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 *ż*; 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].

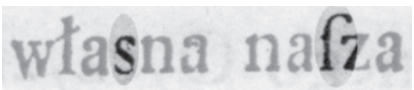
[19] Nie zachowujemy jedynie przypadkowych spacji wewnątrz wyrazu, np. *N áuká* → *Náuká* oraz braku spacji wynikającego z braku miejsca w danej linii.

[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) *mężе* (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