WŁODZIMIERZ GRUSZCZYŃSKI
DOROTA ADAMIEC
MACIEJ OGRODNICZUK

Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 roku) — prezentacja projektu badawczego

Projekt „Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 roku)”[1] został przygotowany przez Pracownię Historii Języka Polskiego XVII i XVIII wieku Instytutu Języka Polskiego PAN we współpracy z Zespołem Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN. Projekt uzyskał finansowanie ze środków Narodowego Programu Rozwoju Humanistyki na lata 2013–2017. Kierownikiem zespołu realizującego prace jest Włodzimierz Gruszczyński.

Projekt ma na celu zbudowanie i udostępnienie korpusu tekstów polskich z XVII i XVIII wieku (do roku 1772) oraz opracowanie narzędzi do dokonywania wyrafinowanych operacji na tym korpusie (wyszukiwanie, filtrowanie, tworzenie zestawień statystycznych itp.)

Elektroniczny korpus tekstów stanowi niezbędne uzupełnienie nowoczesnego warsztatu badawczego współczesnego lingwisty. Prace historycznojęzykowe nad przekrojowym opisem słownictwa polskiego z XVII i XVIII wieku rozpoczęły się w połowie XX wieku. Utworzona została wówczas Pracownia Historii Języka Polskiego XVII i XVIII wieku, której zadaniem było zgromadzenie kartoteki wyrazowej i przygotowanie słownika języka okresu średniopolskiego (oprócz opracowywanej osobno polszczyzny XVI wieku). Obecnie kartoteka zawiera około 2,8 mln kart materiałowych. Ma ona również formę elektroniczną, ponieważ została zdigitalizowana w ramach projektu RCIN[2]. W latach 1999–2004 wydano w formie papierowej pierwszy tom „Słownika języka polskiego XVII i 1. połowy XVIII w.” (w skrócie: SXVII), w którym znalazły się materiały wstępne (obejmujące zasady opracowania, dane bibliograficzne źródeł słownika, historię słownika) i artykuły hasłowe na literę A. W 2005 roku rozpoczęły się prace nad elektroniczną formą SXVII i obecnie kolejne artykuły hasłowe są publikowane wyłącznie w tej postaci[3].

[1] Skrócona nazwa: korpus barokowy.
[2] Repozytorium Cyfrowe Instytutów Naukowych — wielodziedzinowa baza danych udostępniająca publikacje naukowe; projekt finansowany z Programu Operacyjnego Innowacyjna Gospodarka. Kartoteka słownika jest dostępna w Internecie pod adresem: http://rcin.org.pl/dlibra/publication?id=20029.
[3] Słownik jest dostępny w Internecie pod adresem http://sxvii.pl.

Projekt ma charakter heterogeniczny, służy unowocześnieniu metod badań historycznojęzykowych i włącza te badania w nurt językoznawstwa korpusowego. Obecne doświadczenia w tworzeniu korpusów tekstowych języka polskiego ograniczają się w zasadzie do korpusów tekstów współczesnych. Jak wiadomo, prace nad nimi były przez długi czas rozproszone. Ostatecznie jednak w latach 2007–2012 dzięki wspólnej inicjatywie Instytutu Podstaw Informatyki PAN (który był koordynatorem projektu), Instytutu Języka Polskiego PAN, Wydawnictwa Naukowego PWN oraz Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego, zrealizowanej jako projekt badawczy rozwojowy Ministerstwa Nauki i Szkolnictwa Wyższego, powstał Narodowy Korpus Języka Polskiego (skrót: NKJP)[4].

Omawiany projekt ma stanowić chronologiczne i metodologiczne rozszerzenie NKJP. Podobnie jak w NKJP przyjmujemy m.in. szerokie rozumienie pojęcia „korpus tekstów", nie ograniczając się do zbioru tekstów w określonym formacie zgromadzonych na nośniku elektronicznym. Plan prac obejmuje również stworzenie oprogramowania umożliwiającego przeszukiwanie tego zbioru, tworzenie konkordancji żądanych form, tworzenie indeksów występujących w nim słów oraz otrzymywanie informacji o frekwencji słów tekstowych i leksemów, a także o dokładnej lokalizacji poszczególnych słów w tekstach źródłowych.

Korpus barokowy będzie pierwszym polskim korpusem słownictwa historycznego (w powyższym rozumieniu) i zarazem pierwszym krokiem w kierunku zbudowania korpusu narodowego zawierającego teksty ze wszystkich okresów istnienia języka polskiego (w wersji pisanej). Takie założenia sprawiają, że korpus barokowy musi wykorzystywać narzędzia informatyczne zastosowane w NKJP. Konieczne będzie jednak ich zaadaptowanie do specyficznych wymagań historycznego materiału językowego. W tym zakresie tworzenie korpusu przyczyni się też do rozwoju inżynierii lingwistycznej w Polsce.

Objętość „Elektronicznego korpusu tekstów polskich z XVII i XVIII wieku" została zaplanowana na 12 mln segmentów. Nie jest to duży korpus w zestawieniu z NKJP, który w wariancie zrównoważonym obejmuje 300 mln segmentów. Jednak w porównaniu z innymi korpusami historycznymi zakładana objętość korpusu barokowego pozostaje relatywnie duża, np. szwedzkie korpusy tekstów z lat 1520–1850, czyli z okresu zwanego w historii języka szwedzkiego Nysvenska 'nowy szwedzki' zawierają tylko około 1,2 mln segmentów[5].

[4] Korpus jest dostępny w Internecie pod adresem: http://nkjp.pl.
[5] Jest to łączna wielkość czterech podkorpusów różnorakich tekstów z okresu Nysvenska oraz korpusu tekstów Carla Michaela Bellmana (1740–179); por. http://spraakbanken.gu.se/swe/resurser/corpus.

Koniecznym warunkiem funkcjonowania korpusu jako bazy danych wyposażonej w narzędzia informatyczne jest wprowadzenie w tekstach włączanych do korpusu różnego typu znaczników. W projekcie zastosowane zostanie znakowanie (anotacja, tagowanie) w formacie XML zgodnym ze standardem TEI[6] w wersji rozszerzonej w trakcie prac nad NKJP[7] i zaadaptowanej do reprezentacji tekstów średniopolskich. Opracowany zostanie także specjalny formularz służący do wprowadzania tekstów do korpusu, dzięki czemu równocześnie z przepisywaniem tekstu możliwa stanie się jego anotacja w różnych zakresach.

Podstawowe informacje dotyczące kształtu tekstu będą uwzględnione przez znakowanie strukturalne, które umożliwia dokładną lokalizację cytatu. Znaczniki strukturalne służą zgromadzeniu informacji o paginacji, rozdziałach, częściach, księgach itp. Oznakowane zostaną również granice tekstu zasadniczego i różnorakich tekstów pobocznych, jak notki marginesowe, przypisy itp., podpisy pod ilustracjami. W okresie baroku te elementy bywały niezwykle rozbudowane i stanowiły znaczącą część całości tekstu.

Tagowanie zewnątrztekstowe obejmie znaczniki socjolingwistyczne i stylistyczno-genologiczne, dzięki temu w korpusie dostępne będą informacje o autorach, wydawcach, tłumaczach oraz o reprezentowanych gatunkach i stylach literackich. Można więc będzie wyszukiwać jednostki języka z tekstów o z góry ustalonych cechach (np. tylko z tekstów pisanych prozą, których autorzy pochodzili z Kresów Wschodnich).

W ramach znakowania językowego anotowane zostaną wszystkie wtręty obce. Jest to zabieg szczególnie ważny wobec dużego nasycenia ówczesnych tekstów polskich łacińskimi (ale też innymi) cytatami. Dodatkowa informacja szczegółowa o tym, z jakiego języka pochodzą obce elementy, pozwoli również na ich uporządkowanie i umożliwi ewentualną analizę także niepolskiej leksyki występującej w korpusie barokowym. Jednocześnie możliwe będzie przeszukiwanie korpusu z wyłączeniem wszelkich zawartych w nim segmentów obcojęzycznych. Projekt przewiduje również wprowadzenie znakowania typograficznego, obejmującego tagowaniem ligatury, skróty, być może kroje czcionek itp. W ten sposób korpus będzie gromadził informacje dotyczące ówczesnego edytorstwa.

Omówione dotychczas płaszczyzny znakowania można określić mianem pomocniczych. Podstawowym zamierzeniem autorów projektu pozostaje wprowadzenie tagowania morfosyntaktycznego i leksykalnego, które umożliwią funkcjonowanie korpusu barokowego jako nowoczesnego lingwistycznego narzędzia badawczego.

[6] Por. Burnard i Bauman, 2007. Najnowsza wersja elektroniczna stale aktualizowanych wytycznych TEI znajduje się pod adresem http://www.tei-c.org/release/doc/tei-p5-doc/en/html.
[7] Por. rozdział 10 — Znakowanie XML (Przepiórkowski 2012: 169–193).

Zgodnie z założeniami projektu niewielki podkorpus (około 0,5 mln) zostanie oznakowany morfosyntaktycznie przez językoznawców, zaś reszta korpusu zostanie oznakowana automatycznie za pomocą stworzonego w ramach projektu tagera, czyli narzędzia informatycznego do automatycznej anotacji morfoskładniowej. Rozważane jest także wprowadzenie informacji fleksyjnej do wszystkich haseł z „Elektronicznego słownika języka polskiego XVII i XVIII w." na podstawie kartoteki i potraktowanie tej informacji jako swoistego słownika gramatycznego (fleksyjnego) barokowej polszczyzny. Słownik ten byłby podstawą do stworzenia analizatora morfologicznego dla polszczyzny XVII-wiecznej (analogicznie jak „Słownik gramatyczny języka polskiego" jest podstawą dla analizatora Morfeusz, obsługującego NKJP[8]). Opracowanie systemu znaczników morfologiczno-składniowych wiąże się oczywiście z napisaniem swego rodzaju sformalizowanej gramatyki opisowej XVII-wiecznej polszczyzny (ze szczególnym uwzględnieniem fleksji i elementów składni)[9]. Wiarygodność morfosyntaktycznego tagowania automatycznego będzie zapewne mniejsza niż w wypadku NKJP, ze względu na znaczny stopień skomplikowania i braku stabilizacji gramatycznej w języku średniopolskim. Jednak stworzone w korpusie barokowym narzędzia informatyczne mogą zostać w przyszłości udoskonalone.

Oznakowanie leksykalne, czyli lematyzacja, pozwoli na powiązanie słów tekstowych z odpowiednim leksemem. Dzięki temu mechanizmowi będzie możliwe wyszukiwanie w korpusie nie tylko pojedynczych form (słów tekstowych), ale wszystkich wystąpień form danego leksemu. Jeśli okaże się, że analizator morfologiczny będzie wystarczająco wydajny, problem znakowania leksykalnego będzie ograniczony do sytuacji najtrudniejszych, jednostkowych, spowodowanych daleko idącą wariancją ortograficzną charakterystyczną dla tekstów barokowych.

Pierwszym krokiem w planowaniu korpusu tekstów jest określenie kryteriów doboru włączanych tekstów. W korpusie barokowym mają zostać zgromadzone teksty charakteryzujące się zróżnicowaniem gatunkowym właściwym epoce. Korpus obejmie teksty użytkowe (np.: akta sejmikowe, księgi sądowe, inwentarze, poradniki medyczne i gospodarskie, podręczniki, kalendarze, listy, czasopisma, druki ulotne, teksty specjalistyczne), teksty literackie (np.: diariusze, relacje, historie, herbarze, dyskursy,

[8] Por.: http://sgjp.pl/morfeusz.

[9] Podstawą tej gramatyki będą istniejące już hasła w SXVII oraz ustalenia zawarte w monografii pod red. D. Ostaszewskiej (2002).

opowieści, dialogi, dramaty, sielanki, satyry) oraz teksty religijne (np. teksty biblijne, kazania).

Teksty do korpusu pochodzić będą z różnych typów źródeł: z rękopisów, ze starodruków (składanych różnymi odmianami czcionki gotyckiej lub antykwą) oraz z późniejszych (XIX-, XX- i XXI-wiecznych) wydań tekstów napisanych w XVII i XVIII wieku (składanych tradycyjnie bądź na komputerze albo opublikowanych w Internecie). Do korpusu zostaną włączone również teksty dokładnie transliterowane dostępne w postaci elektronicznej, a w szczególności korpus polskiej części międzynarodowego projektu IMPACT[10] zawierający około 1,8 mln segmentów.

Rękopisy z epoki będą przepisywane na komputerze tylko w stosunkowo niewielkiej liczbie (wybrane zostaną albo stosunkowo krótkie, a jednocześnie ważne z językowego punktu widzenia, albo takie, które z jakichś powodów zasługują na publikację, choćby internetową; publikacja taka będzie dodatkowym efektem prac nad korpusem). Ograniczony zakres możliwości wykorzystania rękopisów wynika z trudności we wprowadzeniu tych tekstów na nośnik, gdyż wymagają one czasochłonnego przepisywania przez osobę o wysokich kwalifikacjach. Ewentualne rozszerzenie udziału rękopisów w korpusie barokowym może wyniknąć z wykorzystania rękopisów transliterowanych w pracach magisterskich i doktorskich z zakresu edytorstwa lub historii języka polskiego. Można też rozważyć włączenie do korpusu współczesnych wydań rękopisów po ich porównaniu z oryginałem.

Starodruki w założeniach projektu traktowane są jako podstawowy typ źródeł zgromadzony w korpusie. Starodruki składane gotykiem (lub częściowo gotykiem, a częściowo antykwą) będą wprowadzane na nośnik elektroniczny w postaci tekstowej, czyli przepisane i włączone do korpusu. Przepisywane będą przede wszystkim teksty udostępnione w bibliotekach cyfrowych. Starodruki składane antykwą, a także późniejsze (XIX-, XX- i XXI-wieczne) wydania tekstów XVII- i XVIII-wiecznych składane metodą tradycyjną zostaną wprowadzone na nośnik (w zależności od jakości druku) albo za pomocą programów do automatycznego rozpoznawania kształtów (OCR), albo za pomocą przepisywania (te gorszej jakości). Podjęte będą także próby zastosowania OCR do rozpoznawania tekstów składanych gotykiem[11].

[10] Informacje na temat polskiej części projektu IMPACT (IMProving ACcess to Text) realizowanego w latach 2010–2012 zob.: http://www.man.poznan.pl/online/pl/projekty/117/IMPACT.html.

[11] Wykorzystane zostaną wyniki badań przeprowadzonych w ramach projektu IMPACT opublikowane w formie raportu, por.: M. Heliński, M. Kmieciak, T. Parkoła, Report on the comparison of Tesseract and ABBYY FineReader OCR engines http://lib.psnc.pl/dlibra/docmetadata?id=358&from=publication&showContent=true.

Wydania XIX-wieczne i późniejsze włączymy do korpusu tylko w wyjątkowych wypadkach. Po pierwsze, jeśli są dostępne w postaci tekstu na nośniku elektronicznym i możliwe jest ich uzgodnienie z oryginałem (rękopisem lub starodrukiem). Po drugie, jeśli te wydania są dostępne w postaci papierowej umożliwiającej wykonanie skanu i OCR, a brak ich podstaw z epoki, gdyż rękopisy lub starodruki zaginęły.

Pozornie najwygodniejszym źródłem tekstów do korpusu są publikacje wydawane ostatnio i przygotowywane do druku na komputerze. Pierwszym warunkiem ich wprowadzenia do korpusu jest konieczność uzyskania praw do wykorzystania, zgodnie z obowiązującym prawem autorskim. Niektórzy wydawcy mogą pod pewnymi warunkami udostępnić elektroniczne wersje tekstów opublikowanych wcześniej w postaci książkowej. Takie teksty wymagają jednak, po pierwsze, konwersji z formatu wydawniczego na format tekstowy służący do reprezentacji tekstów w korpusie (z uwzględnieniem konieczności wprowadzenia tagów strukturalnych tekstowych, najlepiej na podstawie oryginału z epoki, o ile jest dostępny), po drugie zaś muszą być wprowadzone do nich zmiany przywracające im cechy językowe (także ortograficzne) oryginałów z epoki, ponieważ w korpusie możliwie wiele tekstów powinno być dostępnych w transliteracji (postać transkrybowana może występować opcjonalnie).

Niewielka część tekstów pochodzić będzie z transkrybowanych publikacji internetowych. Tłumaczy się to tym, że większość tekstów dawnych obecnych w Internecie jest z filologicznego punktu widzenia mało wiarygodna — nie sposób ustalić, która edycja stanowi ich podstawę, jakie są zasady transkrypcji itp. Teksty z Internetu będą jednak wyzyskiwane jako podstawa do wersji transliterowanych, zakładamy bowiem, że szybsze i efektywniejsze będzie uzyskiwanie postaci transliterowanej tekstu, w drodze dokonywania zmian w transkrypcji na podstawie oryginału niż w drodze przepisywania oryginału (w taki sposób zamierzamy doprowadzić do transliterowanej wersji obszernych fragmentów pierwszego wydania „Biblii gdańskiej” z 1632 roku).

Zasadniczy problem przy konstruowaniu „Elektronicznego korpusu tekstów polskich z XVII i XVIII wieku” stanowi kwestia zrównoważonego doboru źródeł do tego korpusu. Za korpus zrównoważony uważa się taki, którego budowa charakteryzuje się dbałością o to, „by żaden ze składników na żadnym z poziomów nie dominował nad innymi” (Przepiórkowski 2012: 26). Realizacja tego słusznego postulatu w wypadku tworzenia korpusu tekstów historycznych trafia na znacznie więcej trudności zarówno teoretycznych, jak i praktycznych niż w wypadku korpusu tekstów współczesnych. Oczywistym i nieusuwalnym problemem pozostaje ograniczona wiedza o strukturze zbioru tekstów

funkcjonujących w Polsce w XVII – XVIII wieku. Historyk języka ma zawsze dostęp jedynie do zachowanego fragmentu dawnej rzeczywistości językowej. Można na tej podstawie określać główne tendencje, wskazywać popularne w danym okresie typy i gatunki tekstów. Niektóre formy piśmiennictwa barokowego, jak druki ulotne, gazety, będące ówcześnie rozpowszechnionym środkiem przekazywania informacji, przetrwały do naszych czasów w bardzo ograniczonych ilościach[12]. Z drugiej strony, zrównoważeniu korpusu mogłoby zagrozić włączenie do niego w całości bardzo obszernych tekstów. Wśród barokowych starodruków licznie występują pozycje znaczące kulturowo o wielkich rozmiarach. We wstępnych szacunkach ustaliliśmy, że „Biblia gdańska" liczy około 742 000 segmentów, zaś „Zielnik" Syreniusza to 468 000 segmentów. Gdyby oba teksty zostały włączone w całości, stanowiłyby około 10 procent planowanego korpusu, co prowadziłoby do braku zrównoważenia. Najlepszym rozwiązaniem wydaje się włączanie do korpusu wybranych obszernych fragmentów tych tekstów stanowiących pewne całości (księgi, rozdziały itp.). Gdyby w przyszłości możliwa stała się rozbudowa korpusu ponad obecnie zaplanowane 12 mln segmentów, nietrudno byłoby uzupełnić obszerne teksty.

Cel podstawowy prezentowanego projektu to zebranie jak największej liczby tekstów i umożliwienie ich przeszukiwania za pomocą wyrafinowanych narzędzi informatycznych, a tym samym dokonanie zasadniczej zmiany w metodach opracowywania języka baroku, a w konsekwencji także literatury i kultury tego okresu. Korpus przyczyni się do zintensyfikowania i przyspieszenia prac nad „Elektronicznym słownikiem języka polskiego XVII i 1. połowy XVIII w.".

Dodatkowym celem projektu jest doprowadzenie do wydania drukiem naukowych (krytycznych) edycji wybranych tekstów, które wejdą w skład korpusu. Będą to teksty o charakterze nieliterackim (literackie wydawane są stosunkowo często przez historyków literatury i ambitniejsze wydawnictwa komercyjne), zwłaszcza teksty naukowe, prawne, administracyjne itp.

Planowanych celów nie udałoby się osiągnąć bez współpracy specjalistów reprezentujących różne dziedziny wiedzy, w szczególności: językoznawców (w tym historyków języka, leksykografów i gramatyków specjalizujących się w formalnym opisie polszczyzny) oraz informatyków (w tym zwłaszcza specjalistów od inżynierii lingwistycznej i metod statystycznych).

[12] Pisał o tym m.in. Konrad Zawadzki: „Według ustaleń Alodii Gryczowej liczba zaginionych XVI-wiecznych pozycji wydawniczych sięga 50% tytułów zachowanych. Wydaje się, bez obawy popełnienia błędu, że co najmniej taki sam odsetek strat można przyjąć dla druków XVII i XVIII w., a dla efemeryd prasowych procent ten należy podwyższyć dwukrotnie" (Zawadzki, 2002:40).

Korpus barokowy zostanie opublikowany w Internecie na zasadzie wolnego dostępu podobnie jak NKJP. Publikacja elektroniczna korpusu wspomoże upowszechnienie społeczne wiedzy o dziedzictwie narodowym tej epoki, szczególnie w zakresie poznawania ewolucji języka ojczystego. Korpus będzie nowym narzędziem badawczym przydatnym w różnych dziedzinach humanistyki, np. dla językoznawców, literaturoznawców, kulturoznawców, historyków, socjologów, dziennikarzy.

Bibliografia

Burnard L., Bauman S. (red.), 2007, Guidelines for Electronic Text Encoding and Interchange (TEI P5). The TEI Consortium.

Lewandowska-Tomaszczyk B. (red.), 2005, Podstawy językoznawstwa korpusowego, Łódź.

Ostaszewska D. (red.), 2002, Polszczyzna XVII wieku. Stan i przeobrażenia, Katowice.

Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B. (red.), 2012, Narodowy Korpus Języka Polskiego, Warszawa.

Słownik języka polskiego XVII i 1. połowy XVIII wieku, 1999–2004, red. K. Siekierska, Kraków.

Zawadzki K., 2002, Początki prasy polskiej. Gazety ulotne i seryjne XVI–XVIII wieku, Warszawa 2002.

SUMMARY

Electronic corpus of 17th and 18th century Polish texts (up to 1772) — presentation of a research project

The article introduces a new research project aimed at the creation of an electronic corpus of 17th and 18th century Polish texts — the first historic corpus of Polish. The paper presents linguistic and computational background of the project. The authors also point out the limitations and difficulties related to completion of the task.