DOI: 10.33896/PorJ.2020.8.2
Krzysztof Opaliński
(e-mail: krzysztof.opalinski@ibl.waw.pl)
ORCID: 0000-0001-8775-4953
Patrycja Potoniec
(e-mail: patrycja.potoniec@ibl.waw.pl)
ORCID: 0000-0002-5911-5422
(Instytut Badań Literackich Polskiej Akademii Nauk, Warszawa)

KORPUS POLSZCZYZNY XVI WIEKU

1. ZAŁOŻENIA

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

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

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

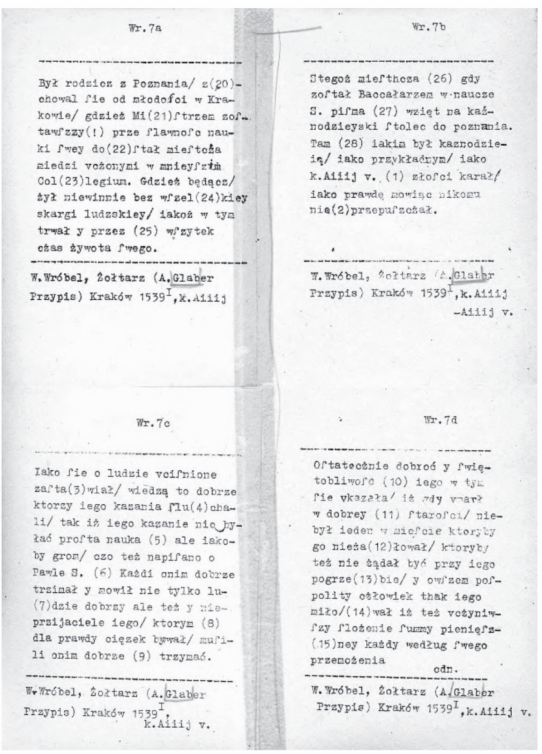
Najoczywistszym rozwiązaniem wydawało się jakiekolwiek utrwalenie materiałów w postaci cyfrowej. Dzięki możliwościom, które daje komputerowe przetwarzanie i zapisywanie danych, droga rozwoju badań językoznawczych (zwłaszcza na dużych zasobach danych) jest od dawna wytyczona: „nowoczesne nośniki informacji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

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

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

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

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



Wr.7a

Był rodzien z Poznania/ z(20)-chował ſie od młodoſci w Krakowie/ gdzieś Mi(21)ſtrzem ſoſ-tawſzy(!) prze ſławnoſc nauki ſwey do(22)ſtał mieſtcza miedzi vczonymi w znieyſzią Col(23)legium. Gdzieś będącz/ żył niewinnie bez wſzel(24)kiey skargi ludzkiey/ iakoż w tym trwał y przez (25) wſzytek czas żywota ſwego.

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

Wr.7b

Stegoż mieſtnecza (26) gdy zoſtał Baccałarzem w nauce S. piſma (27) wziął na kaſnodzieyski ſtolec do poznania. Tam (28) iakim był kaznodzieią/ iako przykładnym/ iako k.Aiiij v. (1) złoſci karaż/ iako prawdę mowiąc nikomu nie(2)przepuſczał.

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

Wr.7c

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

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

Wr.7d

Oſtatecznie dobroć y ſwiętobliwoſć (10) iego w tym ſie vkazała/ iż gdy vmarł w dobrey (11) ſtaroſci/ nie-był ieden w mieſcie ktoryby go nieła(12)ował/ ktoryby/ też nie żądał być przy iego pogrze(13)bie/ y owſzem poſpolity człowiek thak iego miło/(14)wał iż też vczyniwſzy złożenie ſummy pieniężz-(15)ney każdy według ſwego przenoczenia odn.

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

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

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

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

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

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

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

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

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

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

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

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

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

a spodziewany wynik stanowić miało:

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

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

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

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

1.1. Baza źródłowa

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

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

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

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

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

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

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

---

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

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

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

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

1.2. Baza materiałowa

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

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

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

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

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

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

1.3. Zestaw klas i form gramatycznych

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

1.4. Hasłowanie

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

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

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

---

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

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

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

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

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

2. REALIZACJA

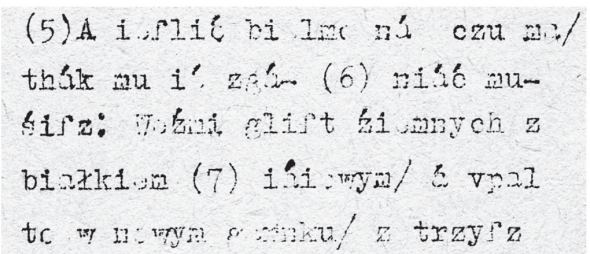
Przy opracowaniu koncepcji korpusu tekstów XVI-wiecznych założenia przyjęte kilka dekad temu dla materiału SPXVI należało zweryfikować na podstawie doświadczeń w pracach redakcyjnych oraz dostosować je do możliwości nowych technologii. Z racji ogromu przedsięwzięcia prace należało podzielić na kilka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

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

2.1. Baza tekstów

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

Ilustracja 3. Fragment matrycy tekstowej egzemplarza archiwalnego.



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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

2.2. Zalążek korpusu

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

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

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

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

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

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

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

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

3. PLANY

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

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

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

Bibliografia

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Corpus of the 16th-century Polish language

Summary

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

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

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

Trans. Monika Czarnecka