

Angelika Peljak-Łapińska

Warszawa

Białoruski N-korpus: W kierunku Białoruskiego Korpusu Narodowego

1. Wprowadzenie

Rozwój lingwistyki korpusowej z każdym rokiem nabiera tempa. Aby analizować język jako całość potrzebne jest szersze spojrzenie – a takie właśnie oferują korpusy językowe – nieporównanie szybsze i bardziej dokładne od ludzkiego mózgu. Zbiory danych różnego typu i różnych rozmiarów tworzy się właściwie we wszystkich ośrodkach badawczych. Wiele krajów może pochwalić się swoim korpusem narodowym – do najlepiej rozwiniętych w Europie należą Brytyjski Korpus Narodowy, Narodowy Korpus Języka Polskiego oraz Narodowy Korpus Języka Rosyjskiego¹. Są to korpusy ogólne, których zadaniem jest reprezentować „w możliwe wyczerpujący sposób wszystkie zjawiska zachodzące w danym języku”².

Ze względu na skomplikowaną sytuację społeczno-polityczną na gruncie białoruskim odnotowano do tej pory raczej małe zainteresowanie lingwistyką korpusową. Mimo że początki tej dziedziny językoznawstwa sięgają lat 60. XX wieku, a pierwszy na świecie korpus o przeznaczeniu ogólnym stworzono ponad 50 lat temu³, to historia lingwistyki korpusowej na Białorusi sięga lat 90. XX wieku, kiedy zaczęto gromadzić pierwsze zbiory tekstów, wyłącznie do celów edukacyjnych. Na początku XXI stulecia rozpoczęto dwa programy – jeden w Białoruskiej Akademii Nauk, drugi na Białoruskim Uniwersytecie Państwowym. Celem pierwszego z projektów było stworzenie wersji demonstracyjnej korpusu, ale wyniki nigdy nie zostały opublikowane. Nie wiadomo, jak skończyło się to przedsięwzięcie. W ramach drugiego projektu powstały dwa korpusy – niestety dostęp do nich miały jedynie małe zespoły badawcze⁴.

¹ E. Hajnicz, *Najbardziej znane korpusy tekstów* [online]
<http://nlp.ipipan.waw.pl/Bib/hajn:11h.pdf> [dostęp: 29.11.2015].

² *Ibidem*, s. 5.

³ *The Brown Corpus. Background and history* [online]
<http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/background.html> [dostęp: 29.11.2015].

⁴ A. Yaskovich, *Praktyczny przewodnik po korpusie języka białoruskiego* [w:] *Praktyczny przewodnik po korpusach języków słowiańskich*, red. M. Hebal-Jeziarska, Warszawa 2014, s. 184–197.

Natomiast w 2010 roku pojawił się pierwszy korpus dostępny dla wszystkich – był to Corpus Albaruthenicum. Jak piszą autorzy projektu, jest to „pierwsza próba stworzenia korpusu specjalistycznego służącego do zapewnienia referencyjnego zbioru danych na użytek językoznawców, nauczycieli, studentów, naukowców i ogółu społeczeństwa”¹. Jak widać z powyższego opisu, Corpus Albaruthenicum zawiera teksty tylko jednego rodzaju, mianowicie naukowe. Nie może służyć zatem do ogólnego opisu języka.

Oprócz opisanego wyżej zbioru danych od 2011 roku rozwija się Rosyjsko-Białoruski i Białorusko-Rosyjski korpus równoległy jako część Narodowego Korpusu Języka Rosyjskiego. W tym przypadku użytkownicy mają do czynienia z zaawansowanymi narzędziami do analizy języka, ale do dyspozycji są jedynie teksty białoruskie przetłumaczone na język rosyjski (czyli pochodzące głównie z literatury pięknej, rzadko z publicystyki) oraz rosyjskie przetłumaczone na białoruski (mogą one służyć głównie analizie przekładu).

Opisane powyżej korpusy języka białoruskiego nie mogą na razie pretendować do miana korpusu narodowego, ponieważ ich rozmiar jest zbyt mały i są ograniczone do jednego lub dwóch stylów. Jednak do formy korpusu narodowego zbliża się powstały zaledwie dwa lata temu Białoruski N-korpus. Do tej pory nie powstała tego typu analiza (częściowo problem ten poruszał Alyaxey Yaskovich), a omawiane zagadnienie jest szczególnie ważne w kontekście białoruskiej lingwistyki korpusowej i rozwoju białoruskiego językoznawstwa w ogóle.

2. Białoruski N-korpus – charakterystyka i omówienie problemów

2.1. Historia korpusu i jego struktura

Mało wiadomo na temat historii Białoruskiego N-korpusu. Pewne informacje można znaleźć na blogu *Беларуская мова*, z którego dowiadujemy się, że korpus powstał w maju 2013 roku. Miał wtedy ok. 15 mln słów (z planowanych 100 mln) i był współtworzony przy udziale wykwalifikowanych językoznawców². Natomiast oficjalna strona korpusu nie podaje żadnych danych dotyczących historii korpusu.

Aż do czerwca 2015 roku było nie wiadomo, ile dokładnie tekstów oraz słów obejmuje cały zbiór danych, ani jakie są procentowe relacje między każdą z części zasobu. Kilka miesięcy temu autorzy opublikowali tabelę przedstawiającą liczbę tekstów i słów z podziałem na poszczególne style oraz z wyróżnieniem tekstów, które nie zostały jeszcze opracowane. Biorąc pod

¹ *Projects* [online] <http://grid.bntu.by/?q=projects> [dostęp: 24.08.2015]; *Corpus Albaruthenicum як частка міжнароднага праекта „BalticGri-II”* [online] <http://innosfera.by/node/822> [dostęp: 29.11.2015].

² *Корпус беларускай мовы* [online] <http://by-mova.livejournal.com/1099892.html> [dostęp: 29.11.2015].

uwagę dane dotyczące liczby słów¹ można obliczyć procentowy udział każdej z części korpusu:

Styl	Procentowy udział w całym zbiorze danych
„urzędowy”	0,007 %
„artystyczny”	3,558 %
„naukowy”	0,142 %
„publicystyczny”	70,148 %
„religijny”	0,008 %
„nieopracowane”	26,137 %

To proste badanie pokazuje kilka interesujących faktów. Po pierwsze, łączna liczba słów w korpusie wynosi nieco ponad 88 mln. Jeśli twórcy zaczęli 2 lata temu z 15 mln i planują 100 mln, to postęp jest widoczny, a autorzy zbliżają się do osiągnięcia planowanej objętości. Analizując skład procentowy Białoruskiego N-korpusu, można zauważyć, że ponad ¼ danych jest opatrzona metadanymi, które niesą przyporządkowane do żadnego stylu. Oprócz tego widoczna jest duża różnica pomiędzy ilością danych z różnych źródeł. Zasadniczą część korpusu stanowią teksty publicystyczne, co jasno pokazuje, że dane nie są zrównoważone; autorzy nie podają wy tłumaczenia takiej dystrybucji tekstów, więc można przypuszczać, że z zebranych zasobów zostanie w przyszłości wydzielony podkorpus zrównoważony.

Aby uporządkować dane w sposób, który odzwierciedla użycie języka przez jego rodowitych użytkowników, potrzebny jest wskaźnik czytelnictwa. Nigdzie w literaturze nie pojawia się informacja na temat badania czytelnictwa na Białorusi. Pewne informacje można czerpać jedynie z ankiet przeprowadzanych od czasu do czasu przez dziennikarzy, np. wśród uczestników wydarzeń kulturalnych, takich jak targi książek². W Polsce działa firma *Polskie Badanie Czytelnictwa*³, która każdego roku sprawdza, co czytają Polacy. Bazując na takim badaniu, przeprowadzonym na szerokim gronie respondentów, można zbilansować korpus tak, żeby odzwierciedlał realne preferencje użytkowników danego języka.

Białoruski N-korpus jest podzielony na pięć głównych części, do których odnośniki znajdują się u góry strony. *Корпус* („Korpus”), *Неапрацаваныя тэксты* („Teksty nieopracowane”), *Граматычная база* („Baza grama-

¹ Słowo *мастацкая* i *мастацкі* znaczą to samo w języku białoruskim, różnica dotyczy rodzaju gramatycznego, dlatego też te dwa style są rozpatrywane jako jeden; *публіцыстычны* jest przymiotnikiem derywowanym od czasownika *публіцыстыка*, dlatego też te dwa style są rozpatrywane jako jeden. Por. informacje na stronie <http://bnkorporus.info/>.

² З. Кустоўскі, *Як шмат і што чытаюць беларусы?* [online] <http://euroradio.fm/yak-shmat-i-shto-chytayuc-belarusy> [dostęp: 29.11.2015].

³ *Polskie Badanie Czytelnictwa* [online] <http://www.pbc.pl/> [dostęp 29.11.2015].

tyczna”), *Спамнаваць* („Pobierz”) i *Пра праект* („O projekcie”). W „Korpusie” i „Nieopracowanych tekstach” możemy wybrać pomiędzy opcjami *Пошук* („Szukaj”), *KWIC* („Słowo kluczowe w kontekście”) i *Класmap* („Klaster”, „Zbitka”). Wydzielenie trzech różnych narzędzi do wyszukiwania pożądanych słów oraz możliwość przeszukiwania nieopracowanych tekstów zostały zaimplementowane dopiero w czerwcu 2015 roku.

2.2. Narzędzia dostępne w korpusie

„Szukaj” oraz „KWIC” (zarówno w „Korpusie”, jak i w „Nieopracowanych tekstach”) oferuje funkcje typowe dla konkordancera, jedyną różnicą jest sposób prezentacji danych. „KWIC”, jak wskazuje sama nazwa, wyświetla słowo kluczowe w kontekście (słowo kluczowe wyświetlane jest na środku linii wraz z kilkoma słowami z lewej i prawej strony), natomiast w oknie „Szukaj” słowo kluczowe wyświetlane jest na początku każdej linii. Dostępnych jest kilka filtrów¹. Istnieje możliwość wyboru autora – wpisanie nawet jednej litery powoduje wyświetlenie listy możliwych nazwisk. Użytkownik może wybrać jeden z pięciu stylów, które były wymienione przy okazji omawiania struktury korpusu („artystyczny”, „publicystyczny”, „naukowy”, „urzędowy”, „religijny”). Wśród każdego ze stylów można dobrać również kilka-kilka-naście gatunków, a w razie niepowodzenia spróbować opcji *іншыя* („inne”). Dodanie stylu religijnego jest jedną z ostatnich zmian, które zaimplementowano w czerwcu 2015 roku. Teksty można też posortować według roku napisania bądź publikacji.

Poniżej dostępnych filtrów widoczne jest okno wyszukiwania. Po wpisaniu słowa można zdecydować o tym, czy wyszukana zostanie jedynie zadana forma, czy też wszystkie formy (*усе словаформы*) oraz jaka część mowy powinna być brana pod uwagę w wyszukiwaniu. W okno zapytań można wpisywać jedynie litery, żadnych znaków (wyjątkiem jest apostrof). Niestety ten fakt może nastroczać pewne problemy, choćby ze względu na niektóre przyimki w języku białoruskim zapisywane przy pomocy dywizu, np. *на-за, з-за*.

W korpusie użytkownik może wyszukiwać również połączenia dwóch lub więcej słów, umożliwia to opcja *дадаць слова* („dodaj słowo”). Do początku czerwca 2015 roku możliwe było szukanie połączenia słów tylko jeśli występowały bezpośrednio obok siebie (*Вызначаны парадак*) lub po prostu w obrębie jednego akapitu (*Любая адлегласць у межах абзаца*). Obecnie dostępna jest trzecia opcja – współwystępowanie w jednym zdaniu (*Любая адлегласць у межах сказа*). Ta możliwość jest najbardziej pomocna w badaniach kolokacji, ponieważ słowa często razem występujące mogą być rozdzielone przez dodatkowe części zdania i wciąż tworzyć kolokację². Wyszukiwanie słów występujących tuż obok siebie daje zbyt zawężone wyniki, za to wyszukiwanie w całym akapicie czyni rezultaty zbyt przypadkowymi.

¹ *Копныс* [online] <http://bnkorporus.info/> [dostęp: 29.11.2015].

² Por. „język, jakim posługuje się na co dzień młody król” zawiera kolokację *posługiwać się językiem*, której poszczególne człony rozdzielone są innymi słowami; źródło: nkjp.pl.

W Narodowym Korpusie Języka Polskiego można określić „okno”, czyli maksymalną odległość między częściami kolokatu, zazwyczaj dla optymalnych wyników przyjmuje się wartość 5 dla tego parametru.

Po ustawieniu wszystkich parametrów następuje wyszukiwanie słowa. Wyszukiwarka wyświetla 100 pierwszych wyników, jeśli u góry strony pojawi się komunikat „Znaleziono więcej niż 100” (*Знойдзена больш за 100*) należy wybrać „Następną stronę” (*Наступная старонка*), aby wyświetlić kolejne 100 wyników. Niestety, ta procedura staje się bardzo uciążliwa, jeśli wyszukiwane słowo pojawia się w korpusie pięć-dziesięć tysięcy razy (oznacza to konieczność klikania w „Następną stronę” nawet tysiąc razy). W oknie pojawiają się wyniki w postaci fragmentów zdań z podświetlonym na czerwono słowem kluczowym. Wybranie opcji „Dokładniej” (*Падрабязней*) pozwala przeczytać całe zdanie, dużo rzadziej akapit.

Jednym z ostatnich udoskonaleń Białoruskiego N-korpusu jest *Класmap*. *Класmap* to narzędzie, które może być rozwinięte tak, żeby służyło jak klasyczny kolokator. Jak dotąd jedyną opcją jest wyszukiwanie słów, które występują bezpośrednio obok słowa kluczowego¹. Możliwy jest wybór liczby słów współwystępujących z prawej lub z lewej strony (*Слоў злева, Слоў справа*). Jak łatwo się domyślić, właściwie tylko wyszukiwanie z kontekstem o rozmiarze jednego słowa przynosi spodziewany efekt. Chociaż trzeba pamiętać, że istnieją kolokacje, które składają się z trzech słów (słowo kluczowe i kolokat składający się z dwóch słów). Jednak, jak wspomniano wyżej, wiele jest wypadków, kiedy poszczególne człony nie występują bezpośrednio obok siebie. Jeśli, podobnie jak w Narodowym Korpusie Języka Polskiego, umożliwiono by wybranie maksymalnego odstępu między słowem kluczowym a jego kolokatem, *Класmap* byłby dużo bardziej przydatnym dla językoznawcy narzędziem.

2.3. Anotacja korpusu

Niezwykle istotnym elementem funkcjonalności korpusu jest jego anotacja. Czysty tekst nie stanowi dobrego materiału do badań, powinien być raczej opisany w sposób zrozumiały dla programu analizującego tekst. Jednym z typów informacji zawartych w Białoruskim N-korpusie są metadane. Wybierając opcję „Dokładniej”, użytkownik otrzymuje informacje o tytule pracy, z której pochodzi wyszukane słowo, imię i nazwisko autora oraz rok publikacji (nie jest on dostępny dla wszystkich tekstów). Autorzy korpusu nie podają żadnej listy źródeł, z których czerpano teksty.

Jak wspomniano wcześniej, korpus jest anotowany gramatycznie. Kliknięcie na dowolne słowo w wynikach wyszukiwania wyświetla informacje z bazy gramatycznej. Na anotację gramatyczną składają się: forma podstawowa (*Lemma*) wraz z zaznaczonym akcentem, określenie części mowy i jej różnych jakości (np. rzeczownik, pospolity, żywotny, forma pełna) oraz

¹ Por. *Класmap* [online]

<http://bnkorpus.info/korpusCluster.html#w0.word=%D0%BC%D1%83%D0%B6%D1%87%D1%8B%D0%BD%D0%B0&w0.allForms> [dostęp: 29.11.2015].

przyporządkowanie do odpowiednich klas gramatycznych (przypadek, rodzaj itd.). Niezwykle interesujący jest fakt, że informacje pojawiają się jako sekwencja kapitalików (pochodzących od słów angielskich) następnie rozwijanych do pełnych nazw białoruskich. W wielu korpusach możliwe jest wyszukiwanie za pomocą wyrażeń regularnych, które pozwalają na komponowanie bardzo złożonych zapytań i dzięki temu, dużo bardziej zaawansowane wyszukiwania. Jak dotąd, w Białoruskim N-korpusie nie jest dostępne takie wyszukiwanie, ale występujące w anotacji kapitaliki można zastosować jako znaczniki do tworzenia wyrażeń regularnych.

Niektóre słowa są anotowane podwójnie, niektóre potrójnie, niektórym przypisanych jest jeszcze więcej anotacji. Taki stan rzeczy sugeruje, że korpus był anotowany automatycznie i niesprawdzony (przynajmniej nie w całości) przez osoby, których zadaniem byłoby jednoznaczne określenie, czy w danym zdaniu słowo występuje jako ta czy inna część mowy; jeśli formy różnych przypadków nie różnią się w zapisie, człowiek może rozpoznać, która występuje w danym przypadku.

Niektóre słowa nie są opatrzone anotacją gramatyczną. Dzieje się tak w przypadku wyrazów nie obecnych w bazie gramatycznej. Są one w większości zapisane *Taraszkewicą* (zwaną też klasycznym wariantem białoruskiej ortografii). Mimo że oficjalnie obowiązuje tzw. *Narkamauka*, wielu Białorusinów (często z powodów ideologicznych) uważa *Taraszkewicę* za poprawną wersję języka białoruskiego i duża część tekstów tworzona jest właśnie w tym wariantcie. W związku z tym analiza w Białoruskim N-korpusie tekstów, np. z Radio Svaboda, jest bardzo problematyczna.

2.4. Zastosowania korpusu

Najbardziej oczywistym zastosowaniem korpusu językowego są badania lingwistyczne. Analiza leksyki wydaje się najłatwiejsza. Korpus umożliwia porównanie częstości występowania słowa kluczowego i jego synonimów w wybranym czasie lub wybranym stylu. W teorii łatwo jest przeprowadzić tego typu badanie w Białoruskim N-korpusie, ale w rzeczywistości nie wszystkie teksty są opisane rokiem wydania, co znacznie utrudnia, np. badanie diachroniczne. Jednakże dane te mogą być łatwo uzupełnione. Dodatkowo, podobnie jak w niektórych korpusach (np. w Narodowym Korpusie Języka Polskiego czy Korpusie Współczesnej Amerykańskiej Angielszczyzny), można by zaimplementować opcję generowania wykresów ilustrujących wyniki – zaoszczędzałoby to czas poświęcony na prezentację efektów pracy.

Jak zauważono wcześniej, słowa znajdujące się w bazie gramatycznej są dobrze anotowane gramatycznie, co stanowi podstawę do wprowadzenia wyrażeń regularnych w wyszukiwaniu. Z kolei wykorzystanie wyrażeń regularnych umożliwia badanie syntaktyki. Za przykład może posłużyć tutaj wyszukiwarka Poliquarp w Narodowym Korpusie Języka Polskiego. Wpisując odpowiednie symbole oznaczające poszczególne klasy gramatyczne i fleksyjne oraz operatory logiczne można utworzyć zapytania dotyczące funkcji

wyrazów w zdaniu czy też zależności między nimi, związku zgody, związku rzędu itd. Sposób tworzenia takich zapytań jest wyczerpująco opisany w dziale „Pomoc” na stronie Narodowego Korpusu Języka Polskiego¹.

Innym możliwym zastosowaniem jest analiza semantyczna. Jednym z podejść jest badanie kolokacji, które „identyfikuje współwystępowanie form językowych w danej próbie języka naturalnego”². Jak zauważono wcześniej, Białoruski N-korpus nie oferuje w pełni funkcjonalnego kolokatora, zatem badanie występowania danego słowa w różnych kontekstach jest bardzo trudne, a analiza dyskursu właściwie niemożliwa, podobnie jak badania z zakresu socjolingwistyki, psycholingwistyki czy stylistyki.

W zakres analizy semantycznej wchodzi badanie podobieństwa semantycznego, które pokazuje stopień podobieństwa między danymi wyrazami przy pomocy reprezentacji każdego z nich w postaci wektora w modelu przestrzeni wektorowej. Kąt pomiędzy badanymi wektorami wskazuje na stopień podobieństwa³. Korzystając z Białoruskiego N-korpusu można zgromadzić dane liczbowe niezbędne do takiego badania, jednakże są one niewystarczające w przypadku neologizmów czy slangu.

3. Białoruski Korpus Narodowy?

Biorąc pod uwagę omówione wyżej aspekty, jasnym jest, że Białoruski N-korpus ma wciąż wiele braków. Należało by uzupełnić metadane tak, żeby umożliwić analizę diachroniczną tekstów. Słowa zapisane przy pomocy *Taraszkewicy* mogłyby zostać uwzględnione w bazie gramatycznej. Niezbędne są także zmiany w interfejsie użytkownika, aby ułatwić korzystanie z narzędzi wyszukiwania. Konieczne jest przeprowadzenie badania czytelności, w celu ustalenia najlepszych proporcji pomiędzy poszczególnymi częściami korpusu. Dla pełnej wiarygodności jako źródła danych, powinny być uzupełnione podstawowe informacje na temat korpusu o informacje o autorach i konsultantach naukowych, listę źródeł danych, informacje na temat ostatnio wprowadzonych i planowanych zmian. Korzystając z dobrej anotacji gramatycznej, można zaimplementować wyszukiwanie za pomocą wyrażeń regularnych, natomiast narzędzie *Класmap* można rozwinąć w pełni funkcjonalny kolokator.

Korzystając regularnie z Białoruskiego N-korpusu łatwo zauważyć, że co kilka miesięcy wprowadzane są pewne korzystne zmiany. Pozwala to domniemać, że korpus będzie się regularnie rozwijał. W przeciwieństwie do *Corpus Albaruthenicum* czy *Równoległego Korpusu Białorusko-Rosyjskiego*

¹ A. Przepiórkowski, J. Wilk, 3. *Język zapytań* [online] <http://nkjp.pl/polijarp/help/plse3.html#x4-50003> [dostęp: 29.11.2015].

² D. Glynn, *Techniques and tools. Corpus methods and statistics for semantics* [w:] D. Glynn, J. Robinson, *Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy*, Amsterdam 2012, s. 307–341.

³ S. Clark, *Vector Space Models of Lexical Meaning* [w:] S. Lapin, C. Fox, *The Handbook of Contemporary Semantic Theory*, 2015, s. 493–522.

i Rosyjsko-Białoruskiego, Białoruski N-korpus posiada zasoby na tyle duże, że wyczerpująco pokazuje niemal wszystkie zjawiska zachodzące w języku białoruskim. Należy przypuszczać, że za kilka lat ten zasób danych przejmie rolę Białoruskiego Korpusu Narodowego.

Angelika Peljak-Łapińska

Belarusian N-korpus: Towards a Belarusian National Corpus

A b s t r a c t

The article examines the current state of the Belarusian N-korpus: its structure, including the range, quality and balance of texts, the type of annotation, data availability, user interface construction, functions available in the search engine as well as documentation for given data. The goal is to answer the question whether Belarusian N-korpus may become the National Corpus. The article argues that even though there are obvious flaws in the discussed resource, it should become the best database of the Belarusian language in a matter of a few years.