istotne są tutaj relacje interpersonalne uczestników procesu komunikacji i intersubiektywna wspólnota komunikacyjna jako zgoda co do przyjętych i wspólnie zaakceptowanych norm współdziałania w ramach aktu mowy, podzielania wiedzy propozycjonalnej i zaufania do subiektywnej szczerości. W modelu wpływu obejmującego perswazję, ale też manipulację, te trzy warunki nie muszą być spełnione (...). Jako zjawisko (i praktyka) perswazja rozciąga się między dowodzeniem i argumentowaniem a manipulowaniem i podporządkowywaniem (zniewalaniem) [Warchala 2019, 343–344, 345].

Zestawienie tych definicji obrazuje to, w jakim kierunku został przeprowadzony opis form perswazji – od ujęcia bliskiemu deskrypcji słownikowej, co nie znaczy, że tożsamemu z nią, kładącemu nacisk na rozszerzenie zakresu perswazji na sferę faktów pozajęzykowych, do ujęcia interdyscyplinarnego, ukazującego płynność granic między perswadowaniem a zniewalaniem. Taki kierunek ewolucji pojęcia perswazji Jacek Warchala szeroko egzemplifikuje zarówno w sferze teorii perswazyjnych, jak i praktycznych działań nakłaniających.

Nie polemizując z zasadniczą autorską koncepcją opisu ewolucji form perswazji w dziejach kultury europejskiej (śródziemnomorskiej), trzeba wnieść – jak sądzę – dwa zastrzeżenia – do przedstawionych definicji. Po pierwsze, zbyt kategoryczne jest stwierdzenie, że perswazja jest „procesem organizującym działanie komunikacyjne", nie wydaje się bowiem, by miała taką funkcję niejako *ex definitione* we wszystkich komunikatach i sytuacjach mowy oraz by była jedynym czynnikiem organizującym działanie komunikacyjne, mamy tu bowiem zawsze „wiązkę determinant". Po drugie, nie jestem pewien, czy właściwe lingwistycznie jest hołdowanie Habermasowskiej teorii wpływu, co umożliwia (podkreślam *umożliwia*, nie zaś *determinuje*) nierozróżnianie perswazji i manipulacji. To, że rzeczywistość w tym wypadku „skrzeczy" nie znaczy, że w postępowaniu klasyfikacyjnym (analityczno-interpretacyjnym) mamy tej rzeczywistości się podporządkowywać.

Publikacja autorstwa Jacka Warchali, poza wstępem, bibliografią i streszczeniem, składa się z dziesięciu merytorycznych rozdziałów: I. *Na początku była retoryka*; II. *Wokół definicji perswazji*; III. *Perswazja a natura języka*; IV. *Argumentowanie a perswazja*; V. *Informowanie jako perswadowanie*; VI. *Manipulacja a perswazja*; VII. *Uwagi o emocjonalności i redefiniowaniu perswazyjnym słowa*; VIII. *O pewnym zastosowaniu metafory jako narzędziu perswazji*; IX. *Perswazja tkwi w opowieści (narracji)*; X. *Podsumowanie: Pragmatyka działania perswazyjnego, czyli o pewnych warunkach perswazji.* W spisie treści i wykładzie treści te rozdziały funkcjonują na zasadzie równorzędnej, choć – moim zdaniem – struktura rozprawy jest zhierarchizowana i *de facto* rozdziały te układają się w trzy segmenty. Pierwszy z nich stanowią rozdziały od I do VI, w których jest zawarta autorsko uporządkowana chronologicznie, metodologicznie i tematycznie historia perswazji wraz z jej interpretacją. Segment drugi stanowią rozdziały od VII do IX, które są swego rodzaju uszczegółowieniem wątków już wcześniej podjętych, w części ich powtórzeniem, w części rozszerzeniem. Może byłoby lepiej, by ich treści były wtopione w rozdziały segmentu pierwszego, ale autor uznał inaczej i należy jego prawa uszanować. Segment trzeci to rozdział X, podsumowanie całości.

Całość wykładu poświadcza gruntowną znajomość podjętej problematyki i niepoślednią erudycję w tym zakresie autora publikacji, który z równą sprawnością analizuje retoryczną teorię antyczną, filozoficzne ujęcia perswazji czasów nowożytnych czy marketingowe i politykierskie gry perswazyjno-manipulacyjne lat najnowszych. Nie jest celem tej recenzji omawianie kolejnych rozdziałów publikacji autorstwa Jacka Warchali, w tym wypadku nakłaniam zainteresowanych do ich indywidualnej lektury. Warto jednak zwrócić uwagę przynajmniej na dwa fragmenty książki, które dobrze oddają jego sposoby analizy i interpretacji.

> Jeśli perswazję potraktujemy jako genetycznie związaną z komunikowaniem, będzie ona właściwością specyficznie sytuacyjnego użycia języka i może wiązać się z każdą z wymienionych przez Jakobsona funkcji szczegółowych. Język żyje i jest kształtowany w działaniu i wszystko wskazywałoby na to, że tak rozumiana nadrzędna jego funkcja była zawsze związana z komunikowaniem jako interakcją społeczną, co nieuchronnie wiąże się z oddziaływaniem i przekształcaniem zastanej sytuacji, ogólnie z konstruowaniem pewnych sytuacji poprzez użycie języka [Warchala 2019, 110].

Przytoczony fragment jest konkluzją rozdziału III *Perswazja a natura języka*. Autor odchodzi w nim od wyjściowej koncepcji perswazji jako „procesu organizującego działania komunikacyjne" [zob. dyskusja wyżej] na rzecz teorii języka w działaniu (*nota bene* bliskiej koncepcji Witolda Doroszewskiego, która nie jest tu przywołana), w której perswazja jest funkcją społecznego charakteru języka, który od starożytności po współczesność bywa akcentowany w różnych koncepcjach lingwistycznych. Takie usytuowanie mechanizmów perswazji przydaje im – bez wątpienia – jedną z prymarnych ról w procesach komunikacji społecznej, chociaż nie jest to rola hegemona w ogóle wobec całokształtu procesów komunikacyjnych.

> Podjęte działania perswazyjne nie zawsze są fortunne, często też spotykamy się ze stwierdzeniem: *starałem się to mu / jej / im wyperswadować, ale byli odporni na moje słowa*. Ale jeśli działania mają mieć charakter perswazji, muszą spełniać, prócz wymienionych powyżej [np. podjęcie uczestnictwa w dyskursie, poddanie się perswazji, utwierdzenie nabytych przekonań – S.D.], jeszcze kilka podstawowych warunków, a mówiąc inaczej – pewne dodatkowe pragmatyczne warunki muszą zostać spełnione, aby akt mowy mógł zostać dopiero określony jako perswazyjny [Warchala 2019, 353].

Jacek Warchala omawia następujące warunki podjęcia aktu perswazji: wolności, polaryzacji sądów, równoważności racji, wspólnoty świata i języka, zgody na sytuację dialogową, afirmatywnego stanowiska, uzgodnienia wspólnego kodu, emocjonalizacji przekazu, budowania właściwego statusu nadawcy, zachowania procedury wywodu („logiki" wywodu), technologizacji perswazji, przyjemności estetycznej (w sumie 12).

Obszerne omówienie tych warunków stanowi *de facto* swego rodzaju „kognitywną" definicję perswazji, czy też – ujmując rzecz inaczej – uszczegółowienie (uzupełnienie) definicji finalnej sformułowanej wcześniej (przytoczonej na wstępie tej recenzji). Jest to zarazem właściwe podsumowanie rozważań autora na temat form perswazji, które zarazem ukazuje główne cechy jego metody badawczej. Jacek Warchala zbiera dane na podstawie kwerendy obszernej literatury przedmiotu (na jego erudycję już zwracaliśmy uwagę), na ich podstawie

formułuje definicje wstępne, które następnie konfrontuje z egzemplami, wynikającymi z oglądu rzeczywistości językowej i pozajęzykowej. Kolejnym etapem jest konstruowanie definicji finalnych, które w ostatnim etapie postępowania badawczego podlegają uszczegółowieniu. Jest to więc metoda spekulatywno-empiryczna, której rezultaty mogą być weryfikowane i w sferze metodologicznej, i w sferze materiałowej.

Jak zaznaczono na wstępie niniejszej recenzji, tematyka podjęta w publikacji Jacka Warchali znajduje różnorakie ujęcia w dziejach myśli humanistycznej, w tym – lingwistycznej, w tym – współczesnej, w tym – polskiej. Z pozycji, które ukazały się ostatnim trzydziestoleciu na gruncie polskiego językoznawstwa jest to – moim zdaniem – opracowanie najwszechstronniejsze. Jest to swego rodzaju „summa wiadomości". Wszyscy, których interesuje problematyka retoryki, perswazji, manipulacji czy dyskursu, nie powinni przechodzić obok tej publikacji obojętnie. Jest to książka, której należy się dobre miejsce w kanonie dzieł współczesnej humanistyki polskiej.

Stanisław Dubisz
(Towarzystwo Kultury Języka, Warszawa,
e-mail: stdubisz@gmail.com)
ORCID: 0000 0002-4784-5399

PROFESOR DR HAB. ANNA KOWALSKA (15.04.1936–25.12.2019)

Profesor Annę Kowalską (dla mnie i dla wielu językoznawców odchodzącego pokolenia po prostu: Hankę Kowalską) znałam od wielu lat jako pełną energii młodszą koleżankę zafascynowaną wykonywaną pracą, zawsze pogodną i otwartą dla otaczającego ją świata. Jej działalność i osiągnięcia naukowe zaznaczyły się na trwale w kilku różnych płaszczyznach i o nich pragnę przypomnieć, żegnając niedawno Zmarłą.

Głównym polem jej zainteresowań była **dialektologia polska**. Badania rozpoczęte pod kierunkiem profesora Witolda Doroszewskiego w zespole opisującym gwary Warmii i Mazur zaowocowały pracą magisterską dotyczącą rybołówstwa w serii „Studiów Warmińsko-Mazurskich" [1963]. Jej dysertacja doktorska *Zróżnicowanie słowotwórcze gwar Mazowsza i Podlasia* [1975; 1979] przyniosła wiele ważnych konstatacji tyczących gwarowych procesów słowotwórczych. Istotny wkład w naszą wiedzę wniósł też dziesięciotomowy *Atlas gwar mazowieckich* [1973–1994], będący w zasadzie dziełem dwóch osób (t. II–X autorstwa Anny Kowalskiej i Aliny Strzyżewskiej-Zaremby), lecz zdecydowana przewaga map, komentarzy i aneksów autorstwa Anny Kowalskiej każą uznać ją za główną twórczynię tego dzieła.

Problematyki mazowieckiej dotyczy również jej praca habilitacyjna *Podziały językowe Mazowsza na tle* podziałów *pozajęzykowych* [1991]. Godna szczególnej uwagi jest tu partia tycząca stosunku dialektologa do etnografii, podkreślająca istotną więź zachodzącą między nazwą a desygnatem. Tu właśnie dobitnie ujawnia się druga sfera zainteresowań badawczych autorki, którą stanowiła szeroko pojęta **etnografia**. Anna Kowalska wiedzę na ten temat zdobyła w ścisłej współpracy z profesor M. Znamierowską-Prűfferową, przygotowując (wspólnie z J. Maciejewskim) koncepcję *Ilustrowanego słownika rybołówstwa tradycyj-*

nego w Polsce [1991]. Przekonanie, że nie można badać nazw bez wiedzy o ich desygnatach, widoczne jest w jej wszystkich pracach tyczących dialektologii polskiej, a także **dialektologii słowiańskiej**. Anna Kowalska bowiem po ukończeniu studiów polonistycznych (1961) podjęła studia slawistyczne ze specjalnością bułgarystyczną, uzyskując z nich w roku 1965 absolutorium na Uniwersytecie Warszawskim. Była w tym czasie zaangażowana w rozpoczynającą się właśnie międzynarodową pracę nad *Językowym atlasem ogólnosłowiańskim* (*Obščeslavjanskij lingvističeskij atlas – OLA*). Wieloletnie uczestnictwo w tych badaniach oraz kontakty z różnorakimi szkołami dialektologicznymi uczyniły z niej slawistkę w pełnym tego słowa znaczeniu. Bezpośrednio już tematyki slawistycznej dotyczą też artykuły pisane wspólnie z Barbarą Falińską: *Polsko-białoruskie i polsko-ukraińskie związki leksykalno-słowotwórcze* [1983], *Paralele zachodnio-południowosłowiańskie obserwowane w gwarach i językach literackich* [1988], *Leksykalne związki języków łużyckich z innymi językami słowiańskimi* [1992], *Analiza porównawcza słownictwa polskiego i czeskiego w aspekcie badań pogranicza* [1993], a także opracowane przez nią liczne mapy i komentarze do poszczególnych tomów *Językowego atlasu ogólnosłowiańskiego*.

Oddzielną dziedziną, którą Anna Kowalska obdarzyła szczególnym zainteresowaniem i której poświęciła niemało twórczego wysiłku, była **logopedia**. Zainteresowania mechanizmami mowy i jej patologiami sprawiły, że ukończyła ona podyplomowe studia logopedyczne na Uniwersytecie Marii Curie-Skłodowskiej w Lublinie, później zaś wieloletnie doświadczenie z pracy w przychodniach logopedycznych zaowocowało wykładami z logopedii i magisterskimi seminariami w Wyższej Szkole Pedagogicznej Towarzystwa Wiedzy Powszechnej w Warszawie.

Anna Kowalska brała ponadto czynny udział w polskim i zagranicznym życiu naukowym. Jej rozległe kontakty i zainteresowania w połączeniu z talentem dydaktycznym czyniły z niej atrakcyjnego dla młodzieży nauczyciela, umiejącego nie tylko przekazać rzetelną wiedzę, lecz również ukazać uroki związanej z tą wiedzą przyszłej pracy.

Hanna Popowska-Taborska
(Warszawa,
e-mail: hanna.taborska@gmail.com)

INFORMACJE DLA AUTORÓW „PORADNIKA JĘZYKOWEGO”

Prosimy Autorów o nadsyłanie artykułów, rozpraw, recenzji publikacji językoznawczych oraz sprawozdań z konferencji, sympozjów i spotkań, ponieważ chcemy, aby „Poradnik Językowy” w szerokim zakresie informował o życiu naukowym w kraju i za granicą.

Uprzejmie prosimy Autorów o przestrzeganie następujących zasad redakcyjnych:

- Objętość artykułu nie powinna przekraczać 14 stron znormalizowanego maszynopisu (ok. 25 000 znaków ze spacjami), objętość recenzji zaś – stron 7 (ok. 12 000 znaków ze spacjami).
- Prosimy o dołączenie do tekstu artykułu krótkiego (10 wersów) streszczenia i słów kluczowych w języku polskim. Streszczenie powinno zawierać: 1) uzasadnienie podjętych badań; 2) prezentację uzyskanych wyników; 3) omówienie zastosowanej metody badawczej.
- W cudzysłowie podajemy tytuły czasopism oraz cytaty – jeżeli nie są wyodrębnione w inny sposób (np. inną wielkością pisma).
- Kursywą wyodrębniamy wszystkie omawiane wyrazy, zwroty i zdania, ponadto tytuły książek i części prac, tzn. rozdziałów i artykułów, oraz zwroty obcojęzyczne wplecione w tekst polski.
- Podkreślenia tekstowe oznaczamy spacją (druk rozstrzelony).
- Znaczenie wyrazów podajemy w łapkach ‘’.
- Prace należy dostarczać w formie papierowej oraz wersji elektronicznej na konto: poradnikjezykowy@uw.edu.pl
- Autorów przysyłających swoje prace po raz pierwszy prosimy o czytelne podanie imienia, nazwiska, tytułu naukowego lub zawodowego, numeru ORCID, nazwy ośrodka naukowego (przy którym chcą afiliować tekst artykułu), adresu prywatnego, adresu e-mail i numeru telefonu. Pliki prosimy przysyłać w formacie edytora MS Word (*doc, *.rtf).
- Autorzy są zobowiązani do złożenia oświadczenia o oryginalności autorskiej tekstów.

Redakcja nie zwraca tekstów niezamawianych.

Cena zł 17,00
w tym VAT 8%

I N F O R M A C J A O P R E N U M E R A C I E

„PORADNIKA JĘZYKOWEGO”

Ceny „Poradnika Językowego” w roku 2020:
prenumerata roczna (10 numerów) – 170,00 zł,
opłata za pojedynczy numer – 17,00 zł.

Zamówienia na pojedyncze egzemplarze pisma można składać bezpośrednio na stronie wydawnictwa: www.elipsa.pl lub kierować na adres e-mail: sklep@elipsa.pl

Prenumerata krajowa i zagraniczna

Zamówienia na prenumeratę w wersji papierowej można składać bezpośrednio na stronie

RUCH SA, www.prenumerata.ruch.com.pl

Ewentualne pytania prosimy kierować na adres e-mail: prenumerata@ruch.com.pl lub kontaktując się z Centrum Obsługi Klienta „RUCH” pod numerami: 22 693 70 00 lub 801 800 803 – czynne w dni robocze w godzinach 7^{00}–17^{00}. Koszt połączenia wg taryfy operatora.

Zamówienia na prenumeratę przyjmują również:

KOLPORTER SA, www.kolporter.com.pl, e-mail: prenumerata.warszawa@kolporter.com.pl
tel. 22 355-04-71 do 75

GARMOND PRESS SA, e-mail: prenumerata.warszawa@garmondpress.com.pl
tel. 22 837-30-08

Subscription orders for all magazines published in Poland available through the local press distributors or directly through:

Foreign Trade Enterprise ARS POLONA SA, ul. Obrońców 25, 03-933 Warszawa
www.arspolona.com.pl, e-mail: arspolona@arspolona.com.pl