Mariusz Leńczuk
Instytut Języka Polskiego PAN w Krakowie
https://orcid.org/0000-0003-0731-4056
mariusz.lenczuk@ijp.pan.pl

Forum Lingwistyczne
nr 7, 2020
ISSN 2450-2758 (wersja elektroniczna)

DOI: http://doi.org/10.31261/FL.2020.07.05

Projektowanie metadanych w korpusie tekstów polskich do 1500 roku – wielopoziomowa struktura informacji*

Metadata Creation in the Language Corpus of Polish Texts until 1500 – a Multi-level Data Structure

Abstract: The subject of research are selected metadata that should characterize the texts collected in the corpus of the oldest attestations of the Polish language. The author of the article compares and analyses the factors affecting the development of the basic data structure used in synchronic and diachronic corpora (author, title, date of the text, text channel, text classification, source of citation). Without those factors taken into account the disambiguation of the object in the database becomes impossible, and the use of grammatical information is unreliable and impractical. The result of the presented analysis is a proposal to extend the level of description for individual markers.

Key words: language corpus, metadata, text, glosses, 13th–15th century

Abstrakt: Przedmiotem badań są wybrane metadane, które powinny charakteryzować teksty zgromadzone w korpusie najdawniejszych zabytków języka polskiego. Autor artykułu porównuje i analizuje czynniki wpływające na rozbudowywanie podstawowej struktury danych stosowanej w korpusach synchronicznych i diachronicznych (autor, tytuł, datacja tekstu, kanał, klasyfikacja tekstu, źródło cytowania), bez których uwzględnienia ujednoznacznienie obiektu w bazie danych staje się niemożliwe, a wykorzystanie informacji gramatycznych – mało wiarygodne i niepraktyczne. Rezultatem przedstawionych analiz jest propozycja rozszerzenia poziomu opisu poszczególnych znaczników.

Słowa kluczowe: korpus językowy, metadane, tekst, glosy, XIII–XV wiek

Pierwsze prace nad strukturą elektronicznego korpusu języka polskiego rozpoczęto ponad ćwierć wieku temu (Przepiórkowski i in., 2012: 5). Od tamtego momentu oprócz udoskonalonego już synchronicznego korpusu języka polskiego powstały również korpusy diachroniczne: *Elektroniczny Korpus Tekstów Staropolskich do 1500 roku*, PolDi – *a Polish Diachronic Online Corpus*, *Korpus polszczyzny XVI wieku*, *the Middle Polish Diachrone Lemmatised Corpus (16th–18th c.)*, *Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 r.)* (KorBa) oraz *Korpus polszczyzny 1830–1918*. Nie wszystkie z nich można nazwać

* Artykuł powstał w ramach projektu badawczego „Baza leksykalna średniowiecznej polszczyzny (do 1500 roku). Fleksja", przeznaczonego do realizacji w latach 2018–2023 pod kierownictwem dr hab. Ewy Deptuchowej, prof. IJP PAN w Krakowie, finansowanego ze środków Narodowego Programu Rozwoju Humanistyki (nr projektu: 0201/NPRH/H11/85/2018).

korpusami w znaczeniu, w którym termin ten używany jest obecnie. Dzieje się tak, ponieważ prace nad nimi wciąż trwają. Część z nich złożona jest wyłącznie z nieoznakowanych tekstów w różnej postaci (transliteracji lub transkrypcji) i w różnym formacie (pdf, xml), inne bazy składają się z tekstów otagowanych podstawowymi znacznikami. W końcu są i takie, które spełniają wymogi stawiane współczesnym korpusom językowym. Niezależnie od stanu bieżących prac wszystkie zespoły tworzące korpusy diachroniczne odwołują się bezpośrednio do jednego wzorca – *Narodowego Korpusu Języka Polskiego*. Jest to podyktowane pragmatyzmem i koherencją. Z jednej strony łatwiej jest budować korpus, mogąc wzorować się na istniejącym i sprawdzonym korpusie, z drugiej strony standardy budowania tego typu baz danych są ściśle określone, co pozwoli w przyszłości połączyć mniejsze korpusy w uspójniony Narodowy Korpus Diachroniczny Polszczyzny (Król i in., 2019: 92–101).

Nawet wstępna analiza struktury metadanych NKJP uwidacznia, że bezpośrednie zastosowanie podobnych rozwiązań w korpusie diachronicznym jest wykluczone. Wynika to głównie z wpływających na tę strukturę historycznych czynników, które nie zawsze są brane pod uwagę podczas opisywania obiektów zamieszczanych w korpusie współczesnym. Porównanie materiału badawczego oraz analiza wspomnianych czynników pozwoli uzasadnić potrzebę modyfikacji gotowej już struktury metadanych, użytej w korpusie synchronicznym, do potrzeb korpusu tekstów polskich do 1500 roku.

Celem niniejszego artykułu jest przedstawienie zróżnicowanej struktury metadanych charakteryzujących teksty pochodzące ze średniowiecznych rękopisów i inkunabułów oraz analiza czynników wpływających na strukturę metadanych. Zanim jednak zostaną omówione szczegółowe zagadnienia związane z porządkowaniem informacji na temat samych tekstów oraz danych źródłowych, przybliżę podstawowe terminy, którymi będę posługiwał się w dalszym ciągu wywodu.

Zgodnie z informacją umieszczoną na stronie *Narodowego Korpusu Języka Polskiego*[1] oraz definicjami słownikowymi (WSJP PAN, SJP PWN) korpusem językowym nazywamy zbiór tekstów uporządkowanych i opracowanych według określonych zasad, który służy do celów naukowych i edukacyjnych. Podstawowymi kryteriami określającymi kompletność korpusu językowego są jego zróżnicowanie, zrównoważenie oraz reprezentatywność. W przypadku korpusu tekstów polskich do 1500 roku, podobnie jak we wszystkich korpusach diachronicznych, kryteria te mogą być spełnione tylko częściowo z uwagi na stan rozpoznania i zachowania zabytków językowych (Gruszczyński, Adamiec, Ogrodniczuk, 2013; Kołodziej, Klapper, 2014; Klapper, Kołodziej, 2015). Pierwszy z czynników ma bezpośredni wpływ na przygotowanie metadanych, drugi – na zasób bazy danych.

Kolejnym pojęciem, które wymaga doprecyzowania, jest tekst. Pierwotnie duże zbiory zabytków językowych (Wydra, Rzepka, 2004) dzielono na teksty rękopiśmienne i drukowane, ale i ta klasyfikacja miała bardziej złożoną hierarchię, w której wyróżniano zabytki prymarne i sekundarne[2]. Zgodnie ze współczesną nomenklaturą możemy powiedzieć, że dawne zbiory składały się z tekstów ciągłych (np. kazań, listów, modlitw, rot przysiąg sądowych, statutów) oraz tekstów nieciągłych (np. glos do tekstów łacińskich, słowników

[1] NKJP: www.nkjp.pl.

[2] Teksty sekundarne nazywane były najczęściej „tekstami drobnymi".

dwujęzycznych, fragmentów zdań czy rymowanek polsko-łacińskich). Podczas komponowania optymalnie zróżnicowanego i zrównoważonego korpusu językowego nie można pominąć żadnej z grup tekstów, ponieważ zaburzyłoby to obraz XV-wiecznego języka polskiego, baza nie odzwierciedlałaby częstotliwości występowania słów, liczby połączeń wyrazowych czy wreszcie struktury gramatycznej języka. Zadaniem artykułu nie jest próba stworzenia definicji tekstu ciągłego i nieciągłego, jednak w kontekście przygotowywania metadanych, zawierających informacje o autorstwie, datacji czy tytule podział na teksty ciągłe (na potrzeby tegoż tekstu określmy je jako samodzielne) i nieciągłe (niesamodzielne)[3] jest niezbędny, ponieważ liczba danych, którymi powinny być opatrzone tak odmienne obiekty, będzie zróżnicowana.

Tytułowe metadane to ustrukturalizowane informacje, które służą do opisu zebranych zasobów, ułatwiają ich identyfikację i znalezienie oraz pozwalają nimi zarządzać. W każdym korpusie językowym, czy to współczesnym, czy historycznym, ich podstawowym zadaniem jest charakterystyka i klasyfikacja tekstu. Dzięki nim użytkownik jest w stanie dowolnie ograniczać zakres poszukiwań oraz tworzyć własne podkorpusy. Można zaryzykować twierdzenie, że w korpusie diachronicznym, którego zasób jest ograniczony liczbą zachowanych obiektów, wynik badań zależy nie tylko od umieszczonych w bazie tekstów, ale również od wiarygodności i przejrzystości zebranych metadanych. Przykładowo, dzięki uściślonym identyfikatorom czasowym będzie można śledzić ilościowe zmiany form gramatycznych w wybranych przedziałach XV wieku (Gruszczyński, Bronikowska, 2015), jeśli jednak dane pozostaną niedoprecyzowane, na przykład znacznik czasowy będzie zawierał ogólną informację pochodzącą od wydawcy, że tekst pochodzi z XV wieku, to otrzymany wynik będzie nieprecyzyjny i z tego względu praktycznie nieprzydatny, ponieważ większość tekstów w tym korpusie będzie pochodziła z tego okresu.

Jakie metadane są niezbędne do jednoznacznej identyfikacji najstarszych zabytków języka polskiego? Do czego mogą być przydatne informacje o tekstach łacińskich, z którymi powiązane są teksty polskie? W jaki sposób można weryfikować i uściślać dotychczasowe ustalenia na temat poszczególnych rękopisów i inkunabułów? Odpowiedzi na te pytania są kluczowe w procesie formułowania zapytań korpusowych. Z uwagi na różnorodność staropolskich źródeł, brak szczegółowych badań oraz ograniczone ramy artykułu zadanie to nie może być w pełni wykonane. Mam jednak nadzieję, że podjęta tu próba przeanalizowania wybranych zagadnień związanych z ustalaniem metadanych w korpusie tekstów polskich do 1500 roku pobudzi do dalszej dyskusji na ten temat, a w konsekwencji przyniesie wymierny efekt w postaci odpowiednio opisanego korpusu.

[3] Zaproponowany doraźny podział tekstów staropolskich służy tu jedynie do zobrazowania odmiennych typów zabytków językowych, wśród których można wyróżnić m.in.: teksty polskie spójne treściowo i gramatycznie; teksty polsko-łacińskie spójne treściowo i gramatycznie; glosy polskie do tekstów łacińskich spójne treściowo i gramatycznie, będące przekładami fragmentów tekstu łacińskiego, bez którego nie można ich zrozumieć; pojedyncze glosy polskie do tekstów łacińskich, które są niezrozumiałe bez łacińskiego kontekstu; słowniczki dwujęzyczne; pojedyncze zdania i zapiski na marginesach i okładkach rękopisów niepowiązane z łacińskimi tekstami. Niezależnie od liczby wydzielonych typów tekstów, można stwierdzić, że jedne są zrozumiałe bez podstawy łacińskiej, drugie zaś funkcjonują wyłącznie w kontekście tekstu podstawowego, jakim jest dla nich tekst łaciński.

Podstawowe metadane, którymi opatrywane są teksty korpusowe, to: autor, tytuł, data, kanał (forma przekazu), lokalizacja cytatu (źródło cytowania) oraz klasyfikacja tekstu. Ponadto zasoby opisywane są również innymi identyfikatorami i odsyłaczami, na przykład: wydawca, data publikacji, drukarnia, miejsce wydania, redaktor, tłumacz, wersja cyfrowa tekstu, podobizna źródła w bibliotece cyfrowej. Nie wszystkie metadane są widoczne dla zewnętrznego użytkownika, część znaczników jest istotna wyłącznie dla autorów korpusu, na przykład kto opracował źródło, kto zrobił korektę, kto był konsultantem. Należy pamiętać, że najważniejszym celem ustrukturalizowania udostępnianych informacji jest ujednoznacznianie obiektów oraz ich typologia. Poniżej opisano czynniki wpływające na strukturę metadanych w korpusie języka polskiego do 1500 roku, które nie zawsze są brane pod uwagę w korpusach synchronicznych.

1. Autor

W korpusach językowych to podstawowy znacznik, dzięki któremu pojedynczy tekst można przypisać konkretnemu autorowi lub autorom. W korpusie najstarszych zabytków języka polskiego taka sytuacja jest rzadkością. Znamy utwory Władysława z Gielniowa, na przykład *Jezusa Judasz przedał* (*Żołtarz Jezusow*), *Anna niewiasta niepłodna* (*De nativitate Marie ista cancio*), *Już się anjeli wiesielą*, czy Andrzeja Gałki z Dobczyna *Pieśń o Wiklefie*, ale są to przypadki odosobnione zarówno z uwagi na ówczesne zwyczaje pisarskie, jak i na fakt, że znakomita większość zachowanych utworów to kopie i odpisy niewiadomego pochodzenia. Z tego też względu takie teksty opatrzone będą informacją – „anonim".

Istnieją również zabytki, które mają autora domniemanego. Za przykład niech posłużą tzw. *Kazania gnieźnieńskie*, których twórcą, według Jerzego Wolnego, mógł być Łukasz z Wielkiego Koźmina (Wolny, 1961). W tym wypadku imię, nazwisko oraz miejsce pochodzenia autora powinny być opatrzone znakiem zapytania lub innym grafemem sygnalizującym ten fakt.

Co jednak zrobić, gdy w łacińskim zbiorze kazań, przypisywanym na przykład Hieronimowi z Pragi, występują polskie modlitwy *Zdrowaś Maria*, *Wierzę* (rękopis CBPP sygn. Lat.F.ch.I.49) lub wezwanie modlitewne *Moc Boga Ojca* (rękopis CBPP sygn. Lat.F.ch.I.240), których Hieronim nie był autorem? W dotychczasowych rozwiązaniach korpusowych metadane autora polskiego tekstu ograniczyłyby się do opisu „anonim", a w korpusie nie zostałaby uwzględniona informacja o autorze kazań. Potrzebny jest zatem dodatkowy poziom opisu: „autor tekstu podstawowego" (np. łacińskiego, niemieckiego, czeskiego, w którym umieszczono tekst polski). Tego rodzaju znacznik będzie miał zastosowanie nie tylko w wypadku powszechnie występujących modlitw codziennych, wezwań do osób świętych oraz tekstów katechizmowych wplecionych w teksty łacińskie, ale przede wszystkim w opisie polskich tekstów nieciągłych, które prawie wyłącznie występują pośród zapisów obcojęzycznych. Zarówno dodatkowy znacznik określający autora tekstu podstawowego, jak i dodatkowy znacznik tytułu tekstu podstawowego umożliwią wydzielenie w korpusie podkorpusu, na przykład kazań określonego autora lub kazań na dany dzień roku liturgicznego czy traktatu o *Pater noster*, co pozwoli przeprowadzić badania porównawcze polskich tekstów w identycznych tematycznie podstawowych tekstach łacińskich.

Charakterystyczną cechą piśmiennictwa średniowiecznego było powielanie pojedynczych utworów (pieśni, modlitw, kazań, traktatów, słowników), a także całych rękopisów. Wśród często kopiowanych dzieł teologicznych można wymienić między innymi kolekcję kazań Piotra z Miłosławia, które były bogato glosowane przez kaznodziejów. Pośród kazań należących do tej kolekcji rozpoznane zostały między innymi łacińskie *sermones* autorstwa Łukasza z Wielkiego Koźmina, Jana Sylwana, Jana Szczekny, Jana Milicza, Mateusza z Krakowa (Bracha, 2007: 65), co sugeruje, że Piotr mógł być kopistą lub kompilatorem zbioru. Proces adaptacji i modyfikacji tekstów popularnych wówczas autorów do indywidualnych potrzeb kaznodziejów był tak powszechny i złożony, że trudno dziś jednoznacznie stwierdzić, czy zachowane w kolekcji Miłosławczyka kazania można jeszcze przypisać wcześniejszym autorom, na przykład Łukaszowi z Wielkiego Koźmina, czy późniejszemu kompilatorowi. Warto się zastanowić, w jakim stopniu przydatne byłoby umieszczenie informacji o pierwotnym autorze łacińskiego tekstu, ewentualnie określenie go jako współautora. Taki zabieg pozwala skolacjonować zachowane odpisy pojedynczych kazań oraz zbadać zachowane w nich polskie glosy i przekłady.

W celu ujednoznacznienia kilku obiektów przypisanych nieznanemu autorowi (anonimowi), posiadających ten sam tytuł (np. *Sermones dominicales*), zachowanych w wielu wariantywnych odpisach, należy dodać jeszcze jeden znacznik, którego nie mają istniejące korpusy diachroniczne, a który wydaje się niezbędny w opisie tekstów średniowiecznych – „pisarz". Warto w tym miejscu zadać pytanie: Czy pisarz mógł być autorem polskich tekstów w zbiorach łacińskich kazań? Pewne jest, że nie każdy. Znaczna część tych tekstów została uwieczniona ręką twórcy manuskryptu, a inne rękami późniejszych właścicieli czy użytkowników. Zdarza się, że w kolofonie skryba pozostawił informację o sobie oraz dniu i roku zakończenia pracy, dzięki czemu możemy z całą pewnością stwierdzić, że wszelkie uwagi sporządziła konkretna osoba, jednak nie wiemy, czy był to kopista, który przepisał polszczyznę z innego egzemplarza, czy był to autor, który samodzielnie wprowadził polskie glosy, komentarze czy teksty ciągłe. Dodatkowa informacja zamieszczona w korpusie pozwoli wyodrębnić zbiór tekstów sygnowanych ręką konkretnego pisarza. Fakt wprowadzenia takiego znacznika nie przesądza o autorstwie, jednak przy braku jakichkolwiek danych znacznik ten umożliwi rozróżnienie kilku odpisów tego samego tekstu oraz pomoże umieścić tekst w określonych ramach czasowych.

Osobnym zagadnieniem, które należy brać pod uwagę, ustalając metadane autora, pozostaje funkcjonowanie w obiegu naukowym wariantywnego nazywania autorów. Przywołany już Hieronim z Pragi znany jest również jako Jan Sylwan, a w katalogach bibliotecznych oraz opracowaniach językowo-literackich i historycznych wymieniany bywa jako Hieronimus (Hieronymus) de Praga, Joannes (Ioannes) Silvanus, a nawet jako Hieronymus Ioannes Silvanus de Praga. Wszystkie możliwe formy jego imienia, nazwiska oraz miejsca pochodzenia powinny trafić do metadanych w znaczniku „autor". Umożliwi to przypisanie tekstów tylko jednej osobie, a nie czterem czy pięciu.

Podsumowując dotychczasowe rozważania na temat pierwszego znacznika, należy stwierdzić, iż niezbędne jest rozgraniczenie: autora tekstu polskiego, autora tekstu podstawowego (obcego) i pisarza. Znacznik autora tekstu powinien mieć rozszerzoną opcję podawania informacji uwzględniającej kilka form zapisu imienia, nazwiska, przydomka oraz miejsca pochodzenia autora.

2. Tytuł

W ustalaniu tytułu średniowiecznego tekstu należy uwzględnić funkcjonujące w literaturze przedmiotu jego warianty. Podobnie, jak w uprzednio omówionym znaczniku, tak też w tym wypadku można mówić o nadmiarze, który wymaga korpusowego ujednoznacznienia. W zależności od środowiska, w którym powstawały edycja lub opracowanie średniowiecznego tekstu, przyjmowano odmienne sposoby identyfikacji obiektu. Przykładowo, polski tekst statutów zapisany w rękopisie nr 1418 należącym do Biblioteki Książąt Czartoryskich w Krakowie określany był jako: *Kodeks Świętosława z Wojcieszyna*, *Kodeks Macieja z Rożana*, *Kodeks (Mikołaja) Suleda*, *Statuty Kazimierza Wielkiego, Władysława Jagiełły i książąt mazowieckich* czy wreszcie *Statuta Regni Poloniae*. Tytuł katalogowy ustalony przez Bibliotekę to *Kodeks Świętosława z Wojcieszyna*. Należy go wiązać z autorem tłumaczenia początkowego fragmentu kodeksu (k. 2r–41v) zawierającym statuty Kazimierza Wielkiego i Władysława Jagiełły. Tłumaczem kolejnej części (k. 43r–57r), zawierającej statuty książąt mazowieckich, był Maciej z Rożana, stąd dodatkowy wariant tytułu. Zwyczajowe określenie funkcjonujące w środowisku twórców *Słownika staropolskiego* to *Kodeks Suleda*. Tytuł ten występuje również w wielu opracowaniach historycznojęzykowych i nawiązuje do osoby pisarza kodeksu. Historycy stosują najczęściej tytuł łaciński, który zawiera informację o zawartości kodeksu.

Przytoczony przykład nie jest odosobniony, większość średniowiecznych tekstów ciągłych posiada przynajmniej dwa tytuły (najczęściej polski i łaciński), które powinny być odnotowane w korpusie. Podczas ustalania tego znacznika warto zadać dodatkowe pytania: Co z tekstami nieciągłymi? Jak je identyfikować? Czy powinny mieć jakiś umowny tytuł? Czy wiązać je z pojedynczymi tekstami łacińskimi, w których wystąpiły, czy raczej z całym kodeksem?

Odpowiedź na pytanie o tytuł może być tylko jedna. Teksty nieciągłe nie mogą mieć polskiego tytułu. Potrzebny jest im jednak punkt odniesienia do łacińskiej podstawy. Glosy w korpusie nie powinny być pozbawione kontekstu, w którym wystąpiły, ani tytułu dzieła, z którego zostały zaczerpnięte, ponieważ uniemożliwiłoby to ich zrozumienie, a co za tym idzie – prawidłowe oznakowanie w korpusie. Niezbędny jest zatem znacznik „tytułu tekstu podstawowego", o którym była mowa już wcześniej. Dzięki niemu możliwa będzie również jakakolwiek klasyfikacja tematyczna tekstów nieciągłych zależna od klasyfikacji tematycznej łacińskiego dzieła, w którym zostały zapisane. Teksty nieciągłe w postaci zdań, powiedzeń, fragmentów czy komentarzy rozpoznawane są tradycyjnie za pomocą incipitów, tzn. początkowych słów utworu, co stosowane jest również w tekstach ciągłych, w sytuacji, gdy utworowi nie przypisano zwyczajowego tytułu.

Ostatnie pytanie, dotyczące wydzielenia tekstów nieciągłych w pojedynczych utworach łacińskich oraz rozgraniczenia tytułów tychże utworów (np. *Sermo in prima dominica quadragesima*) od tytułów całych kodeksów (np. *Sermones*), wydaje się czysto akademickie, ponieważ obecny stan rozpoznania i opracowania średniowiecznych manuskryptów nie pozwala na tak ścisłe ustalenia metadanych, a tym samym skazuje twórców korpusu albo na podjęcie samodzielnych badań tekstologicznych, albo na pozostawienie znacznika „tytuł" – nieuzupełnionego[4].

[4] Osobnym pytaniem pozostaje, jak potraktować w korpusie zgromadzone w kodeksie teksty, które dotychczas w środowisku filologicznym funkcjonowały jako całość (np. źródła *Słownika staropolskiego*

Podsumowując powyższe rozważania na temat drugiego znacznika, niezbędne jest rozgraniczenie: tytułu tekstu polskiego i tytułu tekstu podstawowego (obcego) oraz uwzględnienie kilku funkcjonujących w literaturze przedmiotu wariantów tytułów polskich i obcych, na przykład *Ojcze nasz*, *Modlitwa Pańska*, *Oratio Dominica*.

3. Datacja tekstu

Dotychczasowa praktyka datowania tekstów polskich opierała się na tradycyjnym pozyskiwaniu informacji z rękopisów, wydań tekstów, opisów katalogowych i bibliotecznych, a w nielicznych tylko wypadkach, gdy brakowało tego typu informacji, na współczesnych badaniach kodykologicznych. Najaktualniejszy stan rozpoznania tekstów staropolskich, zawarty w *Opisie źródeł Słownika staropolskiego* (OźSstp, 2005), nie gwarantuje jednak stworzenia wiarygodnej bazy danych, ponieważ duża część scharakteryzowanych tam zabytków języka polskiego zawiera informacje chronologiczne, choć przypisane do poszczególnych tekstów, to jednak odnoszące się do całych kodeksów, bo pochodzące z opisów katalogowych rękopisów, bez szczegółowego datowania fragmentów zawierających polskie zapisy. Istnieje oczywiście grupa tekstów ciągłych, która od dawna znajdowała się w centrum zainteresowania mediewistów, co sprawiło, że informacje na temat ich datacji są aktualne, ale nie dotyczy to niestety tekstów nieciągłych. Większość z nich została zespolona w sztucznie wyodrębnione przez wydawców całostki, na przykład glosy w kazaniach, glosy w rękopisie, i opatrzona datami ogólnymi: XV *in.*, XV *med.*, XV *p. post.*, XV *ex.* lub przybliżonymi, na przykład *ca* 1450, *ca* 1500. Tak szerokie ramy czasowe są niejednoznaczne, a różnice zakresów dat dla poszczególnych tekstów oznaczonych takimi separatorami czasowymi mogą wynosić pięć, dziesięć, a nawet piętnaście lat (w przód i wstecz od wskazanej daty). Wprowadzenie tego typu separatorów do korpusu wiąże się z każdorazową decyzją o wyznaczeniu określonego przedziału czasowego dla pojedynczego tekstu. Z punktu widzenia źródłoznawczego jest to sytuacja idealna, ponieważ powstaje możliwość uporządkowania informacji na temat najstarszych zabytków języka polskiego. Natomiast z punktu widzenia korpusowego praca taka wymagałaby zaangażowania dużego zespołu badawczego oraz wielu lat poszukiwań, co realnie uniemożliwiłoby szybkie opracowanie korpusu. Z uwagi

takie jak: *GlDom*, *GlKazB* I–IV, *GlLek*, *GlPozn*, *GlWroc*). Czy w związku z faktem, że są to historycznie odrębne teksty (powstałe w różnych latach, ale spisane w jednym czasie), należy je wyodrębnić? W wypadku opisanego powyżej rękopisu Biblioteki Książąt Czartoryskich w Krakowie sprawa jest prosta i rozdzielenie tekstów wydaje się uzasadnione. Istnieją jednak manuskrypty zawierające kilkaset odrębnych tekstów nieciągłych, które do tej pory funkcjonowały jako całość – glosy w kodeksie. W rękopisie średniowiecznym mogły znajdować się np. traktaty teologiczne, prawnicze i medyczne, wyciągi z autorów średniowiecznych czy komentarze do ksiąg biblijnych. Niejednokrotnie w takim zbiorze tekstów łacińskich zapisywano fragmenty po polsku datowane na różne lata XV wieku, wpisane przez różnych skrybów, a także różniące się tematyką. Podobnie jest z wydaniami manuskryptów zaginionych, zniszczonych i nierozpoznanych (np. źródła *Słownika staropolskiego* takie jak: *PF* V, *R* XXIV, XXV, *Zab*). Wydania zawierają jedynie skromne wyciągi polskich tekstów nieciągłych z różnych dzieł łacińskich, których nie jesteśmy w stanie zidentyfikować. Jak w takim wypadku opisywać metadane tekstów nieciągłych?

na ograniczenia techniczne nie można pozostawić w korpusie ogólnych i przybliżonych separatorów czasowych bez ich dookreślenia – czy to w odniesieniu do całości tekstów z określonym separatorem, na przykład XV *in.*, czy to dla każdego tekstu osobno. Decyzja ta wpłynie na strukturę chronologiczną tekstów całego korpusu.

Innym zagadnieniem wymagającym przemyślenia jest sposób traktowania kopii dokumentów dawnych, które były oznaczane podwójnymi datami, na przykład: (1490) 1636, co należy rozumieć, że tekst z 1490 roku zaczerpnięto z kopii sporządzonej w 1636 roku. Czasem wydawca pozostawiał jedną datę, na przykład: (1500), i nie informował, z jakiego okresu pochodził dokument, z którego korzystał. Jak traktować tego typu teksty? Czy powinny trafić do korpusu? Jeżeli przyjmiemy, że z uwagi na językową wartość powinny, to należałoby przyjąć takie rozwiązania techniczne, które umożliwią wydzielanie w korpusie tego rodzaju tekstów.

Podsumowując rozważania na temat trzeciego znacznika „data", niezbędne jest rozgraniczenie daty powstania tekstu polskiego i daty powstania kodeksu zawierającego tekst polski, ustalenie przejrzystych ram czasowych dla separatorów przybliżonych i ogólnych typu *circa* (np. ± 10 lat) oraz ustalenie sposobu postępowania z kopiami dokumentów średniowiecznych.

4. Kanał i źródło cytowania (lokalizacja cytatu)

W korpusach językowych „kanał" i „źródło cytowania" to dwa znaczniki, które są od siebie zależne. Kanał to, inaczej mówiąc, miejsce zapisania tekstu, które nie zawsze jest tożsame ze źródłem cytowania. Wynika to z faktu, że teksty w korpusie diachronicznym powinny być podawane w dwóch postaciach ortograficznych: transliteracji i transkrypcji.

W korpusie tekstów do 1500 roku możemy wyróżnić trzy podstawowe kanały cytowania tekstów w transliteracji: rękopis, inkunabuł i książkę[5]. Z uwagi na czas powstania zapisów dwa pierwsze kanały są zrozumiałe i nie wymagają objaśnienia, trzeci obliguje jednak do skomentowania. Pod pojęciem *książki* kryje się takie współczesne wydanie (XIX- lub XX-wieczne) tekstu dawnego, które nie zawiera informacji o tym, skąd wydawca cytował zabytek staropolski (z rękopisu czy inkunabułu). Sytuacja taka mogła zaistnieć zarówno podczas wydawania różnego rodzaju wypisów z zasobów dawnych bibliotek, jak i w szczegółowych omówieniach konkretnego zagadnienia (np. prawnego), gdzie w przypisach, jako ilustracje materiałowe, cytowano fragmenty średniowiecznych dokumentów bez podawania ich lokalizacji (por. np.: *BiblWarsz*, *RocznHist*, *StPPP*). Podobną praktyką wydawniczą było cytowanie w przypisach artykułów naukowych tekstów porównawczych lub wariantów tekstowych z innych odpisów niż omawiany w artykule, bez podawania szczegółowych informacji na ich temat (por. np.: *DłKlejn*, *SalveReg*).

Teoretycznie źródło cytowania powinno być tożsame z kanałem cytowania, gdy polski tekst podawany jest w transliteracji. Wtedy będzie to rękopis, inkunabuł lub książka. Twórcy korpusu mogą zdecydować jednak inaczej: transliteracja tekstu polskiego zostanie podana za wydaniem elektronicznym, które już zweryfikowano, na przykład *Biblioteką za-*

[5] O jeszcze jednym kanale mowa będzie w dalszej części artykułu.

bytków polskiego piśmiennictwa średniowiecznego (płyta DVD) – w takim wypadku należy wydzielić dodatkowy kanał – publikacja elektroniczna.

Źródłem cytowania tekstów podanych w transkrypcji będą najnowsze lub najlepsze (zdaniem twórców korpusu) wydania poszczególnych tekstów. Sytuacja wydaje się prosta i schematyczna, gdy bierzemy pod uwagę teksty ciągłe, które wpisane są w określoną tradycję wydawniczą. Co jednak zrobić z tekstami nieciągłymi, które stanowią większość zasobów staropolskich? Ograniczyć się wyłącznie do tych, które są wydane, czy opracować wszystkie dostępne? Należy pamiętać, że współczesne wydania to niejednokrotnie edycje anachroniczne lub fragmentaryczne, które nie spełniają filologicznych standardów. Zadaniem twórców korpusu będzie zatem opracowanie transkrypcji, a w wielu wypadkach również gruntowne zweryfikowanie dostępnych transliteracji.

Podsumowując spostrzeżenia na temat czwartego i piątego znacznika, warto podkreślić, że korpus diachroniczny powinien wyróżniać cztery kanały: rękopis, inkunabuł, książkę i wydanie elektroniczne, co umożliwia wyseparowanie z bazy danych tekstów pochodzących wyłącznie z jednego kanału. Źródło cytowania zależne będzie od współczesnego stanu edycji tekstów oraz odpowiedniego opracowania transliteracji i transkrypcji poszczególnych tekstów. Może się okazać, że tylko teksty ciągłe będą cytowane w transliteracji i transkrypcji ze znanych wydań, a teksty nieciągłe będą musiały być przygotowane przez zespół korpusowy. W takim wypadku źródłem cytowania będzie opracowanie autorskie lub zespołowe.

5. Klasyfikacja tekstu

Przyjęte w NKJP tematyczna klasyfikacja tekstów i Uniwersalna Klasyfikacja Dziesiętna tekstów zostały zmodyfikowane i dostosowane do potrzeb korpusów dawnych. Znalazło to odzwierciedlenie w KorBie. Zastosowane w nim rozwiązania będą z pewnością wykorzystane w korpusie tekstów polskich do 1500 roku, jednak z pewnymi uzupełnieniami wynikającymi ze specyfiki najdawniejszych zabytków języka polskiego. Przypomnijmy – polskie teksty nieciągłe wymagają uzupełnienia w postaci kontekstu w języku obcym, w którym zostały dopisane. Nie można ustalić klasyfikacji tematycznej polskich glos, ponieważ nie są one tematycznie spójne. Można natomiast ustalić pewne fakty, na przykład że glosy zapisano w łacińskim kazaniu, tłumaczeniu fragmentu biblijnego, traktacie filozoficznym czy bajce. Tego typu informacje pozwalają sklasyfikować tekst podstawowy, a pośrednio – przypisane do niego polskie teksty nieciągłe; można wtedy mówić o klasyfikacji „wtórnej". Jest to propozycja tymczasowego rozwiązania, ponieważ pozostawienie tekstów nieciągłych jako całościowej grupy niesklasyfikowanej tematycznie, gatunkowo i rodzajowo nie pozwoliłoby na prawidłowe wykorzystanie korpusu do badań językowych. Można dostrzec pewne walory takiego rozstrzygnięcia. „Wtórna" klasyfikacja umożliwi zbadanie słownictwa polskiego występującego wyłącznie w jednym rodzaju tekstów, na przykład w kazaniach lub traktatach filozoficznych czy prawniczych. Tego typu badania nie miały dotąd miejsca z uwagi na duże rozproszenie tekstów oraz fragmentaryczność dostępnych wydań.

Rozważania na temat zastosowania ogólnej lub szczegółowej klasyfikacji tekstów staropolskich wymagają uprzedniego zapoznania się z całym jego zasobem. Rozwiązania przyjęte w korpusach synchronicznych mogą być wykorzystane tylko częściowo, gdy mamy

do czynienia z tekstami jednorodnymi: pieśniami, modlitwami, kazaniami, statutami czy listami, jednak większość zachowanych tekstów dawnych należy do pogranicza tematycznego, rodzajowego i gatunkowego. Przykładowo, pojedyncze obiekty korpusu mogą być przypisane jednocześnie do kilku kręgów tematycznych: prawniczych, sądowych, administracyjnych i gospodarczych (np. wyroki sądów kościelnych) lub religijnych, mitologicznych i filozoficznych (np. kazania i traktaty). Podobne rozstrzygnięcia wielokrotnego wyboru trzeba będzie przyjąć przy znacznikach: „język tekstu" (np. *Psałterz floriański*), „rodzaj tekstu" (podobnie jak w KorBie trzeba będzie wydzielić *Biblię* oraz tłumaczenia fragmentów biblijnych), „gatunek tekstu" (np. notatki akademickie zawierające fragmenty kazań, traktatów oraz poradników).

Podsumowując rozważania na temat klasyfikacji tekstów staropolskich, należy podkreślić złożoną strukturę zachowanych zabytków językowych, która wymusza na twórcach korpusu odrębne podejście do tekstów ciągłych i nieciągłych. W przyjętych dotąd rozwiązaniach nie uwzględniono klasyfikacji glos, komentarzy, notatek dwujęzycznych czy fragmentów zdań zapisanych na okładkach manuskryptów. Dostępne opracowania filologiczne oraz bibliologiczne nie pozwalają też jednoznacznie określić rodzaju i gatunku średniowiecznych zabytków. Sklasyfikowanie dużej części tekstów będzie wymagało wprowadzenia pól wielokrotnego wyboru i uznania, że nie są to obiekty jednorodne.

Stworzenie korpusu najdawniejszych zabytków języka polskiego może wydawać się proste, jeżeli bierzemy pod uwagę zasób tekstów znanych z edukacji szkolnej lub akademickiej oraz istniejące narzędzia korpusowe umożliwiające opisanie i anotowanie tekstów. Dla historyków języka, literaturoznawców, historyków czy źródłoznawców zadanie to wydaje się o wiele bardziej skomplikowane, ponieważ nie istnieje aktualny spis wszystkich zabytków języka polskiego do 1500 roku (*Opis źródeł Słownika staropolskiego* zawiera spis wydań dużej części rozpoznanych zabytków, ale nie identyfikuje wszystkich tekstów w rozumieniu korpusowym), nie posiadamy elementarnych informacji (data, autor, tytuł) na temat zachowanych (w różnej postaci) tekstów, a rozproszone w wielu miejscach i w różnej formie transliteracje i transkrypcje tekstów nie spełniają wymogów edytorskich pozwalających na zamieszczenie ich w bazie danych. W tej sytuacji jedynym wyjściem jest stopniowe tworzenie korpusu polskich tekstów do 1500 roku oraz uzupełnianie i weryfikowanie dostępnych już danych na podstawie bezpośrednich badań kodykologicznych, a także najnowszych badań źródłoznawczych i bibliologicznych. Tylko wtedy będzie można przedstawić zbiór uporządkowanych i opracowanych według określonych zasad tekstów, który posłuży do celów naukowych i edukacyjnych.

Źródła

KorBa – *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)* [online: https://korba.edu.pl/query_corpus/; data dostępu: 7.10.2019].

Korpus polszczyzny 1830–1918 [online: http://www.f19.uw.edu.pl/; data dostępu: 7.10.2019].

Korpus polszczyzny XVI wieku [online: http://spxvi.edu.pl/korpus/; data dostępu: 7.10.2019].

Korpus tekstów staropolskich [online: https://ijp.pan.pl/publikacje-i-materialy/zasoby/korpus-tekstow-staropolskich/; data dostępu: 7.10.2019].

NKJP – *Narodowy Korpus Języka Polskiego* [online: http://nkjp.pl/; data dostępu: 7.10.2019].

OźSstp 2005 – Twardzik W., red. we współpracy z E. Deptuchową i L. Szelachowską-Winiarzową. Oprac. Belcarzowa E., Deptuchowa E., Frodyma M., Kalicka K., Leńczuk M., Szelachowska-Winiarzowa L., Wójcikowa Z., 2005: *Opis źródeł Słownika staropolskiego*. Kraków.

Słowniki

SJP PWN – *Słownik języka polskiego PWN*, 2019 [online: https://sjp.pwn.pl/; data dostępu: 7.10.2019].

WSJP PAN – Żmigrodzki P., red., 2008: *Wielki słownik języka polskiego PAN* [online: https://wsjp.pl/index.php?pwh=0; data dostępu: 7.10.2019].

Literatura

Bracha K., 2007: *Nauczanie kaznodziejskie w Polsce późnego średniowiecza. Sermones dominicales et festivales z tzw. kolekcji Piotra z Miłosławia*. Kielce.

Gruszczyński W., Adamiec D., Ogrodniczuk M., 2013: *Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 r.) – prezentacja projektu badawczego*. „Polonica" XXXIII, s. 309–316.

Gruszczyński W., Bronikowska R., 2018: *Tworzenie korpusu tekstów dawnych a korpusu tekstów współczesnych: różnice teoretyczne i warsztatowe (na przykładzie Korpusu tekstów polskich XVII–XVIII wieku)* [online: https://korba.edu.pl/static/documents/publikacje/2015_gruszczynski_bronikowska.pdf; data dostępu: 7.10.2019].

Klapper M., Kołodziej D., 2015: *Elektroniczny Tezaurus Rozproszonego Słownictwa Staropolskiego do 1500 roku. Perspektywy i problemy*. „Polonica" XXXV, s. 87–101.

Kołodziej D., Klapper M., 2014: *Elektroniczny Korpus Tekstów Staropolskich do 1500 r. Perspektywy i problemy*. „Prace Filologiczne" LXV, s. 203–212.

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

Przepiórkowski i in., 2012: Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B., red.: *Narodowy Korpus Języka Polskiego*. Warszawa.

Wolny J., 1961: *Łaciński zbiór kazań Peregryna z Opola i ich związki z tzw. „Kazaniami gnieźnieńskimi"*. W: Lewański J., red.: *Średniowiecze. Studia o kulturze*. T. 1. Warszawa, s. 171–238.

Wydra W., Rzepka W.R., 2004: *Chrestomatia staropolska. Teksty do roku 1543*. Wyd. III. Wrocław.