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, ... svyøto 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 oszadl* ('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, doszwyatczon 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 nagla 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 gotowu ieszteszcie… 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