Uniwersytet Warszawski
Wydział Polonistyki
Instytut Slawistyki Zachodniej i Południowej
Zespół Lingwistyki Korpusowej Języków Słowiańskich

Praktyczny przewodnik po korpusach języków słowiańskich

Redakcja naukowa

Milena Hebal-Jezierska

Warszawa 2014

Wydano nakładem Wydziału Polonistyki Uniwersytetu Warszawskiego
e-mail: wyd.polon@uw.edu.pl

Publikacja finansowana przez Wydział Polonistyki
Uniwersytetu Warszawskiego
i afiliowana przy Wydziale Polonistyki Uniwersytetu Warszawskiego

Recenzenci
Magdalena Derwojedowa
Zygmunt Saloni

Redakcja techniczna, skład i łamanie
Ignacy Doliński

Korekta językowa
Ignacy Doliński
Jerzy Molas
Aleksandra Wieczorek
Autorzy

Projekt okładki
Łukasz Kamiński

ISBN 978-83-6411-04-4

Druk i oprawa: Zakład Graficzny Uniwersytetu Warszawskiego
Zam. nr 655/2013

Printed in Poland

Spis treści

0. Wstęp

Milena Hebal-Jezierska

Przewodnik po korpusach języków słowiańskich jest pierwszą tego typu książką na rynku wydawniczym. Jego celem jest przekazanie Czytelnikowi nie tylko podstawowych informacji o każdym z korpusów, ale także przybliżenie specyfiki pracy z nimi. Przedstawiamy zatem analizę krytyczną zarówno struktury korpusów, jak i narzędzi wbudowanych w dany korpus; prezentujemy możliwości i ograniczenia poszczególnych elektronicznych zasobów tekstów.

Każdy rozdział (oprócz dwóch ostatnich) jest poświęcony korpusowi (lub korpusom) jednego języka słowiańskiego[1]. Autorzy są badaczami wykorzystującymi korpusy w praktyce. Niektórzy z nich są także twórcami elektronicznych zasobów tekstowych. Każdy z rozdziałów naszego przewodnika ma podobną strukturę, żeby łatwiej można było dokonać analizy porównawczej korpusów, które są zróżnicowane pod wieloma względami. Z jednej strony znajdziemy zatem wysoce zaawansowane technologicznie narzędzia, np. Czeski Korpus Narodowy, Słowacki Korpus Narodowy, korpusy języka słoweńskiego, Rosyjski Korpus Narodowy, z drugiej strony – mniej zaawansowane korpusy języków łużyckich, Chorwacki Korpus Narodowy, korpus języka serbskiego, ukraińskiego, białoruskiego oraz korpusy będące dopiero w fazie planowania (korpus języka macedońskiego). Celem niniejszej książki jest także ułatwienie korzystania z korpusów, dlatego Czytelnik znajdzie w niej także sugestie podstawowych zapytań przydatnych do wyszukania materiału, a także wskazówki pozwalające uniknąć pułapek czyhających na użytkownika korpusu.

Rozdziały są zbudowane według pewnego schematu. Należy jednak pamiętać, że każdy badacz pisze z własnego punktu widzenia, zatem rozdziały siłą rzeczy mają po części charakter subiektywny i wybiórczy. Każdy rozdział zawiera krótki zarys historyczny korpusów danego języka, charakterystykę ich struktury tekstowej, jej zalety i wady. W każdym znajduje się także informacja na temat poziomów anotacji wewnętrznej wraz z analizą problemów z jej stosowaniem w pracy badawczej. Podobnie są opisane narzędzia wbudowane w korpusy. Dodatkowo autorzy rozdziałów prezentują przykładowe zapytania

[1] W książce nie zostały omówione zasoby elektroniczne języka kaszubskiego, bośniackiego i czarnogórskiego.

oraz najważniejsze zastosowania naukowe (gramatyki, słowniki), które powstały przy wykorzystaniu określonego narzędzia. Oceniają także możliwości zastosowania korpusowych metod badawczych przy pracy z konkretną bazą tekstową. Po rozdziałach poświęconym poszczególnym językom umieściliśmy dwa rozdziały dotyczące korpusów równoległych. Pierwszy z nich omawia zagadnienia ogólne oraz pokrótce wprowadza w tajniki korpusu ParaSol oraz Korpusu polsko-rosyjskiego UW, drugi zaś poświęcony jest korpusowi InterCorp.

Książka jest przeznaczona dla wszystkich, których zainteresowania są związane z lingwistyką korpusową oraz dydaktyką języków słowiańskich. Publikacja powinna być przydatna dla językoznawców, tłumaczy, lektorów oraz uczestników kursów studiujących języki słowiańskie. Prócz wersji tradycyjnej zostanie opublikowana w internecie, w postaci pliku w formacie PDF, którego treść będzie w miarę możności aktualizowana.

W tym miejscu chciałabym także podziękować Autorom poszczególnych rozdziałów, Korektorom stylistycznym (dr Aleksandrze Wieczorek, dr. hab. Mirosławowi Bańce, dr. Jerzemu Molasowi i dr. Ignacemu Dolińskiemu, który podjął się też redakcji technicznej). Specjalne podziękowania należą się Recenzentom – prof. Zygmuntowi Saloniemu oraz dr hab. Magdalenie Derwojedowej – za życzliwość i ogromny wkład w powstanie tej książki. Dziękuję także swojej Rodzinie (mężowi, mamie oraz dzieciom – Robertowi i Dagnie) za pomoc i wyrozumiałość.

Warszawa, maj 2013

1. Kilka uwag dotyczących terminologii

Milena Hebal-Jezierska
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Ze względu na wielość terminów oraz różnorodność ich znaczeń stosowanych w literaturze przedmiotu, przedstawiamy poniżej terminy wraz z ich synonimami, które używane są w niniejszej książce. Definicje są prezentowane za publikacjami, w których zostały one umieszczone.

Anotacja zewnętrzna – termin używany w niektórych opracowaniach z zakresu lingwistyki korpusowej. Oznacza metainformację o tekście umieszczonym w korpusie. Zawiera m.in. dane o autorze i/lub tłumaczu tekstu, tytuł, wydawnictwo, rok i miejsce wydania itp. (na podstawie Kocek, Kopřivová, Kučera 2000: 6).

Anotacja wewnętrzna – termin używany w niektórych opracowaniach z zakresu lingwistyki korpusowej. Anotacja wewnętrzna zawiera m.in. tokenizację, lematyzację, anotację morfologiczną, anotację syntaktyczną, semantyczną (na podstawie: Kocek, Kopřivová, Kučera 2000: 7). W Narodowym Korpusie Języka Polskiego używa się w miejsce terminu anotacja morfologiczna wyrażenia anotacja morfosyntaktyczna.

Corpus-assisted analysis – jedna z korpusowych metod badawczych polegająca na jednoczesnej analizie danych korpusowych oraz pozakorpusowych (np. wywiady, ankiety itp.) (na podstawie Baker 2010: 8).

Corpus-based approach – jedna z korpusowych metod badawczych polegająca na korzystaniu z korpusu głównie do wyłożenia, sprawdzenia lub egzemplifikacji teorii oraz opisów, które były sformułowane przed dostępem do wielkich korpusów (na podstawie Tognini-Bonelli 2001: 65).

Corpus-driven approach – jedna z korpusowych metod badawczych polegająca na traktowaniu danych korpusowych jako całości. Wnioski wyciągnięte z analizy muszą odzwierciedlać tylko i wyłącznie poświadczenia korpusowe. Badacz powinien starać się zapomnieć o wiedzy, którą posiada na analizowany temat z innych, pozakorpusowych źródeł (na podstawie Tognini-Bonelli 2001: 84).

Hasłowanie – patrz lematyzacja. W niniejszej książce używamy obu terminów.

Kolokacja – stałe połączenie wyrazowe (Lewandowska-Tomaszczyk 2005: 296).

Kolokat – słowo występujące w sąsiedztwie danego wyrazu, tworzące z nim kolokację.

Kolokator – program służący do znajdowania kolokacji.

Konkordancer – program umożliwiający otrzymanie listy poświadczeń (konkordancji).

Konkordancja – zbiór przykładów użycia danej formy wyrazowej, z których każdy jest przedstawiony w swoim własnym środowisku tekstowym (Lewandowska-Tomaszczyk 2005: 296).

Korpus diachroniczny – korpus tekstów, pochodzących z różnych stadiów rozwoju języka (na podstawie: Kocek, Kopřivová, Kučera 2000: 8).

Korpus historyczny – korpus tekstów niewspółczesnych pochodzących z jednego etapu rozwoju języka (Kocek, Kopřivová, Kučera 2000: 8).

Korpus porównywalny – kolekcja tekstów stworzonych przez rodzimych użytkowników w dwóch lub więcej językach i dobranych według takich samych precyzyjnie zdefiniowanych kryteriów. (Lewandowska-Tomaszczyk 2005: 297).

Korpus referencyjny – Korpus referencyjny ma za zadanie dostarczyć kompletnej informacji o języku. Powinien być dostatecznie duży, żeby mógł reprezentować wszystkie relewantne odmiany języka oraz charakterystyczne słownictwo, a także by mógł być użyty jako podstawa gramatyk, słowników, tezaurusów i innych językowych materiałów źródłowych. Model selekcji zwykle określa liczbę parametrów, które zawierają tak wiele socjolingwistycznych czynników, jak to tylko możliwe, oraz określa proporcję każdego typu tekstu reprezentowanego w korpusie. Duży korpus referencyjny może mieć hierarchicznie uporządkowaną strukturę złożoną z części i subkorpusów. (Sinclair 1996).
Jest to zatem korpus zrównoważony i reprezentatywny. (Gałkowski).

Korpus równoległy – zbiór tekstów oraz odpowiadających im przekładów na jeden lub więcej języków. (Lewandowska-Tomaszczyk 2005: 297).

Korpus synchroniczny – korpus tekstów (najczęściej współczesnych użytkownikowi) pochodzących z jednego stadium rozwoju języka.

Korpus uczniowski – zbiór danych językowych pochodzących od osób uczących się danego języka. W zależności od korpusu dane te mogą zostać pozyskane od osób przyswajających język obcy lub uczących się języka ojczystego.

KWIC (Key Word in Context) – najpopularniejsza forma konkordancji, w której wyszukiwany wyraz jest przedstawiony w otaczającym go kontekście z lewej i prawej strony na środku generowanej listy (Lewandowska-Tomaszczyk 2005: 297).

Lematyzacja – przyporządkowanie wyrazowi jego formy podstawowej, tzw. lematu (Kocek, Kopřivová, Kučera 2000: 28). W wielu polskich pracach używa

się w tym znaczeniu także terminu *hasłowanie* (por. Przepiórkowski, Bańko, Górski, Lewandowska-Tomaszczyk 2012).

Lemat (synonimicznie: forma hasłowa, forma podstawowa, forma kanoniczna). Przykładem lematu dla wyrazu *okna* jest *okno*, dla *tnie* – *ciąć*.

Pozycja – patrz token.

Segment – patrz token.

Tag (synonimicznie: znacznik) – patrz tagowanie.

Tag administracyjny (synonimicznie: znacznik administracyjny) – patrz tagowanie.

Tag lingwistyczny (synonimicznie: znacznik lingwistyczny) – patrz tagowanie.

Tag strukturalny (synonimicznie: znacznik strukturalny) – patrz tagowanie.

Tager – program służący do tagowania.

Tagowanie (synonimicznie: znakowanie) – proces, w czasie którego tekstom wchodzącym do korpusu przypisuje się różnorakie informacje. Informacje te są formalnie wyrażone za pomocą tagów (znaczników) administracyjnych, lingwistycznych lub strukturalnych. Przykładem **tagów administracyjnych** są informacje o pochodzeniu, autorstwie, typie i źródle tekstu. **Znaczniki strukturalne** informują o budowie tekstu, np. o akapitach, początku i końcu tekstu. **Tagi lingwistyczne** zawierają dane na temat określonej kategorii językowej, np. morfologicznej (morfosyntaktycznej), semantycznej czy syntaktycznej. Dla wyrazu *kobiety* w zdaniu *Kobiety są piękne* w Narodowym Korpusie Języka Polskiego tagi morfosyntaktyczne są następujące: subst:nom:pl:f (rzeczownik w mianowniku liczby mnogiej, rodzaj żeński). (Opis terminu powstał na podstawie książki Kocek, Kopřivová, Kučera 2000: 25-28 oraz poświadczenia wyszukanego w korpusie NKJP www.nkjp.pl).

Tagset (synonimicznie: system znaczników) – zbiór wszystkich znaczników (Przepiórkowski, Bańko, Górski, Lewandowska-Tomaszczyk 2012: 59).

Token – termin zapożyczony z nauk informatycznych, w których jest rozumiany jako ciąg znaków ograniczony ustalonymi separatorami, powstały na skutek podziału tekstu, czyli tokenizacji (na podstawie Lipiński 2009). W praktyce najczęściej tokenem jest wyraz ortograficzny, którego początek wyznacza pierwsza jego litera, a koniec litera ostatnia. W wielu korpusach za token jest uważany także znak interpunkcyjny, liczba i symbol. W lingwistyce korpusowej języków słowiańskich używa się wielu leksemów nazywających jednostkę powstałą w wyniku podziału tekstu. W Narodowym Korpusie Języka Polskiego jest nią segment (jego definicja jest węższa niż pojęcie wyrazu ortograficznego), w Czeskim Korpusie Narodowym nazywano ją tokenem, od niedawna zaś nazywa się ją pozycją (oryg. *pozice*), w Słowackim Korpusie Narodowym,

w Chorwackim Korpusie Narodowym – tokenem, w Słoweńskim Korpusie Narodowym zaś tokenem/segmentem.
W polskiej literaturze językoznawczej angielski termin *token* jest tłumaczony jako okaz (por. hasło *okaz* opracowane przez Z. Saloniego w EJO). Warto przy tym zauważyć, że w publikacjach dotyczących lingwistyki korpusowej częściej stosuje się anglicyzm token.

Znacznik – patrz tag.

Znakowanie – patrz tagowanie.

Bibliografia

Baker P. (2010). *Sociolinguistics and Corpus Linguistics*, Edinburgh.

Čermák F. (2006). *Kolokace v lingvistice*. W: *Kolokace*, red. Čermák F., Šulc M., Praha.

EJO = (1993). *Encyklopedia językoznawstwa ogólnego*, red. Polański K., Wrocław – Warszawa – Kraków.

Gałkowski B., *Słowniczek terminów korpusowych*. http://korpusy.net/

Kocek J., Kopřivová M., Kučera K. (eds.) (2000). *Český národní korpus. Úvod a příručka uživatele*, Praha.

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

Lipiński M. (2009). *O technologii i organizacji IT*.
http://www.mariuszlipinski.pl/2009_04_01_archive.html

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

Sinclair J. (1996). *EAGLES, Preliminary recommendations on Corpus Typology*, EAG-TCWG-CTYP/P, Version of May.
http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*, Amsterdam/Philadephia.

2. Praktyczny przewodnik po korpusie języka polskiego

Mirosław Bańko
Uniwersytet Warszawski
Instytut Języka Polskiego

Rafał L. Górski
Polska Akademia Nauk
Instytut Języka Polskiego

Narodowy Korpus Języka Polskiego (dalej NKJP) powstał w latach 2007-2011 w ramach projektu rozwojowego nr R17 003 03, sfinansowanego ze środków przeznaczonych na naukę. W projekcie uczestniczyły cztery ośrodki, które już wcześniej gromadziły na własne potrzeby dane korpusowe i zajmowały się tworzeniem narzędzi do ich przetwarzania: Instytut Podstaw Informatyki PAN, Instytut Języka Polskiego PAN, Uniwersytet Łódzki oraz Wydawnictwo Naukowe PWN. W projekcie reprezentowali je: Adam Przepiórkowski (IPI PAN, kierownik projektu), Rafał L. Górski (IJP PAN), Barbara Lewandowska-Tomaszczyk (UŁ), Mirosław Bańko, Marek Łaziński (PWN).

Wymienione instytucje wniosły do wspólnego przedsięwzięcia zebrane przez siebie teksty, dalsze teksty zaś pozyskano w trakcie projektu. W rezultacie udało się stworzyć korpus o objętości ok. 1,8 miliarda segmentów (ok. 1,5 miliarda słów), w tym podkorpus zrównoważony wielkości 300 milionów segmentów (240 milionów słów). Dzięki działającym wcześniej lub skonstruowanym w trakcie projektu narzędziom korpus ten został oznakowany morfosyntaktycznie, co pozwala wyszukiwać w nim wszystkie formy danego leksemu (a nie tylko wybrane) oraz szukać form o określonej charakterystyce gramatycznej.

NKJP obsługują dwie wyszukiwarki: PELCRA i Poliqarp. Pierwsza została wyposażona w przyjazny dla użytkownika interfejs, który pozwala na konstruowanie kwerend przez wypełnianie lub zaznaczanie gotowych rubryk w tabeli. Ponieważ zawiera kolokator oraz narzędzie do tworzenia profili stylistycznych i chronologicznych słowa, jest szczególnie użyteczna do badań leksykalnych. Obsługa drugiej wyszukiwarki wymaga przyswojenia sobie języka zapytań opartego na składni wyrażeń regularnych, nagrodą za ten wysiłek jest jednak możliwość budowania kwerend o znacznej szczegółowości, odwołujących

się zarówno do budowy słów, jak ich charakterystyki funkcjonalnej, przypisanej im przez program znakujący (tager). Właściwość ta czyni Poliqarpa narzędziem szczególnie użytecznym do badań nad systemem gramatycznym języka.

Ponieważ obie wyszukiwarki istniały przed rozpoczęciem prac nad NKJP, w projekcie nie przewidziano budowy nowej wyszukiwarki, która łączyłaby ich funkcje. Ma to swoje zalety (ich dotychczasowi użytkownicy nie muszą się uczyć obsługi nowego narzędzia), ale ma także wady. Uważny użytkownik obu przeglądarek może spostrzec na przykład, że wyniki ich działania nie są identyczne. W szczególności frekwencja szukanej formy różni się w zależności od tego, której wyszukiwarki użyć, nawet jeśli zastosowane kwerendy są równoważne. Ze względu na odmienne zasady działania, różnice tego rodzaju są zrozumiałe, ale budzą niepokój i każą relatywizować wyniki badań do wykorzystanego narzędzia[1].

NKJP jest dostępny pod adresem http://www.nkjp.pl/. Prócz przekierowania na strony wyżej wymienionych wyszukiwarek (PELCRA: http://www.nkjp.uni.lodz.pl/, Poliqarp: http://nkjp.pl/poliqarp/) zamieszczono tam m.in. ogólne informacje o projekcie i jego zespole wykonawczym, listę publikacji i narzędzi korpusowych powstałych w ramach projektu oraz wykaz tekstów, które weszły do NKJP. Najobszerniejsza publikacja dotycząca NKJP (Przepiórkowski, Bańko, Górski, Lewandowska-Tomaszczyk 2012a) opisuje jego historię, strukturę, zasady znakowania, narzędzia i podprojekty oraz przykładowe zastosowania.

Struktura korpusu

Prócz pełnej wersji NKJP dostępnych jest kilka podkorpusów różnej wielkości i o różnym charakterze. Ich objętość podawana jest w segmentach, które na ogół odpowiadają słowom, rozumianym jako napisy literowe ograniczone spacjami, czasem jednak są krótsze (por. niżej). Szczegółowo o zasadach segmentacji informują Przepiórkowski i in. (2012: 60-62). Aby od liczby segmentów przejść do liczby słów, należy tę pierwszą pomnożyć przez 0,8.

Pełny NKJP obejmuje ok. 1,8 mld segmentów, czyli ok. 1,5 mld słów. Podkorpus zrównoważony ma wielkość 300 mln segmentów, czyli 240 mln słów. Prócz nich Poliqarp udostępnia użytkownikom kilka mniejszych podkorpusów, m.in. ręcznie oznakowany pod względem morfosyntaktycznym podkorpus o wielkości 1,2 mln segmentów (ok. 1 mln słów). Natomiast PELCRA daje wybór jedynie między korpusem pełnym a zrównoważonym, przy czym możli-

[1] Decyzja rozwijania w projekcie dwóch wyszukiwarek miała jeszcze jedno uzasadnienie: relacyjne bazy danych lepiej radzą sobie z bardzo dużymi korpusami niż bazy tekstowe. Nie było wiadomo, jak z korpusem obejmującym miliard słów będzie współpracował Poliqarp. Jak się okazało, była to nadmierna ostrożność, gdyż obie wyszukiwarki w zasadzie działają dobrze.

wość ta dotyczy konkordancji, nie kolokacji: wbudowany w nią kolokator działa – skądinąd słusznie – tylko na podkorpusie zrównoważonym.

O strukturze pełnej wersji NKJP ani witryna projektu, ani dostępne publikacje nie informują w szczegółach. Można przyjąć, że jest ona po prostu wynikiem połączenia tekstów, którymi uczestnicy projektu dysponowali w momencie rozpoczęcia prac i które pozyskali w ich trakcie. W książce Przepiórkowskiego i in. (2012a: 36) podano, że najstarsze teksty z korpusie pełnym pochodzą z przełomu XVIII i XIX w. Próbne kwerendy pokazują, że w korpusie pełnym większy udział niż w zrównoważonym mają teksty prasowe i protokoły sejmowe, a także teksty internetowe (co nie jest niespodzianką).

Strukturę wersji zrównoważonej opisano dokładniej: można ją znaleźć w zakładce *Pomoc* wyszukiwarki PELCRA oraz w książce Przepiórkowskiego i in. (2012a), gdzie przedyskutowano też kwestie metodologiczne, związane z ustaleniem struktury korpusu, oraz objaśniono kryteria klasyfikacji tekstów zastosowane w pracy nad NKJP. Szczegóły podane w tab. 2.1. są oparte na informacjach z zakładki *Pomoc* wyszukiwarki PELCRA.

Tab. 2.1. Struktura NKJP: podkorpus zrównoważony.

Rodzaj tekstów	**Udział w NKJP**
Książki	29,0%
z tego:	
Literatura piękna	*16,0%*
Literatura faktu	*5,5%*
Książki naukowo-dydaktyczne	*2,0%*
Książki i prasa informacyjno-poradnikowe	*5,5%*
Prasa	50,0%
z tego:	
Gazety	*26,0%*
Periodyki	*24,0%*
Inne teksty pisane (urzędowe, listy)	4,0%
Internet (blogi, fora, strony www)	7,0%
Teksty mówione (konwersacyjne naturalne i medialne, protokoły sejmowe)	10,0%

Źródło: zakładka „Pomoc" na stronie http://www.nkjp.uni.lodz.pl/, dostęp 18 III 2013.

W strukturze podkorpusu zrównoważonego starano się w zasadzie odzwierciedlić strukturę czytelnictwa w Polsce. Kryterium to nie odnosi się do tekstów mówionych, toteż udział tych ostatnich określono arbitralnie na 10%, co jest i tak wielkością dość dużą, zważywszy na trudności z pozyskiwaniem tekstów tego rodzaju i ich anotacją. Wśród szeroko rozumianych tekstów mówionych dominują stenogramy z posiedzeń Sejmu i jego komisji, obejmujące razem ok.

27 mln segmentów (leżą one w istocie na pograniczu języka mówionego i pisanego). Teksty medialne (transkrypcje audycji radiowych i telewizyjnych) obejmują 0,9 mln segmentów, a teksty konwersacyjne (transkrypcje spontanicznych rozmów) 1,9 mln. Arbitralnie ustalono też udział tekstów, które w korpusach określane są jako „inne" bądź „miscellanea". W NKJP większość z nich to teksty przynależne do gatunków internetowych.

Prócz zrównoważenia stylowo-gatunkowego starano się osiągnąć zrównoważenie tematyczne oraz chronologiczne, dając przewagę – tyleż ze względów merytorycznych, co praktycznych – tekstom najnowszym. Około 80% tekstów włączonych do zrównoważonej części NKJP powstało po 1990 r., 15% w latach 1945-1990 i tylko 5% przed 1945 r. W tej najstarszej części większość stanowią utwory literackie, których żywotność – w sensie recepcji społecznej – jest dłuższa niż innych tekstów. Szkoda jednak, że wielu znaczących twórców literatury polskiej brakuje w NKJP: wynika to z trudności w pozyskaniu praw do wykorzystania ich tekstów.

Teksty książek nie były próbkowane: włączano je do podkorpusu zrównoważonego w całości albo umieszczano poza nim, jeśli nie pasowały do założonej struktury. Próbkowano natomiast teksty gazet i czasopism, aby z jednej strony zapewnić udział różnych tytułów, a z drugiej nie przekroczyć założonych proporcji.

Segmentacja, anotacja, lematyzacja

W NKJP, z punktu widzenia wyszukiwarki Poliqarp, nie istnieje pojęcie wyrazu. Zamiast niego mamy segmenty. Segmentem jest zarówno intuicyjnie rozumiany wyraz, jak również ruchomy morfem, a nawet znak interpunkcyjny. Tak więc sekwencja „*jadłem*," to trzy segmenty (*jadł*+*em*+*,*)[2]. Każdy segment jest lematyzowany i anotowany morfosyntaktycznie[3]; i tak lematy wspomnianej sekwencji to odpowiednio „*jeść*", „*być*" oraz przecinek.

Punktem wyjścia dla zestawu znaczników NKJP jest fleksja. Do jednej „części mowy" jest zaliczany tylko taki zbiór form, który odmienia się przez jednolity zestaw kategorii fleksyjnych. I tak, nie ma „czasowników", są jedynie formy finitywne, imiesłowy przymiotnikowe czynne, rzeczowniki odsłowne itp. Co więcej, *kocha* i *kochał* należą do dwu różnych części mowy: pierwszy segment jest formą finitywną czasownika, drugi tzw. pseudoimiesłowem. Jest tak, ponieważ czasownik w formie czasu nieprzeszłego odmienia się przez osobę, ale już

[2] To nieco zaskakujące potraktowanie ruchomych końcówek ma głębokie uzasadnienie. W przeciwnym wypadku należałoby bowiem przyjąć, że przez osobę odmieniają się również przymiotniki, spójniki etc. por. *głodniście*, *żebym*. Warto dodać, że takie podejście odzwierciedla (oczywiście niezamierzenie) fakt, że ruchoma końcówka jest dawną formą czasownika *być*.

[3] Należałoby tu raczej powiedzieć: morfologicznie, jednak pod wpływem językoznawstwa anglistycznego w lingwistyce korpusowej przyjęło się mówić o anotacji morfosyntaktycznej.

nie przez rodzaj, a w formie czasu przeszłego odmienia się przez rodzaj, ale nie przez osobę[4]. Oznacza to również, że anotacja nie wyróżnia czasu przyszłego. Tzw. czas przyszły prosty (*napiszę*) odmienia się dokładnie tak samo jak czas teraźniejszy (*piszę*), obie formy więc zaliczamy do czasu nieprzeszłego. Formy te dają się jednak rozróżnić, jeśli ograniczymy wyszukiwanie do czasowników spełniających dwa warunki naraz: czas nieprzeszły i aspekt niedokonany (dla praesens) lub czas nieprzeszły i aspekt dokonany (dla futurum).

Interpretacja morfologiczna nie przekracza granic słowa, to znaczy nie rozróżnia np. *być* egzystencjalnego i *być* jako wykładnika czasu bądź strony, a także rozmaitych użyć *się*, takich jak w zdaniach: *Janek budzi się późno*, *Marysia i Basia budzą się nawzajem* i *Poborowych budzi się wcześnie*. Co więcej, skoro *by* – jak wspomniano – ma status segmentu, to nie ma znacznika oznaczającego tryb przypuszczający, choć jest odpowiedni znacznik dla trybu rozkazującego.

Jak widać, w NKJP pewne zjawiska, zwykle traktowane jako morfologiczne, są uważane za składniowe. Skoro jednak korpusu nie da się przeszukiwać przy pomocy anotacji składniowej[5], to zjawiska składniowe musimy – na użytek ich wyszukiwania w korpusie – definiować jako sekwencje segmentów o pewnych cechach morfologicznych, leksykalnych bądź ich kombinacji.

Wracając do zestawu znaczników, dodajmy, że jest on dość rozbudowany, zawiera takie kategorie jak wokaliczność[6], czy poprzyimkowość, o dość ograniczonym zastosowaniu, ale również istotniejsze rozróżnienie na zaimki trzecioosobowe i nietrzecioosobowe, czy wreszcie depreciativa (*chłopy*, *profesory*). Dzięki konsekwentnemu odróżnianiu rodzaju męskoosobowego od męskiego żywotnego i rzeczowego łatwe jest np. wyszukiwanie zdań z nieosobowym agensem. Bardzo nietypową cechą zestawu znaczników jest przypisanie kategorii przypadka przyimkom, warto o tym pamiętać, kiedy szukamy konstrukcji z czasownikami wykazującymi daną rekcję. Zgodnie z intencją jego twórców (Woliński 2003, Przepiórkowski i Woliński 2003) zestaw znaczników miał być maksymalnie rozczłonkowany. W konsekwencji nie ma w nim zasadniczo rzeczowników, ani czasowników, jest bowiem dyskusyjne, czy zaimki rzeczowne bądź rzeczowniki odsłowne są czy też nie są rzeczownikami. Żeby decyzję tę pozostawić użytkownikom korpusu wszystkie te kategorie są oznaczone osobnym znacznikiem. Istnieje jednak kategoria *noun*, obejmująca rzeczowniki (w ujęciu najwęższym), zaimki i rzeczowniki odsłowne, analogicznie istnieje

[4] Przypomnijmy, że wykładnik osoby jest czasu przeszłego, traktowany jako osobny segment.

[5] Źródłowa wersja korpusu zawiera również inne poziomy anotacji, na których wyróżniono grupy syntaktyczne, nazwy własne i znaczenia wybranych wyrazów polisemicznych. Ponieważ jednak konkordancery (przynajmniej na razie) nie pozwalają przeszukiwać tekstów pod tym kątem, nie będziemy tego szerzej omawiali.

[6] Aglutynaty wokaliczne to te, które łączą się z formami zakończonymi na spółgłoskę np. *-em*, niewokaliczne to te, które się łączą z formami na samogłoskę np. *-m*.

kategoria *verb*, obejmująca to, co tradycyjnie rozumiane jest jako czasowniki wraz rzeczownikami odsłownymi.

Metadane udostępniane przez obie wyszukiwarki to: autor, tytuł (książki lub artykułu), źródło (tytuł czasopisma bądź pracy zbiorowej), wydawca, miejsce wydania, data pierwszego wydania lub nagrania tekstu (Poliqarp podaje dodatkowo datę wydania wprowadzonego do NKJP), a także kanał (książka, prasa, spontaniczna konwersacja itp.) oraz typ tekstu (pojęcie zbliżone do stylu funkcjonalnego, ale z nim nietożsame).

Podstawowe zapytania

Jak wspomnieliśmy, użytkownik ma do wyboru dwie wyszukiwarki; decyzja, której użyć, jest przede wszystkim uzależniona od tego, czy jest bardziej zainteresowany leksyką, czy gramatyką. Nie jest naszym celem kompletne przedstawienie składni zapytań, witryna NKJP zwiera bowiem pliki pomocy. Chodzi nam raczej o ukazanie zróżnicowanych możliwości wyszukiwania.

Wyszukiwarka PELCRA

Wyszukiwarka PELCRA pozwala na wygodne formułowanie kwerend przez wpisywanie ich elementów w gotowe rubryki lub wybór ustalonych opcji. Podstawowe narzędzie dostępne jest w zakładce *Konkordancje*. Tworzy ono wykaz wystąpień poszukiwanego słowa lub wyrażenia z uwzględnieniem kontekstów.

Można pytać o pojedyncze słowa, ciągi słów, a także całe leksemy. Można używać symboli zastępczych: gwiazdki, która zastępuje dowolny ciąg znaków, oraz pytajnika, który zastępuje pojedynczy znak. Można budować pytania o postaci alternatywy. Dostępne możliwości zostały wyczerpująco opisane w zakładce *Pomoc*, z której pochodzą niżej wymienione przykłady. Ich przećwiczenie pozwoliło poczynić szereg spostrzeżeń, które przytaczamy; początkujący użytkownicy NKJP zapewne zechcą sami sprawdzić działanie wyszukiwarki na tych lub innych przykładach.

tymianek

Wynikiem kwerendy jest konkordancja obejmująca słowo *tymianek* (bez uwzględnienia odmiany).

dobra wola

Wynikiem kwerendy jest konkordancja z wyrażeniem *dobra wola* (w tym szyku i też bez uwzględnienia odmiany).

tymian*

Wynikiem kwerendy jest konkordancja obejmująca słowa zaczynające się od liter *tymian*, np. *tymianek*, *tymiankowy*, *Tymiankach*.

tymianek**

Wynikiem kwerendy jest konkordancja obejmująca wszystkie obecne w NKJP formy odmiany leksemu *tymianek*, m.in. *tymianek*, *tymiankiem*, *tymianku*. Dwie gwiazdki należy postawić po formie hasłowej leksemu (takiej, która go reprezentuje w typowym słowniku). Wyszukiwarka nie sygnalizuje pomyłki, lecz po prostu nie wyświetla żadnych wyników, można więc dojść do fałszywego wniosku, że NKJP nie zawiera np. słowa *glony*, jeśli zada się zapytanie *glony*** zamiast *glon**. Trzeba też pamiętać, że wyszukiwarka nie uwzględnia odmiany wszystkich wyrazów polskich, jeśli więc kwerenda z użyciem dwóch gwiazdek nie przynosi wyników, warto wykonać analogiczną kwerendę z użyciem jednej gwiazdki.

tymianek**|bazylia**|czosnek**

Wynikiem kwerendy jest konkordancja obejmująca wszystkie obecne w NKJP formy odmiany wymienionych leksemów, m.in. *bazylią*, *czosnku*, *tymiankiem*.

filetow

Wynikiem kwerendy jest konkordancja obejmująca słowa zawierające ciąg liter *filetow*, przed którym i po którym mogą wystąpić inne litery, m.in. więc słowo *odfiletować*. Zamiast gwiazdki, która zastępuje zero lub więcej liter, można użyć pytajnika, który zastępuje dokładnie jedną literę. Kwerenda *filet?** przynosi więc konkordancję ze słowami *fileta*, *filetami*, *filetował*, *filetyczna* i in., ale nie ze słowem *filet*.

*essa**

Wynikiem kwerendy jest konkordancja obejmująca słowa o zakończeniu *essa*, z uwzględnieniem ich odmiany, m.in. więc słowa *bessa*, *bessy*, *hostessę*, *Odessie*, ale także nazwisko *Baroness* (które wyszukiwarka traktuje mylnie jako dopełniacz liczby mnogiej rzeczownika *baronessa*).

Jak widać, wyszukiwarka pomija różnicę między wielką a małą literą, a wyszukiwanie z uwzględnieniem odmiany wyrazów daje czasem wyniki nadmiarowe. Ponadto nie jest brane pod uwagę znaczenie wyrazów, w konkordancji mogą zatem wystąpić konteksty ilustrujące wiele znaczeń, a nawet słów homonimicznych. Na przykład kwerenda *bale* przynosi i *bale dębowe*, i *bale sylwestrowe*, a w wynikach dla zapytania *bal*** jest zarówno dopełniacz liczby mnogiej *bali*, jak i równobrzmiąca forma czasownikowa *bali* (*się*).

PELCRA pozwala wygodnie szukać ciągów wyrazowych, w szczególności związków frazeologicznych. Po pierwsze, można wyłączyć opcję *Zachowaj szyk*, która jest zaznaczana domyślnie: zależnie od wyboru kwerenda *pieróg** leniwy*** zwróci wśród wyników *leniwe pierogi* albo nie. Po drugie, można wybrać inną niż domyślna wartość w polu *Maks. odległość*, co pozwoli uwzględnić wyrażenia nieciągłe: np. kwerenda *moja żona* przy odległości ustalonej na 1 obejmuje nie tylko ciąg *moja żona*, ale też *moja była żona*. Po trzecie, można użyć potrójnego znaku podkreślenia do rozdzielenia grup wariantywnych składników ciągu, np.

łza**|łezka**___oko**___kręcić**|zakręcić**

Powyższa kwerenda przy maksymalnym odstępie ustalonym na 2 i bez zaznaczonej opcji *Zachowaj szyk* uwzględni zwroty *kręcą się łzy w oczach*, *kręcą się w oczach łzy*, *kręciła się łza w oku*, *łezka kręciła się w oku* i inne:

Konkordancje PELCRY można sortować: służy o tego rozszerzony formularz, który ukazuje się po wyklikaniu linku *Zaawansowane*. Dostępne jest sortowanie według słowa kluczowego, lewego albo prawego kontekstu, a także według źródła, daty i kanału przekazu, którym może być prasa, książka, internet lub mowa (podział na kanały, alternatywny wobec stylowo-gatunkowej klasyfikacji tekstów, jest wykorzystywany także przez funkcję tworzenia profili, o czym niżej w punkcie *Dodatkowe funkcje*). Na życzenie sortowanie może być dwustopniowe, np. konkordancję można uporządkować według daty powstania utworu, a w drugiej kolejności według kontekstu.

Inną użyteczną funkcją, którą oferuje rozszerzony formularz, jest grupowanie wyników. Aby uniknąć nadmiaru jednorodnych danych, można zażądać, aby np. z jednego tekstu wybranych zostało nie więcej niż x wystąpień szukanego słowa (gdzie x przybiera osiem wartości do wyboru z przedziału od 1 do 100). Analogicznie można ograniczyć liczbę wyników z określonego roku, kanału lub źródła (rozróżnienie między tekstem a źródłem jest istotne zwłaszcza w wypadku publikacji prasowych, gdzie tekstem jest artykuł, a źródłem – czasopismo lub gazeta).

Ponadto rozszerzony formularz wyszukiwania pozwala selekcjonować teksty za pomocą klasyfikacji stylowo-gatunkowej i podziału na kanały, a nawet przez określenie daty powstania tekstu, jego tytułu i tytułu źródła. Można np. szukać słowa *aborcja* tylko w literaturze pięknej (jest ono tutaj dość rzadkie), albo tylko prasie (znacznie częstsze), można wybrać tytuł prasowy (jeden lub kilka) albo wykluczyć określone źródła z przeszukiwanego zbioru. Dodatkowo można zawęzić wyniki przez określenie słów, które powinny lub które nie mogą wystąpić w tym samym akapicie co słowo szukane.

PELCRA wyświetla domyślnie 100 wyników na stronie, wartość tę można skokowo zmniejszać (do 10) lub zwiększać (do 10 000). Po przejrzeniu wyników

można obejrzeć następną stronę itd. Warto pamiętać, że sortowanie działa tylko w obrębie danej strony, tzn. poszczególne strony są sortowane niezależnie od siebie. Również liczba wystąpień słowa w NKJP podawana jest w odniesieniu do danej strony, np. kwerenda *aborcja*** przy liczbie wyników na stronie ograniczonej do 100 przynosi 101 (!) kontekstów i aby sprawdzić łączną frekwencję tego leksemu w NKJP, trzeba odpowiednio zwiększyć liczbę wyników na stronie. W danym przykładzie wystarczający okazuje się próg 5000 wyników (wyszukiwarka zawiadamia o znalezieniu 4233 wystąpień, które wyświetla na stronie), ale w wypadku bardzo częstych słów nawet próg 10 000 wystąpień może się okazać za mały. Frekwencję bardzo częstych słów trzeba więc sprawdzać za pomocą wyszukiwarki Poliqarp.

Cennym udogodnieniem oferowanym przez PELCR-ę jest możliwość pobrania wyników w formacie arkusza kalkulacyjnego MS Excel, dzięki temu użytkownik może je dalej sortować lub edytować narzędziami przeznaczonymi do obsługi arkuszy kalkulacyjnych. Innym udogodnieniem jest mechanizm generowania hipertekstowych odsyłaczy do wyników: w tym celu po utworzeniu listy wyników należy wyklikać przycisk z napisem *URL*, aby tuż powyżej listy zobaczyć skompresowany odsyłacz, w którym zakodowane są wszystkie informacje o wybranych opcjach wyszukiwania. Można umieścić go np. w publikacji albo wysłać pocztą elektroniczną, bez potrzeby oddzielnego zapisywania wszystkich opcji, ponieważ każda osoba po wyklikaniu go lub wklejeniu w pasek adresów przeglądarki internetowej zobaczy ekran wyszukiwania ze wszystkimi szczegółami, tak jakby sama zadała zapytanie.

O innych funkcjach udostępnianych przez wyszukiwarkę PELCRA, mianowicie o profilach stylistycznych i szeregach czasowych, o kolokatorze, wyszukiwarce tekstów mówionych i słowach dnia – zob. niżej w części *Dodatkowe funkcje*.

Wyszukiwarka Poliqarp

Najprostszym zapytaniem jest po prostu wpisanie w okienko wyszukiwania słowa w żądanym kształcie graficznym. Jeśli wpiszemy słowo *pies*, to wyszukiwarka znajdzie wystąpienia tego napisu literowego, ale nie znajdzie słowa *psu*, *psom* ani nawet *Pies*. Aby znaleźć wystąpienia wszystkich form tego leksemu, pisanych wielką lub małą literą, należy wpisać zapytanie:

$$[\mathsf{base} = \mathsf{pies}]^{7}$$

gdzie *base* wprowadza wyraz hasłowy, czyli nazwę szukanego leksemu. Kwerendę tę można dalej ograniczyć do tych form leksemu *pies*, które spełniają

[7] Spacje w kwerendach są opcjonalne.

określone warunki morfologiczne, np. mają wartość celownika liczby pojedynczej lub mnogiej:

[base = pies & cas = dat]

Można określić też warunki graficzne, np. jeśli zależy nam jedynie na wystąpieniach słowa pisanego wielką literą (znaczenie kropki i gwiazdki wyjaśnimy później):

[base = pies & cas = nom & nmb = sg & orth = P.*]

Warto zauważyć, że o to samo można zapytać na różne sposoby – poprzednie zapytanie jest równoważne następującemu:

[orth = "Psu|Psom"][8]

Naturalnie można we wzorcu wyszukiwania wpisać jedynie znaczniki:

[pos = subst & cas = acc & nmb = sg]

przy czym kolejność kategorii jest dowolna, ten sam wynik daje kwerenda:

[cas = acc & nmb = sg & pos = subst]

W tym miejscu należy objaśnić użycie wyrażeń regularnych w tworzeniu zapytań. Znak | oznacza alternatywę, np. kwerenda:

pies | kot

znajduje wystąpienia słów *pies* i *kot*, a kwerenda:

[base = pies | base = kot]

wyszukuje wszystkie formy leksemów *pies* i *kot*. Warto dodać, że wyszukiwarka akceptuje nawet 100 i więcej razy powtórzony znak alternatywy, np. zapytanie o 100 jednostek leksykalnych oddzielonych znakiem |.

Kropka zastępuje dowolny znak. Dowolną liczbę powtórzeń oznacza się gwiazdką. Tak więc zapytanie:

[orth = P.*]

[8] W niektórych kwerendach obligatoryjne są cudzysłowy; szczegóły można znaleźć w plikach pomocy.

wyszuka wszystkie segmenty zaczynające się od wielkiego P, w tym samo „P", ponieważ dowolną liczbą powtórzeń jest w także zero (aby otrzymać wyniki ze znakiem powtórzonym co najmniej raz, należy po nim umieścić znak + zamiast znaku *).

Tego rodzaju zapytania miewają praktyczne zastosowanie: wyobraźmy sobie, że chcemy odnaleźć wszystkie możliwe przedrostki, które biorą udział w derywowaniu czasowników pochodzących od *mówić*. Można to zrobić za pomocą zapytania, choć w praktyce ograniczenie zwracanych wystąpień do 2000 sprawi, że niekoniecznie będzie to pełna lista:

[base = .+mówić]

Co prawda, zapewne znany jest nam ten inwentarz, ale celem badań korpusowych jest uzupełnienie i weryfikacja danych pochodzących z kompetencji językowej badacza, jeśli więc wpiszemy do zapytania przewidywane przedrostki, to nie uda się nam zweryfikować naszej wiedzy[9]. Podobnie można próbować znaleźć czasowniki derywowane od rzeczownika *ząb*:

[base = ".*z(ę|ą)b.*" & pos=verb]

Powyższe zapytanie wyszuka wszystkie słowa (ściślej: segmenty), które zaczynają się od dowolnego ciągu liter, następnie zawierają literę *z*, potem *ę* bądź *ą*, potem *b*, po czym następuje znowu dowolny ciąg liter.

Można ściśle określić minimalną i maksymalną liczbę wystąpień za pomocą ciągu {n,m} (gdzie n – wielkość minimalna, m – maksymalna), np.:

[orth = się] [ ] {2,4} [base = budzić]

co oznacza: wyszukaj wystąpienia słowa *się* odgraniczone dwoma, trzema bądź czterema dowolnymi segmentami od wyrazu *budzić*. Z kolei kwerenda:

([pos = adj] [base = '["','"]']) {3,}

każe wyszukać sekwencje: przymiotnik i przecinek następujące po sobie nie mniej niż trzy razy, czyli wyszukuje ciągi w rodzaju *(sprzęt) fotograficzny*, *filmowy*, *radiowy*, *elektroakustyczny*.

Wymienione operatory znajdują zastosowanie nie tylko do szukania słów, ale i ciągów słów. Załóżmy, że chcemy wyszukać użycia czasownika zwrotnego *budzić się*, innymi słowy – chcemy, by wyszukiwarka znalazła wystąpienia czasownika zwrotnego i równocześnie nie znalazła wystąpień czasownika niezwrotnego. Zapytanie:

[9] Dość powiedzieć, że autorom nie był znany czasownik *podmówić*, dopóki, przygotowując ten tekst, nie wpisali omawianego zapytania do wyszukiwarki.

[base = budzić] [orth = się]

będzie bardzo skuteczne, jeśli chodzi o eliminację użyć niezwrotnych, ale pominie znaczną liczbę użyć zwrotnych, np. *budziłam się* (pomiędzy segmentami *budził* i *się* jest segment *am*), *budziłabym się*, *budził ci się*. Dokładniejsze jest zapytanie:

[base= budzić] [ ] {,3} [orth = się]

czyli: znajdź sekwencję segmentów, z których pierwszy to dowolna forma czasownika *budzić*, ostatni to *się* i które bądź to sąsiadują ze sobą, bądź są rozdzielone nie więcej niż trzema innymi segmentami. Ale zauważmy, że kwerenda ta wyszuka również ciąg *budził sensacji pojawiając się*, a z drugiej strony zaś nie znajdzie kontekstów, w których *się* występuje przed czasownikiem. Aby objąć zapytaniem większą liczbę przykładów, musimy ją rozbudować:

[base = budzić] [ ] {,3}[orth = się] | [orth = się] [base = budzić]

Oczywiście można i to zapytanie rozbudowywać, pozwalając, by pomiędzy *się* i *budzić* mogły się pojawić inne segmenty.

Oprócz wspomnianych operatorów szerokie zastosowanie ma jeszcze jeden, mianowicie ! – oznaczający wykluczenie. Kwerenda:

[base = pies & !cas = nom]

oznacza: wyszukaj wszystkie wystąpienia form leksemu *pies* z wyjątkiem mianowników. Podobnie zapytanie:

[!orth = się] [base = budzić] [!orth = się]

pozwoli wyeliminować sporą część zwrotnych użyć czasownika *budzić*.

Dodatkowe funkcje

Wyszukiwarka PELCRA

Informacja o frekwencji daje podstawową wiedzę o zakresie użycia słowa, ale nie informuje o jego dystrybucji stylistycznej. Aby dowiedzieć się, w których odmianach języka jakieś słowo jest używane i z jaką częstością, należy skorzystać z funkcji *Profil*. Wyświetla ona dwie tabele: jedna pokazuje, jak często szukane słowo (wyrażenie, leksem) występuje w poszczególnych rodzajach tekstów wyodrębnionych ze względu na kryterium stylistyczne (w publicystyce, beletrystyce, tekstach konwersacyjnych itp.), druga podaje analogiczne wyniki

w podziale na kanały (książka, prasa, internet, teksty mówione, a nawet rękopis, choć trudno znaleźć słowo, dla którego frekwencja w kanale *rękopis* nie byłaby zerowa). Frekwencja podawana jest w przeliczeniu na milion słów, co w dużym stopniu uniezależnia wyniki od struktury korpusu. Tabelom towarzyszą poglądowe wykresy, wystarczy więc jedno spojrzenie, aby zdobyć orientację w charakterystyce stylistycznej słowa. Niestety, nie objaśniono tu skrótów oznaczających poszczególne rodzaje tekstów, rozwiązania symboli takich, jak *publ*, *nd*, *qmow*, trzeba więc szukać w zakładce *Pomoc*.

Prócz dystrybucji słowa między odmiany stylistyczne tekstów i kanały użytkownik NKJP może poznać przybliżoną częstość użycia słowa w poszczególnych latach w okresie 1988-2010. Służy do tego funkcja *Czas*. Wyniki są, jak poprzednio, przedstawiane w tabeli i na wykresie, przy czym częstości są określane tu nie w przeliczeniu na milion słów, lecz w przeliczeniu na tysiąc akapitów. Analiza szeregów czasowych – jak napisano w zakładce *Pomoc* – „ukazuje, iż niektóre słowa, frazy, idiomy, nazwy własne i zwroty zyskują znacznie na popularności w krótkim czasie, odzwierciedlając tym samym nośność danego tematu w dyskursie publicznym". Inne możliwe zastosowanie funkcji *Czas* to badanie adaptacji wyrazów obcych (ciekawe obserwacje nasuwa np. porównanie danych dla słów *dealer* i *diler*). Ze względu na strukturę korpusu głębokość analizy diachronicznej jest jednak niewielka – nie sięga w głąb poza koniec lat 80. XX w.

Dla badaczy słownictwa użytecznym narzędziem jest wbudowany w wyszukiwarkę PELCRA kolokator, dostępny w głównym menu. Generuje on automatycznie listę typowych połączeń wyrazowych ośrodka kolokacji (słowa, ciągu słów lub leksemu), selekcjonowanych i porządkowanych ze względu na wartość testu chi-kwadrat. Domyślnie wyszukiwane są wszelkie kolokacje, ale można ograniczyć wybór kolokatów do jednej z czterech części mowy: rzeczowników, czasowników, przymiotników (z uwzględnieniem imiesłowów przymiotnikowych) lub przysłówków. Można uwzględnić wyłącznie kolokaty lewo- albo prawostronne, można też domyślny kontekst, równy 1, zwiększyć do 2, aby uwzględnić kolokaty nie leżące w bezpośrednim sąsiedztwie ośrodka kolokacji.

Ze względu na znaczną złożoność obliczeniową, kolokacje są generowane na podstawie niepełnej liczby wystąpień ośrodka kolokacji: domyślnie wynosi ona 10 000, użytkownik może ją podnieść ją do 50 000, ale w praktyce nie przynosi to widocznego efektu. Można też zmienić minimalną liczbę współwystąpień ośrodka kolokacji i kolokatu, przy której kolokacje są uwzględniane w wynikach. Domyślna wartość to 5 (przy mniejszej wyniki byłyby mało wiarygodne), wyższe wartości (7, 11 lub 13) warto wybrać, jeśli lista wyników jest bardzo długa i badacz chce się ograniczyć do analizy kolokacji najczęstszych. Podobnie jak konkordancer kolokator udostępnia skompresowane odsyłacze do wyników (po wyklikaniu przycisku *URL*).

Jeszcze jednym narzędziem PELCR-y, dostępnym z głównego menu, jest wyszukiwarka tekstów mówionych. W pewnym stopniu dubluje ona konkordancer, który też pozwala ograniczyć obszar przeszukiwania do tekstów mówionych, ale oferuje nieco więcej możliwości, np. pozwala sortować wyniki według wieku, wykształcenia i płci rozmówców.

Swego rodzaju ciekawostką – luźno związaną z projektem NKJP – jest codzienne automatyczne porównywanie względnej frekwencji słów w czterech tytułach prasowych z ich frekwencją tamże w okresie porównawczym, obejmującym miniony rok. Serwis ten, dostępny w głównym menu przeglądarki PELCRA, nazywa się *Słowa dnia* i może służyć obserwacji wydarzeń zachodzących w świecie i relacjonowanych przez media, odzwierciedlających się w okresowych wzrostach i spadkach popularności poszczególnych słów. O założeniach tego projektu i jego potencjalnych zastosowaniach informuje szczegółowo jeden z rozdziałów w książce Przepiórkowskiego i in. (2012).

Wyszukiwarka Poliqarp

Również wyszukiwarka Poliqarp pozwala zapisać wyszukane konkordancje w formacie – do wyboru – .HTML lub .xls (format arkuszy kalkulacyjnych programu Excel). Konkordancja ma długość maksymalnie 1000 wierszy, jeśli więc szukamy częstych słów, to wynik będzie bardzo niereprezentatywny: znalezione wystąpienia mogą pochodzić z zaledwie kilku źródeł. Aby otrzymać próbkę reprezentatywną, należy wyklikać *Ustawienia* i na ekranie, który się otworzy, zaznaczyć *Losowa próbka rozmiaru*, następnie zaś, po zapisaniu ustawień, wykonać kwerendę. To samo należy zrobić, jeśli chcemy znać dokładną liczbę szukanego słowa lub wyrażenia. Określając bowiem rozmiar losowej próbki, sprawiamy, że Poliqarp oprócz konkordancji wyświetli informację o całkowitej liczbie wystąpień szukanego ciągu w korpusie. Nawiasem mówiąc, jeśli interesuje nas jedynie liczba wystąpień, warto wpisać możliwie małą wielkość próbki, aby szybciej uzyskać odpowiedź[10].

Ekran *Ustawienia* pozwala ponadto określić sposób sortowania i wyświetlania wyników. Domyślnie wyświetlane w kontekście są tylko szukane segmenty, ale też ich formy podstawowe i znaczniki, co sprawia, że wyniki są mało czytelne. Można to zmienić, żądając, by konkordancer wyświetlał np. tylko segmenty. Niestety nie ma możliwości wyświetlania tylko znaczników, bądź tylko lematów, co bywa przydatne, jeśli konkordancja ma być dalej opracowywana. Ponadto można zaznaczyć w ustawieniach, by konkordancje zawierały znaczniki i lematy w prawym i lewym kontekście. Niestety ustawień tych nie da się zapisać, trzeba je wprowadzać na nowo podczas każdej sesji.

[10] W tym miejscu warto zasygnalizować, że jeśli nie włączymy opcji „losowej próby", rzeczywista liczba wystąpień pojawi się dopiero na jednym z kolejnych ekranów.

Niezwykle przydatną, a rzadko spotykaną funkcją jest podział wzorca wyszukiwania na dwie części, z których każda może być sortowana. Wyobraźmy sobie, że interesują nas struktury typu *naiwny jak dziecko*, *głupi jak but*. Wyszukujemy je zapytaniem:

[pos = adj] [base = jak] [pos = subst]

Dopóki interesują nas przymiotniki w ogóle, niezależnie od obecnego w porównaniu rzeczownika, wystarczy posortować wyniki według początku wzorca wyszukiwania. Jeśli będziemy chcieli jednak wiedzieć, jakie może być *dziecko* czy *but*, to takie sortowanie nic nie da. Do tego celu służy znak ^, zapytanie:

[pos = adj] [base = jak] ^ [pos = subst]

podzieli konkordancję na cztery kolumny (lewy kontekst, prawa część dopasowania, lewa część dopasowania, prawy kontekst) i pozwoli sortować według prawego dopasowania. Szczególne znaczenie ma ta opcja, jeśli w zapytaniu określamy negatywne ograniczenie wyszukiwania. Przypuśćmy, że interesują nas wystąpienia rzeczownika w mianowniku w pozycji innej niż na początku zdania. Zapytanie takie ma formę[11]:

[!base = "[.]"] [pos = subst & cas = nom]

Tyle tylko, że nie interesuje nas pierwszy segment, bo znaczy on: wszystko, byle nie kropka. Dodanie znaku ^ pozwoli posortować konkordancję według tego, co nas rzeczywiście interesuje, czyli rzeczownika:

[!base = "."] ^ [pos = subst & cas = nom]

Podobnie jak PELCRA, Poliqarp pozwala ograniczyć wyszukiwanie za pomocą metadanych, np.:

[base = pies] meta author = Libera

Na koniec należy podkreślić, że jakkolwiek NKJP nie jest udostępniany inaczej niż przez Internet, to istnieje wersja stacjonarna konkordancera Poliqarp, o identycznej składni zapytań, którą można zainstalować komputerze i używać do przeglądania korpusu IPI PAN o objętości 250 milionów segmentów, dostępnego do pobrania pod adresem http://korpus.pl. Zaletą takiego rozwiązania, poza uniezależnieniem się od kaprysów sieci, jest to, że przeglądarka pozwala

[11] Przypomnijmy, że kropka oznacza dowolny znak, jeśli więc chcemy szukać kropki, musimy ująć ją w cudzysłów.

wyszukać do 500 tysięcy wystąpień i dysponuje modułem pozwalającym grupować kolokacje.

Zastosowania

Mimo że NKJP jest korpusem stosunkowo nowym, liczba prac naukowych, które zeń czerpią dane językowe, jest znaczna. Prace te należą do różnych obszarów badań nad językiem, np. słowotwórstwa (Burkacka 2010), frazeologii (Przybylska 2009), dyskursu politycznego (Kołodziej 2012). Sposób wykorzystania korpusu jest w nich oczywiście różny: w niektórych przytaczane są jedynie przykłady użycia dla zilustrowania powziętych skądinąd hipotez, w innych tezy badawcze formułowane są w wyniku analizy danych korpusowych.

Z autorów tego artykułu R. Górski (2008) używał wyszukiwarki Poliqarp do kwerend w korpusie IPI PAN (dostępnym w witrynie NKJP) w badaniach nad diatezą. Ponieważ system znaczników korpusu IPI PAN (podobnie jak NKJP) nie odróżnia słów pomocniczych od słów autosemantycznych, konieczne było formułowanie kwerend, które by możliwie skutecznie wyszukiwały wszystkie wystąpienia strony biernej oraz zdań typu: *Jeśli uważnie kroi się ogórek, spod palców rozchodzi się miła woń* (J. Sosnowski, *Linia nocna*). Przeglądarka Poliqarp daje jednak ogromne możliwości budowania zapytań łączących warunki gramatyczne (odnoszące się do anotacji morfosyntaktycznej) z warunkami czysto formalnymi (odnoszącymi się do budowy literowej słów).

Drugi autor używał Poliqarpa do automatycznej selekcji dwuznacznych zdań typu *Tramwaj wyprzedził samochód*, co posłużyło nie tylko do weryfikacji wcześniejszych badań nad stopniem rozpowszechnienia tego typu struktur, ale też do ukazania szerokich możliwości wyszukiwarki (Bańko 2013). Doświadczenie wynikające z tego rodzaju prac pokazuje, że choć często nie da się sformułować kwerendy idealnej – obejmującej wszystko, czego badacz potrzebuje, i nic ponadto – to jednak można osiągnąć zadowalające wyniki za pomocą kolejnych przybliżeń, osiąganych metodą prób i błędów.

Wśród projektów naukowych opartych na NKJP przedsięwzięciem najważniejszym, bo zakrojonym na najszerszą skalę, jest obecnie *Wielki słownik języka polskiego* IJP PAN, zob. http://wsjp.pl/. W pracach nad nim NKJP służy nie tylko jako źródło cytatów, ale też jako podstawa analizy znaczeń, charakterystyki stylistycznej słów, ich obrazów kolokacyjnych, schematów składniowych itp. Jako jedyny słownik współczesnej polszczyzny o tym stopniu szczegółowości opisu, w dodatku dostępny bezpłatnie w internecie, WSJP adresowany jest nie tylko do językoznawców, lecz do ogółu zainteresowanych użytkowników.

W bardziej akademickim, a zarazem podstawowym (nie użytkowym) projekcie badawczym NKJP służy jako źródło danych do analizy adaptacji wyrazów obcych w języku polskim, zob. http://www.approval.uw.edu.pl. Jedno

z zadań przewidzianych do wykonania w tym projekcie polega na porównawczej analizie 100 par wyrazowych typu *helikopter – śmigłowiec* oraz typu *jazz – dżez*, przy czym różnic między członami każdej pary szuka się m.in. w ich obrazach kolokacyjnych, korzystając z wyszukiwarki PELCRA, obsługującej NKJP.

Z zastosowań najbardziej popularnych można wymienić na koniec bardzo częste powoływanie się na dane z NKJP w Poradni Językowej Wydawnictwa Naukowego PWN (Przepiórkowski i in. 2012b), zob. http://poradnia.pwn.pl. W połowie kwietnia 2013 r. w publicznie dostępnym archiwum poradni NKJP był przywoływany ponad 200 razy, co jest liczbą znaczną, zważywszy na fakt, że wcześniej w odpowiedziach poradni przytaczane były dane z Korpusu Języka Polskiego PWN i dopiero od 2010 r. podstawowym korpusem referencyjnym poradni PWN stał się NKJP – znacznie większy i przede wszystkim lepiej oprogramowany od korpusu PWN. W tym miejscu warto też podkreślić sam fakt korzystania z danych korpusowych w poradnictwie językowym, w którym nieraz za podstawę rozstrzygnięć służyły jedynie słowniki i wydawnictwa normatywne, a uzus, czyli zwyczaj językowy zaświadczony w korpusie, bywał traktowany nieufnie.

Bibliografia

Bańko M. (2013). *Zdania typu* Rury przedziurawiły dzieci *we współczesnej prasie polskiej*. „Poradnik Językowy", nr 4, s. 34-44.

Burkacka I. (2010). *O sufiksach* -ni(a) *i* -eri(a) *w funkcji wykładników nazw miejsc*. „LingVaria" V, nr 1, s. 31-38.

Górski R.L. (2008). *Diateza nacechowana w polszczyźnie. Studium korpusowe*. Kraków.

Kaliska A. (2012). *Co-occurrence des ADV (Instr) figés dans les constructions V ADV (Instr) libres en polonais*. W: *31 e Colloque International sur le Lexique et la Grammaire*.

Kołodziej J.H. (2010). *Democracy as part of the Polish collective identity – the socio-linguistic perspective*. W: Góra M., Mach Z. (red.), *Collective Identity and Democracy. The Impact of EU Enlargement*. Oslo, s. 307-339.

Przepiórkowski A. i Woliński M. (2003). *The Unbearable Lightness of Tagging. A Case Study in Morphosyntactic Tagging of Polish*. W: *The Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, s. 109-116.

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B., red. (2012a). *Narodowy Korpus Języka Polskiego*. Warszawa. http://www.nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

Przepiórkowski A., Bańko M., Łaziński M., Górski R.L., Lewandowska-Tomaszczyk B., Pęzik P. (2012b). *Practical applications of the National Corpus of Polish*. „Prace Filologiczne" LXIII, s. 231-239.

Przybylska R. (2009). *Dwie rodziny frazeologiczne z komponentem kapelusz.* „LingVaria" IV, nr 2, s. 59-68.

Woliński M. (2003). *System znaczników morfosyntaktycznych w korpusie IPI PAN.* „Polonica" XXII-XXIII, s. 39-55.

3. Praktyczny przewodnik po korpusie języka czeskiego

Milena Hebal-Jezierska
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Informacje ogólne

Projekt Czeskiego Korpusu Narodowego (dalej CzKN) jest realizowany w Instytucie Czeskiego Korpusu Narodowego na Uniwersytecie Karola w Pradze. Pierwotnie miał powstawać w Instytucie Języka Czeskiego Czeskiej Akademii Nauk, ale ze względu na niezrozumienie idei przez ówczesną dyrekcję całe przedsięwzięcie zostało przeniesione na Wydział Filozoficzny Uniwersytetu. W 1994 r. został tam założony Instytut Czeskiego Korpusu Narodowego, natomiast w 2000 r. udostępniono pierwszy korpus referencyjny. Inicjatorem tego przedsięwzięcia, a także wieloletnim dyrektorem Instytutu był prof. František Čermák. Nowym dyrektorem, wybranym pod koniec 2012 r., został przedstawiciel młodego pokolenia lingwistów, dr Václav Cvrček. Szczegółowe informacje na temat CzKN można znaleźć na stronie www.korpus.cz

Dostęp do korpusu jest dwojaki. Z niepełnej wersji (bez dostępu do większości funkcji) mogą korzystać wszyscy chętni bez konieczności rejestrowania się. Wystarczy wejść na stronę internetową http://ucnk.ff.cuni.cz/verejny.php. Pełnej wersji korpusu można używać po spełnieniu określonych warunków, m.in. zobowiązania się do niewykorzystywania danych korpusowych do celów komercyjnych. Dostęp do niego jest także bezpłatny; należy wysłać odpowiedni formularz pocztą lub elektronicznie (https://www.korpus.cz/toolbar/signup.php). Po otrzymaniu loginu i hasła trzeba program obsługujący korpus (Bonito 1) przegrać na dysk lub zalogować się na stronie internetowej i korzystać online z korpusu obsługiwanego przez środowisko NoSketch Engine. Tuż przed oddaniem książki do druku do dyspozycji użytkowników CzKN został oddany, tymczasowo w wersji testowej, kolejny program obsługujący korpus. Aby skorzystać z KonText-u, należy zalogować się na stronie internetowej. Nie ma możliwości korzystania z korpusu za pomocą innego oprogramowania, np. Poliqarpa.

Charakterystyka korpusu

Czeski Korpus Narodowy jest jednym z największych zintegrowanych korpusów języków słowiańskich. Prezentuje światowy poziom pod względem przygotowania zarówno korpusów, jak i narzędzi potrzebnych do ich obsługi. Tworzy go w tym momencie 14 korpusów języka pisanego (synchronicznych), 6 korpusów języka mówionego, 2 korpusy diachroniczne oraz kilkanaście korpusów równoległych zintegrowanych w projekcie InterCorp. Tuż przed oddaniem tekstu do druku zostały udostępnione następujące korpusy: SYN2013PUB (korpus tekstów publicystycznych z lat 2005-2009), JEROME (korpus przekładów), SKRIPT2012 (korpus szkolnych prac pisemnych), ORAL2013 (reprezentatywny korpus czeszczyzny mówionej).

Na czternaście korpusów zawierających teksty współczesnego pisanego języka czeskiego jedenaście stanowią korpusy referencyjne, a zaledwie trzy z nich należy zaklasyfikować jako niereferencyjne. Do korpusów niereferencyjnych należy korpus SYN, który jest połączeniem wszystkich korpusów współczesnych tekstów pisanych mających w nazwie skrót SYN. Do omawianych tu zasobów należy także korpus uczniowski nierodzimych użytkowników języka (CZESL--PLAIN) oraz korpus specjalistycznych tekstów lingwistycznych (LINK). Oprócz wyżej wymienionych do baz niereferencyjnych należą wszystkie dostępne korpusy diachroniczne oraz korpusy równoległe.

Tab. 3.1. Struktura Czeskiego Korpusu Narodowego[1]
http://www.korpus.cz/struktura.php

	Liczba słów	**Lematyzacja**	**Anotacja morfologiczna**	**Rok wydania**	**Charakterystyka**
SYN	1300 mln	Tak	Tak	2010	Niereferencyjny, połączenie wszystkich synchronicznie pisanych korpusów typu SYN
SYN2010	100 mln	Tak	Tak	2010	Korpus gatunkowo zrównoważony, przeważają teksty z lat 2005-2009
SYN2009PUB	700 mln	Tak	Tak	2010	Korpus tekstów publicystycznych lat 1995-2007

[1] Tabela nie zawiera korpusów udostępnionych tuż przed oddaniem tekstu do druku: SYN2013PUB (korpus tekstów publicystycznych z lat 2005-2009), JEROME (korpus przekładów), SKRIPT2012 (korpus szkolnych prac pisemnych), ORAL2013 (reprezentatywny korpus czeszczyzny mówionej).

[ciąg dalszy tabeli ze strony poprzedniej]

	Liczba słów	Lematyzacja	Anotacja morfologiczna	Rok wydania	Charakterystyka
SYN	1300 mln	Tak	Tak	2010	Niereferencyjny, połączenie wszystkich synchronicznie pisanych korpusów typu SYN
SYN2010	100 mln	Tak	Tak	2010	Korpus gatunkowo zrównoważony, przeważają teksty z lat 2005-2009
SYN2009PUB	700 mln	Tak	Tak	2010	Korpus tekstów publicystycznych lat 1995-2007
SYN2006PUB	300 mln	Tak	Tak	2006	Korpus tekstów publicystycznych lat 1989-2004

Tab. 3.1A. Korpusy języka mówionego w Czeskim Korpusie Narodowym.

Nazwa korpusu	Liczba słów	Lematyzacja	Tagowanie	Rok publikacji	Charakterystyka
ORAL2008	1 mln	Nie	Nie	2008	Korpus czeszczyzny mówionej, socjolingwistycznie zrównoważony
ORAL2006	1 mln	Nie	Nie	2006	Korpus czeszczyzny mówionej
SCHOLA2010	790 000	Nie	Nie	2010	Korpus języka mówionego używanego w czasie lekcji zarówno przez nauczycieli, jak i uczniów
PMK	675 000	Nie	Nie	2001	Praski korpus mówiony
BMK	490 000	Nie	Nie	2002	Brneński korpus mówiony

Tab. 3.1B. Korpus diachroniczny w Czeskim Korpusie Narodowym.

Nazwa korpusu	Liczba słów	Lematyzacja	Tagowanie	Rok publikacji	Charakterystyka
DIAKORP	1,95 mln	Nie	Nie	2005	Część diachroniczna CzKN

Tab. 3.1C. Korpus równoległy w Czeskim Korpusie Narodowym.

Nazwa korpusu	**Liczba słów**	**Lematyzacja**	**Tagowanie**	**Rok publikacji**	**Charakterystyka**
InterCorp	92 mln	Częściowo	Częściowo	2008	Korpus równoległy (niereferencyjny)

Z korpusów zróżnicowanych gatunkowo można utworzyć podkorpusy zawierające np. jeden gatunek tekstu. Żeby to zrobić, należy wejść w menu Bonito 1 w zakładkę *Dotaz* 'Pytanie', a następnie *Vytvoření subkorpusu* 'Utworzyć subkorpus'. W NoSketch Engine natomiast budujemy podkorpus, klikając link *Subcorpus/Subkorpus*, znajdujący się po prawej stronie w części *Expert options/Pokročilá nastavení* 'Ustawienia zaawansowane'. Można także zbudować podkorpus, wybierając opcję *CQL* oraz *vložit „within"* (więcej zob. rozdział *Praktyczny przewodnik po korpusie równoległym InterCorp* Elżbiety Kaczmarskiej i Alexandra Rosena). W programie KonText sporządzenie podkorpusu jest możliwe po wyborze zakładki *Subcorpora/Subkorpusy*, a następnie opcji *Create new subcorpus/ Vytvořit nový subkorpus*.

Uwagi do pracy z poszczególnymi korpusami

Struktura wybranych korpusów

Na wzmiankę zasługuje przede wszystkim struktura korpusów gatunkowo zrównoważonych. Sposób doboru tekstów w zasobie SYN2000, na który złożyły się teksty w następujących proporcjach: beletrystyka – 15%, literatura specjalistyczna – 25%, publicystyka – 60%, był w środowisku czeskim głośno krytykowany. Negatywnie oceniono zwłaszcza zbyt dużą przewagę publicystyki nad małą liczbą tekstów specjalistycznych oraz beletrystycznych. W związku z tym w następnych korpusach zrównoważonych (SYN2005 oraz SYN 2010) struktura przybrała kształt o wiele bardziej proporcjonalny. 40% zajmują w nich teksty literackie, 27% – literatura specjalistyczna, 33% – publicystyka.

Moja ocena tych struktur nie jest jednoznaczna. W badaniach morfologicznych języka literackiego wyniki uzyskane na korpusie SYN2000, czyli bazie danych o składzie nieproporcjonalnym, są bardziej reprezentatywne. Jest to spowodowane tym, że język w tekstach beletrystycznych dość często odbiega od normy języka literackiego. Jednakże przy badaniach semantycznych (stereotypów narodowościowych) lepiej pracowało się na korpusach o strukturze bardziej zbilansowanej. W tekstach literackich bowiem znajduje się więcej połączeń utrwalonych, pozwalających na rekonstrukcję stereotypu.

Wielkość korpusów

Najtrudniej pracuje się z zasobami małymi, nieanotowanymi, często także niereferencyjnymi. Takim przykładem może być korpus równoległy InterCorp, używany do badań translatologicznych. Badacz ciągle zmaga się z małą ilością materiału, nierzadko także z brakiem anotacji nieczeskiego tekstu.

Ogromnego nakładu pracy wymaga także analiza językowa wykonywana przy wykorzystaniu korpusów tekstów mówionych. Oprócz małego rozmiaru tych zasobów czynnikiem znacznie utrudniającym przeprowadzenie badania jest brak anotacji. Wyszukanie materiału jest bardzo czasochłonne, a rezultaty ekscerpcji bardzo często nie spełniają oczekiwań użytkownika. Nierzadko niemożliwe jest porównanie języka pisanego z mówionym.

Niereferencyjność i referencyjność korpusu

Przy korzystaniu z zasobów niereferencyjnych należy pamiętać, że służą one przede wszystkim do monitorowania konkretnych zjawisk językowych. Struktura tych korpusów ulega częstym zmianom, dodawane składowe nie są względem siebie ilościowo równoważone. Wykonywanie zatem przy ich pomocy badań ilościowych może prowadzić do błędnych wniosków, a uzyskane wyniki mogą być niereprezentatywne (szerzej na ten temat Hebal-Jezierska 2013). Szczególnie niebezpieczna jest w tym wypadku praca z oprogramowaniem Bonito, w którym nie ma możliwości sprawdzenia częstości formy pod względem jej występowania na 1 milion słów. W związku z tym, że ulegająca zmianom struktura tego korpusu nie jest udostępniana publiczności, nie ma też możliwości dokonania samodzielnych wyliczeń. Dla przykładu, śledząc występowanie formy *vúdú* w tekstach (nie znając przy tym struktury niereferencyjnego zasobu SYN), otrzymujemy informację, że rzeczona postać ortograficzna jest dość typowa dla czeskiej publicystyki (stanowi aż 60% wszystkich poświadczeń), rzadsza dla beletrystyki (w tym wypadku jej częstość wynosi ok. 37%). Tymczasem wynik ten jest konsekwencją dużej liczby tekstów prasowych. Kwestia ta jest częściowo rozwiązana w NoSketch Engine. Pomimo tego, że aktualna struktura niereferencyjnego korpusu SYN nie jest podawana, mamy do dyspozycji opcję *Text Types*, która wylicza częstość danej formy w 1 milionie pozycji/tokenów. Z tego wynika, że analizowana tu forma *vúdú* jest typowa dla beletrystyki (wynosi 0,9 na 1 milion pozycji/tokenów), a nie jak mogło się wydawać – dla publicystyki (wynosi 0,1 na 1 milion pozycji/tokenów). Ze względu jednak na swoistość korpusów niereferencyjnych, przede wszystkim dobór tekstów bez spełnienia kryteriów reprezentatywności (dokładane są teksty z kolejnych korpusów), nie należy na nich przeprowadzać badań ilościowych, nawet mając do dyspozycji narzędzia typu *Text Types*.

Przy formułowaniu wniosków bazujących na pracy z referencyjnymi korpusami wielogatunkowymi trzeba pamiętać, że teksty są do nich dodawane w różnych proporcjach. Należy zatem brać pod uwagę częstość występowania badanej formy względem liczebnej reprezentacji danego typu tekstu. W tym wypadku jednak rzecz jest o wiele łatwiejsza, ponieważ struktura jest podana użytkownikowi do wiadomości, a teksty są dobierane według odpowiednich kryteriów. Korzystając z programu Bonito, wyliczeń trzeba dokonać samemu, natomiast, używając oprogramowań NoSketch Engine i KonText, można skorzystać z opisywanej już funkcji *Text Types*. Poniżej prezentujemy sposób obliczeń. Postać *vúdú* w referencyjnym zrównoważonym gatunkowo korpusie SYN2010 występuje 61 razy, z tego w tekstach beletrystycznych – 43 razy. SYN2010 liczy ok. 100 mln słów, natomiast beletrystyka zajmuje w nim 40% całej struktury (czyli ok. 40 mln), zatem rzeczona postać występuje w jednym milionie słów 1,075 raza (jest to rezultat podzielenia liczby 43 mln przez 40 mln (liczba słów występująca w tekstach beletrystycznych). Liczba 43 000 000 jest natomiast iloczynem liczby wystąpień badanej formy (43) i jednego miliona (liczymy wystąpienia formy na 1 mln słów). Warto zwrócić uwagę, że program NoSketch Engine tych samych obliczeń dokonuje, korzystając z liczby pozycji, a nie słów. Pozycją jest każde wystąpienie, czyli wyraz, znak interpunkcyjny, liczba. Jest ich zatem w korpusie więcej niż słów. Dlatego wynik będzie się nieznacznie różnić od rezultatu liczonego przy użyciu wyrazów. Niemniej jednak różnica ta nie wpływa na wynik badania.

Anotacja zewnętrzna

Dane zawarte w tzw. anotacji zewnętrznej są w Czeskim Korpusie Narodowym dosyć szczegółowe, choć ich zakres nie jest jednakowy we wszystkich korpusach nawet tego samego typu.

W korpusach zrównoważonych, synchronicznych, wielogatunkowych jest to kwestia dosyć zróżnicowana. W SYN2000 oraz w FSC2000 wyświetlają się informacje o autorze, tytule w postaci nazwy skróconej, typie tekstu, także oznaczonym kodem (ich wyjaśnienia dostępne są na stronie http://wiki.korpus.cz/doku.php/cnk:syn2000) i roku wydania. W SYN2005 oraz w SYN2010 metadane są znacznie bardziej szczegółowe. Możemy zatem uzyskać informacje o autorze, tytule, wydawnictwie, miejscu i roku wydania, typie tekstu wraz z jego podtypem, języku źródłowym, autorze przekładu, numerze ISBN/ISSN (najczęściej jednak to pole pozostaje puste). W korpusie niereferencyjnym SYN elementy składowe anotacji zewnętrznej są prezentowane w ten sam sposób, co w korpusach SYN2005 oraz SYN2010. Dotyczy to także zasobu SYN2000, w którym w korpusie pojedynczym, jak zostało wyżej wspomniane, są wyświetlane mniej szczegółowe dane. Dodatkowym elementem umieszczonym w bazie

SYN jest informacja wskazująca korpus, z którego pochodzi konkretne poświadczenie. W zasobie LINK anotacja zewnętrzna jest oparta na podobnej zasadzie jak w korpusach SYN2005 oraz SYN2010. Dodatkowym parametrem jest dyscyplina językoznawstwa, np. lingwistyka ogólna. W korpusach ORWELL i ORW-mte[2] metadane są ograniczone do numerów, jakie zajmuje w kolejności dane słowo oraz zdanie.

Bardzo szczegółową anotację zewnętrzną posiada korpus korespondencji KSK-DOPISY. Znajdziemy tu numery identyfikacyjne (sygnatury, numer z archiwum, numer identyfikacyjny), dane o nadawcy (płeć, wiek, wykształcenie, miejsca i czas długotrwałego pobytu osoby piszącej list), stosunek adresata do nadawcy, informacje o adresacie (płeć, wiek, wykształcenie), dane o liście (rok i forma napisania listu - ręcznie, pisany na maszynie, wydrukowany z komputera).

W korpusach mówionych BMK oraz PMK są wyświetlane informacje na temat osoby wypowiadającej tekst: płeć (Z oznacza kobietę, M – mężczyznę), wiek (I 20-35, V powyżej 35), wykształcenie (B – wykształcenie podstawowe najwyżej maturalne, A – wyższe włącznie ze studentami) oraz określany jest stopień oficjalności rozmowy (F – oficjalna, N – nieoficjalna). Wadą tej anotacji jest zbyt małe zróżnicowanie wieku oraz wykształcenia. Szczególnie podział od 35 lat wzwyż wydaje się zbyt ogólny. Podobnie wykształcenie podstawowe powinno być oddzielone od średniego.

Oznaczenia stosowane w omawianych tu korpusach języka mówionego są wykorzystane w zasobach ORAL2006 oraz ORAL2008. Dodatkowo znajdują się tam także dane na temat zaszeregowania dialektalnego oraz liczby respondentów, ich wieku podanego w latach oraz skrót ukończonej przez nich szkoły.

Bardzo szczegółowe metadane są dostępne w korpusie SCHOLA2010. Dotyczą one zarówno przeprowadzonej sondy, jak i szkoły, klasy oraz lekcji, w której została nagrana, a także osoby uczestniczącej w badaniu (razem 46 parametrów). Co ważne, można tworzyć subkorpusy pozwalające na filtrowanie informacji według poszczególnych kryteriów, np. wybrać teksty mówione tylko przez kobiety lub nagrywane wyłącznie w Pradze. Żeby otrzymać wypowiedzi spełniające określone warunki, należy w menu Bonito 1 rozwinąć zakładkę *Korpus*, następnie *Vytvoření subkorpusu* 'Sporządzenie korpusu', dalej w okienku z napisem *Jméno subkorpusu* 'Nazwa subkorpusu' wpisujemy konkretny parametr, np. *zeny*, w polu oznaczonym *Značka* odpowiedni skrót, w tym wypadku *sp*, a w miejscu określonym jako *Podmínka* 'Warunek', np. pohlavi="Z". Do filtrowania danych można także użyć opcji filtru pozytywnego lub negatywnego lub odpowiedniej sekwencji wyrazów regularnych. Dokładniejsze informacje na temat selekcji materiału są dostępne pod adresem: https://ucnk.ff.cuni.cz/schola_vyhled.php. W programie NoSketch Engine budowa subkorpusu jest o wiele

[2] Korpus znakowany tagami wypracowanymi w projekcie EU Multext-East.

prostsza. Po kliknięciu zakładki *Subcorpus/Subkorpus* można wybrać kolumnę z danymi, których potrzebujemy do przeprowadzenia analizy, lub wykorzystać funkcję *CQL- vložit „within"*.

W niereferencyjnym korpusie uczniowskim (CzESL-PLAIN) anotacja zewnętrzna jest bardzo skąpa; właściwie otrzymujemy tylko nazwę dokumentu, typ tekstu (podział obejmuje eseje, teksty specjalistyczne oraz prace Romów zamieszkujących tereny zagrożone wyłączeniem społecznym).

Metadane korpusów diachronicznych zawierają następujące elementy: autor, tytuł, rok tekstu, strona oraz data włożenia źródła do DIACORP-u. W przypadku, gdy nie można określić dokładnego roku, wpisywany jest wiek.

Elementy składowe anotacji wewnętrznej nie pojawiają się w programie Bonito 1 bez włączenia odpowiedniej funkcji. Aby uzyskać dostęp do tych danych, należy wybrać w menu Bonito 1 zakładkę '*Zobrazení*', a następnie kliknąć napis *Zdroje* 'Źródła'. Po zastosowaniu tej procedury pokażą nam się po lewej stronie każdego poświadczenia zaznaczone przez nas dane. Możemy użyć innego sposobu, czyli po prostu kliknąć poświadczenie prawym przyciskiem myszy. W dolnym okienku pojawią się wtedy dane źródła, z którego pochodzi fragment tekstu. Zaletą opcji wyboru z menu jest to, że można zaznaczyć, których danych potrzebujemy. Bardzo przydatna jest funkcja tworzenia subkorpusów, pozwalająca na wyszukiwanie materiału spełniającego określone przez badacza warunki, np. wyszukiwania danego gatunku tekstu lub tekstów z określonego roku czy też tekstów oryginalnych[3]. Podczas wielu lat pracy z tymi korpusami nie zauważyłam rażących błędów w wyświetlaniu i poprawności metadanych.

W oprogramowaniach NoSketch Engine i KonText automatycznie wyświetla się po lewej stronie każdego poświadczenia tytuł utworu. Jeżeli klikniemy go dwa razy lewym przyciskiem myszy, w ramce pojawią się na dole strony wszystkie dane bibliograficzne. Jeśli natomiast potrzebujemy tych danych dla wszystkich poświadczeń lub chcemy wyświetlać tylko część z nich, to w NoSketch Engine należy nacisnąć zakładkę z nazwą *View/Zobrazení* (*Custom/Vlastní*, natomiast w programie KonText trzeba wybrać opcję *View options/Možn. zobrazení* (*Attributes, structures and references/Atributy, struktury a metainformace*).

Anotacja wewnętrzna

Anotacja wewnętrzna (lematyzacja oraz tagowanie morfologiczne) jest obecna we wszystkich korpusach typu SYN. Tagset oparto na gramatyce tradycyjnej, co oznacza, że grupa liczebników i zaimków jest tu podzielona według schematu tradycyjnego. Liczebniki porządkowe będą zatem oznaczane jako liczebniki,

[3] Warto zwrócić uwagę, że w Czeskim Korpusie Narodowym znajdują się też teksty tłumaczone na język czeski z innych języków.

a nie, jak w polskim korpusie, jako przymiotniki. Wyróżnia się więc: rzeczownik, czasownik, przymiotnik, przysłówek, liczebnik, zaimek, przyimek, spójnik, wykrzyknik, partykułę. Dodatkowo oznaczone są wyrazy nierozpoznane przez tager oraz znaki interpunkcyjne. Tagi występują w następującej kolejności: część mowy, rodzaj, liczba, przypadek, osoba, stopień, czas, negacja, strona, nacechowanie stylistyczne, aspekt (niedostępny w SYN2000 oraz ORWELL-u). Szczegóły dostępne są na stronie http://wiki.korpus.cz/doku.php/seznamy:tagy. W żadnym z korpusów zintegrowanych pod nazwą Czeski Korpus Narodowy nie ma anotacji składniowej ani semantycznej. Poprawność zarówno lematyzacji, jak i tagowania morfologicznego jest najsłabsza w pierwszym z korpusów, należących do grupy SYN, czyli SYN2000. W pozostałych część błędów została poprawiona, ale trzeba uważać zwłaszcza na wyrazy homonimiczne i formy synkretyczne. Zalecam sprawdzanie manualne wyników otrzymanych z automatycznego wyszukiwania. Im wyższy numer korpusu, tym lepsza anotacja wewnętrzna. Porównajmy hasłowanie leksemu *optimismus*. W tabeli zostały zestawione ze sobą wyniki z korpusów SYN2000, SYN2005, SYN2010. Jak widać, w pierwszym z wymienionych zasobów leksem *optimismus*, niezależnie od postaci ortograficznej (z *-ismus-* lub *-izmus-*), w jakiej występuje w poświadczeniu, jest lematyzowany zawsze w formie *optimizmus*. Wprowadza to badacza w błąd, dlatego, że rzeczywistych wariantów z *-z-* jest o wiele mniej niż z *-s-*. W korpusach SYN2005 i 2010 ten błąd został naprawiony. Lemat odpowiada formom ortograficznym wyrazów.

Tab. 3.2. Postać lematu form ortograficznych leksemu *optimismus*/*optimizmus* w poszczególnych korpusach SYN.

SYN 2000			**SYN 2005**		
Lemat	**Słowo**	**Frekwencja**	**Lemat**	**Słowo**	**Frekwencja**
optimizmus		2030	optimismus		1415
	optimismus	890		optimismus	630
	optimismu	592		optimismu	429
	optimismem	370		optimismem	257
	Optimismus	147		Optimismus	79
	OPTIMISMUS	8		Optimismem	9
	Optimismem	7		Optimismu	6
	Optimismu	6		OPTIMISMUS	4
	optymizmu	4		OPTIMISMU	1
	optimizmus	3	optimizmus		12
	OPTIMISMEM	1		optimizmus	7
	OPTIMISMU	1		optymizmu	3
	Optymizmem	1		Optimizmus	1
				optymizmem	1

[ciąg dalszy tabeli ze strony poprzedniej]

SYN2010		
Lemat	**Słowo**	**Frekwencja**
optimismus		1296
	optimismus	569
	optimismu	396
	optimismem	251
	Optimismus	60
	Optimismu	5
	Optimismem	4
	OPTIMISMU	3
	optimism	3
	OPTIMISMEM	2
	OPTIMISMUS	2
	Optimisme	1
optimizmus		13
	Optimizmus	6
	Optymizmu	4
	Optymizmem	2
	Optimizmus	1

Nie oznacza to jednak, że wszystkie niekonsekwencje w tym zakresie zostały wyeliminowane, np. forma ortograficzna *voodoo* jest zlematyzowana we wszystkich korpusach jako *vúdú*. Źle oznaczone tagi także mogą wprowadzić w błąd badacza, który nie sprawdzi wyników otrzymanych automatycznie.

Poniżej podaję przykład formy *matky*, która powinna być oznaczona jako rzeczownik rodzaju żeńskiego w dopełniaczu liczby pojedynczej (NNFS2), tymczasem tag wskazuje na biernik liczby mnogiej.

"Řekl jsi jméno své <matky/NNFP4-----A----->," vysvětlil Cadfael
[Petersová E., Pokání bratra Cadfaela, Praha 2004]
'Wypowiedziałeś imię swojej matki, wyjaśnił Cadfael'

Korpus FSC2000 nie jest wprawdzie tagowany morfologicznie, ale ma lematyzację pod względem poprawności na wyższym poziomie niż SYN2000.

Zarówno lematyzacji, jak i tagowania brakuje we wszystkich korpusach języka mówionego, korpusach diachronicznych, w korpusie uczniowskim oraz w korpusie korespondencyjnym. W korpusie równoległym informacje te są dostępne zawsze po stronie czeskiej, po stronie języka obcego zaś nie zawsze i nie w całości. Należy pamiętać, że brakuje ujednoliconego dla wszystkich języków tagsetu. Przy wyszukiwaniu warto także pamiętać, że pozostałe języki mogą mieć system znaczników oparty na odmiennej gramatyce niż język czeski.

Program rozróżnia małe i duże litery, więc jeśli chcemy, żeby wyszukał wyrazy w obu wariantach ortograficznych, musimy to zaznaczyć w zapytaniu

(np. [Mm]atky lub [lc="matky"]). Jeśli wyszukujemy według komendy typu [lemma="matka"], program znajdzie wyrazy pisane zarówno od dużej, jak i małej litery. Leksemy złożone zawierające w sobie dywiz lub apostrof są lematyzowane w zależności od tego, czy pomiędzy wyrazem a danym znakiem występuje spacja czy nie. Leksemy bez spacji są najczęściej, ale nie zawsze, traktowane jako całość wraz ze znakiem, wyrazy ze spacją – oddzielnie. Dlatego słowa z łącznikiem i dywizem są nierzadko lematyzowane jako jedna jednostka.

Tab. 3.3. Postać lematu leksemów zawierających dywizy, apostrofy, ukośniki itp.

Wyraz/liczba	**Lemat(y)**
1. Połączenia bez spacji	
Wyrazy pospolite z łącznikiem	
oranžově-růžově-květináčově	oranžově-růžově-květináčově
Nazwiska podwójne	
Müllerová-Blekotová	Müllerová-Blekotová
Czasownik z partykułą -li	
Bude–li	být – li (3 oddzielne lematy)
Poprzyimkowe formy nieakcentowane	
Doň	Doň
Formy z użyciem apostrofu	
Sister's	Sister's
Skróty i skrótowce	
ČNK	ČNK
str	str
Inne połączenia bez spacji	
8/5	8/5
2. Połączenia ze spacją	
50 – 51	50 - 51 (3 oddzielne lematy)
8/5	8 / 5 (jw.)

Warto wspomnieć o tokenizacji tekstów w korpusie. Token zwany tu jest pozycją. Samodzielnym tokenem/pozycją jest zatem nie tylko wyraz, ale także liczba oraz znak interpunkcyjny. Z tego względu liczba słów w korpusie różni się od liczby tokenów.

Ani lematyzacja, ani tagowanie nie wyświetla się automatycznie. W Bonito 1 należy wejść w *Zobrazení – Atributy* i tam zaznaczyć interesujące nas opcje. W NoSketch Engine te same funkcje znajdziemy w opcji *View*, w oprogramowaniu KonText – w zakładce *View options*.

Podstawowe zapytania

Zapytania tworzymy za pomocą wyrażeń regularnych oraz wyrazów umownych wpisywanych w okienko programu z napisem *Nový dotaz* 'Nowe zapytanie'. Dokładny spis używanych symboli wraz z ich znaczeniem dostępny jest

na stronach: http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy. W wyrażeniach regularnych używa się znaku kropki, gwiazdki, plusa, znaku zapytania, nawiasu kwadratowego, nawiasu okrągłego, pionowej kreski, odwrotnego (lewego) ukośnika (backslash).

Poniżej przedstawiamy przykładowe zapytania.

Tab. 3.4. Przykładowe zapytania.

Typ zapytania	**Przykłady zapytań używanych w oprogramowaniu Bonito 1** (Zapytania wpisujemy do okienka z napisem *Nový dotaz*)	**Przykłady zapytań używanych w oprogramowaniach NoSketch Engine oraz KonText**	**Tłumaczenie**
Zapytanie o formę wyrazową o określonej postaci ortograficznej, np. pisane wielką literą	Knížky Otrzymamy poświadczenia pisane wielką literą.	Knížky Aby otrzymać poświadczenia o określonej postaci ortograficznej (w tym przypadku pisane wielką literą), należy wpisać wyraz, zaznaczając wcześniej opcję *Phrase/Fráze*.	Książki
Zapytanie o formę wyrazową z wariantami ortograficznymi	[Kk]nížky Otrzymamy poświadczenia pisane zarówno wielką, jak i małą literą.	Jeśli uprzednio zaznaczymy opcję *Basic/Základní* lub *Word form/Slovní tvar*, to możemy wpisać wyraz Knížky lub knížky. Program w obu wypadkach znajdzie słowa kluczowe pisane od dużej i małej litery.	Książki książki
Zapytanie o lemat	[lemma="knížka"]	Najpierw zaznaczamy opcję *Lemma*, a następnie wpisujemy leksem, np. knížka.	książka
Zapytanie o część wyrazu	.*knížka kníž.* .*kníž.*	.*knížka kníž.* .*kníž.* W zależności od tego, czy chcemy wyniki otrzymać w formie kanonicznej lub w jakiejkolwiek formie gramatycznej, musimy uprzednio zaznaczyć opcję *Lemma*, *Phrase/Fráze*, *Word Form/Slovní Tvar*. Należy pamiętać, że zaznaczenie zakładek *Basic* lub *CQL* zwróci zbiór pusty.	.*książka książ.* .*książ.*

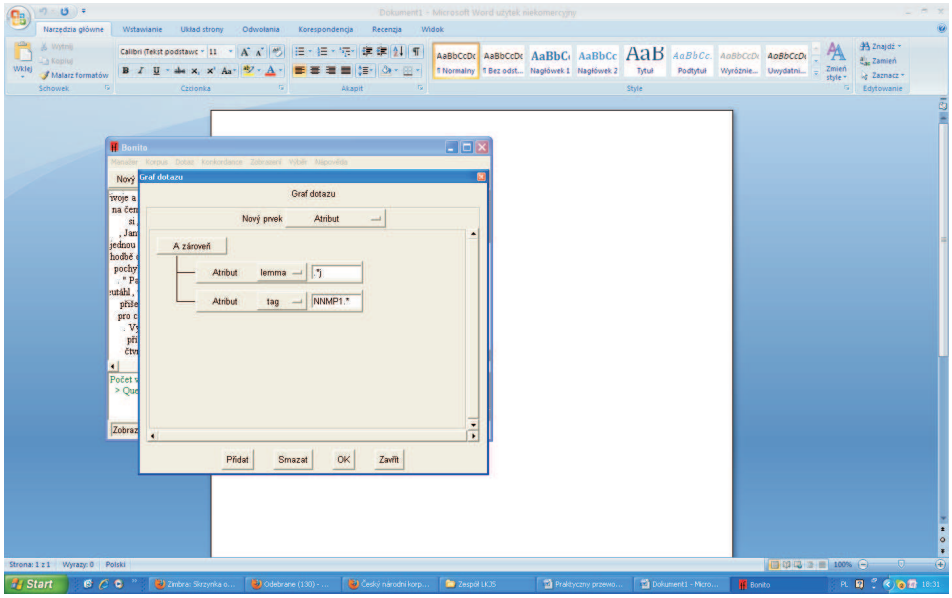
[ciąg dalszy tabeli ze strony poprzedniej]

Typ zapytania	**Przykłady zapytań używanych w oprogramowaniu Bonito 1** (Zapytania wpisujemy do okienka z napisem *Nový dotaz*)	**Przykłady zapytań używanych w oprogramowaniach NoSketch Engine oraz KonText**	**Tłumaczenie**
Zapytanie o grupę wyrazów	tlusté knížky	tlusté knížky Najpierw należy zaznaczyć opcję *Phrase/Fráze*.	grube książki
Zapytanie o grupę lematów	[lemma="tlustý"][lemma="knížka"]	Należy wybrać opcję *CQL* i wpisać [lemma="tlustý"][lemma="knížka"]	gruby książka
Zapytanie o kategorię gramatyczną	[tag="NNMP1.*"]	Wybieramy opcję *CQL*, a następnie *vložit tag*. Kolejne symbole wybieramy z listy rozwijanej. Program sam formułuje zapytanie.	Rzeczownik męski żywotny w mianowniku liczby mnogiej
Zapytanie o lemat w określonej kategorii gramatycznej	[(lemma=".*j")& &(tag="NNMP1.*")]	Należy wybrać opcję *CQL* i wpisać [(lemma=".*j")& &(tag="NNMP1.*")]	Leksem zakończony na *-j* w mianowniku liczby mnogiej

Jest jeszcze druga możliwość zadania pytania – graficzna (Bonito 1). Znaleźć ją można w menu w zakładce *Dotaz* 'Zapytanie', dalej *Grafické vytváření* 'Tworzenie graficzne'. Funkcja ta jest użyteczna dla osób, które mają problemy z zapamiętaniem sekwencji i typów znaków do utworzenia zapytania. Na stronach korpusu oraz serwisu Youtube jest dostępna krótka prezentacja na temat tworzenia zapytań w sposób graficzny: https://ucnk.ff.cuni.cz/bonito/graficky.php.

Przykład:
Chcemy znaleźć wyrazy kończące się na *j*, a zarazem występujące w mianowniku liczby mnogiej. Możemy zatem wpisać do okienka z zapytaniem komendę: [(lemma=".*j")&(tag="NNMP1.*")] lub posłużyć się opcją tworzenia pytania w sposób graficzny i ułożyć pytanie tak jak poniżej na ilustracji.

Rys. 3.1.



Po wciśnięciu przycisku *OK* program automatycznie wpisze odpowiednie pytanie do okienka.

Funkcje dostępne w programie obsługującym CzKN

Wszystkie korpusy z wyjątkiem korpusów równoległych zintegrowanych pod nazwą InterCorp są obsługiwane przez program Bonito, który umożliwia korzystanie z wielu funkcji ułatwiających proces wyszukiwania i analizy. Natomiast przez programy NoSketch Engine są oraz KonText obsługiwane wszystkie korpusy włącznie z paralelnymi.

Wybór liczby poświadczeń

Liczba wyświetlanych poświadczeń jest ustalana przez użytkownika programu Bonito 1 w zakładce *Zobrazení – Rozsah*. Dodatkowo można wybrać, które z nich chcemy obejrzeć (pierwsze, środkowe, ostatnie, przypadkowe). Jeśli interesują nas wszystkie dane, to program je wyświetli. Trzeba jednak pamiętać, że przy bardzo dużej ilości materiału może się on zawiesić w trakcie pracy. Aby tego uniknąć, należy przeglądać poświadczenia stopniowo, wykorzystując opcję *Skok na řádek* 'Skok na wiersz', który umożliwia przejście do poświadczenia o określonym numerze. Oprócz tego możemy zaznaczyć interesujące nas poświadczenia, stosując kombinację klawiszy CTRL+ALT+INSERT, i skopiować je do pliku, w którym pracujemy.

W NoSketch Engine ani w programie KonText niestety nie ma funkcji umożliwiającej przechodzenie do dowolnego poświadczenia. Liczbę poświadczeń można określić, używając opcji *View/Zobrazení*. W podkategorii *Custom/Vlastní* w jednym okienku wybieramy oprócz liczby poświadczeń także szerokość kontekstu oraz wyświetlanie lematów i tagów słów kluczowych lub wszystkich tokenów. Natomiast z lewej strony w menu rozwija się wybór próbki poświadczeń *Sample/Vzorek*[4] oraz filtrowania *Filter/Filtr*[5]). Oprócz omówionej kategorii *Custom* w menu *View/Zobrazení* dostępna jest ponadto opcja KWIC (wyraz kluczowy znajduje się pośrodku wiersza) oraz *Sentence* (pokazuje zdania bez wyrównania do KWIC). W oprogramowaniu KonText liczbę poświadczeń można nastawić w zakładce *View options/Možn. zobrazení*.

Ustawianie kontekstu

Kontekst dla wyszukanego materiału można ustawić w zakładce *Zobrazení – Kontext*. Maksymalna liczba znaków na lewo i/lub na prawo od KWIC wynosi 500. W przypadku konieczności rozszerzenia kontekstu wybranego poświadczenia, należy kliknąć dwa razy interesujące nas zdanie; jego poszerzoną wersję znajdziemy w dolnym okienku. Druga z opcji jest wygodniejsza dla użytkownika, któremu długość kontekstu przeszkadza w jednoczesnym sprawdzaniu źródeł poświadczeń. Ma jednak jedną wadę. Z okienka, w którym pojawia się zdanie w szerszym kontekście, nie można kopiować tekstu do pliku, w którym pracujemy.

W NoSketch Engine kontekst ustawiamy w opcji *View/Zobrazení*, a w oprogramowaniu KonText wybieramy zakładkę o nazwie *View options/Možn. zobrazení*.

Filtry

W Bonito 1 istnieje możliwość oddzielenia zbędnych poświadczeń od poświadczeń od potrzebnych do dalszego badania. Służą do tego dwie funkcje: tzw. filtr pozytywny oraz negatywny. Przy korzystaniu z pierwszego z nich w zapytaniu powinno znaleźć się wyrażenie potrzebne do dalszej analizy. Przy użyciu filtru negatywnego natomiast w zapytaniu należy wpisać to, co ma zostać usunięte z listy wyników. Posłużmy się następującym przykładem. W okienko zapytań wpisaliśmy komendę *.*proti*. Otrzymujemy szereg wyrazów np. *naproti*, *oproti*, *proti*. Jednakże do dalszego badania potrzebny jest nam tylko wyraz *oproti*. Klikając zatem filtr pozytywny, wpisujemy w okienko zapytań leksem *oproti*. Jeśli natomiast chcemy poświadczenia leksemu *oproti* usunąć, to po wpisaniu

[4] Ta opcja umożliwia wyświetlenie poświadczeń pierwszych w kolejności, środkowych, ostatnich oraz przypadkowych.

[5] Filtr pozytywny oraz negatywny.

go w okienko zapytań, zaznaczamy filtr negatywny. Funkcja ta jest niezwykle użyteczna przy homonimach.

W oprogramowaniach NoSketch Engine i KonText ta sama funkcja jest dostępna w opcji *Filter/Filtr*.

Wybór i redukcja poświadczeń

W Bonito 1 mamy możliwość zaznaczenia i usunięcia konkretnych poświadczeń. Można także dokonać tzw. inwersji zaznaczonych rzędów, wtedy program wyróżnia kolorem nieoznaczone poświadczenia, natomiast usuwa oznaczenie wierszy uprzednio wyróżnionych. W NoSketch Engine oraz w oprogramowaniu KonText na razie tych funkcji brakuje, co znacząco utrudnia analizę językową. Trwają prace nad ich udostępnieniem.

Sortowanie poświadczeń

W Bonito 1 możemy posortować poświadczenia według alfabetu *a fronte* lub *a tergo*. Żeby to zrobić, należy wejść w *Konkordance* 'Konkordancja' – *jednoduché třídění* 'Sortowanie zwykłe'. Alfabetycznie mogą zostać uporządkowane wiersze według wyrazu wyszukiwanego lub według danego leksemu leżącego w określonej przez nas odległości od KWIC. Jeśli szukaliśmy materiału na podstawie zapytania:

.*tunel.*

i zaznaczyliśmy opcję *lewostronny KWIC*, to otrzymamy rzędy poświadczeń w kolejności następującej: *Albulatunel*, *Citrofortunella*, *ekotunel*, *Eurotunel* itp.

Dostępne jest także bardziej zaawansowane sortowanie poświadczeń znajdujące się w funkcjach *Obecné třídění* 'Sortowanie ogólne' (umożliwia czytanie poświadczeń posortowanych według wyrazów poprzedzających słowo kluczowe lub następujące po nim) oraz *Třídění podle Skupin* 'Sortowanie według grup' (po kliknięciu konkordancji możemy oznaczyć cyframi 1-9 wybrane poświadczenia, tworząc grupy. Wpisanie cyfry 0 spowoduje usunięcie pozostałych liczb).

W NoSketch Engine mamy dodatkowo do wyboru sortowanie wielopoziomowe dostępne w opcji *Sort/Třídění*. Opcja *Sort/Třídění* pozwala na otrzymanie konkordancji posortowanej według wyrazu lub wyrazów znajdujących się przed po albo samym KWIC. W podkategorii *Custom* możemy skonfigurować program stosownie do naszych potrzeb. Po wyborze opcji *Left/Vlevo* program sortuje alfabetycznie według pierwszego wyrazu poprzedzającego KWIC, po wyborze opcji *Right/Vpravo* sortuje alfabetycznie według wyrazu następującego po KWIC. Naciskając *Node/KWIC*, otrzymujemy poświadczenia posortowane alfabetycznie według KWIC, przyciskając *References/Reference* – posegregowane alfabetycznie według tytułów tekstów. Funkcja *Shuffle/Promíchat*

umożliwia otrzymanie poświadczeń wymieszanych. W programie KonText zarówno sortowanie zwykłe, jak i wielopoziomowe znajdziemy w zakładce *Concordance/Konkordance* (*Complex sorting options/Úplné možnosti třídění*).

Funkcje statystyczne

We wszystkich oprogramowaniach do dyspozycji użytkownika jest moduł ekstrakcji kolokacji. W Bonito 1 można go znaleźć w zakładce *Konkordance* 'Konkordancja' – *Statistiky* 'Statystyki' – *Kolokace* 'Kolokacje'. W NoSketch Engine oraz w programie KonText natomiast wejście do modułu znajduje się na pasku z napisem *Collocations/Kolokace*. Kolokator działa bez zastrzeżeń. Wyszukiwanie kolokatów jest możliwe według różnych atrybutów (lematu, formy fleksyjnej, tagu) w dowolnej odległości od słowa kluczowego W Bonito 1 wyniki pokazują się w tabeli wraz z następującymi testami statystycznymi: informacja wzajemna, test T, frekwencja względna kolokatu, frekwencja całkowita kolokatu. W zależności od ustawienia możemy otrzymać wyniki pokazujące kolokaty o najwyższej wartości informacji wzajemnej lub testu T. Poniżej przedstawiamy wyniki wyszukiwania lewostronnych kolokatów dla leksemu *vlk* 'wilk' (pozycja -1) (zob. także poniżej rozdział Jerzego Molasa pt. *Praktyczny przewodnik po korpusie języka chorwackiego*).

Tab. 3.5. Lewostronne kolokaty leksemu *vlk*.

Kolokaty w postaci lematu	**MI-score (Informacja wzajemna)**	**T-score (Test T)**	**Rel. F [%] (Frekwencja względna wyrażona w procentach)**	**Abs. F (Frekwencja całkowita)**
vačnatý	14,82	3	50	9
vysočinský	12,1	2,645	7,609	7
vytí	11,49	4,241	4,972	18
stepní	11,45	2,999	4,839	9
nažraný	11,38	1,731	4,615	3
apeninský	11,1	1,731	3,797	3
smečka	10,87	5,193	3,241	27
výt	10,62	3,315	2,723	11
předhodit	10,38	2,234	2,315	5
krvelačný	10,06	1,998	1,852	4
vyhladovělý	10,03	1,998	1,81	4
hladový	9,935	4,686	1,695	22
mořský	9,913	8,935	1,669	80
zat.	9,437	1,73	1,2	3
sežrat	8,685	2,639	0,7128	7
šedý	8,281	5,083	0,5385	26
osamělý	8,275	3,73	0,5364	14
polární	7,236	1,721	0,2611	3
výskyt	6,907	2,624	0,2078	7

[ciąg dalszy tabeli ze strony poprzedniej]

Kolokaty w postaci lematu	**MI-score (Informacja wzajemna)**	**T-score (Test T)**	**Rel. F [%] (Frekwencja względna wyrażona w procentach)**	**Abs. F (Frekwencja całkowita)**
vlk	6,776	1,982	0,1898	4
říše	6,487	2,797	0,1553	8
zlý	6,383	3,277	0,1445	11

Tabela wyświetlana w Bonito 1 zawiera 5 kolumn. W pierwszej znajdują się kolokaty w formie uprzednio wybranej przez użytkownika (np. słowo, lemat, tag). Następna kolumna, oznaczona w tabeli symbolem MI-score, przedstawia wartości testu informacji wzajemnej, pozwalającego na ocenę statystycznej istotności podanych połączeń wyrazów. Im wyższa jego wartość, tym większe prawdopodobieństwo, że dane wyrażenie/zwrot jest statystycznie istotną kolokacją. Połączenia wyrazowe o wartości co najmniej 7 są często uważane za potencjalnie spełniające warunek kolokacyjności. Druga kolumna przedstawia wyniki alternatywnego testu, tzw. testu T-score. I w tym wypadku znamienna jest wysoka wartość liczbowa: im wyższa, tym kolokacja bardziej istotna. W trzeciej kolumnie znajdują się informacje o tzw. frekwencji względnej wyrażonej w procentach, natomiast czwarta kolumna zawiera dane na temat frekwencji całkowitej konkretnego połączenia wyrazowego.

Najbardziej reprezentatywne wyniki otrzymamy, kierując się wartościami informacji wzajemnej. Warto je porównać z wynikami testu T, a przede wszystkim wartościami frekwencji całkowitej kolokatu, aby otrzymać pełny obraz statystyczny wybranego połączenia. Korzystając z wartości informacji wzajemnej, trzeba być świadomym, że najwyższe wyniki osiągają leksemy z niską frekwencją. Warto zatem zabezpieczyć się przed tym, ustalając dolną granicę częstości występowania wyrazów w trakcie pracy z modułem ekstrakcji kolokacji.

Odnalezione przez program połączenia należy sprawdzić, wpisując je kolejno w okienko zapytań. Zdarza się bowiem, że wyszukane przez program kolokacje tworzą połączenia pozorne. Dochodzi do tego zwłaszcza wtedy, gdy leksem pochodzący z prasy stanowi np. tytuł lub podtytuł tekstu. Program może połączyć taki wyraz ze słowem zaczynającym zdanie.

O wiele lepszy kolokator jest dostępny w oprogramowaniach NoSketch Engine oraz KonText. Wyróżnia się większą liczbą testów pozwalających na zbadanie łączliwości wyrazów. Oprócz tych dostępnych w Bonito (informacja wzajemna, test T, frekwencja względna, frekwencja całkowita) kolokaty są tu udostępniane także przy użyciu między innymi logarytmu Dice, innych typów informacji wzajemnej (MI3), testu wiarygodności. Na szczególną uwagę zasługuje logarytm Dice, dzięki któremu jakość wyszukanych kolokacji jest wyższa niż przy użyciu innych testów, w tym informacji wzajemnej. Zwraca on o wiele

mniej przypadkowych połączeń, kolokacje są trafniej dobrane. Moje spostrzeżenia potwierdza także literatura przedmiotu (por. Curran 2004). Dodatkowym udogodnieniem jest możliwość wyboru zarówno testów badających łączliwość leksemów, jak i testu według którego kolokaty będą porządkowane.

Tab. 3.6. Testy dostępne w module ekstrakcji kolokacji dostępnym w NoSketch Engine na przykładzie kolokatów leksemu *vlk* uporządkowanych według wartości logarytmu Dice'a.

	Freq	**T-score**	**MI**	**MI3**	**log likeli-hood**	**min. sensi-tivity**	**logDice**	**MI.log_f**	**relative freq.** [%]
mořský	80	8,935	9,913	22,556	943,920	0,017	8,570	43,560	1,669
smečka	28	5,289	10,923	20,537	369,326	0,013	8,286	36,780	3,361
nažrat	19	4,358	12,885	21,381	304,179	0,009	8,111	38,601	13,103
výt	21	4,581	11,551	20,336	295,623	0,010	8,098	35,706	5,198
vytí	20	4,471	11,639	20,283	284,043	0,009	8,052	35,437	5,525
hladový	23	4,791	9,999	19,046	273,518	0,011	7,790	31,777	1,772
samotář	14	3,740	11,330	18,945	192,634	0,007	7,566	30,683	4,459
vačnatý	9	3,000	14,817	21,157	172,434	0,004	7,117	34,118	50,000
stepní	9	2,999	11,448	17,788	125,322	0,004	7,007	26,360	4,839
sežrat	12	3,459	9,463	16,633	133,667	0,006	6,992	24,272	1,222
šedý	26	5,083	8,281	17,682	247,091	0,005	6,941	27,292	0,539
vysočinský	7	2,645	12,101	17,716	104,004	0,003	6,705	25,164	7,609
předhodit	7	2,644	10,870	16,485	91,742	0,003	6,626	22,603	3,241
osamělý	14	3,730	8,275	15,890	132,858	0,005	6,604	22,409	0,536
bál	6	2,447	9,698	14,868	68,787	0,003	6,283	18,872	1,439
vlk	8	2,816	7,776	13,776	70,376	0,004	5,959	17,087	0,380
krvelačný	4	1,998	10,063	14,063	47,888	0,002	5,818	16,195	1,852
vyhladovělý	4	1,998	10,029	14,029	47,703	0,002	5,815	16,142	1,810
zavýt	4	1,998	9,852	13,852	46,709	0,002	5,797	15,856	1,600
hřivnatý	3	1,732	12,358	15,528	45,682	0,001	5,522	17,132	9,091
nažraný	3	1,731	11,380	14,550	41,475	0,001	5,500	15,776	4,615
apeninský	3	1,731	11,099	14,268	40,280	0,001	5,491	15,386	3,797

Oprócz wymienionych funkcji statystycznych w programach Bonito, NoSketch Engine oraz KonText jest też dostępny rozkład częstości (*Frequency/ Frekvenční distribuce*), który także można ustawić według poszczególnych atrybutów (forma fleksyjna, lemat, tag). Funkcja ta jest niezmiernie przydatna. Pokazuje w tabeli liczbę poświadczeń dla danego typu wyrazu. Sprawdza się zwłaszcza przy komendach pozwalających na wyszukanie różnych wyrazów. Przykładem może być zapytanie .*samot.* Pierwsza tabela przedstawia wyrazy, które wyszukał program, od najczęstszych do najrzadszych, druga tabela natomiast zawiera tę samą treść uporządkowaną według lematów. Przy okazji możemy zweryfikować poprawność lematyzacji. Pod koniec pierwszej tabeli widać kilka słów błędnie zlematyzowanych.

Tab. 3.7. Rozkład częstości formy .*samot.*. W tabeli po lewej wyniki są uporządkowane według form fleksyjnych, po prawej – według lematów.

Lemat	**Frekwencja**
samotný	13 665
samota	2 951
samotář	265
samotářský	177
samotka	129
samotářsky	30
samotinký	23
samotářství	17
samotářka	16
polosamota	6
samotíž	5
samotvrdnoucí	4
samotok	4
samotřetí	3
samotařit	3
vsamotném	2
samotěsnicí	1
samotně	1
vlka-samotáře	1
samotažnice	1
samotížky	1
strýček-samotář	1
samotěžbu	1
diváci-samotáři	1
samotbroumovského	1
osamotného	1
samotnú	1
samotřetím	1
vlků-samotářů	1
samotah	1
samotestující	1

Lemat	**Słowo**	**Frekwencja**
samotný		13 665
	samotné	3 198
	samotného	2 352
	samotný	1 920
	samotnou	1 260
	samotná	1 123
	samotném	978
	samotných	841
	samotným	780
	samotnému	557
	samotní	373
	samotnými	198
	samotného	36
	samoten	18
	samotnej	13
	samotnemu	6
	samotni	5
	samotna	4
	samotnej	2
	samotnu	1
samota		2 951
	samotě	1 607
	samoty	540
	samota	301
	samotu	291
	samotou	109
	samot	54
	samotách	34
	samotami	7
	samotám	5
	samoto	3

[ciąg dalszy tabeli ze strony poprzedniej]

Lemat	Frekwencja
samotrávení	1
samotřejmě	1
jachtařem-samotářem	1
sasamotorcycle	1
osamotňovat	1
samotín	1
osamotněl	1
vlk-samotář	1
dvousamotě	1
potápka-samotářka	1
samottéma	1
samotížný	1
samotného/jí	1
samotvorba	1
samotého	1
samotnost	1
samotářův	1
ksamotné	1
samotnouopustit	1
samotrávících	1

Lemat	Słowo	Frekwencja
samotář		265
	samotář	140
	samotáři	48
	samotáře	46
	samotářem	14
	samotářů	12
	samotářům	4
	samotářích	1
samotářský		177
	samotářský	56
	samotářské	30
	samotářských	18
	samotářského	18
	samotářským	17
	samotářská	17
	samotářští	10
	samotářském	3
	samotářskou	3
	samotářskej	2
	samotářskému	2

W menu *Konkordance*[6] – *Statistiky* – *Rozložení* można także sprawdzić równomierność rozkładu danego elementu językowego w tekstach. Wadą tej funkcji jest to, że po najechaniu na graf nie pokazuje typu tekstu. Pozostaje domyślić się, że jeśli poświadczenia są skupione np. na początku grafu, to pochodzą z tekstów literackich, jeśli na końcu, to z publicystycznych. Szczegółowa identyfikacja źródła jest wprawdzie możliwa po kliknięciu czerwonej linii (program odsyła wtedy do konkretnych poświadczeń), nie zastępuje to jednak braku oznaczeń typów tekstu na grafie. W NoSketch Engine oraz w programie KonText opcja wyszukiwania według rozkładu częstości jest bardziej rozbudowana. Oprócz prostego wyszukiwania leksemów według częstości ich występowania można także sprawdzić frekwencję wyrazów występujących na konkretnej pozycji w lewo lub w prawo od słowa kluczowego. Przy pomocy rozkładu częstości można także szukać słowa kluczowego według wybranych metadanych, np. aby poznać typ tekstu, tytuł, autora i język oryginału.

Funkcja *Frequency/Frekvenční distribuce* pozwala na korzystanie z rozbudowanych bardziej niż w Bonito 1 funkcji statystycznych. Opcja *Custom/ Vlastní* umożliwia, tak jak we wcześniej omówionych elementach wyszukiwarki, ustawienie dostępnych funkcji według potrzeb (do wyboru *Multilevel frequency*

[6] Słowo kluczowe może być wyszukiwane w lemacie, tagu lub w konkretnej formie fleksyjnej.

distribution/ Víceúrovňová frekvenční distribuce 'Wielopoziomowy rozkład częstości' (pozwala na znalezienie frekwencji nie tylko konkretnego leksemu, ale także wyrazów go poprzedzających lub po nim następujących), *Text type frequency distribution/ Frekvenční distribuce* 'Rozkład częstości a typ tekstu' (pozwala na znalezienie frekwencji leksemu/formy wyrazowej/taguów w wybranych składowych metainformacji). W części statystycznej są dostępne także jeszcze następujące opcje: *Nodetags/ Značky* (frekwencja całkowita i względna tagów wyszukiwanego leksemu), *Nodeforms/ Slovní tvary* (frekwencja całkowita i względna form wyszukiwanego leksemu), *Doc Ids/ Dokumenty* (frekwencja całkowita oraz liczba wystąpień na 1 milion poświadczeń w konkretnym tekście), *TextTypes/ Typy textu* (frekwencja całkowita oraz liczba wystąpień na 1 milion poświadczeń w konkretnym typie tekstu).

NoSketch Engine

Ze względu na to, że oprogramowanie NoSketch Engine, zdecydowaliśmy się zaprezentować je także całościowo, a nie tylko w oddzielnych częściach, umieszczonych w różnych miejscach niniejszego rozdziału (zob. także *Praktyczny przewodnik po korpusie równoległym InterCorp* autorstwa Elżbiety Kaczmarskiej i Alexandra Rosena).

Jak już zostało nadmienione na początku artykułu, praca z korpusem jest możliwa także w przestrzeni wirtualnej (po zalogowaniu) w NoSketch Engine. Korzystanie z tego programu jest wygodniejsze pod wieloma względami od oprogramowania Bonito 1. Przede wszystkim nie trzeba zgrywać całego oprogramowania na twardy dysk. Oprócz tego opcje są skoncentrowane w mniejszej liczbie zakładek. Nie trzeba np. otwierać oddzielnych okienek do ustawienia kontekstu, liczby poświadczeń, metadanych, atrybutów. Wystarczy otworzyć zakładkę *View/ Zobrazení*. Przestrzeń ta stawia użytkownikom o wiele mniejsze wymagania w zakresie znajomości terminologii korpusowej oraz tworzenia składni zapytań. Przejściową niedogodnością jest to, że nie wszystkie funkcje są na razie dostępne.

Spróbujmy zatem przedstawić możliwości interfejsu NoSketch Engine. Na początku pracy wybieramy korpus (automatycznie jest wybierana wyjściowa zakładka *Concordance/ Konkordance*), z którym chcemy pracować, następnie typ pytania (*query type/ Typ dotazu*). Tutaj mamy do wyboru: *Basic/ Základní* (wyszukuje konkretną formę wyrazową pisaną zarówno małą, jak i wielką literą; jeśli wpisane słowo będzie miało kształt formy kanonicznej leksemu, to program znajdzie poświadczenia we wszystkich formach wyrazowych), *Phrase/ Fráze* (zwraca konkretny ciąg wyrazów we wpisanych formach fleksyjnych), *Lemma* (wyszukuje wszystkie formy fleksyjne konkretnego leksemu. W okienko zapytań należy wpisać lemat, czyli formę kanoniczną leksemu np. okno). *Word form/ Slovní tvar* (szuka konkretnej formy wyrazowej np. oknu, okna),

Charakter/Podřetězec (zwraca ciąg znaków), *CQL* (szuka wyników za pomocą języka zapytań, którego częścią składową są wyrażenia regularne; można między innymi wpisywać pytania o konkretne tagi)[7]. Więcej informacji na temat tworzenia zapytań można znaleźć w części *Tworzenie zapytań*. Następnie w okienku *Query/Dotaz* umieszczamy zapytanie oraz wciskamy przycisk *Search/Hledat*. Po lewej stronie na niebieskim pasku wyświetlone są następujące opcje menu *New query/Nový dotaz* 'Nowe pytanie', *Word list/Seznam slov* 'Lista słów', *Expert options/Pokročilá nastavení* 'Ustawienia zaawansowane', *Context/Kontext* 'Kontekst' *Subcorpus/Subkorpus* 'Subkorpus', *User/Uživatel* 'Użytkownik', *Password change/Změna hesla* 'Zmiana hasła'. Należy pamiętać, że zaznaczamy opcję, która jest dostępna w danym korpusie. Warto wiedzieć, że tworzenie zapytania w opcji CQL jest bardzo przystępne. Przy budowie składni zapytań, która ma za zadanie szukać wyników według określonej sekwencji tagów, możemy skorzystać z podpowiedzi, która pojawia się postaci listy rozwijanej. Jeżeli zatem potrzebujemy znaleźć rzeczowniki rodzaju męskiego, zaznaczamy kolejno rzeczownik, rodzaj, a program sam wprowadza odpowiednie symbole.

Po wyświetleniu konkordancji z lewej strony ukazuje się menu, które pozwala na zaawansowaną pracę z poświadczeniami. Składa się z kilku głównych kategorii: *Concordance/Konkordance* 'Konkordancja', *Word List/Seznam slov* 'Lista słów', *Save/Uložit* 'Zapisać', *View/Zobrazení* 'Widok', *Sort/Třídění* 'Sortowanie', *Frequency/Frekvenční distribuce* 'Rozkład częstości', *Collocations/Kolokace* 'Kolokacje', *Conc.Desc./Popis dotazu*. Funkcja *Word List/Seznam slov* pozwala na stworzenie listy słów z korpusu, który samodzielnie załączymy, lub z określonego podkorpusu. Opcja *Save/Uložit* umożliwia zapisanie dokumentu w formacie tekstowym lub HML. Można też zaznaczyć konkretne poświadczenia i za pomocą kombinacji CTRL + C skopiować je do pliku, w którym się pracuje.

Następny element menu to *View/Zobrazení* 'Widok'. Tutaj znajdują się już omawiane w podrozdziale *Wybór liczby poświadczeń* zakładki *Custom/Vlastní*, *KWIC*, *Sentence/Věta*. Przy wyborze opcji *Custom/Vlastní* na niebieskim pasku menu pojawia się możliwość ustawienia próbki poświadczeń (*Sample/Vzorek*) oraz filtrowania danych (*Filter/Filtr*). Prócz tego dostępne są już omówione opcje *Sort/Třídění* (zob. *Sortowanie poświadczeń*), *Frequency/Frekvenční distribuce* oraz kolokator (zob. *Funkcje statystyczne*).

Ostatni element menu – oznaczony jako *ConcDesc* (ang. *concordance description*) – jest zwięzłym opisem konkordancji zawierającym liczbę poświadczeń oraz typ zapytania.

[7] Szczegóły na temat funkcji „within" są opisane w rozdziale Elżbiety Kaczmarskiej i Alexandra Rosena *Praktyczny przewodnik po korpusie równoległym InterCorp*.

KonText

Oprogramowanie KonText, podobnie do programu NoSketch Engine, jest dostępne po zalogowaniu ze strony internetowej. Menu zawiera następujące pozycje: *Query/Dotaz* (*New query/Nový dotaz*, *Recent queries/Předchozí dotazy*, *Word list/Seznam slov*), *Subcorpora/Subkorpusy* (*Create new subcorpus/Vytvořit nový subkorpus*, *My subcorpora/Mé subkorpusy*), *Save/Uložit* (*CSV*, *XML*, *TXT*, *Custom/Vlastní*), *Concordance/Konkordance* (*Current concordance/Aktuální konkordance*, *Sorting/Třídění*, *Shuffle/Promíchat*, *Sample/Vzorek*, *Query overview/Popis dotazu*, *Undo/Zpět*), *Filter/Filtr* (*Positive/Pozitivní*, *Negative/Negativní*), *Frequency/Frekvenční distribuce* (*Node tags/Značky*, *Node forms/Slovní tvary*, *Doc IDs/Dokumenty*, *Text types/Typy textu*, *Custom/Vlastní*), *Collocations/Kolokace* (*Custom/Vlastní*) *View options/Možnosti zobrazení* (*KWIC*, *Sentence/Věta*, *Attributes, structures and references/Atributy, struktury a metainformace*, *General concordance view options/Obecné volby zobrazení konkordance*), *Help/Nápověda* (*User manual/Uživatelská příručka*, *Linguistic terms/Lingvistické pojmy*, *Available corpora/Dostupné korpusy*). Więcej na temat programu KonText można przeczytać w rozdziale *Praktyczny przewodnik po korpusie równoległym InterCorp* Elżbiety Kaczmarskiej i Alexandra Rosena.

Zastosowania

Czeski Korpus Narodowy, ze względu na wielość swoich podkorpusów, jest zasobem, z którego można korzystać do badań z użyciem znanych metod lingwistyki korpusowej, zarówno corpus-based, jak i corpus-driven. Nie wszystkich metod można użyć, stosując którykolwiek z korpusów, np. badanie profili kolokacyjnych często nie udaje się na mniejszych korpusach.

W Czechach wydano wiele publikacji językowych, które powstały przy wykorzystaniu korpusów językowych. Nie sposób tu wszystkich wymienić. Warto jednak wspomnieć o serii *Korpusová lingvistika* 'Lingwistyka korpusowa', wydawanej przez Instytut Czeskiego Korpusu Narodowego oraz wydawnictwo *Lidové noviny*. Składa się na nią wiele publikacji różnego typu: monografii, tomów zbiorowych, tomów pokonferencyjnych. Oprócz tego poza serią wydawane są książki i słowniki. Wśród nich warto wymienić kontrowersyjną pierwszą korpusową gramatykę języka czeskiego (*Mluvnice současné češtiny*, red. V. Cvrček), słowniki języka Karla Čapka, Bohumila Hrabala, słowniki komunizmu, słownik realiów komunistycznych, słowniki frekwencyjne. Lista publikacji Instytutu jest dostępna na stronie: http://ucnk.ff.cuni.cz/publikace.php.

Drugim ośrodkiem, gdzie są wydawane publikacje korpusowe, jest Instytut Języka Czeskiego Czeskiej Akademii Nauk. Instytut ma na swoim koncie przed wszystkim tomy zbiorowe oraz księgi pokonferencyjne obejmujące zagadnienia

lingwistyki korpusowej, szczególnie z zakresu czeskiej gramatyki. Warto tu wymienić liczącą kilkaset stron książkę dotyczącą wybranych zagadnień gramatyki czeskiej (*Kapitoly z české gramatiky*, pod red. F. Štíchy). W Instytucie Języka Czeskiego trwają także prace nad akademicką gramatyką języka czeskiego (obecnie nad tomem poświęconym słowotwórstwu). Studia językowe tworzone na podstawie pracy z korpusem powstają także, choć w mniejszym zakresie, na Uniwersytecie Masaryka w Brnie oraz w Ołomuńcu.

Bibliografia

Cvrček V. a kol. (2010). *Mluvnice současné češtiny*. Praha.

Čermák F., ed. (2007). *Frekvenční slovník mluvené češtiny*. Praha.

Čermák F., ed. (2007a). *Slovník Karla Čapka*. Praha.

Čermák F., Blatná, R., eds. (2005). *Jak využívat Český národní korpus*. Praha.

Čermák F., Blatná, R., eds. (2006). *Korpusová lingvistika: Stav a modelové přístupy*. Praha.

Goláňová H., *Sociolingvistické značky a charakteristiky v korpusu SCHOLA2010*. http://www.korpus.cz/schola-popis.php (Dostęp: III 2013).

Hajič J., *Popis morfologických značek – poziční systém*. http://www.korpus.cz/bonito/znacky.php (Dostęp: III 2013).

Hebal-Jezierska M. (2008). *Wariantywność końcówek fleksyjnych rzeczowników męskich żywotnych w języku czeskim*. Warszawa.

Hebal-Jezierska M. (2013). *Podstawowe zasady korzystania z korpusów przy badaniu języka*. W: Chlebda W. (red.) *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*, Opole, s. 17-30.

Kocek J., Kopřivová M., Kučera K., eds. (2000). *Český národní korpus – úvod a příručka uživatele*. Praha.

Kopřivová M. *Struktura korpusu ORAL2006*. https://www.korpus.cz/struktura_oral06.php (Dostęp: III 2013).

Kopřivová M., Waclawičová M. (2006). *Representativeness of Spoken Corpora on the Example of the New Spoken Corpora of the Czech Language*. W: *Proceedings of the international conference "Corpus linguistics – 2006"*. St. Petersburg, s. 174--181.

Křen M., *Dotazovací jazyk korpusového manažeru Bonito*. http://www.korpus.cz/diakorp.php (Dostęp: III 2013).

Kučera K. (1998). *Diachronní složka Českého národního korpusu: obecné zásady, kontext a současný stav. [I]*. „Listy filologické", Vol. 121., No. 34., s. 303-314.

Kučera K., Stluka M., *Uvedení do diachronní složky ČNK*. https://www.korpus.cz/diakorp.php (Dostęp: III 2013)

Statistics used in the Sketch Engine Lexical Computing Ltd., https://trac.sketchengine.co.uk/wiki/SkE/Help/PageSpecificHelp/collocconc

Štícha F., ed. (2011). *Kapitoly z české gramatiky*. Praha.

Šulc M. (1999). *Korpusová lingvistika: první vstup*. Praha.

4. Praktyczny przewodnik po korpusie języka słowackiego

Milena Hebal-Jezierska
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Informacje ogólne

Początki Słowackiego Korpusu Narodowego (SNK) sięgają roku 2002, w którym z inicjatywy ministra szkolnictwa Republiki Słowacji rząd uchwalił ustawę powołującą jednostkę naukową Słowackiego Korpusu Narodowego. Dokument zakładał m.in. zapewnienie przestrzeni oraz zaplecza technologicznego do kontynuowania prac nad korpusem języka słowackiego. Takim miejscem stał się zakład Słowackiego Korpusu Narodowego z siedzibą w Instytucie Języka Słowackiego Słowackiej Akademii Nauk. Jednostka ta prowadzi stronę internetową dostępną pod adresem www.korpus.sk. Nie należy zrażać się ograniczoną funkcjonalnością tej witryny (brak szczegółowych opisów poszczególnych zasobów oraz ich użycia); sam korpus jest bez wątpienia wart polecenia. Można także liczyć na pomoc pracowników Instytutu, którzy chętnie udzielają wszelkich informacji. Mieli oni także swój udział w opracowaniu niniejszej publikacji, przekazując informacje na temat zastosowań SNK oraz uzupełniając brakujące dane na temat struktur poszczególnych korpusów. W tym miejscu dziękuję za pomoc p. R. Garabíkowi oraz dyrektor SNK p. M. Šimkovej.

Do uzyskania ograniczonego dostępu do korpusu nie jest konieczna rejestracja, należy natomiast wyrazić zgodę na przestrzeganie zasad korzystania, zamieszczonych pod adresem http://www.korpus.sk/webinterface.html. Po ich zaakceptowaniu użytkownik zostaje przekierowany na odpowiednią stronę. Pełny dostęp do zasobów korpusowych uzyskujemy dopiero po rejestracji, której należy dokonać na stronie http://www.korpus.sk/registration.html. Warunkiem otrzymania loginu i hasła jest wysłanie pocztą podpisanego dokumentu, zamieszczonego na stronie, oraz e-maila. Dopiero po wykonaniu tych czynności otrzymamy instrukcję zawierającą nie tylko informacje na temat logowania, lecz także szczegóły techniczne dotyczące instalacji programu Bonito 1, obsługującego Słowacki Korpus Narodowy. Można także skorzystać z programu

NoSketch Engine, również dostępnego po rejestracji z witryny https://bonito.korpus.sk/index.html.

Charakterystyka korpusu

Słowacki Korpus Narodowy jest projektem integrującym wiele różnych regularnie aktualizowanych korpusów języka pisanego i mówionego. Nowe opracowanie danego korpusu zastępuje jego poprzednią wersję, która przestaje być dostępna dla użytkowników. W razie potrzeby skorzystania ze starszego zasobu trzeba wystosować odpowiednią prośbę do pracowników SNK. W poniższej tabeli zaprezentowane są korpusy języka słowackiego składające się na Słowacki Korpus Narodowy.

Tab. 4.1. Korpusy języka pisanego (tabela na podstawie informacji zamieszczonych na stronie www.korpus.sk).

Nazwa	**Rok aktualizacji**	**Struktura**	**Liczba tokenów / liczba słów**[1]
prim-6.1-public--all	2013	68,8% publicystyka, 13,9% teksty literackie, 15,3% teksty specjalistyczne, 2% inne	829 771 945 / 655 572 511
prim-6.1-public--sane		bez tekstów z nieprawidłową diakrytyką, sprzed 1955 r., z terenów spoza Słowacji i z czasopism lingwistycznych	773 493 137 / 610 493 493
prim-6.1-public--vyv		podkorpus zrównoważony (33,3% publicystyka, 33,3% teksty literackie, 33,3% teksty specjalistyczne)	317 496 718 / 251 519 537
prim-6.1-public--inf		podkorpus tekstów publicystycznych	540 812 859 / 425 325 094
prim-6.1-public--prf		podkorpus tekstów naukowych, specjalistycznych i popularnonaukowych	105 886 349 / 83 885 837
prim-6.1-public--img		podkorpus tekstów literackich	113 820 575 / 90 714 140
prim-6.1-public--sk		podkorpus oryginalnych słowackich tekstów	558 261 948 / 440 708 351
prim-6.1-public--img-sk		podkorpus oryginalnych słowackich tekstów literackich	35 283 156 / 28 260 019
r55az89-3.0		korpus tekstów z lat 1955-1989 (11,9% publicystyka, 55,5% teksty literackie, 24,1% teksty specjalistyczne, 8,5% pozostałe)	62 885 729 tokenów
r-mak-3.0		korpus ręcznie anotowany morfologicznie (44,3% teksty literackie, 36,7% publicystyczne, 19,0% teksty specjalistyczne)	1 207 813 tokenów

[1] W korpusie języka słowackiego podstawową jednostką jest token.

[ciąg dalszy tabeli ze strony poprzedniej]

Nazwa	**Rok aktualizacji**	**Struktura**	**Liczba tokenów / liczba słów**
prim-6.0	2013 (dostępny na żądanie)	77,8% publicystyka, 9,8% teksty literackie, 11% specjalistyczne, 1,4% pozostałe	1155 mln tokenów
prim-5.0	2011 (dostępny na żądanie)	73% publicystyka, 14% teksty literackie, 12% specjalistyczne, 1% pozostałe	719 mln tokenów
prim-4.0	2009 (dostępny na żądanie)	65% publicystyka, 17% teksty literackie, 16% specjalistyczne, 2% pozostałe	526 mln tokenów
prim-3.0	2007 (dostępny na żądanie)	57% publicystyka, 21,5% teksty literackie, 18,5% specjalistyczne, 3% pozostałe	350 mln tokenów
prim-2.1	2006 (dostępny na żądanie)	63% publicystyka, 20% teksty literackie, 12% teksty specjalistyczne, 5% pozostałe	300 mln tokenów
prim-2.0[2]	2005 (dostępny na żądanie)	71,90% publicystyka, 13,21% teksty literackie, 4,45% teksty specjalistyczne	250 mln tokenów
prim-1	2004 (dostępny na żądanie)	93,39% publicystyka, 3,69% teksty literackie, 1,8% teksty specjalistyczne	182 mln tokenów

Tab. 4.2. Korpusy języka mówionego (tabela na podstawie informacji zamieszczonych na stronie www.korpus.sk).

Nazwa	**Rok aktualizacji**	**Uwagi**
s-hovor-1.0	2008	
s-hovor-2.0	2010	
s-hovor-3.0	2011	
s-hovor-4.0	2012	2611 tys. (353 godziny nagrań)
s-hovor-4.0-upn	2012	subkorpus s-hovor-4.0 tylko przepisane wypowiedzi świadków z projektu *Oral History* Instytutu Pamięci Narodowej Republiki Słowackiej
s-hovor-4.0-sane	2012	subkorpus s-hovor-4.0

[2] Dziękuję panu dr. Radovanowi Garabíkowi z Instytutu Słowackiego Korpusu Narodowego za przekazanie informacji na temat struktury korpusów prim-2.0 oraz prim1.

Tab. 4.3. Korpusy profilowane (tabela na podstawie informacji zamieszczonych na stronie www.korpus.sk).

Nazwa	**Rok aktualizacji**	**Struktura**	**Liczba tokenów**
web-2.0	2012	korpus tekstów internetowych	1 045 558 148
web-1.0	2011	korpus tekstów internetowych	952 095 260
legal-1.0	2011	przepisy prawne	146 899 704
		korpus języka krymskotatarskiego	423 472

Tab. 4.4. Korpus historyczny (tabela na podstawie informacji zamieszczonych na stronie www.korpus.sk)

Nazwa	**Rok aktualizacji**	**Struktura**	**Liczba tokenów**
hks-1.0	2012	korpus historyczny języka słowackiego	370 758

Tab. 4.5. Korpusy równoległe (tabela na podstawie informacji zamieszczonych na stronie www.korpus.sk)[3].

Nazwa	**Rok aktualizacji**
Korpus równoległy słowacko-francuski	2012
Korpus równoległy słowacko-rosyjski	2014
Korpus równoległy słowacko-czeski	2014
Korpus równoległy słowacko-łaciński	Nie podano
Korpus równoległy słowacko-angielski	Nie podano
Korpus równoległy słowacko-bułgarski	Niedostępny
Korpus równoległy terminów komputerowych	Niedostępny

Warto pamiętać, że teksty zawarte w słowackich korpusach publicystycznych często dotyczą sytuacji na terenie Czech, nie zaś Słowacji, co odzwierciedla specyfikę słowackiej prasy i publicystyki. Pokazały to np. badania niektórych stereotypów językowych dotyczących nazw narodowości (np. obrazu Wietnamczyka) z wykorzystaniem słowackich korpusów publicystycznych. Szczególnie analiza profili kolokacyjnych wymagała dużej ostrożności i dokładnego przejrzenia poświadczeń, z których pochodziły znalezione kolokaty, znaczna część poświadczeń dotyczyła bowiem sytuacji czeskiej, a nie słowackiej.

[3] Tabela nie obejmuje udostępnionego przed oddaniem książki do druku korpusu równoległego słowacko-węgierskiego.

Anotacja zewnętrzna

We wszystkich korpusach typu *prim* oraz w zasobie *r55az89-3. 0* metadane są bardzo szczegółowe. Ich wadą jest nieczytelność niektórych skrótów dla zwykłego użytkownika. Przedstawiamy zatem wszystkie skróty w kolejności wyświetlania[4]: doc.bogo (kombinacja pierwszych liter imienia i nazwiska autora oraz liczby porządkowej), doc.cong (imię nazwisko autora oraz tytuł tekstu), doc.date (data wydania), doc.orgd (data pierwszego wydania), doc.name (tytuł tekstu), doc.orgn (oryginalny tytuł tekstu), doc.auth (imię i nazwisko autora), doc.orga (imię i nazwisko autora w pisowni oryginalnej), doc.lang (język tekstu), doc.auts (płeć autora), orgl (język oryginału), trnn (czy tekst był przetłumaczony), trnr (nazwisko tłumacza), trns (płeć tłumacza), isbn, isnn, rhym (czy tekst jest rymowany), type (typ tekstu), subt (podtyp tekstu), genr (gatunek tekstu), subg (gatunek tekstu), domn (dziedzina specjalistyczna), subd (bliższe określenie dziedziny specjalistycznej), medi (nośnik tekstu, np. internet, radio), vari (wariant języka słowackiego: literacki, nieliteracki), para (czy tekst jest podzielony na akapity), emph (czy tekst zawiera informację o oryginalnych wyróżnieniach w tekście), dcrt (prawidłowość diakrytyki użytej w tekście), comn (uwagi), corr (czy tekst był poddany korekcie), bibl (dane bibliograficzne), id (dane identyfikacyjne). Więcej na ten temat można znaleźć na stronie http://www.korpus.sk/bibstyle.html.

Bardzo szczegółowe są także metadane dla korpusów mówionych. Dotyczą one zarówno samej wypowiedzi, jak i relacji między uczestnikami rozmowy, a także nadawcy wypowiedzi. Do opisu tekstu można tu wykorzystać niemal 40 parametrów. Ich zaletą jest czytelność zastosowanych symboli, natomiast wadą jest to, że rzadko są wypełnione konkretnymi wartościami. Parametry dotyczące wypowiedzi to: numer identyfikacyjny wypowiedzi (id), temat wypowiedzi (doc.topic), spontaniczność (doc.spontaneous) oraz formalność (doc.formal). Uczestnicy rozmowy są charakteryzowani poprzez określenie stopnia znajomości (doc.familiar) oraz wzajemnej hierachii (doc.equal). Oprócz tego anotacja obejmuje uwagi anotatorów (doc.comment), datę (doc.date) i miejsce (doc.place) nagrywania rozmowy, liczbę jej uczestników (doc.speakers). Metadane o respondencie zawierają informacje na temat płci (spk.sex m – mężczyzna, a – kobieta), wieku (spk.age), zawodu, wykształcenia (podstawowe, zawodowe, średnie, wyższe, brak), miejsca urodzenia (spk.birthplace), miejsca najdłuższego pobytu (spk.liveplace), miejsca obecnego pobytu (spk.place), języka ojczystego (spk.L1), języków, którymi respondent się posługuje (spk.languages), dialektu używanego przez ankietowanego (spk.dialect). Wśród parametrów znajduje się też informacja, czy respondent został wcześniej powiadomiony o nagraniu (spk.informed). Na uwagę zasługuje szerokie spektrum parametrów dotyczących wieku i wykształcenia ankietowanych.

[4] Informacje na ten temat pochodzą ze strony http://www.korpus.sk/bibstyle.html.

O wiele mniej szczegółowo opracowane są metadane w korpusach: *r_ mak3.0*, *x-legal-1.0*, *web-2.0*. W pierwszym z wymienionych zasobów informacje zawężone są danych bibliograficznych. W korpusie *x-legal-1.0.* znajdziemy natomiast szczegóły dotyczące nazwy dokumentu oraz daty wydania. W zasobie *web-2.0* dodatkowo prezentowane są dane na temat strony internetowej oraz typu tekstu. W korpusie historycznym *hist-1.0* anotacja zewnętrzna obejmuje tytuł, datę, miejsce, komentarze, dane bibliograficzne.

Dla wszystkich korpusów informacje na temat tekstu nie wyświetlają się od razu, jeśli używamy wyszukiwarki Bonito 1. Można je włączyć, wybierając odpowiednią zakładkę z menu albo klikając poświadczenie prawym przyciskiem myszy. W pierwszym przypadku mamy możliwość zaznaczenia, których danych potrzebujemy, w drugim przypadku automatycznie pokażą się wszystkie dostępne informacje. W NoSketch Engine tzw. doc.bogo jest wyświetlane przy każdym poświadczeniu. Aby otrzymać pozostałe metadane, należy skorzystać z zakładki *View options*.

Anotacja wewnętrzna

Wszystkie synchroniczne korpusy języka słowackiego są lematyzowane i znakowane morfologiczne, natomiast brakuje w nich anotacji syntaktycznej oraz semantycznej. Obie anotacje są dostępne także w korpusach równoległych, a dotyczy to wszystkich wersji językowych. Natomiast *historyczny korpus języka słowackiego* jest pozbawiony zarówno lematyzacji, jak i anotacji morfologicznej, semantycznej oraz składniowej.

Tokenizacja i lematyzacja

Token w SNK nie zawsze odpowiada wyrazowi ortograficznemu. Dotyczy to np. wyrazów zapisywanych z łącznikiem, których każda część (także łącznik) jest osobnym tokenem. Co ważne, jako samodzielne tokeny są oznaczane także znaki interpunkcyjne. Z tym podziałem związana jest lematyzacja: znaki interpunkcyjne, nawet te umieszczone w środku wyrazu, są hasłowane oddzielnie. Przykłady zamieszczamy w tabeli poniżej.

Tab. 4.6. Postać lematu wyrazów i liczb zawierających łączniki, apostrofy, ukośniki itp.

Wyraz/liczba	**Lematy**
historicko-jazykovo-kultúrne	historicko - jazykovo - kultúrne
-ár	- ár
Nazwiska podwójne Turečková-Čambalová	turečková - čambalová
Czasownik z partykułą -li bude-li	byť - li
Poprzyimkowe formy nieakcentowane doň	do_ono, do_on
Formy z użyciem apostrofu sister's	sister ' s
8/5	8 / 5

Warto zauważyć, że lematy pisane są małą literą. Dotyczy to także nazw własnych, np. *Tatry* mają formę kanoniczną określoną jako *tatra*, nazwisko *Turečková* jest zapisane w lemacie jako *turečková*.

Jak w większości korpusów, tak i w Słowackim Korpusie Narodowym zdarzają się błędy w zakresie hasłowania. Dobrze jest więc zawsze sprawdzić, czy wyniki otrzymane automatycznie są wygenerowane na podstawie poprawnej lematyzacji. Największą wadą hasłowania jest przypisywanie różnych form kanonicznych tym samym leksemom. Przykładem mogą być nazwiska żeńskie, które raz są hasłowane w formie rodzaju męskiego, innym razem – żeńskiego, np.:

Premiéra sa vydarila, hoci <Turečková/turečková -/- Čambalová/čambalová> neuvádza presný dátum.
<Turečková/turečkový -/- Čambalová/čambalový> však s pohnutím spomína, že hneď [...]

Sposób tokenizacji i lematyzacji jest niezwykle ważny dla badacza przeprowadzającego badania językowe na SNK, a to ze względu na to, że przekłada się na sposób wpisywania komend do okienka wyszukującego materiał. Jeśli np. wpiszemy wyraz z łącznikiem zgodnie z obowiązującą ortografią (np. *historicko-jazykovo-kultúrne*), to nie otrzymamy stosownych poświadczeń. Tego typu leksemy należy wpisywać, tak jakby każda ich część włącznie z dywizem była osobnym wyrazem, czyli *historicko - jazykovo - kultúrne*.

Tagset korpusu języka słowackiego jest dokładnie opisany na stronie http://www.korpus.sk/morpho.html. Za podstawę posłużył tradycyjny podział na części mowy, który został poddany pewnym modyfikacjom. Oznacza to, że z jednej strony wyróżnia się podstawowe części mowy, tj.: rzeczownik, przymiotnik, liczebnik (przy czym liczebniki odmieniające się przymiotnikowo są

oznaczane jako liczebniki o paradygmacie przymiotnikowym), zaimek, przysłówek, wykrzyknik, partykuła, spójnik, z drugiej strony dodatkowo oddzielnie oznacza się np. liczebnik wyrażony liczbowo, który jest określony jako liczba. Z innych jednostek indeksowanych oddzielnymi kodami należy wymienić: partykuły refleksywne *sa* i *si*, imiesłowy, partykułę warunkową *by*, skróty, nazwę własną, znaki interpunkcyjne, znak nierozpoznawalny przez program, zapis błędny, inne elementy niebędące słowami. Każda z wymienionych kategorii jest dookreślana poprzez charakterystyczne dla niej cechy, np. typ paradygmatu, do którego należy, rodzaj, liczbę, stronę itp.

Przy pracy z wykorzystaniem znakowania warto pamiętać o sprawdzaniu jego poprawności, szczególnie w miejscach, gdzie mamy do czynienia z synkretyzmem lub homonimią. Oto jedno z błędnie otagowanych poświadczeń:

[...] *potom ďalšia a ďalšia, všetky plné <chlapov/chlap/SSmp4*[5] *>*
s kalašnikovmi na krku.

Podstawowe zapytania

Zapytania można tworzyć na kilka sposobów. Pierwszy to wpisanie odpowiedniej sekwencji wyrażeń regularnych oraz symboli umownych do okienka zapytań, np. [lemma="matka"]. Szczegółowy spis używanych symboli można znaleźć na stronie http://korpus.juls.savba.sk/usage.html.

Przykładowe zapytania prezentuję w poniższej tabeli oddzielnie dla oprogramowania Bonito i NoSketch Engine.

Tab. 4.7. Wpisywanie i tworzenie zapytań w programie Bonito oraz NoSketch Engine.

Zapytanie o	**Bonito 1** (Zapytania wpisujemy do okienka z napisem *Nový dotaz*)	**NoSketch Engine**
formę wyrazową	Matky	Matky Aby otrzymać poświadczenia o określonej postaci ortograficznej (w tym przypadku pisane wielką literą), należy zaznaczyć opcję *Phrase*.
formę wyrazową z wariantami ortograficznymi	[Mm]atky Otrzymamy poświadczenia pisane zarówno wielką, jak i małą literą.	Jeśli uprzednio zaznaczymy opcję *Simple* lub *Word form*, to możemy wpisać wyraz *Matky* lub *matky*. Program w obu wypadkach znajdzie formy pisane wielką i małą literą.
Lemat	[lemma="matka"]	Najpierw zaznaczamy opcję *Lemma*, a następnie wpisujemy leksem, np. *matka*.

[5] Rzeczownik występuje tutaj w dopełniaczu liczby mnogiej, a nie w bierniku w liczbie pojedynczej, jak informuje o tym tag.

[ciąg dalszy tabeli ze strony poprzedniej]

Zapytanie o	**Bonito 1** (Zapytania wpisujemy do okienka z napisem *Nový dotaz*)	**NoSketch Engine**
Lemat wyrazów pisanych wielką literą	[lemma="bratislava"] Nazwy własne wpisujemy przy wyszukiwaniu ich lematu małą literą!	Najpierw zaznaczamy opcję *Lemma*, a następnie wpisujemy leksem, np. bratislava Nazwy własne wpisujemy przy wyszukiwaniu ich lematu małą literą!
Lemat wyrazów zawierających łącznik, apostrof, ukośnik itp.	[lemma="historicko"][lemma="-"][lemma="jazykovo"] Każda z części wyrazu jest oddzielnym lematem!	Należy wybrać opcję *CQL* i wpisać [lemma="historicko"][lemma="-"][lemma="jazykovo"] Każda z części wyrazu jest oddzielnym lematem!
Część wyrazu	.*tka matk.* .*tk.*	.*tka matk.* .*tk.*
grupę wyrazów	dobré matky	dobré matky Najpierw należy zaznaczyć opcję *Phrase*
grupę lematów	[lemma="dobrý" [lemma="matka"]	Należy wybrać opcję *CQL* i wpisać lemma="dobrý"] [lemma="matka"]
kategorię gramatyczną	[tag="SSfp1"]	Należy wybrać opcję *CQL* i wpisać [tag="SSfp1"]
kategorię gramatyczną	[(lemma=".*j")&(tag="SSfp1")]	Należy wybrać opcję *CQL* i wpisać [(lemma=".*j")&(tag="SSfp1")]

W interfejsie Bonito 1 odpowiednie zapytanie można sformułować także za pomocą opcji zwanej *Implicitný atribút* (znajduje się on w menu w zakładce *Korpus*). Wówczas nie musimy formułować skomplikowanego zapytania. Dla zapytania o lemat lub tag zaznaczamy stosowną opcję w zakładce *Implicitný atribút*, a następnie wpisujemy dany wyraz bądź tag do okienka zapytań, np. matka lub SSfp1. Przy wyszukiwaniu lematu należy pamiętać, żeby wyraz wpisywać małą literą, natomiast przy leksemach zawierających łączniki, apostrofy itp. – żeby wpisywać każdą jego część oddzielnie. Warto pamiętać o tym, że raz zaznaczony atrybut pozostaje niezmienny aż do następnej modyfikacji dokonanej przez użytkownika. Z jednej strony może to przysparzać problemów w wypadku wyszukiwania według różnych atrybutów, z drugiej strony przy kwerendach polegających na pracy z jednym atrybutem pozwala na zaoszczędzenie czasu.

Funkcje dostępne w programach obsługujących SKN – Bonito oraz NoSketch Engine

W związku z tym, że NoSketch Engine jest programem obsługującym zarówno Słowacki Korpus Narodowy, jak i Czeski Korpus Narodowy, odsyłamy Czytelnika do zapoznania się z odpowiednim fragmentem tekstu w rozdziale traktującym o Czeskim Korpusie Narodowym. W tym miejscu warto zwrócić uwagę na drobne różnice w funkcjonowaniu oprogramowania zaaplikowanego dla języka słowackiego. W *NoSketch Engine* dla SNK brakuje opcji pozwalającej na wpisanie tagu przy pomocy rozwijanej listy oraz nie ma możliwości tworzenia korpusów i podkorpusów przy pomocy warunku within (zob. *Praktyczny przewodnik po korpusie języka czeskiego*). Brakuje także osobnej podkategorii typu *Custom*, umożliwiającej ustawienie dostępnych funkcji zgodnie z potrzebami użytkownika. W opcji prezentującej częstość wystąpień w konkretnym tekście (Doc Ids) dostępna jest na razie tylko frekwencja całkowita, a nie jak w przypadku korpusu języka czeskiego częstość występowania danej formy na 1 milion tokenów. Nie ma także możliwości automatycznego sprawdzenia frekwencji szukanego słowa w różnych typach tekstu (brak opcji *Text Type*). W kolokatorze brakuje modułu obliczającego frekwencję względną.

Zastosowania[6]

Różnorodność korpusów zintegrowanych w SNK pozwala na przeprowadzenie badań językowych z użyciem znanych metod korpusowych corpus-driven, corpus-based oraz corpus-assisted.

Dane z SNK zostały wykorzystane w różnym stopniu zarówno do publikacji leksykograficznych (jako pomoc w publikacjach: Balážová 2005, Považaj, Kačala, Pisárčiková 2003, jako materiał główny w pracach: Buzássyová, Jarošová 2006, Buzássyová, Jarošová 2011, Nižníková 2006), jak i pozycji z zakresu gramatyki, szczególnie z morfologii i składni (walencji) (Sokolová, Ivanová 2006, Sokolová, Ološtiak, Ivanová 2007, Sokolová 2007). Oddzielne artykuły poświęcone są wielu dziedzinom języka, co pokazuje, jak różnorodne są możliwości wykorzystania korpusu w celach badawczych. Znajdziemy wśród nich analizy słowotwórcze (Šimková 2010), fonologiczne (Karčová 2011), a także badania z zakresu językowego obrazu świata (Šimková 2011) oraz dyskursu (Karčová 2010).

Na koniec warto wspomnieć o projekcie wspierającym i rozwijającym nowoczesne technologie META-NET http://www.meta-net.eu/whitepapers/vol-

[6] Podrozdział na temat zastosowań SNK w poszczególnych projektach i publikacjach powstał dzięki informacjom przekazanym przez panią dr M. Šimkovą, dyrektor Słowackiego Korpusu Narodowego.

umes/slovak, którego jednym z produktów jest tzw. „biała księga Słowacji" *The Slovak Language in the Digital Age*.

Bibliografia

Balážová Ľ. et al. (2005). *Slovník cudzích slov – akademický*. Bratislava.

Buzássyová K., Jarošová A. (2006). *Slovník súčasného slovenského jazyka A–G*. Bratislava.

Buzássyová K., Jarošová A. (2011). *Slovník súčasného slovenského jazyka H–L*. Bratislava.

Garabík R., Gianitsová L., Horák A., Šimková M. (2004). *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu*, http://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf (Dostęp: III 2013).

Garabík R., Karčová A., Šimková M., Gajdošová K. (2008). *Hovorený korpus slovenčiny*. W: M. Kopřivová, M. Waclawičová (ed.) *Čeština v mluveném korpusu*. Praha, s. 227-233.

Karčová A. (2010). *Analýza jazykového prejavu v televíznej relácii (na báze Slovenského hovoreného korpusu*. W: *Varia 19. Zborník plných príspevkov z XIX. kolokvia mladých jazykovedcov (Trnava – Modra – Harmónia 18.-20. 11. 2009)*. Zost. J. Hladký – Ľ. Rendár. Trnava, s. 139-146.

Karčová A. (2011). *Výslovnostné realizácie dvojhlásky ô v súčasných hovorených prejavoch na báze Slovenského hovoreného korpusu*. W: *Korpusová lingvistika Praha 2011. 2. Výzkum a výstavba korpusů*. Ed. F. Čermák – K. Kučera – V. Petkevič. Praha, s. 240-251.

Nižníková J. (2006). *Valenčný slovník slovenských slovies (Na korpusovom základe)*. Vol. 2. Prešov.

Pisárčiková M., Považaj M. (2004). *Synonymický slovník slovenčiny*. Bratislava (korpus wykorzystany przy ostatnim wydaniu).

Považaj M., Kačala J., Pisárčiková M. (2003). *Krátky slovník slovenského jazyka*. Bratislava (SNK wykorzystany przy dwóch ostatnich wydaniach).

Slovenský národný korpus http://korpus.sk

Sokolová M. (2007). *Nový deklinačný systém slovenčiny*. Prešov.

Sokolová M., Ivanová M. eds. (2006). *Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli*. Prešov.

Sokolová M., Ološtiak M., Ivanová M., et al. (2007). *Slovník koreňových morfém slovenčiny*. Prešov.

Šimková M. (2006). *Slovak National Corpus – History and Current Situation*. W: Šimková M. (red.), *Insight into Slovak and Czech Corpus Linguistics*, Bratislava, s. 151-159, http://korpus.juls.savba.sk/files/zbornik/insight.pdf

Šimková M. (2008). *Korpusová lingvistika na Slovensku.* „Jazykovedný časopis" 59. 1-2, s. 11-24.

Šimková M. (2010). *Produktívne a neproduktívne formanty pri tvorení názvov vlastností v slovenčine a češtine (korpusovolingvistická analýza).* W: *Slovo – Tvorba – Dynamickosť. Na počesť Kláry Buzássyovej.* Ed. M. Šimková. Bratislava, s. 492-504.

Šimková M. (2011). *Vianoce a čo s nimi súvisí v našom jazykovom obraze sveta.* W: „Romboid", roč. 46, č. 9-10, s. 130-132, Bratislava.

Šimková M. (red.) (2006a). *Insight into Slovak and Czech Corpus Linguistics.* Bratislava.

http://korpus.sk/Metad%28c3a1%29ta%2820%29o%2820%29respondentovi.html (Dostęp: 28 IV 2013)

http://korpus.sk/Metad%28c3a1%29ta%2820%29o%2820%29nahr%28c3a1%29vke.html

http://korpus.sk/Znacky_pouzivane_v_SHK.html

http://www.korpus.sk/bibstyle.html

5. Praktyczny przewodnik po korpusie języka dolnołużyckiego

Fabian Kaulfürst
Chóśebuske wótźělenje Serbskego instituta z.t.
– Chociebuski Oddział Instytutu Łużyckiego

Tłum. Marcin Szczepański
Chóśebuske wótźělenje Serbskego instituta z.t.
– Chociebuski Oddział Instytutu Łużyckiego

Informacje ogólne

Opisywany korpus języka dolnołużyckiego to *Dolnoserbski tekstowy korpus*, (w skrócie: Dotko), reprezentujący szeroko rozumianą literaturę dolnołużycką[1]. Powstał i jest rozwijany[2] w Chociebuskim Oddziale Instytutu Łużyckiego (Serbski institut)[3].

Prace nad korpusem rozpoczęto wkrótce po założeniu Oddziału w 1992 r. pod kierunkiem ówczesnego pracownika Instytutu dr. Hana Steenwijka, który z pomocą studentów przystąpił do digitalizacji tekstów dolnołużyckich. Od niego opiekę nad rozbudową korpusu przejął dr Hauke Bartels, który pozyskując środki finansowe ze źródeł zewnętrznych, zorganizował digitalizację starszych tekstów poza Instytutem. Umożliwiło to względnie szybkie wprowadzenie do korpusu m.in. kompletu roczników czasopisma *Bramborski (Serbski) Casnik* (lata 1848-1933; w tym okresie był to praktycznie jedyny dolnołużycki periodyk) oraz różnych wydań Biblii. Obecnym koordynatorem prac związanych z digitalizacją tekstów i rozwijaniem korpusu jest dr Fabian Kaulfürst.

[1] W chociebuskim oddziale Instytutu powstaje równocześnie korpus języka mówionego dialektów dolnołużyckich. Projekt ten jest finansowany ze źródeł zewnętrznych przez fundację Volkswagen Stiftung w ramach programu wspierającego dokumentację języków zagrożonych (DoBeS – Dokumentation Bedrohter Sprachen).

[2] Należy podkreślić, że obecnie korpus znajduje się w fazie rozwoju. Sytuacja personalna Oddziału nie pozwala na realizację w krótkim czasie wielu zamierzonych etapów. Dalsza wieloaspektowa rozbudowa korpusu jest planowana, zależy jednak od faktycznych możliwości w przyszłości.

[3] Informacje o Instytucie Łużyckim i jego chociebuskim oddziale znaleźć można w Internecie pod adresem http://www.serbski-institut.de.

Do roku 2009 Dotko mógł być używany tylko wewnętrznie. W kwietniu tego roku nawiązano kontakt z pracownikami[4] Instytutu Czeskiego Korpusu Narodowego (Ústav Českého národního korpusu, ÚČNK)[5]. W kolejnych miesiącach Marcin Szczepański i Fabian Kaulfürst ujednolicili kodowanie (unikod), strukturę (XML) oraz metainformacje zawarte w tekstach pochodzących z różnych okresów digitalizacji. Tak przygotowany korpus dolnołużycki[6] opublikowano w internecie[7] w grudniu 2010 r. Współpraca Oddziału z ÚČNK daje użytkownikom korpusu Dotko możliwość korzystania z niektórych funkcji praskich aplikacji korpusowych. Jednocześnie korpus jest dostępny ze strony WWW Oddziału[8] przez aplikację przygotowaną przez M. Szczepańskiego, która za pośrednictwem uproszczonego interfejsu graficznego pozwala kierować zapytania do korpusu na ÚČNK. Na stronie Oddziału znajdują się także informacje o korpusie, przewodnik po wyrażeniach regularnych oraz opis ortograficznego zróżnicowania opublikowanych tekstów (http://www.dolnoserbski.de/korpus/).

Korpus jako całość zawiera praktycznie wszystkie teksty dolnołużyckie powstałe od znanych początków piśmiennictwa do roku 1848[9], większość tekstów z okresu 1848-1945[10] oraz znaczną część dzieł nowszych, m.in. niektóre roczniki czasopisma *Nowy Casnik*. W związku ze względnie ograniczoną liczbą zasobów piśmiennictwa dolnołużyckiego w porównaniu z „większymi" językami, można się spodziewać, że wkrótce Dotko stanie się pełnym korpusem dolnołużyckiego języka pisanego. Obecnie zawiera około 27 mln tokenów, z czego ok. 12 mln jest dostępnych przez internet.

Charakterystyka korpusu

Korpus języka dolnołużyckiego to korpus ogólny, diachroniczny, obejmujący całość piśmiennictwa. Zawiera teksty od 1. poł. XVI w. do czasów współczesnych. Z uwagi na specyficzną sytuację piśmiennictwa dolnołużyckiego Dotko nie jest i nie może być korpusem zrównoważonym (por. niżej). Pracownicy Oddziału

[4] Dziękujemy w tym miejscu dr. Michalowi Křenowi z ÚČNK za pomoc, doradztwo i koordynację współpracy z czeskiej strony.

[5] Pomysłodawcą współpracy i autorem internetowej wersji poglądowej korpusu był ówczesny pracownik Oddziału, dr Ruprecht v. Waldenfels.

[6] Z przyczyn prawnych udostępniono publicznie jedynie teksty wydane do roku 1937.

[7] http://www.korpus.cz/dotko.php

[8] http://dolnoserbski.de/korpus

[9] Rok 1848 – początek regularnego wydawania pierwszego dolnołużyckiego czasopisma „Bramborski Casnik" – wprowadza do literatury dolnołużyckiej publicystykę i tym samym otwiera nowy, ważny okres w łużyckim piśmiennictwie (por. Bartels 2010: 9 n.) Oceniając kompletność dokumentacji piśmiennictwa do tej daty, można stwierdzić, że brakuje jedynie rękopisów, których edycje drukiem nie są dostępne.

[10] Ten okres to dla języka dolnołużyckiego czas najintensywniejszego rozwoju literatury i najpełniejszej konsolidacji języka pisanego.

na bieżąco uzupełniają korpus zarówno tekstami starszymi[11], jak i nowymi wydawnictwami oraz aktualnymi rocznikami czasopism. Dotko nie jest zatem korpusem referencyjnym, ale w dużym stopniu korpusem monitorującym.

Za pewne rozwiązanie powyższych ograniczeń można uznać istniejące zróżnicowane formy dostępu do materiału zawartego w korpusie. Chodzi m.in. o możliwość definiowania bardziej lub mniej synchronicznych podkorpusów obejmujących materiał z danego przedziału czasowego, np. wybrane roczniki czasopisma *Bramborski (Serbski) Casnik*. Do badań nad rozwojem dolnołużyckiego języka literackiego można zestawić jednojęzyczny korpus równoległy obejmujący różne wydania Starego i Nowego Testamentu, z których większość jest nowymi redakcjami wydań poprzednich[12].

Biorąc pod uwagę gatunki tekstów zawarte w korpusie oraz jego cel, jakim jest objęcie całości piśmiennictwa dolnołużyckiego, można stwierdzić, że korpus Dotko oddaje gatunkowy charakter tego piśmiennictwa. Przeważają teksty prasowe (ok. 71%), znaczny jest udział tekstów biblijnych (ok. 11%), resztę stanowią teksty bardziej zróżnicowane gatunkowo, przy czym także wiele spośród nich ma charakter religijny.

Cechą specyficzną piśmiennictwa dolnołużyckiego jest ograniczona liczba autorów. *Bramborski (Serbski) Casnik* był przez dziesięciolecia niemalże jedynym medium regularnie oddziałującym na język literacki. Dlatego też jego poszczególni redaktorzy mieli wyjątkowo silny osobisty wpływ na rozwój języka dolnołużyckiego. Dotyczy to zwłaszcza Kita Šwjeli, który przez ponad pół wieku (1864-1915) był praktycznie jedynym autorem materiałów prasowych. Ponadto aż do roku 1945 większość tekstów wychodziła spod pióra pastorów, a więc osób mających bliski kontakt ze stylem biblijnym. Okoliczności te zostawiły wyraźny ślad w rozwoju języka dolnołużyckiego do drugiej wojny światowej. Innym istotnym faktem jest to, że znacząca liczba zawartych w korpusie dzieł beletrystycznych to przekłady, głównie z języka niemieckiego, górnołużyckiego, czeskiego i słowackiego. Także część tekstów prasowych to tłumaczenia (często niejawne) artykułów z prasy niemieckiej.

Nie ulega wątpliwości, że niereferencyjność korpusu, dominacja niewielkiej grupy autorów oraz podobne czynniki mają istotny wpływ na badania językowe, na co zwrócił uwagę Bartels (2010). Badany język w pewnych okresach był silnie kształtowany np. przez rodzimy dialekt autora, jego subiektywne wybory czy też stosunek do puryzmu językowego (por. Bartels 2006, 2008a).

[11] Szczególnych problemów przysparza wyszukiwanie i pozyskiwanie tekstów oryginalnych. Książki były wydawane w niewielkich nakładach. Dlatego pewne teksty najprawdopodobniej nie trafią do korpusu, choć są znane z cytatów i innych nawiązań. Cyfryzacja katalogów i zasobów bibliotecznych na świecie pozwala jednak mieć nadzieję, że w przyszłości uda się odnaleźć książki uważane obecnie za utracone (por. Kaulfürst 2010).

[12] Przykład badań na podstawie takiego korpusu to studium Kaulfürst 2012: 73-104.

Anotacja zewnętrzna

Zewnętrzny opis tekstów korpusu obejmuje następujące informacje[13]: 1) niepowtarzalny identyfikator danego tekstu, 2) dane bibliograficzne – autor, redaktor, tłumacz, wydawca, numer wydania, miejsce wydania, rok wydania, różne postaci tytułu (pełny tytuł, część główna tytułu, podtytuł, tytuł skrócony), 3) gatunek tekstu[14], 4) rodzaj pisma użytego w druku – łacinka lub fraktura, 5) uwaga nt. (nie)zachowania oryginalnej ortografii źródła. Ponadto każdy tekst zawiera elementy opisu struktury wewnętrznej: 1) pozycja w tekście źródłowym – numer strony, adres biblijny, w rocznikach czasopism numer wydania itp., 2) identyfikacja elementów formalnej struktury tekstu – nagłówki, stopki, podpisy ilustracji, objaśnienia trudnych słów itp., 3) oznaczenia elementów oryginału, które pominięto podczas digitalizacji – ilustracje, większe fragmenty w języku niemieckim, powtarzające się reklamy, 4) elementy, które mogą nie być brane pod uwagę przy wyszukiwaniu, np. fragmenty obcojęzyczne, niemieckie odpowiedniki przy słowach dolnołużyckich[15], 5) dane służące wewnętrznej dokumentacji opracowania korpusu i digitalizacji tekstów – dostępność oryginału, dostępność, format, źródło i miejsce przechowywania obrazów cyfrowych, stopień integracji tekstu do korpusu od wstępnego odpisu do kompletnego pliku XML, komentarze, liczba tokenów, data wprowadzenia do korpusu[16].

Stopnia poprawności większości danych nie można określić, ale można założyć, że metainformacje dotyczące całości tekstu są w dużej mierze rzetelne. Problem natomiast może stanowić np. numeracja stron, której systematyczna kontrola nie była możliwa; wykonano jednak liczne testy automatyczne, co istotnie ograniczyło liczbę błędów.

Niektóre z wymienionych metainformacji są widoczne dla użytkowników interfejsu ÚČNK. Są to: identyfikator (document.id), pełny tytuł (document.full_title), autor (document.author), miejsce wydania (document.place), rok wy-

[13] Informacje te rozmieszczone są w: a) plikach XML z tekstami, b) zewnętrznej bazie danych powiązanej z odpowiednimi tekstowymi za pomocą identyfikatorów. Jakość adnotacji poszczególnych tekstów jest różna – w niektórych zbiór metainformacji jest kompletny, niektóre zaś zawierają jedynie dane podstawowe. Przyczyną takiego stanu jest m.in. bardzo długi okres tworzenia korpusu, zmieniające się cele i warunki. Z powodu braków personalnych problemy te nie zostały do dziś w pełni rozwiązane. Istotną poprawę jakości korpusu (w różnych aspektach) mogłaby przynieść w przyszłości realizacja projektu historyczno-dokumentującego systemu informatycznego dolnołużyckiej leksyki (Bartels 2013a i 2013b).

[14] Obecnie używane wartości to: „publicystyka", „beletrystyka", „tekst biblijny". Oprócz tego wprowadzono również wartość „antologia", m.in. dla wydawnictw o treści mieszanej.

[15] Jest to często stosowana forma wprowadzania i popularyzacji nowej leksyki w języku mniejszości.

[16] Część danych wspomnianych w pkt. 2, 3, 7-10 w niektórych tekstach nie ma zastosowania lub znanej wartości, np. w tekstach prasowych na ogół brak autorów, tłumaczy czy wydawców.

dania (document.year), rodzaj pisma drukarskiego w oryginale (document.orig_font) oraz numer strony w źródle lub adresacja innego typu (anchor.name)[17].

Użytkownicy interfejsu w serwisie dolnoserbski.de widzą skrót tytułu, będący odnośnikiem do listy dostępnych tekstów, która z kolei zawiera kompletne dane bibliograficzne całości tekstu (ale nie precyzyjną lokalizację danego wystąpienia w konkordancji)[18].

Anotacja wewnętrzna

Anotację wewnętrzną korpus Dotko zawiera obecnie w stopniu minimalnym. Podział na zdania i tokeny wykonano na ÚČNK za pomocą tamtejszych metod i narzędzi. W korpusie brak znakowania części mowy; możliwość jego przeprowadzenia zależy od środków i sytuacji Chociebuskiego Oddziału Instytutu Łużyckiego w najbliższych latach. Powyższe dotyczy także znakowania morfologicznego, morfosyntaktycznego, syntaktycznego i semantycznego, a także lematyzacji, przy których szczególnym utrudnieniem jest zróżnicowanie (orto)graficzne i diachroniczne tekstów.

Podstawowe zapytania

Możliwości wykonywania zapytań zależą od wyboru jednego z trzech scenariuszy. 1) Badanie na całym korpusie, po wprowadzeniu wszystkich tekstów do wybranego programu analizującego. Funkcje i formy działań zależą przy tym od użytego menedżera lub edytora korpusu. Taki model postępowania jest przewidziany tylko wewnątrz Instytutu, ponieważ z przyczyn prawnych nie wszystkie teksty mogą zostać opublikowane. 2) Dostęp przez stronę internetową ÚČNK[19]. Mechanizmy tamtejszych korpusów języka czeskiego można zasadniczo stosować także do korpusu Dotko (zob. *Praktyczny przewodnik po korpusie języka czeskiego* Mileny Hebal-Jezierskiej). Z przyczyn oczywistych niedostępne są funkcje wykorzystujące znaczniki, których w korpusie dolnołużyckim (jeszcze) brak[20]. Obecnie użyteczne są zapytania typu Query Type Basic, Phrase, Word Form, Character, CQL[21], ale nie Query Type Lemma i Tag. 3) Uproszczony dostęp przez stronę internetową Oddziału[22]. Tu głównym ograniczeniem

[17] Jedną z informacji można wybrać jako standardową (najlepszym kandydatem wydaje się identyfikator). Będzie ona widoczna w konkordancji z lewej strony wystąpień, a po kliknięciu na nią otworzy się u dołu okna przegląd pozostałych informacji związanych z danym tekstem.

[18] http://dolnoserbski.de/korpus/zredla

[19] Dostęp tylko przez aplikację WWW. Aplikacja systemowa Bonito 1 nie działa.

[20] Podobne ograniczenie dotyczy obecnie także innych korpusów diachronicznych, np. DIAKORP (zob. „Praktyczny przewodnik po korpusie języka czeskiego").

[21] Funkcjonalności języka CQL nie można jednak w pełni wykorzystać – z powodów opisanych w artykule niemożliwe jest np. dopasowywanie form podstawowych czy znaczników.

[22] http://dolnoserbski.de/korpus

jest brak możliwości wykonywania zapytań wielowyrazowych. Można stosować wyrażenia regularne; na stronie podano objaśnienia i przykłady użycia[23].

Brak lematyzacji czy oznaczeń części mowy oraz różnorodność systemów (orto)graficznych wymagają od użytkownika dobrej znajomości historycznych wariantów pisowni. Pomocą w tym wypadku może być podane na stronie ogólne zestawienie wariantów ortograficznych[24]. Dodatkowym warunkiem skutecznego odpytywania jest znajomość różnych (np. historycznych czy dialektalnych) możliwych form i wariantów morfologicznych i fonetycznych.

Aby wyszukać (prawie) wszystkie wystąpienia abstrakcyjnej formy wyrazowej *serbski* 'serbołużycki', należy skonstruować zapytanie:

[Ssſß]+er+b?[sſß]+k[ijy]+

Dopasuje ono następujące formy tekstowe (tu według częstości wystąpień): *ßerbſki*, *ßerski*, *serbski*, *Sserbſki*, *ßerſki*, *Serbski*, *Serbſki*, *ßerbski*, *Sſerbſki*, *Sserski*, *serski*, *Sſerſki*, *Sserſki*, *Serski*, *Sſerski*, *ſerbski*, *ſerbſki*, *ßerßki*, *Serſki*[25]. Oto konkretny przykład z korpusu (niżej w pisowni współczesnej):

Schkorzy ſaßej na lipach ßejźe, ßerbſki a nimſki ſpiwaſch wéże
[Bramborske Nowiny 1881, 9]
współcz.: Škórce zasej na lipach sejźe, serbski a nimski spiwaś wěże
'Szpaki znów siedzą na lipach, umieją śpiewać po łużycku i niemiecku'

Wyszukiwanie wszystkich form tekstowych formy podstawowej *ruka* 'ręka' wymaga bardziej złożonego zapytania. Wyrażenie:

[Rr]u[ct]?[kcz][aeiou](ch|m[ia]?|wu?)?

zostanie dopasowane do (prawie) wszystkich wystąpień formy podstawowej *ruka* we wszystkich współcześnie regularnych formach fleksyjnych z uwzględnieniem możliwych historycznych wariantów (orto)graficznych[26]. W korpusie można jednak znaleźć również przestarzałą bezkońcówkową formę dopełniacza l.mn. *ruk*, do której powyższe wyrażenie nie zostanie dopasowane. Zapytanie:

[Rr]+uc?k

przynosi, po usunięciu wyników homonimicznych, dwa przykłady:

[23] http://dolnoserbski.de/korpus/regularne_wuraze

[24] http://dolnoserbski.de/korpus/psawopisne_warianty

[25] Analiza wyników pokazuje, że podwójne <rr> nie pojawia się w tym kontekście; brak także pewnych kombinacji dopasowywanych przez podwyrażenia <[sſß]+> i <[ijy]+>. Należy pamiętać, że teoretycznie taka pisownia byłaby możliwa. Interesującym zjawiskiem jest dopasowywanie przez aplikację ÚČNK także form tekstowych *ßerbşki*, *ßerşki*, *Sferşki*, chociaż zapytanie nie zawiera znaku <ş>.

[26] Tak zbudowane wyrażenie dopasowuje, niestety, także nieoczekiwane formy tekstowe, jak *Rutka*, *rukawu*, które przed dalszym opracowaniem trzeba z wyników usunąć.

Tym ſłożejam pak ńej wele drogotkow do ruk padnuło
[Serbski Casnik 1931, 37]
współcz.: Tym złożejam pak njej' wjele drogotkow do ruk padnuło
'Złodziejom jednak nie wpadło w ręce wiele kosztowności'

To wiźecy wugjarnu knecht rukawy, plunu do ruk a da se na togo młożeńca, wuflinknu jomu jadnu za wuśy, aż ten se ned do kopicki kidnu.
[Pratyja za dolnych Serbow na lěto 1932]
'Widząc to, parobek zakasał rękawy, splunął w dłonie i ruszył na młodzieńca, walnął go w ucho, aż się tamten przewrócił'

Wystąpienia formy dopełniacza l.podw. *rukowy* (zamiast *rukowu*) można znaleźć jedynie wtedy, gdy, znając ją z dialektów przejściowych, świadomie się jej pośród wyników oczekuje[27]:

Běda nam, gaby tu móz do rukowy krydnuli!
[Serbski Casnik 1932, 31]
współcz.: Běda nam, gaby tu móc do rukowu krydnuli!
'Biada nam, jeśli taką moc otrzymamy do rąk!'

Inne dostępne funkcje w oprogramowaniu

Oprócz wyżej wymienionych można wskazać jeszcze kilka udogodnień interfejsu dostępnego na stronach Chociebuskiego Oddziału Instytutu Łużyckiego. Jednym z nich jest klawiatura ekranowa pomocna przy wprowadzaniu liter ze znakami diakrytycznymi (np. ǵ, ʃ́, ẃ). W konkordancjach dopasowana forma tekstowa jest wyróżniona i umieszczona w osi listy; po jej kliknięciu ukazuje się szerszy kontekst danego wystąpienia. Konkordancję tworzą wszystkie dopasowania, przy czym jednocześnie na stronie pokazanych jest 50.

Dostęp do korpusu Dotko przez stronę ÚČNK daje wiele dodatkowych możliwości analogicznych do tych oferowanych w korpusach języka czeskiego (zob. rozdział *Praktyczny przewodnik po korpusie języka czeskiego* Mileny Hebal-Jezierskiej). Obejmują one np. określanie kontekstu, szukanie kolokacji, filtrowanie wyników i stosowanie funkcji statystycznych.

Zastosowania

Korpusu języka dolnołużyckiego Dotko można używać na wiele sposobów. Możliwe są zarówno badania typu corpus-based, corpus-driven, jak i ich połączenia.

[27] Wystarczy wtedy zmienić podwyrażenie <wu?> na <w[uyiü]?>.

Czynnikiem utrudniającym prowadzenie badań jest brak znaczników części mowy czy morfologicznych. Skuteczność wyszukiwania i reprezentatywność wyników zależą zatem w dużej mierze od tego, w jaki sposób użytkownik korzysta z korpusu, oraz od jego znajomości pisowni historycznych.

Ponieważ Dotko jest dostępny w internecie dopiero od grudnia 2010 r., a odpowiedzialni zań pracownicy Chociebuskiego Oddziału Instytutu Łużyckiego kierują swoje wysiłki przede wszystkim na rozbudowę korpusu, liczba badań językowych na nim przeprowadzonych jest jeszcze ograniczona. Jest używany – na razie niesystematycznie – głównie jako wsparcie przy projektach Oddziału oraz tamże przygotowywanych posiedzeniach Dolnołużyckiej Komisji Językowej. Najważniejszym projektem jest obecnie budowa aktywnego słownika niemiecko-dolnołużyckiego[28]. Planowany jest historyczno-dokumentujący system informatyczny leksyki dolnołużyckiej (zob. Bartels 2013a, 2013b), w którym korpus odegra dużą rolę[29]. Można oczekiwać, biorąc pod uwagę sytuację językową[30], że w badaniach nad językiem dolnołużyckim znaczenie korpusu w najbliższej przyszłości wzrośnie.

Do opracowań z zakresu leksykologii dolnołużyckiej powstałych z uwzględnieniem danych korpusowych należą m.in. Bartels 2006, 2009a i 2009b. Tematów gramatycznych dotyczą artykuły Bartels 2008a, 2008b i Bartels/Spieß 2012. Z kolei Kaulfürst 2012 bada historyczny rozwój języka na podstawie przykładów o charakterze ortograficznym, leksykalnym, morfologicznym i składniowym wybranych z podkorpusu równoległego obejmującego trzy redakcyjnie zależne przekłady Biblii.

Ważne dla charakterystyki korpusu Dotko jest jego intensywne wykorzystanie w celach niejęzykoznawczych, czyli np. jako źródło (często jedyne istotne) do badań historycznych i etnologicznych albo pomoc przy tworzeniu materiałów prasowych. Wobec tego często bezpośrednio wpływa na współczesną recepcję języka dolnołużyckiego.

Bibliografia

Bartels H. (2006). *Wordowaś: Wó ranych wopytach wutłocowanja pózyconki.* „Lětopis" 53, 2, s. 90-103.

Bartels H. (2008a). *Konkurrierende Passivkonstruktionen in der niedersorbischen Schriftsprache. Ein Beispiel für Sprachwandel durch Purismus.* W: Kempgen S. et al. (red.) *Deutsche Beiträge zum 14. Internationalen Slavistenkongress Ohrid 2008.* Die Welt der Slaven – Sammelbände 32, München, s. 27-38.

[28] http://dolnoserbski.de/dnw

[29] Jeżeli realizacja projektu się rozpocznie, dalsze opracowanie korpusu będzie jednym z głównych warunków jej powodzenia.

[30] Najmłodsi rodzimi użytkownicy języka dolnołużyckiego to z nielicznymi wyjątkami osoby w wieku 70-75 lat, a ostatni rodzimi językoznawcy kończą właśnie swoją aktywność zawodową (por. Bartels 2010: 7).

Bartels H. (2008b). *Sekundäre Prädikation im Niedersorbischen.* W: Schroeder Chr., Hentschel G., Boeder W. (red.) *Secondary predicates in Eastern European languages and beyond.* Studia Slavica Oldenburgensia 16, Oldenburg, s. 19-39.

Bartels H. (2009a). *Loanwords in Lower Sorbian, a Slavic Language of Germany.* W: Haspelmath M., Tadmor U. (red.) *Loanwords in the World's Languages. A Comparative Handbook.* Berlin – New York, s. 304-329.

Bartels H. (2009b). *Lehnwörter im Niedersorbischen. Ergebnisse aus einem internationalen Forschungsprojekt.* „Lětopis" 56, 1, s. 53-80.

Bartels H. (2010). *Das (diachrone) Textkorpus der niedersorbischen Schriftsprache als Grundlage für Sprachdokumentation und Sprachwandelforschung.* W: Hansen B., Grković-Major J. (red.) *Diachronic Slavonic Syntax. Gradual Changes in Focus.* Wiener Slawistischer Almanach. Sonderbände 74, München – Berlin – Wien, s. 7-18.

Bartels H. (2013a). *Zur Konzeption eines historisch-dokumentierenden Wortschatz--Informationssystems des niedersorbischen. Pläne zur Behebung eines drängenden Forschungsdesiderats.* W: Kempgen S., Franz N., Jakiša M., Wingender M. (red.) *Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013.* München – Berlin, s. 37-46.

Bartels H. (2013b). *Ku koncepciji historisko-dokumentěrujucego informaciskego systema dolnoserbskego słowoskłada. Plany k wótpóranju nuznego slěźeńskego deziderata.* „Lětopis" 60, 1, s. 16-26.

Bartels H., Spieß G. (2012). *Restrictive relative clauses in the Sorbian languages.* W: *Sorbian in typological perspective. STUF – Language Typology and Universals*, vol. 65, Issue 3, s. 221-245.

Kaulfürst F. (2010). *Eksemplar lažy w Americe.* „Rozhlad" 61, 11, s. 17-18.

Kaulfürst F. (2012). *Dolnoserbska biblija w tśich redakcijach. Pśinosk k wuwiśeju dolnoserbskeje pisneje rěcy.* W: Motornyj V., Scholze D. (red.) *Pytannja sorabistyky: materialy XIII Mižnarodnogo sorabistyčnogo seminaru 2010.* L'viv, s. 73-104.

6. Praktyczny przewodnik po korpusie języka górnołużyckiego

Fabian Kaulfürst
Chóśebuske wótźělenje Serbskego instituta z.t.
– Chociebuski Oddział Instytutu Łużyckiego

Tłum. Marcin Szczepański
Chóśebuske wótźělenje Serbskego instituta z.t.
– Chociebuski Oddział Instytutu Łużyckiego

Informacje ogólne[1]

Korpus języka górnołużyckiego – Hornjoserbski tekstowy korpus, w skrócie Hotko – reprezentuje piśmiennictwo górnołużyckie od połowy XIX w. do dziś. Za jego rozwój odpowiada Oddział Językoznawczy Instytutu Łużyckiego[2] w Budziszynie (Serbski institut), działający w kooperacji z Centralną Biblioteką Łużycką (Serbska centralna biblioteka)[3].

Prace nad korpusem rozpoczął w 1996 r. ówczesny pracownik Instytutu, prof. dr Edward Wornar, a pięć lat później w internecie opublikowano pierwszą wersję systemu. Po objęciu przez Wornara katedry w Instytucie Sorabistyki Uniwersytetu w Lipsku odpowiedzialność za korpus przejęła kierownik Oddziału Językoznawczego, dr hab. Sonja Wölkowa. W ciągu kilku lat z pomocą personelu naukowo-technicznego Instytutu wykonano digitalizację wielu tekstów; obecnie Hotko zawiera ok. 36 mln tokenów (wraz ze znakami interpunkcyjnymi 44 mln).

W lecie 2010 r., wzorem Oddziału Chociebuskiego, postanowiono nawiązać współpracę z Wydziałem Czeskiego Korpusu Narodowego (Ústav Českého národního korpusu, ÚČNK)[4]. Konieczne ujednolicenie struktury tekstów i metainformacji przeprowadzono z pomocą współpracowników z Dolnych Łużyc.

[1] Za przygotowanie informacji o historii i aktualnym stanie korpusu dziękuję dr hab. Sonji Wölkowej.

[2] Szczegółowe informacje o Instytucie można znaleźć w Internecie pod adresem http://www.serbski-institut.de.

[3] http://www.serbski-institut.de/cms/os/24/biblioteka

[4] W tym miejscu należą się podziękowania dla dr. Michala Křena z ÚČNK za pomoc, doradztwo i koordynację współpracy z czeskiej strony.

Od r. 2013 Hotko jest dostępny w sieci WWW pod adresem http://www.korpus.cz/hotko.php. W przyszłości planowany jest także dostęp przez własny interfejs graficzny ze strony http://www.serbski-institut.de. Obecnie można tam znaleźć informacje o korpusie, instrukcje wykorzystania wyrażeń regularnych, zestawienie wariantów pisowni i inne pomocne materiały[5].

Charakterystyka korpusu

Korpus górnołużycki ma charakter zdecydowanie diachroniczny i obejmuje źródła z okresu dwóch stuleci, przy czym ponad połowa (54%) tekstów pochodzi z czasów najnowszych (po zjednoczeniu Niemiec w l. 1989/1990). Z przyczyn praktycznych główny nacisk położono na wprowadzenie ważniejszych dzieł literatury górnołużyckiej. Równoważenie względem gatunków i innych kryteriów odgrywa rolę drugorzędną.

Hotko nie jest korpusem referencyjnym, zatem jest nieustannie uzupełniany. Z jednej strony digitalizuje się starsze dzieła, z drugiej – we współpracy z wydawnictwem *Domowina* (Ludowe nakładnistwo Domowina)[6] i Centrum Językowym *Witaj* (Rěčny centrum WITAJ)[7] – wprowadza się na bieżąco najnowsze teksty górnołużyckie. Wynika stąd, że Hotko jest korpusem monitorującym.

Zakres czasowy źródeł można określić przez budowę podkorpusów synchronicznych dla wybranych okresów.

Dominującymi typami tekstów są publicystyka (57%) i beletrystyka (23%, z dużym udziałem poezji). Na pozostałe źródła składają się różne gatunki. Zamierzony brak zrównoważenia gatunkowego korpusu ma oczywisty wpływ na badania językowe. Np. do badań składni należałoby ze źródeł wykluczyć słowniki i zestawienia terminów, poza tym trzeba by zwrócić szczególną uwagę na większe prawdopodobieństwo nacechowania szyku wyrazów w utworach poetyckich.

Anotacja zewnętrzna

Użytkownik interfejsu ÚČNK ma dostęp do następujących metainformacji zapisanych w anotacjach zewnętrznych: jednoznaczny identyfikator (document.id), autor (document.author), tłumacz (document.translator), wydawca lub wydawnictwo (document.publisher), miejsce (document.place) i rok wydania (document.year), rok powstania dzieła (document.year_orig)[8], pełny tytuł (docu-

[5] Dokładny adres to http://www.serbski-institut.de/cms/os/48/hornjoserbski.

[6] http://www.domowina-verlag.de

[7] http://www.recny-centrum-witaj.de

[8] Rozróżnienie między rokiem wydania a rokiem powstania ma zastosowanie w przypadku znacznej odległości czasowej między datami i dotyczy np. późniejszych edycji czy

ment.full_title), tytuł uproszczony (document.title), podtytuł (document.subtitle), tytuł skrócony (document.short_title)[9], rodzaj tekstu (document.type)[10], informacja o typie pisma drukarskiego (fraktura lub łacinka) użytego w oryginale (document.orig_font)[11].

Fragmenty w językach innych niż górnołużycki nie zostały oznaczone jako obcojęzyczne. Należy na to zwrócić uwagę zwłaszcza w przypadku tekstów lub ich fragmentów w blisko spokrewnionych językach słowiańskich, np. dolnołużyckim[12]. Warunkiem poprawności wyników jest wtedy kontrola i samodzielne oczyszczenie uzyskanych konkordancji.

Anotacja wewnętrzna

Anotację wewnętrzną wykonano w stopniu minimalnym. Podział na zdania i tokeny przeprowadzono na ÚČNK za pomocą tamtejszych metod i narzędzi. Brak natomiast znaczników części mowy, elementów morfologicznych, syntaktycznych czy morfosyntaktycznych. Poza tym korpus nie jest lematyzowany, należy zatem wyszukiwać poszczególne możliwe formy tekstowe badanego leksemu bądź morfemu.

Podstawowe zapytania

Jeśli chodzi o odpytywanie, Hotko znajduje się w podobnej sytuacji jak korpus dolnołużycki. Zasadniczo istnieją te same możliwości co w przypadku pozostałych korpusów udostępnianych przez ÚČNK[13]. Ograniczenia wynikają z braku znaczników i lematyzacji[14]. Dodatkowym utrudnieniem są występujące w korpusie górnołużyckim różne tradycje ortograficzne[15]. Problem ten jest

niezmienionych wznowień. W badaniach diachronicznych należy zatem uwzględnić rok powstania. Niestety zdarzają się sytuacje, gdy podany rok prowadzi do fałszywych wyników – np. w r. 1995 w serii „Lětopis" (nr 42/2) wydano opracowanie XVII-wiecznego rękopisu, sam artykuł powstał również w tym roku, w związku z czym cytowane w nim dzieło oryginalne będzie przez korpus datowane na rok 1995.

[9] Umożliwia to przyporządkowanie poszczególnych wyników do odpowiednich tekstów. Dokładniejsze określenie lokalizacji (np. numeru strony) nie jest jeszcze możliwe.

[10] Obecnie używane wartości to: „beletrystyka", „literatura religijna", „słowniki i terminologie", „literatura naukowa" oraz „publicystyka". W ramach publicystyki określono dodatkowo poszczególne periodyki (document.pub_subtype).

[11] Dane nie zawsze są kompletne lub nie występują we wszystkich dokumentach; np. artykuły prasowe nie mają określonego autora.

[12] Język dolnołużycki pojawia się często w artykułach czasopisma „Rozhlad".

[13] W przeciwieństwie do korpusów czeskich dostęp do korpusów łużyckich ograniczony jest do aplikacji WWW. Instalowany lokalnie menedżer Bonito 1 nie działa.

[14] Podobne ograniczenie dotyczy obecnie także innych korpusów diachronicznych, np. DIAKORP. Obecnie Hotko obsługuje następujące klasy zapytań: Query Type Basic, Phrase, Word Form, Character, CQL (ale nie Query Type Lemma oraz Tag).

[15] Objaśnienie wariantów pisowni znajduje się pod adresem http://www.serbski-institut.de/cms/os/479/prawopisne-warianty.

jednak, w porównaniu z korpusem dolnołużyckim, mniej uciążliwy, a to dzięki wcześniejszej i konsekwentniejszej normalizacji języka górnołużyckiego, chociaż przez długi czas w jednoczesnym użyciu pozostawały dwa lub trzy warianty.

W celu pełnej ekscerpcji zawartych w korpusie informacji należy dobrze znać zbiór możliwych form badanego słowa oraz koleje rozwoju języka literackiego. Jako przykład można podać wyszukiwanie wszystkich form dopełniacza i celownika zaimka *wón* 'on'. Trzeba m.in. zwrócić uwagę na równolegle obecne w piśmiennictwie tradycje (Faska 1998: 167-177) oparte na różnych dialektach języka – w tzw. ewangelickim wariancie języka literackiego występują formy *jeho*, *jemu*, natomiast w wariancie katolickim odpowiadają im formy *joho*, *jomu*[16]. Takie różnice są wyraźnie potwierdzane przez wyniki przeszukiwania korpusu; zapytanie:

[jJ][eo]ho

wskazuje na 86 178 wystąpień formy *jeho*/*Jeho* (85%) i 15 033 formy *joho*/*Joho* (15%), a zapytanie:

[jJ][eo]mu

odpowiednio 36 157 wystąpień *jemu*/*Jemu* (84%) i 6963 *jomu*/*Jomu* (16%)[17].

Innym przykładem jest niekonsekwentne przeprowadzenie w wygłosie wyrazu typowej górnołużyckiej innowacji fonetycznej, polegającej na przejściu samogłoski *a* do *e* między miękkimi (por. Kaulfürst 2012: 111-113). To zjawisko widoczne jest także w języku literackim i materiale korpusowym. Zapytanie:

[Jj]and[źž]el[ea]j

dopasowuje formę mianownika liczby podwójnej leksemu *jandźel* 'anioł' i daje (po usunięciu form homonimicznych) 19 wystąpień formy zakończonej na *-ej* (70%) i 8 formy na *-aj* (30%). Wystąpienia na *-aj* pochodzą sprzed 1915 r. i pojawiają się głównie wśród autorów z kręgu katolickiej tradycji literackiej.

Dostępne funkcje w oprogramowaniu oprócz wyżej wymienionych

Dostęp do Hotko przez stronę internetową ÚČNK pozwala korzystać ze wszystkich dodatkowych narzędzi korpusów czeskich, takich jak określanie kontekstu,

[16] Przedstawiony tu obraz jest uproszczony. Zależnie od czasu lub autora, formy typu *joho* mogą pojawić się także w tekstach należących bardziej do ewangelickiego kręgu językowego i odwrotnie.

[17] Dodatkowym urozmaiceniem są cytowania starszych zapisów w publikacjach naukowych. W korpusie można znaleźć formy: *ieho* (24×), *ioho* (2×), *Ioho* (1×), *yeho* (4×), *yoho* (11×), *yóho* (5×), *jého* (2×) oraz *iemu* (4×), *yemu* (1×), *yomu* (1×), *jému* (1×). Ponadto dwa wystąpienia formy *iomu* pochodzą z fragmentu w języku dolnołużyckim.

szukanie kolokacji, filtrowanie wyników czy stosowanie funkcji statystycznych (zob. *Praktyczny przewodnik po korpusie języka czeskiego* Mileny Hebal-Jezierskiej).

Zastosowania

W ogólnym ujęciu górnołużycki korpus tekstów nadaje się zarówno do badań typu corpus-based, jak i typu corpus-driven. W obu przypadkach utrudnieniem jest brak oznaczeń części mowy i anotacji morfologicznych. Sposób wykorzystania korpusu zależy od celu badań i inwencji badającego. Wiedza o historycznych odmianach pisowni i zróżnicowaniu dialektalnym odgrywa tu rolę mniejszą niż w przypadku korpusu dolnołużyckiego. Mimo to dokładna znajomość rozwoju różnych aspektów językowych piśmiennictwa górnołużyckiego pozostaje warunkiem koniecznym uzyskania reprezentatywnych i kompletnych wyników.

Badań opartych na górnołużyckim korpusie tekstów przeprowadzono dotychczas niewiele. Generalnie należy stwierdzić, że wykorzystywany jest głównie jako środek pomocniczy w projektach językoznawczych Instytutu oraz jako źródło informacji przy przygotowaniu materiałów dla Górnołużyckiej Komisji Językowej.

Wśród publikacji powstałych w oparciu m.in. o korpus Hotko dominuje tematyka leksykograficzna. Należą do nich słowniki i opracowania takie jak: Bura 2003, Ivčenko, Wölke 2004 i 2009, Jentsch, Pohontsch, Schulz 2006, Wölkowa 2011, Wornar, Strauch 2007, oraz poradniki, np. Pohončowa, Šołćina, Wölkowa 2009.

Korpus wykorzystano w badaniach z dziedziny słowotwórstwa i morfologii: Pohončowa 2011, Werner 2003, Wölke 2011.

Z korpusu Hotko czerpali również autorzy dwukrotnie wznawianego podręcznika do samodzielnej nauki języka górnołużyckiego (Schulz, Werner 2000, 2002[2], 2012[3]).

Bibliografia

Bura R. (2003). *Polska, czeska i górnołużycka frazeologia pochodzenia biblijnego a Nowy Testament Jakuba Wujka, Biblia kralicka oraz Nowy Testament Michała Frencla*. Kraków.

Faska H. (1998). *Wuwiće a změny hornjo- a delnjoserbskeje spisowneje rěče a jeju normow*. W: Faska H., *Serbšćina. Najnowsze dzieje języków słowiańskich*. Opole, s. 167-177.

Ivčenko A., Wölke S. (2004). *Hornjoserbski frazeologiski słownik / Obersorbisches phraseologisches Wörterbuch / Верхнелужицкий фразеологический словарь*. Budyšin.

Ivčenko A., Wölke S. (2009). *Wo wćipnych Jěwach a nócnych hawronach. Rěčne wobroty za dźěći a starših.* Budyšin.

Jentsch H., Pohontsch A., Schulz J. (2006). *Deutsch-obersorbisches Wörterbuch neuer Lexik.* Budyšin.

Kaulfürst F. (2012). *Studije k rěči Michała Frencla.* Spisy Serbskeho instituta 55. Budyšin.

Pohončowa A. (2011). *Wosebitosće při wužiwanju słowotwórbnych srědkow w publicistice Jakuba Barta-Ćišinskeho.* W: Scholze D., Schön F. (red.), *Jakub Bart-Ćišinski (1856-1909). Erneuerer der sorbischen Literatur / Wobnowjer serbskeje literatury.* Spisy Serbskeho instituta 54, Budyšin, s. 224-232.

Pohončowa A., Šołćina J., Wölkowa S. (2009). *Z labyrinta serbšćiny. Bjesady wo rěči.* Budyšin.

Schulz J., Werner E. (2000). *Obersorbisch im Selbststudium / Hornjoserbšćina za samostudij.* Budyšin. (2. wyd. 2002, 3. wyd. 2012).

Werner E. (2003). *Die Verbalaffigierung im Obersorbischen.* Spisy Serbskeho instituta 34, Budyšin.

Wornar E., Strauch M. (2007). *Jendźelsko-hornjoserbski šulski słownik. English Upper Sorbian Learner's Dictionary.* Budyšin.

Wölke S. (2011). *Perifrastiske paradigmy z 'esse' a 'habere' w hornjoserbšćinje.* W: Тополињска З. (red.), *Перифрастични конструкции со 'esse' и 'habere' во словенските и во балканските јазици.* Скопје, s. 93-107.

Wölkowa S. (2011). *Frazeologizmy w rěči Jakuba Barta-Ćišinskeho.* W: Scholze D., Schön F. (red.), *Jakub Bart-Ćišinski 1856-1909. Erneuerer der sorbischen Literatur / Wobnowjer serbskeje literatury.* Spisy Serbskeho instituta 54, Budyšin, s. 233-243.

7. Praktyczny przewodnik po korpusie języka chorwackiego

Jerzy Molas
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Informacje ogólne

Obecnie istnieją dwa korpusy języka chorwackiego. Pierwszy, *Hrvatski nacionalni korpus* (dalej: HNK; http://www.hnk.ffzg.hr), tworzony jest od 1998 r. przez zespół Marka Tadicia związanego z Instytutem Lingwistyki Wydziału Filozoficznego Uniwersytetu w Zagrzebiu (Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu). Natomiast drugi, *Hrvatski jezični korpus* (dalej: HJK; http://riznica.ihjj.hr/index.hr.html) stworzony został pod kierownictwem Dunji Brozović-Rončević z Instytutu Języka Chorwackiego i Językoznawstwa w Zagrzebiu (Institut za hrvatski jezik i jezikoslovlje). Opracowanie tego korpusu rozpoczęło się w 2006 r. w ramach projektu chorwackiego Ministerstwa Nauki i Sportu, jednak jego realizacja nie wyszła poza etap przygotowania interfejsu wyszukiwawczego oraz załadowania tekstów z sieci lub udostępnionych przez włączone do projektu wydawnictwa. Z tego względu przedmiotem szczegółowego opisu w przewodniku jest pierwszy z korpusów, drugi omówiony został skrótowo na końcu tekstu.

Twórca HNK, Marko Tadić, od początku kariery naukowej zajmuje się problemami lingwistyki komputerowej i związany jest z Wydziałem Filozoficznym Uniwersytetu w Zagrzebiu. W 1994 r. uzyskał na tej uczelni doktorat na podstawie dysertacji *Računalna obradba morfologije hrvatskoga književnoga jezika* (Komputerowe opracowanie morfologii chorwackiego języka literackiego), od 2001 r. stoi na czele Katedry Lingwistyki Algebraicznej i Informatycznej (Katedra za algebarsku i informatičku lingvistiku), a w 2004 r. uzyskał stanowisko profesora nadzwyczajnego Uniwersytetu w Zagrzebiu. Marko Tadić jest autorem i współautorem książek: *Hrvatski jezik u digitalnom dobu* (Język chorwacki w epoce cyfryzacji; Heidelberg 2012; wraz z D. Brozović-Rončević i D. Kapetanoviciem), *Research Infrastructures in the Digital Humanities* (Strasbourg 2011; wraz C. Moulini i in.), *Jezične tehnologije i hrvatski jezik* (Technologie

językowe i język chorwacki; Zagreb 2003) oraz *Hrvatski čestotni rječnik* (Chorwacki słownik frekwencyjny; Zagreb 1999; wraz z M. Mogušem i M. Bratanić), jak również kilkudziesięciu artykułów i innych publikacji naukowych poświęconych problemom komputerowego opisu i analizy języków naturalnych, ze szczególnym uwzględnieniem chorwackiego.

Początki chorwackich prac, które łączyć można z językoznawstwem korpusowym, sięgają jeszcze przełomu lat sześćdziesiątych i siedemdziesiątych XX wieku, a związane są z tym samym Instytutem Lingwistyki, w którym powstał później HNK. Prekursorami tych działań byli Željko Bujas, który w 1975 r. opublikował, utworzoną w 1967 r., konkordancję komputerową leksyki poematu *Osman* Ivana Gundulicia (Ž. Bujas, *Ivan Gundulić „Osman". Kompjutorska konkordancija*, Zagreb, Sveučilišna naklada Liber 1975, t. 1-2), oraz Rudolf Filipović. Pod jego kierownictwem prowadzone były intensywne prace nad wykorzystaniem korpusów równoległych do kontrastywnej analizy języków serbsko-chorwackiego i angielskiego (R. Filipović, *Jugoslavenski projekt za kontrastivnu analizu srpskohrvatskog i engleskog jezika. Organizacija i zadaci projekta*, Zagreb, Institut za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, 1968).

W latach siedemdziesiątych i osiemdziesiątych XX w. ukazały się zestawiane komputerowo konkordancje języka dawnych dzieł literatury chorwackiej (por. M. Moguš, Ž. Bujas, *Kompjutorska konkordancija Razvoda istarskoga*, Zagreb, Institut za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, 1976; tychże, *Kompjutorska konkordancija Marulićevih djela*, Zagreb, Zavod za lingvistiku, 1980; tychże, *Kompjutorska konkordancija „Balada Petrice Kerempuha" Miroslava Krleže*, Zagreb, Zavod za lingvistiku Filozofskog fakulteta u Zagrebu, 1982). Na podstawie tych opracowań powstał w 1996 r. *Jednomilijunski korpus hrvatskoga jezika* opracowany pod kierownictwem Milana Moguša (tzw. *Mogušev korpus*; niedostępny w sieci), który z kolei stanowił punkt wyjścia dla pierwszego chorwackiego słownika frekwencyjnego (*Hrvatski čestotni rječnik*, red. M. Moguš, M. Bratanić, M. Tadić, Zagreb, Zavod za lingvistiku Filozofskog fakulteta, Školska knjiga, 1999).

W tym samym czasie i w tym samym środowisku dojrzewała równocześnie koncepcja stworzenia wielomilionowego reprezentatywnego korpusu języka współczesnego. HNK stanowi więc kontynuację wcześniej prowadzonych prac oraz korzysta z doświadczeń ich autorów. Dowodzi tego fakt, że w pierwszej wersji projektu, która ujrzała światło dzienne w grudniu 1998 r., obok docelowo trzydziestomilionowego reprezentatywnego korpusu tekstów powstałych po 1990 r., a więc po rozpadzie wspólnoty jugosłowiańskiej (także językowej) i uzyskaniu przez Chorwację niepodległości, udostępniono również użytkownikom archiwum HETA (Hrvatski elektronički tekstni arhiv). Obejmowało ono teksty starsze, sprzed 1990 r., oraz te, które mogły naruszać reprezentatywność głównego korpusu. Archiwum HETA miało stać się swego rodzaju praktycznie

nieograniczonym repozytorium oraz korpusem monitorującym języka chorwackiego. Pierwsza wersja korpusu funkcjonowała w sieci do końca 2004 r.

Od stycznia 2005 r. udostępniona została wersja 2.0 HNK. Różniła się ona znacznie od swej poprzedniczki. Przede wszystkim zmianie uległ interfejs wyszukiwawczy. Autorzy nawiązali współpracę z czeskimi twórcami oprogramowania korpusowego i do obsługi korpusu zastosowali program Bonito. Od wiosny 2008 r. dostępna jest wersja 2.5 (beta) korpusu. Obejmować ma ona, według deklaracji autorów, 101,3 mln lematyzowanych i anotowanych jednostek oraz powinna umożliwiać przeszukiwanie zasobów według kategorii i form gramatycznych. Problem polega na tym, że zgodnie z informacjami na stronie HNK wersja ta od 2008 r. jest dopiero zapowiadana, a jej możliwości można sprawdzić jedynie, posługując się niewielkim podkorpusem cw2000. W rzeczywistości rozbieżności między deklaracjami autorów a realnymi możliwościami korpusu są większe i to one będą przede wszystkim przedmiotem opisu w przewodniku.

Charakterystyka struktury HNK

HNK 2.0. obejmował następujące podkorpusy zintegrowane:

Tab. 7.1.

Nazwa podkorpusu	**Źródło materiału**	**Lata powstania tekstów**	**Numery**	**Liczba jednostek korpusowych**
cw1998-2000	Croatia Weekly	1998-2000	5-118	1 600 000
cw2000	Croatia Weekly	2000	112--118	118 000
dv2005	Dubrovački vjesnik	2005		880 000
fo2003-2005	Fokus	2003-2005		2 700 000
gs2002-2005	Glas Slavonije	2002-2005		17 000 000
hr51-55	Hrvatska revija	2001-2005		1 400 000
Klasici	*Klasici hrvatske književnosti I. – epika, romani, novele*, CD 1, Bulaja naklada 2000.	1556-1943		3 600 000
Marulic	Dzieła Marka Marulicia, prawdopodobnie również z wydawnictwa *Klasici hrvatske književnosti* (*Judita; Suzana; Vartal; Dijaloški tekstovi; Poslanice Marka Marulića Katarini Obirtić*)	II poł. XV w. – I poł. XVI w.		81 000
Nacional108--265	Nacional	1998-2000	108--265	6 900 000
nn1990-2005	Narodne novine	1990-2005		18 000 000

[ciąg dalszy tabeli ze strony poprzedniej]

Nazwa podkorpusu	**Źródło materiału**	**Lata powstania tekstów**	**Numery**	**Liczba jednostek korpusowych**
vi210-278	Vijenac	2002-2004	210--278	15 000 000
vj2000-2003	Vjesnik	2000-2003		46 600 000
vl1999_03--04	Večernji list	III-IV 1999		2 200 000
				116 079 000

Dostępna obecnie wersja HNK 2.5. beta zawiera prawdopodobnie ten sam zestaw, chociaż w menu brak podkorpusów fo2003-2005 oraz vi210-278. Materiały w nich zawarte wyświetlane są jednak podczas wyszukiwania. Według informacji podawanych przez Bonito, korpus liczy obecnie 101 884 103 jednostek. Mylące dla użytkownika jest dołączenie do listy wyświetlanych podkorpusów angielskojęzycznych zbiorów o nazwie HZN (Hrvatski zavod za norme), liczącego ponad 30 000 jednostek, oraz podkorpusu susanne (więcej informacji: http://www.grsampson.net/SueDoc.html), liczącego 150 000 jednostek. Podkorpusy te nie są przeszukiwane podczas pracy z HNK.

Docelowo HNK osiągnąć ma następującą strukturę tekstów:

Tab. 7.2.

Rodzaj tekstu	**%**
Teksty informacyjne, w tym:	74
Gazety	37
dzienniki	22
tygodniki	9
dwutygodniki	6
Czasopisma, prasa kolorowa	*16*
tygodniki	9
miesięczniki	4
dwumiesięczniki, kwartalniki	3
Książki, broszury, korespondencja...	*21*
publicystyka	4
teksty popularne	3,5
korespondencja, druki efemeryczne	0,5
nauka i wiedza	13
Beletrystyka, w tym:	23
powieści	*13*
opowiadania, nowele, szkice	*5*
eseistyka	*4*
dzienniki, (auto)biografie...	*1*
Teksty heterogeniczne	3

Inspiracją jest prawdopodobnie budowa ČNK, jednak autorzy nie podają zasad, jakimi się kierowali przy ustalaniu takich właśnie proporcji. W praktyce nie ma to większego znaczenia, gdyż HNK w obecnej postaci daleki jest od zaplanowanej struktury. Porównując dane liczbowe z pierwszej tabeli, łatwo wyliczyć, że teksty prasowe stanowiły w wersji 2.0 97% materiału. Reszta to beletrystyka o bardzo zróżnicowanym językowo i chronologicznie charakterze. HNK uznać więc można za korpus tekstów prasowych z przełomu XX i XXI wieku, „zanieczyszczony" utworami literatury pięknej powstałymi na przestrzeni ostatnich pięciuset lat. Autorzy tłumaczą taką sytuację chęcią jak najszybszego udostępnienia całości materiału użytkownikom, którzy do chwili osiągnięcia planowanej struktury zadowolić się muszą zbiorem niereprezentatywnym i niezrównoważonym, co jednak umożliwia już w chwili obecnej rozpoczęcie badań nad językiem chorwackim z wykorzystaniem narzędzi korpusowych. Niepokojący jest fakt, iż sytuacja nie uległa zmianie od roku 2008.

Instalacja i podstawowe funkcje wyszukiwawcze

Do przeglądania zasobów HNK służy program Bonito[1], wykorzystywany przez Czeski Korpus Narodowy. Program instalacyjny odpowiedni do posiadanego systemu operacyjnego znaleźć można pod adresem: http://www.textforge.cz/download. Po uruchomieniu zainstalowanego programu pamiętać należy o konieczności zmiany adresu serwera (*Server address*) na *hnk2.ffzg.hr*. W polu *User name* wpisać trzeba *gost*, a pole *Password* pozostawić puste.

Krótka i niezbyt dokładna instrukcja obsługi w języku chorwackim dostępna jest pod adresem: http://www.hnk.ffzg.hr/pretraga.html. Oryginalną instrukcję obsługi w języku czeskim znaleźć można pod adresem: http://ucnk.ff.cuni.cz/bonito/index.php.

Lista dostępnych, wymienionych wyżej, korpusów wyświetlana jest po kliknięciu szarego przycisku w prawym górnym rogu okna głównego programu. HNK 2.5 beta nie umożliwia tworzenia podkorpusów użytkownika.

Przeszukiwanie proste polega na wpisaniu poszukiwanej formy wyrazowej w polu *New query* i potwierdzeniu Enterem. Wyszukiwanie może obejmować jeden lub kilka oddzielonych spacjami wyrazów. Jeżeli nie jesteśmy pewni, jakiej formy wyrazowej szukamy, można skorzystać z funkcji *Word List* dostępnej w menu *Corpus*. Po wskazaniu prawdopodobnej sekwencji znaków (możliwe jest wykorzystanie wyrażeń regularnych – patrz niżej) i kliknięciu przycisku *Find Pattern* wyświetlone zostaną dostępne w korpusie wyrażenia spełniające

[1] Tuż przed oddaniem książki do druku udostępniono nową wyszukiwarkę dla HNK. Jest nią NoSketch Engine. O jej obsłudze można przeczytać w rozdziale traktującym o Czeskim Korpusie Narodowym. Nowa wersja korpusu HNK 3.0 obsługiwana przez tę przeglądarkę nie różni się pod względem struktury i funkcji od omówionej tu wersji 2.5, która jest nadal dostępna po zainstalowaniu programu Bonito.

zadane kryterium wraz z liczbą konkordancji. Wystarczy wówczas wybrać interesującą nas formę i kliknąć przycisk *Make Concordance*. Program wyświetli wyniki wyszukiwania dla wskazanej formy.

Program umożliwia też stosowanie wyrażeń regularnych, czyli zastępczych symboli wieloznacznych. Pamiętać należy, że różnią się one od znaków stosowanych w popularnych programach, np. edytorach. Można je stosować na początku, w środku i na końcu poszukiwanego wyrażenia. W tabeli przedstawiono podstawowe wyrażenia regularne używane w HNK i przykłady ich zastosowania. Pełny opis w języku czeskim dostępny jest pod adresem: http://ucnk.ff.cuni.cz/bonito/regular.php.

Tab. 7.3.

Symbol	**Znaczenie**	**Przykład**	**Przykładowy rezultat**
.	zastępuje dokładnie jeden znak	kuć.	kuća, kuće, kući
.*	zastępuje zero lub więcej znaków	student.*	student, studenti, studentske
.+	zastępuje jeden lub więcej znaków	student.+	studente, studenti, studentske
(abc)+	zastępuje ujętą w nawiasy część wyrażenia lub jej powtórzenie	(ta)+	ta, tata
(a|b)	oznacza alternatywę znaków *a* lub *b*	kuć(a|e|i|u|o|om|ama)	kuća, kuće, kući, kuću...
?	wyszukiwanie wyrażenia ze znakiem, po którym następuje ?, lub bez tego znaku	z?gaziti	gaziti, zgaziti
"a" | "b"	wyszukuje wyrażenie a lub b	"voz" | "vlak"	voz, vlak
(?i)	wyszukiwanie wyrażeń niezależnie od wielkości liter[2]	(?i)prsten	prsten, Prsten, PRSTEN
[ab]	zbiór liter, które mogą być podstawiane w wyrażeniu	ku[chm]a	kuca, kuha, kuma
[^ab]	wyszukiwanie wyrażeń, które nie zaczynają się od umieszczonych w nawiasach znaków	[^k]uma	guma, Huma, šuma
[a-c]	zakres liter, które mogą być podstawiane w wyrażeniu	ku[c-n]a	kuca, kuha, kula, kuma, kuna
.{x}	x to liczba zastępowanych znaków poprzedzających nawias	kuć.{2}	kućne, kućni, kućom

[2] W HNK nie działa funkcja *lowercase*, która sprawia, że program ignoruje przy wyszukiwaniu wielkość liter.

[ciąg dalszy tabeli ze strony poprzedniej]

Symbol	**Znaczenie**	**Przykład**	**Przykładowy rezultat**
.{x,y}	x, y to dopuszczalny zakres liczby znaków poprzedzających nawias	kuć.{1,3}	kuća, kućom, kućnom, kućama
\	umożliwia wyszukiwanie znaków funkcyjnych używanych w wyrażeniach regularnych jako znaków interpunkcyjnych	\.	.

Możliwe jest także przeszukiwanie fragmentów tekstu z oznaczeniami strukturalnymi. Polecenie *<head>* 'wyraz' służy do wyszukiwania wyrazu na początku tytułu, 'wyraz' *</head>* wybiera wyraz na końcu tytułu, a 'wyraz' *within <head/>* – wewnątrz tytułu. W wyszukiwaniu z oznaczeniami strukturalnymi można stosować wyrażenia regularne.

Edycja wyników wyszukiwania

Do ustalenia rozmiaru kontekstu towarzyszącego wyszukiwanej frazie służy polecenie menu *View* > *Context* lub klawisz skrótu F7. Najbardziej praktyczne jest operowanie liczbą znaków (*Characters*), pozycji tekstowych (*Positions*) i zdań (<s>) po lewej i prawej stronie wyszukiwanego wyrażenia. Kryteria wybiera się po naciśnięciu przycisków okna dialogowego, ukazującego się po wybraniu polecenia z menu lub użyciu klawisza skrótu.

Do sortowania alfabetycznego wyników służy polecenie menu *Concordance* > *Simple sort* lub kombinacja klawiszy Ctrl+S. Najpierw należy określić liczbę sortowanych pozycji (*Number of positions to sort*). Sortowanie może odbywać się według kolejności alfabetycznej lewego (*Left context*) lub prawego kontekstu (*Right context*), kolejności alfabetycznej wyszukiwanego wyrażenia (*KWIC from left*), odwróconej kolejności alfabetycznej (*KWIC from right*). Po zaznaczeniu *KWIC from left* oraz opcji *Backword* otrzymujemy indeks *a tergo*. Opcja *Ignore case* służy wyłączeniu odrębnego sortowania wyrazów pisanych wielkimi i małymi literami. Szczegółowe informacje na temat sortowania wyników w Bonito dostępne są pod adresem: http://ucnk.ff.cuni.cz/bonito/trideni.php.

Przed zapisaniem wyników wyszukiwania można usunąć niektóre wersy. W tym celu należy kliknąć wybrane wiersze (zostaną podświetlone na niebiesko), wybrać polecenie menu *Cocordance* > *Delete selected* i potwierdzić wybór, naciskając Enter. Liczbę wierszy można też zredukować za pomocą okna dialogowego wyświetlanego po wybraniu polecenia menu *Concordance* > *Reduce*. Po kliknięciu pierwszego klawisza wyświetlana jest lista, na której wybieramy, czy pojawić się mają pierwsze (*First*), środkowe (*Middle*), końcowe (*Last*) czy

wybrane losowo (*Random*) rekordy, Następnie określamy liczbę wyświetlanych jednostek, a przyciskiem po prawej stronie ustalamy, czy polecenie dotyczyć ma wierszy, procenta wierszy czy 1/100 procenta.

Tak opracowane efekty wyszukiwania można wydrukować za pomocą polecenia menu *Concordance* > *Print*. Rozwiązanie to jest o tyle niewygodne, że program Bonito wybiera pierwszą zainstalowaną drukarkę i pliku nie można np. przekształcić w format .pdf. Bardziej praktyczne jest zapisanie wyników wyszukiwania w postaci pliku, który można następnie opracować w dowolnym edytorze tekstów. Służy do tego polecenie menu *Concordance* > *Save to file* lub klawisz skrótu F2. Po wybraniu tego polecenia ukazuje się okno dialogowe, w którym należy zmienić sposób kodowania (*Character encoding*) z domyślnego ustawienia *ascii* na *utf-8*. W ten sposób zachowujemy znaki diakrytyczne charakterystyczne dla danego alfabetu, a kodowanie to jest czytelne dla wszystkich używanych obecnie edytorów. Następnie określamy, czy zapisywany plik ma obejmować tylko wyświetlone na danym ekranie linie (pole wyboru *displayed lines only*), czy też wszystkie rekordy w danym wyszukiwaniu (*all lines*). Kolejne rekordy można opatrzyć numerami (opcja *line numbering*) oraz ustawić wielkość kontekstu (przycisk *Context*); opcja *align KWIC*, która ma służyć do wyrównywania tekstu według poszukiwanego wyrażenia, w praktyce nie działa, podobnie jak przycisk *Header*, służący w zamyśle autorów programu do ograniczenia lub poszerzenia informacji wyświetlanych w nagłówku rekordu. Zapisywanie dużych zbiorów danych może trwać, nawet przy szybkim połączeniu z internetem, dość długo. Podczas zapisywania pliku za przyciskiem wyboru podkorpusu w prawym górnym rogu okna głównego programu przesuwać się będzie napis *STOP* – jego zniknięcie oznacza, że transfer pliku został zakończony. Wtedy dopiero można kontynuować pracę.

Anotacja zewnętrzna

Anotacja zewnętrzna bazuje również na możliwościach programu Bonito (polecenie z menu *View* > *References* lub klawisz skrótu F4), jednakże konsekwentnie wypełnione są tylko pola oznaczające numer jednostki (*Token number*), numer dokumentu korpusowego (*doc*) oraz rodzaj dokumentu (*doc.type*) i symbol pliku (*doc.file*). Użytkownik ma więc dostęp wyłącznie do informacji o dacie powstania tekstu i/lub jego pochodzeniu. Informacje te pojawiają się na początku wyszukanego rekordu, okno programu należy więc przewinąć za pomocą suwaka u dołu okna maksymalnie w lewo.

#10490689,doc#19913,tip=article,izv=gs20030330hr19034 *Iz svega je očito, unatoč činjenici da se nitko o tome ne želi otvoreno i jasno izjašnjavati...*

'Widać z tego wyraźnie, mimo że nikt otwarcie i jasno nie chce się na ten temat wypowiedzieć...'

Cytat to fragment artykułu (*tip=article*) z gazety *Glas Slavonije* z dnia 30 marca 2003 r. (*izv=gs20030330*), jest to jednostka numer 10490689 z dokumentu 19913. Objaśnień dokładniejszej lokalizacji cytatu (hr19034) autorzy nie podają. W kolejnym przykładzie mamy tylko informację o numerze jednostki i dokumentu, typie tekstu oraz jego pochodzeniu – *Hrvatska revija* numer 54. Brakuje informacji o roku powstania tekstu:

#23365733,doc#48143,tip=article,izv=HR540201A *Drugi način na koji se priprema kuhana šunka...*
'Inny sposób przygotowania gotowanej szynki...'

W przypadku tekstów literackich oprócz numerów jednostki i dokumentu podaje się tylko skrót nazwiska autora i tytułu. W poniższym przykładzie jest to Petar Hektorović i jego *Ribanje i ribarsko prigovaranje* (1556). Dokładniejsze dane bibliograficzne znaleźć można pod adresem: http://www.hnk.ffzg.hr/Izvori_Klasici.html:

#23989872,doc#48257,tip====NONE===,izv=hektorov_rib *Po gorah najliše gdino su ravnine </l> <l> I stine najviše i guste planine ; </l> <l> Odkud pak ističu bistrimi vodami , </l> <l> K moru se zamiču barzimi rikami . </l> <l> Kakono tko < kuha> naprišno u sudu , </l> <l> Često oganj duha ne prašćajuć trudu.*

Co ciekawe, ten sam fragment wyekscerpowany bezpośrednio z podkoprusu Klasici opatrzony został informacją o rodzaju literackim, jaki reprezentuje:

#95484,doc#5,vrst=poetry,izv=hektorov_rib *Po gorah najliše gdino su ravnine I stine najviše i guste planine ; Odkud pak ističu bistrimi vodami , K moru se zamiču barzimi rikami . Kakono tko < kuha> naprišno u sudu , Često oganj duha ne prašćajuć trudu .*

Różne numery jednostek oraz dokumentów korpusowych, jak również brak oznaczeń początku i końca wiersza w podkorpusie każą przypuszczać, że zawartość korpusu głównego w stosunku do jego składowych nie odzwierciedla dokładnie ich struktury i została w nieco odrębny sposób opracowana.

Anotacja wewnętrzna

Lematyzacja opracowana została konsekwentnie jedynie w podkorpusie cw2000. Mylące dla użytkownika może być to, że polecenie *lemma* działa także w korpusie głównym. Jednak uzyskiwane tam wyniki są w znacznej mierze spaczone. Na przykład po wpisaniu:

[lemma="poljski"]

uzyskujemy nie tylko jednostki odnoszące się do przymiotnika 'polski', lecz również teksty zawierające nazwę kraju 'Polska'. W lematyzowanym podkorpusie cw2000 wartości te są już rozróżniane. Niestety, korpus ten liczy zaledwie 118 529 jednostek. Trudno więc stwierdzić, w jaki sposób autorzy poradzili sobie z polisemią i homonimią, gdy np. przymiotnik *poljski* oznacza nie tylko 'polski', lecz także 'polny' i 'polowy'. W korpusie głównym rozróżnienie takie oczywiście nie występuje. Podejrzewać jednak można, iż nie uwzględniono go również w lematyzowanym podkorpusie. Do wniosku takiego skłania zastosowany tam sposób opisu homonimów. Na przykład polecenie [lemma="dug"] zwraca wartość zero. Dopiero rozróżnienie *dug1* i *dug2* powoduje, że otrzymujemy wyniki dla znaczeń, odpowiednio, 'dług' oraz 'długi'. Ponieważ nie ma w podkorpusie odrębnych jednostek *poljski1* i *poljski2*, wnioskować można, że autorzy ograniczyli się jedynie do rozróżnienia lematów o odrębnej charakterystyce morfologicznej.

Pomijając tego typu problemy, które pozostają nierozwiązane w większości korpusów, przyznać należy, że lematyzacja w podkorpusie cw2000 wykonana została dość starannie. Autorzy poradzili sobie z problematycznymi zwykle skrótowcami. Na przykład ze skrótem partii HDZ (Hrvatska demokratska zajednica 'Chorwacka Wspólnota Demokratyczna') związane są trzy lematy. Jeden odnosi się do samej partii (HDZ, czyt. hadeze), drugi do przymiotnika dzierżawczego utworzonego od skrótu (HDZ-ov, czyt. hadezeov), a trzeci do członka tej partii (HDZ-ovac, czyt. hadezeovac).

Opis morfologiczny zgodny jest z systemem MULTEXT-East Morphosyntactic Specifications (wersja 3.0 z 10 maja 2004 r.; szczegółowy opis składni zapytań pod adresem: http://nl.ijs.si/ME/Vault/V3/msd/html/). Anotacja konsekwentnie stosowana jest tylko w korpusie testowym cw2000. Dane wyświetla się za pomocą polecenia *View* > *Attributes* lub klawisza skrótu F5. Można opatrzyć nimi cały tekst:

*Suradnja/**Ncfsn** s/**Spsi** Makedonijom/**Npfsi** <radi /**radi1**/**Spsg**> ulaska /**Ncmsg** u/**Spsa** EU/**Ynmsa***
'Współpraca z Macedonią w celu wejścia do UE'

lub ograniczyć je tylko do wyszukiwanego słowa:

*Što bi, usporedbe <radi/**radi2**/**Stsg**>, učinila majka...*
'Co, dla porównania, uczyniłaby matka...'

Mylący dla użytkownika jest fakt, że – wbrew informacjom autorów – anotacja morfologiczna pojawia się także w głównym korpusie HNK. Jednak najwyraźniej nie została ona zweryfikowana, a jest jedynie rezultatem znakowania

automatycznego, w którym roi się od podstawowych błędów. Na przykład w poniższym, losowo wybranym zdaniu czasowniki zostały opisane jako rzeczowniki i to w dodatku w różnych przypadkach.

Muškarac/Ncmsn je/Vcip3s glava/Ncfsn kuće/Ncfsg ,/Z a/Ccs za/Spsa ženu/Ncfsa je/Vcip3s da/Css ***<kuha/Ncfsn> ,/Z pere/Ncfsg*** *i/Ccs odgaja/Vmip3s djecu/Ncfsa./*
'Mężczyzna jest głową domu, a do kobiety należy, by gotowała, prała i wychowywała dzieci'

Tak więc, mimo że procent poprawnie oznaczonych form znacznie przewyższa odsetek błędów, to jest ich wystarczająco wiele, by podważyły sens korzystania z anotacji morfologicznej w korpusie głównym HNK. Natomiast korpus cw2000 jest tak mały, że dane – chociaż bardzo dokładne i starannie opracowane – nie wystarczały do wyciągania ogólniejszych wniosków na temat zjawisk morfologicznych w języku chorwackim.

Ponownie również pozostaje wyrazić żal, że znakowanie nie objęło całego korpusu, gdyż w korpusie testowym udało się wychwycić bardzo subtelne różnice kategorialne. O ile nie dziwi na przykład fakt, że przyimek *s* opisywany jest jako łączący się z dopełniaczem *s1/Spsg* oraz *s2/Spsi* w połączeniach z narzędnikiem, to już rozróżnienie przyimka *radi* występującego w pre- (*radi1/Spsg*) i postpozycji (*radi2/Stsg*; ta druga sytuacja stosunkowo rzadko) budzić może uznanie.

Kolokacje

Do ustalenia, czy mamy do czynienia z przypadkowym połączeniem wyrazów, czy też charakteryzuje się ono pewną stałością, regularnością, a więc można mówić o kolokacji, konieczna jest interpretacja czterech parametrów. Dostęp do nich uzyskujemy za pomocą polecenia menu *Concordance* > *Statistics* > *Collocations* lub przy użyciu klawisza skrótu Ctrl+L. Pojawia się wówczas okno dialogowe, w którym określamy zakres wyszukiwanych połączeń wyrazowych (*In the range from ... to ...*) z prawej (wartości dodatnie parametru) lub lewej strony (wartości ujemne) wyrażenia stanowiącego centrum kolokacji. W celu uniknięcia zafałszowania parametrów MI-score i frekwencji względnej (por. niżej) ustalić należy minimalną liczbę wystąpień danej formy w całym korpusie (*Minimum frequency in corpus*) oraz w wyszukiwanym zakresie (*Minimum frequency in given range*). Na końcu ustalamy maksymalną liczbę wyświetlanych połączeń (*Maximum number of diplayed lines*) oraz sposób prezentacji danych – według frekwencji względnej lub bezwzględnej liczby wystąpień danej formy. Po kliknięciu przycisku OK ukazuje się zestawienie form występujących w zadanym zakresie wraz z parametrami *MI-score*, *T-score*, *Rel.f.[%]* oraz *Abs.f.*

Frekwencja względna (*Rel.f.[%]*) i parametr MI-score odzwierciedlają zależność pomiędzy liczbą wystąpień uwzględnianych wyrażeń a liczbą ich połączeń. Innymi słowy, możemy zakładać, że im wyższy parametr MI-score oraz frekwencja względna, tym częściej dwa wyrażenia występują łącznie w stosunku do ogólnej liczby ich wystąpień. Wadą tych parametrów jest to, że najwyższe wyniki uzyskujemy przy słowach, które w korpusie występują sporadycznie, na przykład raz lub dwa, ale zawsze razem. W tym właśnie celu ustala się minimalną liczbę wystąpień wyrażenia. Wywodzący się ze statystyki parametr T-score odzwierciedla zależność między rozłożeniem liczby wystąpień poszczególnych wyrażeń i ich połączeń a losowym rozkładem słów w korpusie. Im wyższa wartość T-score, tym większe prawdopodobieństwo, że mamy do czynienia z kolokacją. Frekwencja absolutna oznacza liczbę połączeń danego wyrażenia z centrum kolokacji w określonym zakresie.

Oto początek listy prawostronnych kolokacji przymiotnika *srpski*:

word	**MI-score**	**T-score**	**Rel. f[%]**	**Abs. f**
paravojsku	11.89	1.732	60	3
paramilitarce	11.89	1.732	60	3
koncentracijskih	11.86	9.846	58.79	97
artiljeriji	11.78	2.235	55.56	5
paradržava	11.77	4.581	55.26	21
paravojske	11.73	5.998	53.73	36
odmetničkih	11.62	1.999	50	4
paravlasti	11.62	2.828	50	8
krajinama	11.62	1.732	50	3
nacionalnosti	11.61	29.28	49.4	858
paradržave	11.41	5.914	43.21	35
šoviniste	11.4	1.731	42.86	3
rezervistima	11.28	3.315	39.29	11
paravojska	11.21	2.999	37.5	9

Na pierwszym miejscu znajduje się forma biernikowa rzeczownika *paravojska*, którym określano w czasie wojny lat dziewięćdziesiątych nieregularne serbskie formacje wojskowe. Frekwencja względna wskazuje, że wystąpił on w połączeniu z przymiotnikiem *srpski* trzy razy, co stanowiło 60% wszystkich wystąpień formy *paravojsku*. Znalazło to odzwierciedlenie w wysokiej wartości parametru MI-score, jednak ze względu na dużą liczbę potwierdzeń przymiotnika *srpski* w korpusie parametr T-score nie należał do najwyższych. Ze względu na niską frekwencję bezwzględną można by pominąć to słowo przy poszukiwaniu kolokacji, jednak zwraca uwagę fakt, że na kolejnych pozycjach znajdują się formy *paravojske* i *paravojska* z wyższą frekwencją bezwzględną oraz wartościami T-score. Tak więc zestawienie *srpska paravojska* w różnych formach przypadków gramatycznych uznać należy za dość stabilną i często pojawiającą się w korpusie kolokację. Wpisuje się ona w wojenny kontekst większości pozostałych połączeń. Na trzecim miejscu z wysokimi wartościami wszystkich

parametrów uplasowała się forma *koncentracijskih*, stanowiąca część zestawienia *srpski koncentracijski logori* 'serbskie obozy koncentracyjne'. W podobną tonację powojennego dyskursu prasy chorwackiej wpisują się połączenia *srpska artilerija* 'serbska artyleria', *srpska paradržava* 'serbskie quasi-państwo', *srpski odmetnički* 'serbski zdradziecki', *srpske paravlasti* 'serbskie quasi-władze', *srpske krajine* 'serbskie regiony (zamieszkałe przez Serbów, którzy ogłosili secesję)'. Największą frekwencję bezwzględną oraz parametr T-score ma połączenie *srpske nacionalnosti*, jest fragment stosowanego w prasie chorwackiej określenia mniejszości serbskiej *građani srpske nacionalnosti* 'obywatele narodowości serbskiej', występującego niezwykle często w wielu kontekstach, a zwłaszcza w związku z problematyczną kwestią powrotu uchodźców serbskich na zajęte przez Chorwatów terytoria.

Jak więc widać, mimo że HNK nie ułatwia ani wyszukiwania według lematów, ani analizy morfologicznej, możliwe jest już w obecnej postaci wykorzystanie korpusu do opisu wybranych aspektów chorwackiego dyskursu prasowego.

Kolejną użyteczną funkcją statystyczną umożliwiającą stwierdzenie odchyleń ilościowych rozkładu wystąpień poszukiwanego wyrażenia jest możliwość graficznego przedstawienia średniej jego wystąpień w poszczególnych miejscach korpusu. Służy do tego polecenie menu *Concordance* > *Statistics* > *Distribution overview*. Po jego wywołaniu pojawia się wykres przedstawiający frekwencję wyrażenia w poszczególnych częściach korpusu. Program nie umożliwia co prawda określenia na wykresie, w jakim dokumencie poszukiwana jednostka występuje częściej lub rzadziej niż w pozostałych materiałach, ale po kliknięciu interesującego nas miejsca na grafie, program przewija okno główne do odpowiedniego dokumentu, który można zidentyfikować dzięki informacjom zawartym w anotacji zewnętrznej.

Możliwości wykorzystania HNK

Największe mankamenty HNK, uniemożliwiające wykorzystanie go jako pełnowartościowego źródła do opisu systemu gramatycznego współczesnego języka chorwackiego to niezrównoważona i niereprezentatywna struktura materiałów oraz brak anotacji morfologicznej i lematyzacji. Do czasu usunięcia tych braków korpus może być wykorzystywany jedynie jako źródło pomocnicze, służące do weryfikacji wcześniej przyjętych założeń badawczych. Biorąc pod uwagę wspomniane ograniczenia HNK, niemożliwe z jego udziałem są nie tylko jakiekolwiek badania wywodzące się podejścia corpus-driven, ale nawet w badaniach corpus-based dane korpusowe należałoby weryfikować za pomocą innych źródeł, np. tradycyjnych opisów gramatycznych lub potraktowanych jako swoisty korpus monitorujący zasobów chorwackiego internetu.

Stosunkowo reprezentatywny zestaw tekstów literackich, powstałych na przestrzeni pięciu wieków, zawarty w podkorpusie Klasici pozwala natomiast

na dość dokładne obserwacje diachroniczne określonych zjawisk gramatycznych i leksykalnych. Podobnie jednak jak w przypadku opisów synchronicznych, może on służyć wyłącznie do dodatkowego potwierdzenia wcześniej postawionych i zweryfikowanych hipotez.

Jak wspomniano wyżej, w punkcie *Kolokacje*, HNK całkiem nieźle sprawdza się natomiast jako narzędzie do analizy i opisu chorwackiego dyskursu prasowego przełomu wieków. Stosunkowo duża liczba zróżnicowanych gatunkowo, treściowo i ideowo materiałów pozwala nie tylko na weryfikowanie założeń opartych na przesłankach pozakorpusowych, ale przy pewnej dozie ostrożności pozwala na *stricte* korpusowe badania semantyczne bazujące na podejściu corpus-driven. Jako przykłady możliwości wykorzystania badawczego HNK posłużyć mogą podane w bibliografii do artykułu prace: Molas 2009a, 2009b, 2010a, 2010b.

Hrvatski jezični korpus

Wszelkie informacje na temat drugiego korpusu języka chorwackiego, czyli HJK, wynikają wyłącznie z praktycznych działań, jakie wykonać można na stronie korpusu http://riznica.ihjj.hr/index.hr.html, ponieważ dokumentacja tego projektu nie jest dostępna. Na wymienionej stronie działają tylko dwa hiperłącza umożliwiające przełączenie się na stronę przeszukiwania podkorpusu literatury (http://riznica.ihjj.hr/philologic/Riznica.whiz-bang.form.hr.htl), podkorpusu prasowego (http://riznica.ihjj.hr/philologic/Tiskovi-ne.whiz-bang.form.hr.html) lub całego korpusu (http://riznica.ihjj.hr/philologic/Cijeli.whiz-bang.form.hr.html). Nie działają hiperłącza odsyłające do spisu frekwencji jednostek korpusowych oraz alfabetycznego indeksu tych jednostek. Ostatnie aktualności pochodzą z października 2006 r. Objętość korpusu jest prawdopodobnie większa niż HNK, np. dla formy *kuća* 'dom' w HNK otrzymujemy ponad 8800 potwierdzeń, a w HJK ponad 10 000; dla przymiotnika *hrvatski* 'chorwacki' w HNK ponad 27 000 potwierdzeń, ale w HJK już ponad 47 000. Niestety, nie jest znana struktura korpusu. Anotacja zewnętrzna wskazuje, że składa się on z dzieł literatury chorwackiej, przekładów oraz zbioru tekstów prasowych, m.in. Vjesnik z lat 2000-2006. Korpus nie jest lematyzowany ani anotowany morfologicznie. Obecny interfejs nie wskazuje, by autorzy w ogóle planowali wprowadzenie tego rodzaju funkcji i właściwości. Bardzo rozbudowane są natomiast możliwości wyszukiwania bibliograficznego. Niezwykle interesująco wygląda też kolokator, reprezentujący wyniki w postaci przejrzystej i łatwej do dostosowania do indywidualnych potrzeb tabeli. Ze względu na brak odniesień statystycznych są one jednak zupełnie bezużyteczne.

Bibliografia

Molas J. (2009a). *Zamjenice njezin i njen – norma i uzus suvremenoga hrvatskoga jezika.* W: Pranjković I., Silić J., Badurina L. (red.), *Jezični varijeteti i nacionalni identiteti. Prilozi proučavanju standardnih jezika utemeljenih na štokavštini.* Zagreb, s. 347-356.

Molas J. (2009b). *Semantičke promjene u rječniku političkih termina.* W: Kryżan-Stanojević B. (red.), *Lice i naličje jezične globalizacije.* Zagreb, s. 119-129.

Molas J. (2010a). Strelica je pogodila grješnika – *o pewnym problemie pisowni chorwackiej w świetle danych komputerowego korpusu języka.* „Slavia Meridionalis" 10, s. 223-234.

Molas J. (2010b). *Językowy obraz Polski i Polaków w materiałach prasowych komputerowego korpusu języka chorwackiego.* W: Bem-Wiśniewska E., Dubisz S., Porayski-Pomsta J., Wroczyński T. (red.), *Zabawy pożyteczne prozą. Z przyjacielstwa Andrzejowi K. Guzkowi uczynione i Jemuż dedykowane.* Warszawa, s. 306-325.

8. Praktyczny przewodnik po korpusie języka serbskiego

Jacek Duda
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Wstęp

Korpusy języka serbskiego są jednym z wielu nowoczesnych sposobów opisu systemu gramatycznego oraz słownictwa języka serbskiego. Jednak w porównaniu ze słownikami elektronicznymi znajdują się jeszcze na bardzo niskim stopniu rozwoju (por. Krstev 2008, 11). Korpusy języka serbskiego są obecnie dalece niedoskonałe i prace nad nimi – z różnych przyczyn – posuwają się bardzo powoli, wobec czego nie można ich porównywać z o wiele bardziej rozbudowanymi korpusami języków słowiańskich, takimi jak Czeski Korpus Narodowy czy Narodowy Korpus Języka Polskiego. Z tego powodu analiza zagadnień związanych z serbską lingwistyką korpusową będzie niepełna i nie obejmie wszystkich aspektów funkcjonowania korpusów, pracy z nimi czy szczegółów technicznych.

Pomimo istnienia dwóch korpusów języka serbskiego, w niniejszym rozdziale szerzej omówiony zostanie jeden z nich – Korpus Współczesnego Języka Serbskiego (Korpus savremenog srpskog jezika, dalej: KSSJ). Jest to bowiem korpus kilkadziesiąt razy większy, dużo lepiej opracowany i wyposażony w więcej narzędzi niż drugi z korpusów (Korpus srpskog jezika, dalej: KSJ)[1]. Ponadto KSSJ bada język w ujęciu synchronicznym, podczas gdy KSJ prezentuje ujęcie diachroniczne. Nie oznacza to jednak, że zostanie on w opracowaniu całkowicie pominięty.

Informacje ogólne

Spośród dwóch korpusów języka serbskiego, chronologicznie pierwszy jest Korpus srpskog jezika, powstający od roku 1957 pod kierownictwem Đorđa Kosticia oraz jego syna Aleksandra, reprezentujących Instytut Fonetyki Eksperymentalnej i Patologii Mowy w Belgradzie. Od roku 1996, kiedy w proces jego

[1] http://www.serbian-corpus.edu.rs, 21 III 2013.

tworzenia włączyli się pracownicy Laboratorium Psychologii Eksperymentalnej Uniwersytetu w Belgradzie, jego zbiory są zdigitalizowane i dostępne w internecie pod adresem http://www.serbian-corpus.edu.rs. KSJ liczy 11 milionów słów i – jak już wspomniano – jest korpusem diachronicznym, tj. obejmuje zasoby w języku serbskim (serbsko-chorwackim, chorwackim[2]) pochodzące z całej jego historii, od wieku XII do czasów współczesnych. Zawarte w korpusie źródła podzielone są chronologicznie według kryterium okresu, w którym powstawały. Największa część korpusu obejmuje materiał współczesny i zawiera ok. 7 z 11 milionów słów. Pozostałe cztery części (XII–XVII w.; wiek XVIII i początek XIX w. bez dzieł Vuka Karadžicia; dzieła Vuka Karadžicia; druga połowa XIX w.) zawierają po około milion słów każda. Należy w tym miejscu zauważyć, że digitalizacja materiałów w przypadku tego korpusu polegała jedynie na przeniesieniu zbiorów do postaci cyfrowej bez wyposażenia ich w nowoczesne narzędzia wyszukiwawcze. Dopiero od roku 2002 trwają prace (jeszcze niezakończone) nad zakodowaniem zawartości korpusu w języku XML. Dlatego też nie jest możliwe obecnie korzystanie z korpusu *online*, co czyni go niemal nieużytecznym z perspektywy współczesnej lingwistyki korpusowej. Fragment przykładowego arkusza zdigitalizowanego tekstu obrazuje rysunek poniżej.

Rys. 8.1. Próbka materiału z KSJ.

Word entry	text	ptc	ptc word type	gramm. form	numeric. code	# graph	# syl	phonol. structure
велик	Вељим		п	2 инс Ј му	202611	5	2	10101
део	делом		и	инс Ј му	100611	5	2	10101
свој	свога		з	присв свп г Ј му	428211	5	2	11010
ток	тока		и	г Ј му	100211	4	2	1010
река	река		и	н Ј ж	100112	4	2	1010
Дрина	Дрина		и	н Ј ж	100112	5	2	11010
протицати	протиче		гл	през ЗЛ Ј	521310	7	3	1101010

Źródło: http://www.serbian-corpus.edu.rs/pdf/andric.pdf

Drugim z korpusów, chronologicznie dużo późniejszym, jest Korpus Współczesnego Języka Serbskiego, który traktowany jest w zasadzie jako narodowy korpus języka serbskiego. Prace nad nim rozpoczęły się w roku 2002 na Wydziale Matematyki Uniwersytetu w Belgradzie, pod kierunkiem Ljubomira Popovicia, Duško Vitasa i Cvetany Krstev. Obecnie obejmuje on około 112 milionów słów. W porównaniu z korpusem diachronicznym jego struktura jest inna. Źródła nie są podzielone chronologicznie, a według kategorii takich jak serbska literatura piękna, beletrystyka zagraniczna, prasa, literatura naukowo-techniczna i inne. Korpus zawiera liczne podkorpusy, takie jak pochodzący z 1985 roku niewielki korpus języka serbsko-chorwackiego, korpus tekstów

[2] Obecność tych tekstów wynika z czasu powstania korpusu, traktowanego jako reprezentacja wspólnego języka Serbów, Chorwatów, Bośniaków i Czarnogórców.

prasowych poświęconych kryzysowi politycznemu roku 2000, korpus przysłów Vuka Karadžicia oraz dwa korpusy równoległe – serbsko-francuski oraz serbsko-angielski. Dostępna jest również otagowana i nieotagowana wersja korpusu z roku 2003. Korpus znajduje się w internecie pod adresem http://www.korpus.matf.bg.ac.rs. Od strony technicznej obsługiwany jest przez system IMS Corpus Workbench, stworzony na Uniwersytecie w Stuttgarcie.

W przeciwieństwie do pierwszego z wymienionych korpusów, drugi jest w całości dostępny w internecie dla zarejestrowanych użytkowników. Rejestracja jest bezpłatna i można jej dokonać, wysyłając wiadomość elektroniczną pod jeden z podanych na stronie internetowej adresów.

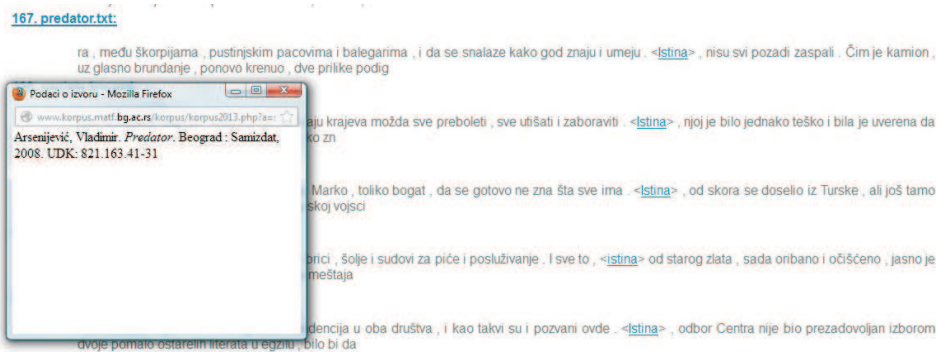
Charakterystyka Współczesnego Korpusu Języka Serbskiego

Pomimo względnie rozwiniętej struktury, omawiany korpus posiada wiele luk, które utrudniają wykorzystanie w pełni jego zasobów. Są to zarówno niedostatki techniczne, jak i obiektywne problemy wynikające z ogólnej charakterystyki języka serbskiego.

Struktura korpusu nie jest zbilansowana: 57% źródeł stanowią dzienniki, 13% innego typu prasa, 15% to literatura piękna, 9% stanowią inne teksty związane z kulturą, 7% to literatura naukowa, a jedynie 1% – teksty techniczne (Krstev 2008: 139). W rezultacie niemożliwe jest prowadzenie wszystkich typów badań na różnorodnych rodzajach tekstów. Korpus przydatny jest przede wszystkim do prac nad współczesną prasą serbską (źródła obejmują dzienniki wydawane od 1993 roku). Nie pokazuje on jednak pełnego obrazu i stanu języka.

Korpus jest bardzo dobrze anotowany zewnętrznie. Przy każdym rezultacie wyszukiwania pojawia się w osobnym oknie krótki opis bibliograficzny źródła, często z sygnaturą, pod którą źródło indeksowane jest przez Bibliotekę Narodową w Belgradzie. Przykład pokazuje Rys. 8.2.

Rys. 8.2. Anotacja zewnętrzna KSSJ.



Źródło: http://www.korpus.matf.bg.ac.rs

Korpus, w przeciwieństwie do KSJ, nie jest anotowany morfologicznie. Jako przyczynę takiego stanu rzeczy autorzy korpusu podają niską skuteczność anotacji morfologicznej dla tekstów serbskojęzycznych, co wiąże się z dużą liczbą występujących w języku serbskim form synkretycznych pośród różnych części mowy (np. forma tekstowa *duše* może być zarówno formą dopełniacza liczby pojedynczej oraz mianownika i biernika liczby mnogiej rzeczownika *duša* 'dusza', jak również formą 3.os. l.mn. czasownika *dušiti* 'dusić'). To z kolei skutkuje zaledwie 55-60% skutecznością anotacji, a poziom taki jest uważany przez autorów za zbyt niski (Vitas, Krstev 2012). Dlatego KSSJ nie jest anotowany wewnętrznie. Natomiast KSJ (zob. Rys. 8.1.) jest anotowany morfologicznie za pomocą systemu ponad 800 tagów określających miejsce każdego leksemu w systemie gramatycznym serbszczyzny[3].

Synkretyzm form nie jest jedynym zjawiskiem, które nastręcza problemów zarówno twórcom korpusu, jak i jego użytkownikom. Inne cechy charakterystyczne dla języka serbskiego, które wpływają znacząco na postać korpusu, a także na tempo prac nad nim, to:

– równoległe użycie dwóch alfabetów (cyrylicy i łacinki), przy czym w alfabecie łacińskim używane są charakterystyczne grafemy, które powinny być zakodowane w sposób niewymagający instalacji dodatkowych czcionek,

– stosowanie przez autorów tekstów elektronicznych różnych sposobów kodowania tekstu, co powoduje podobne problemy jak w przypadku odmiennych metod zapisu znaków narodowych,

– obecność dwóch wariantów wymowy – ekawskiej i ijekawskiej,

– skomplikowany najnowszy okres historii języka serbskiego, który przez kilkadziesiąt lat oficjalnie tworzył wraz z chorwackim jeden język serbsko-chorwacki (Krstev 2008).

Problem dwóch alfabetów został rozwiązany za pomocą transkrypcji tekstów cyrylickich na zmodyfikowaną dla potrzeb korpusu łacinkę, przy czym charakterystyczne dla języka serbskiego grafemy zostały oddane w następujący sposób:

Tab. 8.1. Specjalne znaki alfabetu serbskiego w KSSJ.

Cyrylica	ћ	ч	ђ	џ	љ	њ	ш	ж
Łacinka	ć	č	đ	dž	lj	nj	š	ž
Korpus	cx	cy	dx	dy	lx	nx	sx	zx

Źródło: C. Krstev, *Processing of Serbian*, s. 138.

Wymaga to opanowania nowego dla użytkownika sposobu zapisu, zmniejsza jednak możliwość pomyłki czy niewłaściwego wyświetlenia rezultatów zapytania spowodowanego innym kodowaniem strony.

[3] Zob. http://www.serbian-corpus.edu.rs/ns/tagging/new/sifre_nove.htm, 16 IV 2013.

Większym problemem jest występowanie dwóch wariantów wymowy. Co prawda, za dominujący w języku serbskim standard uznaje się zapis wariantu wymowy ekawskiej, gdzie kontynuantem prasłowiańskiego **ě* jest *e*, w wielu tekstach jednak, zwłaszcza starszych i gwarowych, występuje wymowa ijekawska, gdzie kontynuantem tym jest *(i)je*. W nieanotowanym morfologicznie korpusie sprawia to problem, ponieważ przed zapytaniem należy znać i uwzględnić w wyszukiwaniu formę obowiązującą w zapisie ijekawskim (np. *hljeb* zam. *hleb*). Na temat wyszukiwania poszczególnych fraz w korpusie więcej informacji można znaleźć w dalszej części rozdziału.

Jak wspomniano powyżej, korpus jest jeszcze nieanotowany morfologicznie, dlatego nie można mówić o systemie tagowania; prace nad anotacją morfologiczną i syntaktyczną są prowadzone przez zespół pod kierunkiem Miloša Utvicia. Autorzy podają, że do tagowania zostanie użyty program TnT (Statistical Part-of-Speech Tagging). Pierwsze próby pracy z otagowanym tekstem zostały już podjęte. Zespół Utvicia przygotował otagowany morfosyntaktycznie tekst powieści *1984* George'a Orwella.

Zestaw narzędzi służących lub mogących służyć do anotacji morfologicznej tekstów w języku serbskim można znaleźć również na stronie http://nl.ijs.si/ME/, na której znajduje się powstały w połowie lat 90. standaryzowany opis języka serbskiego na platformie MULTEXT. Jak zaznaczają wszakże eksperci, nie jest to jedyny taki opis i zestaw, który powstał w ciągu ostatnich dwudziestu lat[4].

Korzystanie z korpusu – podstawowe zapytania

Wyszukiwanie zapytań w korpusie omówione jest na stronie internetowej korpusu pod adresem http://www.korpus.matf.bg.ac.rs/prezentacija/uputstvo.html. Poniżej podane zostaną podstawowe formy zapytań, najczęściej używane przez korzystających z korpusu.

Język serbski jest językiem fleksyjnym, w związku z czym wyszukiwany leksem posiada wiele form, które bardzo często podlegają obocznościom fonetycznym i ortograficznym, a odmiana implikuje występowanie wielu rozmaitych końcówek. Na przykład leksem oznaczający 'dziewczynę' może mieć postać ekawską *devojka*, ijekawską *djevojka* oraz dialektalną *đevojka*. Do tego należy dodać oboczności tematu powstające podczas odmiany, np. ruchome *a* w dopełniaczu liczby mnogiej (*devojaka*) czy palatalizacja *k* do *c* w celowniku liczby pojedynczej (*devojci*). Aby w pełni wykorzystać dane korpusowe, a wyniki wyszukiwania uwzględniały wszystkie możliwe formy i oboczności danego leksemu, wprowadzono specjalne operatory ułatwiające zadawanie pytań. Są to:

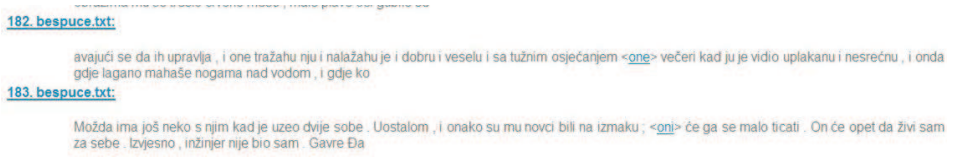
– kropka (.), która zastępuje dowolny znak graficzny. I tak na zapytanie:

[4] Zob. http://www.korpus.matf.bg.ac.rs/rec/SrpFSD.html, 17 IV 2013.

on.

otrzymamy wszystkie słowa trzyliterowe zaczynające się od ciągu znaków *on-*, a zatem *ona*, *one*, *oni*, *ono*, *onu* itd.

Rys. 8.3. Wyniki wyszukiwania z użyciem kropki.



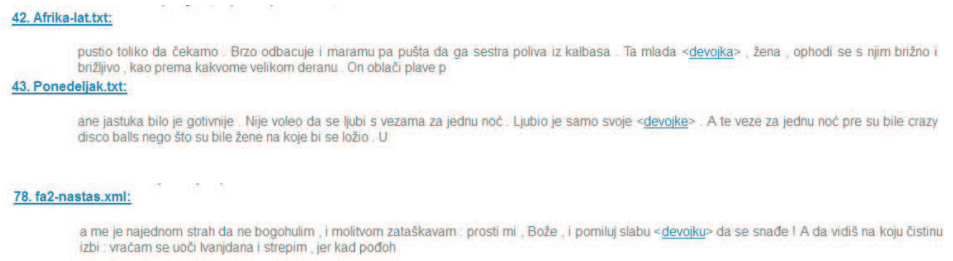
Źródło: http://www.korpus.matf.bg.ac.rs

– nawias kwadratowy ([]) – znaki wstawione w nawias kwadratowy tworzą tzw. klasę znaków, co ułatwia otrzymanie oczekiwanych wyników. Na przykład zapytanie:

devojk[aeu]

powoduje wyświetlenie wszystkich form leksemu *devojka*, kończących się na *-a*, *-e* i *-u*, tj. M, D, B l.p. oraz M i B l.mn.

Rys. 8.4. Wyniki wyszukiwania z użyciem nawiasu kwadratowego.



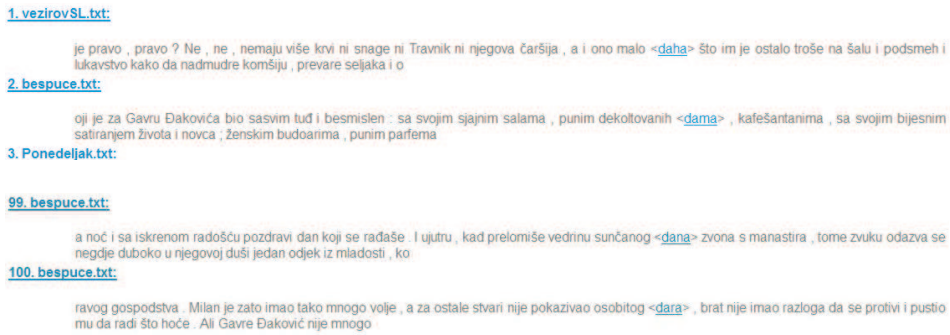
Źródło: http://www.korpus.matf.bg.ac.rs

– łącznik (-) – w połączeniu z nawiasem kwadratowym oraz odpowiednią sekwencją znaków pozwala zastępować dowolny znak z wyznaczonego w ten sposób zakresu. Na przykład zapytanie:

da[a-z]a

powoduje wyświetlenie wszystkich form wszystkich leksemów czteroliterowych, mających początek na *da-* zakończonych na *-a*, gdzie trzecim znakiem jest dowolna litera z zakresu od *a* do *z*, jak np. *dana* ‘dnia’, ‘dni’; *data* ‘dana’; *dara* ‘daru, podarunku’ itp.

Rys. 8.5. Wyniki wyszukiwania z użyciem nawiasu kwadratowego oraz łącznika.



Źródło: http://www.korpus.matf.bg.ac.rs

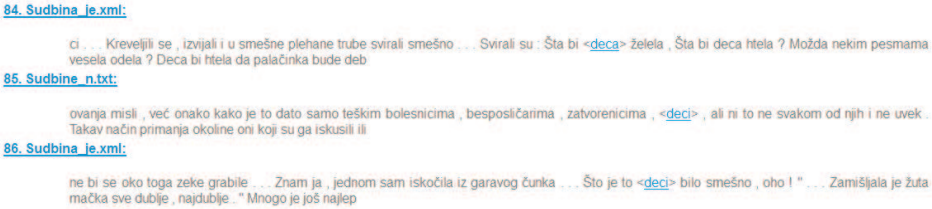
Jeszcze bardziej precyzyjnym narzędziem są operatory wyszukiwania. Są to:

– alternatywa (|), umożliwiająca wyszukanie kilku konkretnych ciągów znaków. Na przykład zapytanie o ciąg znaków:

deca|deci

zaowocuje otrzymaniem rezultatów zawierających jedynie te dwie formy rzeczownika zbiorowego *deca* 'dzieci'[5].

Rys. 8.6. Wyszukiwanie z użyciem alternatywy.



Źródło: http://www.korpus.matf.bg.ac.rs

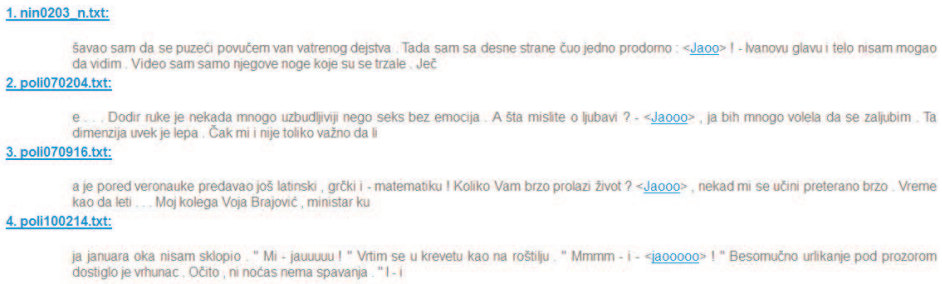
– znak plus (+) służy wyszukiwaniu ponowionych ciągów tego samego znaku pod warunkiem, że znak ten występuje co najmniej dwa razy. Zapytanie o wykrzyknik *jao* 'ojej' ze znakiem plus na końcu:

jaoo+

zwróci rezultaty zawierające dowolną liczbę liter *o* na końcu.

[5] Synkretyzm form wspomniany w części poświęconej problemom korzystania z korpusu powoduje, że system odnajdzie również leksem *deci* oznaczający '0,1 litra'. Problemu tego nie byłoby przy wyszukiwaniu formy ijekawskiej, która w przypadku rzeczownika zbiorowego 'dzieci' w D. brzmi *djeci*, zaś w przypadku miary pojemności brzmi tak jak ekawska (*deci*).

Rys. 8.7. Wyszukiwanie z użyciem znaku +.



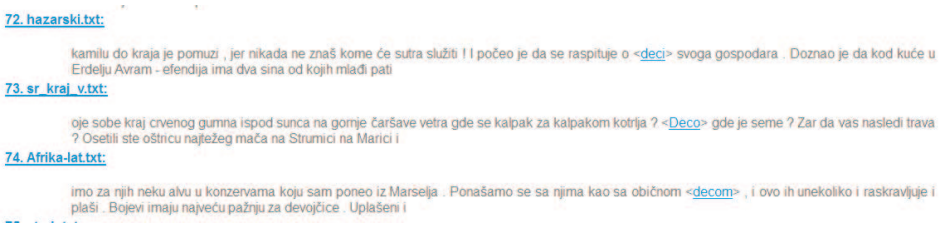
Źródło: http://www.korpus.matf.ac.rs

– nawias okrągły () to operator grupowania, pozwalający na szybkie wyszukanie wszystkich form danego leksemu. Na przykład wyszukiwanie ciągu znaków:

dec(a|e|i|o|u|om)

pozwoli na uzyskanie wszystkich form przypadkowych rzeczownika *deca* ‘dzieci’.

Rys. 8.8. Wyszukiwanie z użyciem nawiasu okrągłego.



Źródło: http://www.korpus.matf.ac.rs

Używanie kombinacji tych operatorów i znaków graficznych pozwala użytkownikom na zadawanie bardziej skomplikowanych zapytań, pozwalających na przykład wyszukać różne postaci wspomnianego wyżej leksemu *devojka* ‘dziewczyna’. Aby wyszukać wszystkie możliwe formy, należy podać wyczerpujące wszystkie kombinacje zapytanie. Będzie miało ono następujący kształt:

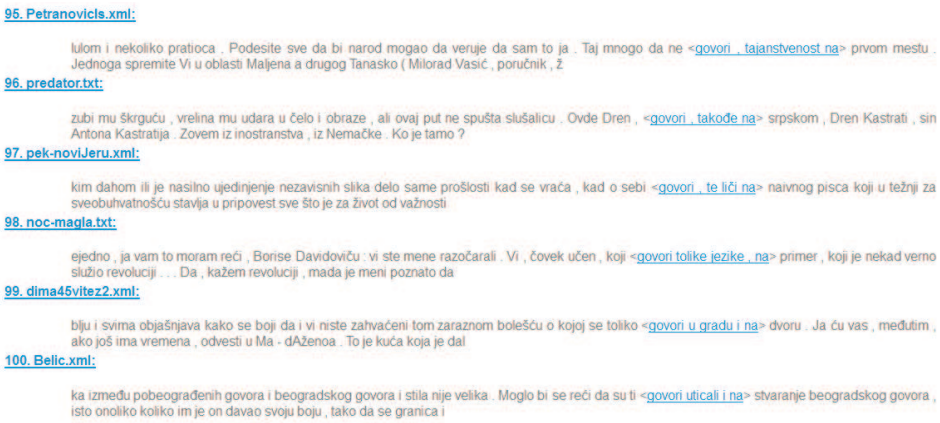
d(j|x)?(e|i)voja?(k|c)[a-z]+

Za pomocą KSSJ można wyszukiwać również większe fragmenty tekstu. Służy do tego operator w postaci podwójnego nawiasu kwadratowego [] oraz nawiasów klamrowych. Nawiasy kwadratowe służą do zaznaczania połączenia dwóch leksemów, a liczby w nawiasach klamrowych definiują najmniejszą i największą odległość między obydwoma leksemami. I tak na przykład zapytanie:

govori []{0,3} na

zwróci rezultaty zawierające ciągi wyrazów, zaczynające się od wyrazu *govori* 'mówi', a kończące na wyrazie *na*, przy czym oba wyrazy mogą występować najdalej trzy leksemy od siebie.

Rys. 8.9. Wyszukiwanie grup wyrazów w KSSJ.



Źródło: http://www.korpus.matf.bg.rs

Konkordancje, kolokacje, narzędzia statystyczne

Największą wadą KSSJ jest niemal zupełny brak narzędzi służących do analizy konkordancji. Korpus pozwala tylko na sporządzenie spisu konkordancji. Ze spisem uzyskanych konkordancji można pracować jedynie ręcznie, samodzielnie ekscerpując interesujące przypadki, co zdecydowanie spowalnia pracę z korpusem, jeśli nie czyni jej niemożliwą w przypadku bardziej popularnych leksemów. Nie istnieje również żadne oprogramowanie pozwalające na statystyczną analizę otrzymanych wyników.

Metodologia badań korpusowych z użyciem korpusów języka serbskiego

W przypadku pierwszego z omawianych korpusów, KSJ, trudno jest określić, jakie badania można przeprowadzać na tym materiale. Przypomina on bowiem bardziej zbiór zdigitalizowanych i przygotowanych do dalszej elektronicznej obróbki fiszek, które bez obudowania ich elektronicznymi narzędziami, w tym choćby wyszukiwarką lematów, przypominają raczej słownik elektroniczny niż współczesny korpus językowy.

Nieco większe możliwości – choć, jak wspomniano już powyżej, niewielkie – daje KSSJ. Można z jego wykorzystaniem prowadzić badania zarówno używając podejścia corpus-based, jak i corpus-driven. Niemniej jednak będą one

znacznie ograniczone z powodu braku narzędzi, które mogłyby ułatwić analizę wyszukanego materiału czy ocenę jego istotności z punktu widzenia statystyki.

Serbska lingwistyka korpusowa w porównaniu z badaniami nad korpusami innych języków słowiańskich jest jeszcze w stadium wczesnego rozwoju i obecnie powstaje niewiele opracowań poświęconych korpusom samym w sobie. Są to w przeważającej większości niewielkie fragmenty większych prac naukowych, zbiorów pokonferencyjnych czy artykuły naukowe, skupiające się na wybranych problemach związanych z tworzeniem korpusów, które chyba jednak wciąż jeszcze możemy określić jako *in statu nascendi*.

Jedyną pozycją książkową zawierającą informacje na temat tworzenia i stanu korpusów języka serbskiego jest opracowanie *Processing of Serbian. Automata, Texts and Electronic Dictionaries* autorstwa Cvetany Krstev (Krstev 2008). Pozycja ta koncentruje się raczej na słownikach elektronicznych, zawiera jednak również rozdziały poświęcone korpusom i niektórym zagadnieniom związanym z ich budową, przy czym wartymi zauważenia są zwłaszcza rozdziały i fragmenty poświęcone tworzeniu korpusów równoległych z użyciem KSSJ. Tworzony jest właśnie Równoległy Korpus Języków Francuskiego i Serbskiego, niestety, jest on jeszcze niedostępny w wersji *online*.

Podsumowanie

Korpusy języka serbskiego są dzisiaj tworami pośrednimi pomiędzy słownikami elektronicznymi a korpusami językowymi i z uwagi na to, że cały czas nie mają wystarczającej objętości oraz, co ważniejsze, wciąż tworzony jest aparat mający na celu ułatwienie pracy ze zgromadzonymi zasobami, należy stwierdzić, iż moment, kiedy będzie można wykorzystywać je w pełni w pracy naukowej, jest jeszcze dość odległy. Należy jednak mieć nadzieję, że w przyszłości oba omawiane zbiory staną się prawdziwym przedmiotem, a także narzędziem badań korpusowych nad współczesnym i historycznym językiem serbskim.

Bibliografia

Krstev C. (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Belgrade.

Krstev C., Vitas D. (2012). *Processing of corpora of Serbian using electronic dictionaries*. „Prace Filologiczne" LXIII, s. 279-292.

9. Praktyczny przewodnik po korpusie języka słoweńskiego

Adam Sewastianowicz
(badacz niezależny)

Nova Beseda

Pierwszym w założeniu referencyjnym korpusem języka słoweńskiego był korpus Nova Beseda (na wcześniejszych etapach rozwoju projektu pod nazwą CORTES), opracowany przez zespół Instytutu Języka Słoweńskiego im. Frana Ramovša Słoweńskiej Akademii Nauk i Sztuk, udostępniony bezpłatnie za pośrednictwem strony internetowej w roku 1999 (obecnie pod adresem http://bos.zrc-sazu.si/s_beseda.html). Korpus ten początkowo zawierał wyłącznie teksty literackie opublikowane w latach 1854-1996 (zbiór nadal dostępny jako korpus (Stara) Beseda pod adresem http://bos.zrc-sa-zu.si/main_si_l2.html) i był w kolejnych latach uzupełniany o teksty współczesne reprezentujące także inne gatunki. Pomimo znacznego wzrostu objętości (od 3 milionów słów w 1999 r. do 318 milionów obecnie) Nova Beseda pozostaje daleka od ideału korpusu zrównoważonego i reprezentatywnego; dość wspomnieć, że aż 67% wszystkich tekstów pochodzi z jednego źródła – dziennika *Delo*. Korpus ten pozbawiony jest także lematyzacji i jakiejkolwiek anotacji lingwistycznej. Te niewątpliwe jego słabości wynikają zapewne z faktu, że Nova Beseda w zamyśle miała być poddana dalszemu opracowaniu i docelowo stać się jednym z filarów przyszłego zintegrowanego narodowego korpusu języka słoweńskiego, do którego powstania jednak nigdy nie doszło (por. http://bos.zrc-sazu.si/a_about_si.html). Wszystko to stanowi o ograniczonej użyteczności tego korpusu zarówno pod względem referencyjności, jak i funkcjonalności.

FidaPLUS

Największym i oferującym najszerszy zestaw funkcji korpusem języka słoweńskiego jest korpus FidaPLUS. Prace nad nim ukończono w roku 2006, od tego czasu dostępny jest pod adresem www.fidaplus.net (dostęp bezpłatny w celach badawczych i dydaktycznych, wymagana jest rejestracja użytkownika). Korpus

ten jest owocem współpracy Wydziału Humanistycznego Uniwersytetu w Lublanie jako instytucji koordynującej projekt, Wydziału Nauk Społecznych tego uniwersytetu oraz Działu Technologii Wiedzy instytutu naukowo-badawczego Jožef Stefan. Pracami zespołu korpusowego kierował dr Marko Stabej. Projekt budowy korpusu FidaPLUS stanowił kontynuację i rozszerzenie realizowanego w latach 1997-2000 projektu, w ramach którego opracowano znacznie mniejszy korpus FIDA, dziś już niedostępny w internecie.

Struktura tekstowa

FidaPLUS jest jednojęzycznym korpusem tekstów pisanych, zawiera 621 milionów słów i z założenia ma mieć charakter referencyjny i synchroniczny, tj. stanowić kompletne odwzorowanie języka, aktualne w przybliżeniu w czasie jego powstawania (Arhar, Gorjanc 2007: 98; o korpusie FIDA także Erjavec i in. 1998: 1). Jeśli idzie o dobór tekstów wchodzących w jego skład, dają się zauważyć istotne różnice w stosunku do korpusu Nova Beseda, przekładające się na większy potencjał referencyjny korpusu FidaPLUS. Przytłaczająca większość tekstów (97,79%) pochodzi z lat 1991-2006, z czego 73% z lat 2000-2006, postulat synchroniczności jest więc niewątpliwie spełniony.

Struktura tekstowa korpusu została odzwierciedlona w klasyfikacji tekstów na podstawie czterech przecinających się kryteriów, a wyszukiwarka korpusu umożliwia przeszukiwanie podkorpusów wydzielonych na podstawie dowolnej ich kombinacji.

Pierwsze kryterium stanowi czas ukazania się tekstu (wybrany rok lub dowolnie zdefiniowany okres kilku lat). Dla korpusu, którego niemal 3/4 tekstów pochodzi z okresu zaledwie sześciu lat, kryterium to wydaje się nie mieć kluczowego znaczenia.

Drugim kryterium klasyfikacji jest „typ" tekstu (odpowiadający w ogólnym zarysie „kanałowi" w NKJP). Korpus FidaPLUS jest tu nieco bardziej zrównoważony i zdecydowanie bardziej reprezentatywny niż Nova Beseda. Chociaż nadal dominują w nim teksty gazetowe, a ich całkowity udział jest tylko nieznacznie niższy (65,26%), to większa jest liczba uwzględnionych tytułów. Ze względu na niezwykle bogate i głębokie zróżnicowanie regionalne języka słoweńskiego (nawet jeśli nie dotyczy ono bezpośrednio odmiany literackiej języka) pozytywnie należy ocenić włączenie do korpusu tekstów pochodzących z gazet redagowanych w różnych regionach Słowenii. Na stronie korpusu brak informacji na temat proporcji ilościowych pomiędzy poszczególnymi tytułami, jednak analiza wyników dla kilku próbnych zapytań o formy wyrazowe o wysokiej frekwencji pozwala wysnuć wniosek, że wśród trzech największych źródeł tekstów obok stołecznych dzienników *Delo* i *Dnevnik* znalazł się wydawany w Mariborze (Styria) *Večer*, zaś wyraźnie zauważalny udział w wynikach mają

także *Gorenjski glas* (Kranj, Górna Kraina), *Dolenjski list* (Novo mesto, Dolna Kraina) i *Primorske novice* (Koper, Przymorze). Lista wszystkich źródeł gazetowych oprócz innych tytułów ukazujących się w wymienionych regionach obejmuje także m.in. *Vestnik* (Murska Sobota, Prekmurje) i *Notranjske novice* (Logatec, Kraina Wewnętrzna). Takie proporcje wydają się w przybliżeniu odpowiadać potencjałowi ludnościowemu poszczególnych regionów.

Kolejnymi źródłami tekstów były czasopisma (23,26%), wydawnictwa książkowe (8,74%) oraz – po raz pierwszy w słoweńskich korpusach – materiały opublikowane w internecie (1,24%), a także „inne" (1,5%).

Stosunkowo niski odsetek tekstów książkowych jest podyktowany dwoma czynnikami. Po pierwsze przyjęcie perspektywy ściśle synchronicznej mogło spowodować, że w dwumilionowej Słowenii w założonych wąskich ramach czasowych nie ukazało się wystarczająco dużo książek, aby dało się z nich pozyskać do korpusu więcej tekstów. Po drugie pierwotnie zaplanowano korpus 300-milionowy, w którym ta sama liczba tekstów książkowych miałaby udział znacznie bliższy ideałowi korpusu zrównoważonego i reprezentatywnego (ponad 17%), jednak już na zaawansowanym etapie prac zespół korpusowy pozyskał dodatkowe finansowanie umożliwiające ponad dwukrotne zwiększenie objętości opracowanych tekstów, która jest także jednym z kryteriów reprezentatywności korpusu (Arhar, Grojanc 2007: 97).

Uwagę zwraca także niewielka liczba tekstów internetowych (pomimo ich dostępności) i znikomy udział tekstów mówionych (Nova Beseda zawiera 42-milionowy podkorpus stenogramów z posiedzeń Zgromadzenia Państwowego). Niemal całkowity brak tekstów mówionych usprawiedliwiony jest faktem, że już na etapie prac nad korpusem FIDA w samym kierownictwie zespołu korpusowego powstała idea stworzenia osobnego referencyjnego korpusu mówionego języka słoweńskiego (Stabej 1998: 100; Stabej, Vitez 2000: 79).

Teksty korpusu zostały też sklasyfikowane pod względem „gatunku" (odpowiednik „typu" w NKJP) zgodnie z poniższym zestawieniem (w drugiej kolumnie podano procentowy udział odpowiednio tekstów określonego gatunku w całym korpusie lub danego podgatunku w zbiorze tekstów reprezentujących gatunek nadrzędny):

Literatura piękna		3,47%
w tym:	teksty poetyckie	1,70%
	teksty prozatorskie	93,55%
	teksty dramatyczne	2,23%
	niesklasyfikowane	2,25%

[ciąg dalszy tabeli ze strony poprzedniej]

Teksty nieliterackie		96,41%
w tym:	specjalistyczne	10,36%
	humanistyczne i społeczne	b.d.
	przyrodnicze i techniczne	b.d.
	niespecjalistyczne	89,55%
	niesklasyfikowane	0,08%
Teksty niesklasyfikowane		0,11%

Jest to podział stosunkowo mało szczegółowy, lecz główną jego wadą jest brak informacji o kryteriach przyporządkowania tekstów do poszczególnych gatunków i podgatunków (zwłaszcza o wyznacznikach „specjalistyczności") oraz o sposobach rozstrzygnięcia kwestii ważnych przypadków granicznych, takich jak np. literatura faktu, co może w pewnych wypadkach utrudniać uzyskanie żądanych rezultatów. Przegląd wyników wyszukiwania w zbiorze tekstów zakwalifikowanych jako specjalistyczne dla zapytania o przykładową formę wyrazową *omogoča* 'umożliwia' wykazał, że specjalistyczność jest tu rozumiana szeroko, ponieważ większość poświadczeń pochodziła z wydawnictw prasowych adresowanych do szerokiego kręgu czytelników, a o ich specjalistyczności świadczył przede wszystkim deklarowany profil tematyczny wydawnictwa inny niż ogólny społeczno-polityczny. Wewnętrzny podział tekstów specjalistycznych na humanistyczne i społeczne z jednej strony oraz przyrodnicze i techniczne z drugiej wydaje się natomiast wystarczająco jasny i niewymagający dookreślenia.

Ostatnią płaszczyzną klasyfikacji tekstów, również umożliwiającą zawężenie wyszukiwania, jest podział na teksty poddane i niepoddane korekcie. Rozróżnienie to jest dosyć nietypowe, jednak w wysoce specyficznej słoweńskiej sytuacji językowej okazało się na tyle istotne, aby zastosować je w korpusie (Arhar, Gorjanc 2007: 99). Należy bowiem uwzględnić, że podstawowe zróżnicowanie współczesnych odmian języka słoweńskiego nie sprowadza się do wyodrębnienia w ramach jednego systemu językowego zestawu odmian funkcjonalno-stylistycznych, lecz w znacznej mierze opiera się także na kryteriach systemowych, przy czym różnice fonetyczne, morfologiczne, składniowe i leksykalne pomiędzy wyróżnianymi odmianami lub ich grupami są na tyle głębokie, że należałoby mówić o osobnych pokrewnych systemach (Toporišič 2000: 13), których przynależność do jednego języka narodowego, choć nie podlega dyskusji, bazuje tyleż na pokrewieństwie, co na szeregu czynników pozajęzykowych. Liczba wyodrębnianych odmian (literacka, postulowana literacko-potoczna, ogólna potoczna, regionalne potoczne i dialektalne) sięga przy tym kilkudziesięciu. W takiej sytuacji zarówno niezwykle trudne w realizacji, jak i niecelowe byłoby tworzenie korpusu uwzględniającego wszystkie lub nawet tylko niektóre współcześnie funkcjonujące odmiany, bez względu na ich żywotność i niekiedy szeroki geograficzny i funkcjonalny zakres użycia. Naturalne jest

więc, że korpus FidaPLUS, choć jego nazwa brzmi *Korpus slovenskega jezika*, bez dodatkowych kwalifikatorów, jest w istocie korpusem słoweńskiego języka literackiego jako odmiany dominującej w piśmiennictwie, jedynej będącej przedmiotem nauczania w szkołach (choć już nie zawsze medium nauczania), której przypisuje się funkcje konstytuowania i reprezentowania narodu. Ponieważ zaś typowym aktywnym użytkownikiem tej odmiany (twórcą jej tekstów) nie jest statystyczny Słoweniec ani nawet osoba ogólnie wykształcona, lecz osoba zawodowo zajmująca się tworzeniem, tłumaczeniem lub weryfikacją tekstów przeznaczonych do publikacji lub emisji za pośrednictwem mediów elektronicznych (Toporišič 2000: 15), to zasadne było wprowadzenie rozróżnienia na teksty powstałe w typowy dla tej odmiany „zawodowy" sposób i pozostałe teksty. Sytuacja ta tłumaczy także niewielki udział w korpusie tekstów internetowych, jako częściej przekraczających granice odmiany literackiej lub reprezentujących inne odmiany. Słoweński język literacki (rozumiany jako idiom określony przez zestaw cech systemowych, nie zaś tylko jako styl lub rejestr) jest przy tym językiem prymarnie i przede wszystkim pisanym, kształt i samo istnienie jego formy mówionej (w sytuacjach innych niż odczytywanie lub recytowanie z pamięci tekstu przygotowanego w postaci pisemnej) pozostaje zagadnieniem dyskutowanym (por. np. Srebot Rejec 1998), co z kolei wyjaśnia wyłączenie tekstów mówionych, w których wyraża się całe bogactwo słoweńskich odmian językowych innych niż literacka, do osobnego korpusu.

Anotacja zewnętrzna

W wynikach wyszukiwania każde poświadczenie poprzedzone jest dwoma odsyłaczami. Informacje o tekście, z którego dany kontekst pochodzi, dostępne są po kliknięciu pierwszego z nich. Pewne informacje zakodowane są już w samym odsyłaczu za pomocą kolorów: niebieski oznacza gazetę, zielony – czasopismo, różowy – książkę, zaś pomarańczowy – źródło internetowe. Dodatkowo dla pism periodycznych odsyłacz zawiera tytuł lub jego skrót. Pełna anotacja zewnętrzna, dostępna po kliknięciu odsyłacza, obejmuje kolejno: wartości przypisane im zgodnie z klasyfikacją według czterech opisanych kryteriów (data, typ/kanał, gatunek, korekta: tak/nie), informacje o długości tekstu (liczba akapitów, zdań, słów i form wyrazowych) oraz – dla tekstów publikowanych – informacje przeniesione z centralnego internetowego słoweńskiego systemu bibliograficznego COBISS (tytuł, podtytuł, autor/autorzy, tłumacz, wydawca, cechy fizyczne, struktura treści, grupa docelowa, autor układu graficznego/fotografii, seria wydawnicza, dla pism periodycznych także początek ukazywania się i wcześniej stosowane tytuły, oraz szereg innych). O ile informacje dotyczące klasyfikacji i długości tekstu są przedstawione w jasny sposób, o tyle wyczerpujące informacje pozyskane z systemu COBISS są bardzo nieprzejrzyste. Nadmiar tych informacji, mała czcionka i błędy w wyświetlaniu liter ze znakami diakrytycz-

nymi sprawiają, że odszukanie informacji istotnych w pracy z korpusem jest znacznie utrudnione.

Anotacja wewnętrzna

Segmentacja

Materiały opublikowane przez zespół korpusowy nie poruszają zagadnienia segmentacji, pewne informacje na ten temat można było jednak uzyskać za pomocą próbnych zapytań.

Formy złożone pisane z łącznikiem (w ich zakres nie wchodzą dwuczłonowe nazwiska w języku słoweńskim pisane rozłącznie bez łącznika) traktowane są jako ciągi osobnych segmentów, np. *družbeno*|-|*politični* 'społeczno-polityczny'. To samo dotyczy skrótowców z końcówką fleksyjną oddzieloną łącznikiem: *DVD*|-|*jev* 'płyt DVD', *WC*|-|*jema* '(dwiema) toaletami' *(na) MTV*|-|*ju* 'na MTV', *BMW*|-|*ji* 'BMW (M l.mn.)' oraz słów obcych zawierających apostrof: *d*|'|*Artagnan*, *rock*|'|*n*|'|*roll*. Końcówki sygnalizujące odmianę skrótowców zostały przy tym błędnie przypisane odpowiednio: końcówka *-jev* do lematu JEV (przym.), rzadziej JEV (rzecz.), pomimo że słowa takie w języku słoweńskim nie istnieją, końcówki *-jem* i *-jema* do JESTI 'jeść', *-ju* do ON, zaś *-ji* do ONA.

Formy poprzyimkowe zaimków typu *-nj*, *-me*, *-se* tworzą jeden segment wraz z poprzedzającymi je przyimkami: *skozenj* 'przezeń', *vame* 'we mnie', *predse* 'przed siebie'. Segmenty te przypisano do lematów odpowiednich zaimków (ON, JAZ, SE), pomijając ich powiązanie z przyimkami.

Jako jeden segment potraktowano także te zaprzeczone formy czasowników, które pisane są łącznie z partykułą przeczącą i w których zaszły zmiany fonetyczne utrudniające – zwłaszcza w przypadku formy *ni* – wyodrębnienie partykuły i czasownika: *nisem* (*ne* + *sem*) 'nie jestem', *ni* (*ne* + *je*) 'nie jest' *nočeta* (*ne* + *hočeta*) 'nie chcecie (l.podw.)' *nimajo* (*ne* + *imajo*) 'nie mają'. Takie samo rozwiązanie przyjęto dla pisanej łącznie z partykułą przeczącą partykuły *bi* 'by': *nebi* w *nebi brali* 'nie czytaliby' stanowi jeden segment.

Partykuła *bi* 'by' oraz odpowiadające polskim ruchomym końcówkom czasowników formy czasownika *biti* 'być' występujące w funkcji słowa posiłkowego czasu przeszłego (*sem*, *si*, *je*; *sva*, *sta*, *sta*; *smo*, *ste*, *so*) pisane są zawsze rozdzielnie od czasownika głównego, w związku z czym są także oznaczone jako osobne segmenty.

Skróty zawierające kropki zostały zinterpretowane jako ciąg segmentów, np. *d*|.|*o*|.|*o*|. 'sp. z o.o.'.

Zaimki nieokreślone z cząstką *(-)koli* '-kolwiek', dla których norma ortograficzna dopuszcza pisownię zarówno łączną, jak i rozdzielną, np.: *kdorkoli* i *kdor koli* 'ktokolwiek', *kakršnegakoli* i *kakršnega koli* 'jakiegokolwiek', zostały podzielone na segmenty zgodnie z zapisem: formy pisane łącznie stanowią jeden segment, zaś pisane osobno – dwa segmenty. Odmienna segmentacja dwóch

wariantów zapisu wyrażeń funkcjonalnie całkowicie równoważnych prowadzi do różnej ich interpretacji w zakresie części mowy. Formy typu *kdorkoli* oznaczone są jako zaimek nieokreślony, zaś typu *kdor koli* – jako zaimek względny + przysłówek.

Opisane przypadki wskazują, że zastosowano segmentację ściśle powiązaną z zapisem, tzn. segmentem jest zawsze ciąg znaków innych niż odstępy i znaki interpunkcyjne, niezależnie od konsekwencji, jakie dla prawidłowej anotacji może mieć zastosowanie w segmentacji wyłącznie kryterium formalnego.

Tagset

Przyjęty podział na części mowy obejmuje następujące klasy słów (w kolejności stosowanej w materiałach pomocowych korpusu): rzeczowniki, czasowniki, przymiotniki, przysłówki, przyimki, spójniki, partykuły, wykrzyknienia, liczebniki, skrótowce i zaimki.

Niektóre z tych klas są heterogeniczne pod względem zestawu kategorii gramatycznych przysługujących słowom wchodzącym w ich skład. Dla przykładu, klasa zaimków obejmuje zarówno odmienne przez przypadek i liczbę zaimki rzeczowne, jak i niepodlegające odmianie zaimki przysłowne. W zestawie kategorii przypisanych poszczególnym klasom uwzględniono nie tylko kategorie gramatyczne, ale również określenia podklas (np. podział klasy czasowników na pełnoznaczne, modalne i pomocnicze), kategorie semantyczne (nazwy pospolite i własne w ramach klasy rzeczowników), a nawet określenia sposobu zapisu (liczebniki zapisane słownie, cyframi arabskimi i cyframi rzymskimi). W ramach jednej klasy może być przeprowadzonych kilka niezależnych, przecinających się podziałów na podklasy. Przysługujące danej klasie słów kategorie nie są przy tym zhierarchizowane – np. kategoria przypadka przypisana jest do całej klasy zaimków na równi ze wszystkimi pozostałymi kategoriami, nie zaś tylko wybranym odmiennym przez przypadek podklasom zaimków.

Prowadzi to do występowania redundancji i pustych miejsc w składni zapytań. Na przykład w zapytaniu o zaimek przysłowny należy obowiązkowo wypełnić miejsca przewidziane na określenie kategorii osoby, rodzaju, liczby, przypadka, liczby obiektu posiadanego, rodzaju obiektu posiadanego i żywotności, które to kategorie zaimkom tym nie przysługują, w związku z czym miejsca te trzeba wypełnić odpowiednią liczbą znaków <?>, nadając tym kategoriom wartość nieokreśloną. Z kolei oznaczenie w zapytaniu podklasy zaimków osobowych nie zwalnia z konieczności określenia wartości dla kategorii „funkcja składniowa" (dostępne wartości: rzeczowne/przymiotne/przysłowne), choćby przez ustawienie wartości nieokreślonej, pomimo że osobowość zaimka implikuje jego rzeczowność. Taka konstrukcja tagsetu pozwala też na tworzenie zapytań poprawnych z punktu widzenia składni, lecz pozbawionych sensu, np.

o zaimek 2.os. l.poj. przysłowny nieżywotny; oczywiście zapytanie takie nie zwróci żadnego wyniku.

Z heterogenicznością wyróżnionych klas wiąże się także przyporządkowanie słów o takiej samej charakterystyce fleksyjnej do różnych klas. Na przykład słowa *nizozemski* 'holenderski', *tisti* 'tamten, ten' i *tretji* 'trzeci' zostały przypisane odpowiednio do klasy przymiotników, zaimków i liczebników, choć lista przysługujących im kategorii gramatycznych i form odmiany jest dla wszystkich trzech identyczna. Ponieważ jednocześnie wyszukiwarka korpusu wymaga określenia w zapytaniu o formę gramatyczną także części mowy, to zapytanie np. o D l.mn. wszystkich wyrazów o odmianie przymiotnikowej musi się składać z trzech ciągów.

Imiesłowy przymiotnikowe czynny i bierny zostały przypisane do klasy czasowników, której nie przysługuje kategoria przypadka, dlatego nie jest możliwe zapytanie o te imiesłowy w określonym przypadku gramatycznym.

Pełną listę klas słów i przysługujących im kategorii oraz ich symboli przedstawiono na stronie korpusu w zakładce *Pomoč*.

Lematyzacja

Formy hasłowe przypisane słowom włączonym do poszczególnych klas odpowiadają w ogólnym zarysie tradycyjnym formom słownikowym dla odpowiednich części mowy. Poniżej omówiono jedynie przypadki szczególne.

Formy M r.m. liczebnika głównego oznaczającego 'jeden': samodzielną *eden* i przyrzeczownikową *en* przypisano do dwóch osobnych lematów: EDEN i EN, przy czym wszystkie formy przypadków zależnych w rodzaju męskim oraz formy rodzaju żeńskiego i nijakiego przypisano do lematu EN. Każdy liczebnik porządkowy, mnożny i należący do podklasy „inne" (wieloraki, wielokrotny i itd.) posiada własny lemat, osobny od lematu liczebnika głównego. Odrębne formy hasłowe przypisano także liczebnikom zapisanym cyframi arabskimi i rzymskimi, np.:

Tab. 9.1. Lematy przykładowych liczebników zapisanych cyframi arabskimi i rzymskimi.

Segment	**Lemat**
štiri 'cztery (r.ż. i nij.)'	ŠTIRJE
četrtega 'czwartego'	ČETRTI
dvojnim 'podwójnym'	DVOJEN
trikratni 'trzykrotny (okr.)'	TRIKRATEN
4	4
VI	VI

Wszystkie zaimki osobowe przypisano do trzech lematów: JAZ 'ja', TI 'ty', ON 'on', np.:

Tab. 9.2. Lematy zaimków osobowych.

Segment	Lemat
mene 'mnie' *naju* 'nas (l.podw.)' *me* 'my (r.ż. l.mn.)'	JAZ
teboj 'tobą' *vidva* 'wy (r.m. l.podw.)' *nam* 'nam (l.mn.)'	TI
ono 'ono' *ju* 'ich (l.podw.)' *njimi* 'nimi (l.mn.)'	ON

Inaczej potraktowano zaimki dzierżawcze, tworząc osobne lematy nie tylko dla form każdej osoby, ale także dla każdej z trzech liczb osoby posiadającej, a w 3.os. l.poj. również dla form rodzajowych. W poniższej tabeli cyframi arabskimi oznaczono liczbę gramatyczną posiadacza, zaś rzymskimi – obiektu posiadania (1/I – l.poj., 2/II – l.podw., 3/III – l.mn.); wszystkie podane przykłady mają formę narzędnika:

Tab. 9.3. Lematy zaimków dzierżawczych.

Segment	Lemat
mojo 'moją' *mojima* 'moimi (1, II)'	MOJ
tvojimi 'twoimi (1, III)'	TVOJ
njegovo 'jego (1, I r.ż.)'	NJEGOV
njenima 'jej (1, II)'	NJEN
vajinimi 'waszymi (2, III)'	VAJIN
njihovim 'ich (3, I r.m.)'	NJIHOV

Lematami skrótów niezawierających kropek lub tylko zakończonych kropką są one same (bez ew. końcowej kropki), nie zaś ich rozwinięcia. Skróty pisane z wewnętrznymi kropkami są – jak wspomniano w sekcji „Segmentacja" – interpretowane jako ciąg segmentów, z których każdemu przypisano osobny lemat, np.:

Tab. 9.4. Lematy skrótów i skrótowców.

Segment	Lemat(y)
NLB (Nova Ljubljanska banka)	NLB
mag. (magister) 'magister'	MAG
t.i. (tako imenovani) 'tzw.'	T, I

Lematyzacja korpusu była całkowicie automatyczna i opierała się na z góry zadanym słowniku form wyrazowych. Jeżeli program nie odnalazł w słowniku formy odpowiadającej dokładnie danemu segmentowi, analizował go pod względem możliwości wystąpienia w nim określonych błędów (literowych, ortograficznych i gramatycznych). Dzięki temu np. błędne (potoczne lub dialektalne) formy *stricom* zam. *stricem* 'wujem', *reku* zam. *rekel* 'powiedział' czy *gledama* zam. *gledava* 'patrzymy (l.podw.)' zostały przynajmniej w części wystąpień prawidłowo przypisane odpowiednio do lematów STRIC 'wuj', REČI 'powiedzieć' i GLEDATI 'patrzeć'. W przypadku, gdy procedura ta nie dała rezultatu, lematyzator próbował utworzyć lemat na podstawie interpretacji zakończenia segmentu. W ten sposób dla segmentu *Pomurci* prawidłowo stworzył i przypisał lemat POMUREC. W korpusie znalazły się jednak także błędnie skonstruowane lematy. Dla przykładu, segment *online* program uznał za przymiotnik i przypisał mu formę podstawową ONLIN. Jeżeli również próba stworzenia lematu kończyła się niepowodzeniem, segment pozostawał nieprzypisany do żadnej formy podstawowej. Dotyczyło to głównie skrótowców, ciągów nieliterowych i nazw obcych (Arhar, Gorjanc 2007: 102).

Duża trudność w automatycznej lematyzacji była związana z występowaniem w języku słoweńskim obok homonimów (np. *tkiva* 'tkanki' lub 'tkajmy (l.podw.)') szczególnie dużej liczby homografów. Wynika to z faktu, że standardowa grafia słoweńska nie tylko nie uwzględnia zjawisk prozodycznych – miejsca akcentu oraz długości (i ew. tonu) samogłosek akcentowanych, ale także do zapisu kilku (do trzech) fonemów stosuje niekiedy jeden znak, np. *e* dla |e|, |ẹ|, |ə|; *l* dla |l|, |ụ|[1]. W związku z tym wiele słów różnie wymawianych zapisywanych jest jednakowo, np. *pojem* [ˈpọ:jəm] 'pojęcie', [ˈpo:jem] 'śpiewam' i [poˈjẹ:m] 'zjadam'; *predal* [preˈda:l] 'szuflada' i [preˈda:u̯] 'przekazał', *roka* [ˈro:ka] 'ręka' i [ˈrọ:ka] 'terminu'. Z tego powodu większym problemem w korzystaniu z korpusu FidaPLUS niż nierozpoznane słowa są przypadki błędnego przypisania lematów. Jak obliczono, w całym korpusie segmentów nieprzypisanych są 2%, zaś błędnie przypisanych 13% (Erjavec, Krek 2008: 322). Ponieważ korpus zawiera wiele segmentów, których prawidłowa interpretacja nie nastręcza trudności, to odsetek błędów dla segmentów posiadających homografy jest więc zapewne znacznie wyższy. Na przykład w przypadku dwuznacznego segmentu *lahko*, który mógł być przypisany zarówno do lematu przymiotnikowego LAHEK 'lekki', jak i przysłówkowego LAHKO 'lekko' (przysłówek ten występuje także w konstrukcji modalnej: *lahko pridem* 'mogę przyjść'), odsetek błędnych przypisań lematu w próbie stu losowych poświadczeń dla zapytania o ten segment jako formę przymiotnikową wyniósł 68%.

Z tego względu wyszukiwarka korpusu oferuje trzy grupy „kanałów" wyszukiwania (o sposobach ich użycia będzie jeszcze mowa). Kanały 1 i 2 służą

[1] [ụ] jest wariantem pozycyjnym fonemu |v|.

do standardowego wyszukiwania – wyniki uzyskane przy ich użyciu są zgodne z lematyzacją segmentów. Zastosowanie kanału 5 i 6 powoduje natomiast uwzględnienie w wynikach wszystkich segmentów, które potencjalnie mogłyby być przypisane do lematu, o który sformułowano zapytanie – w tym wypadku homograficzne segmenty, które zostały (często błędnie) przypisane do innych lematów, zostaną także wzięte pod uwagę. Kanały 3 i 4 stanowią rozwiązanie pośrednie – oprócz segmentów przypisanych do danego lematu uwzględniają również te segmenty, co do których istnieje duże prawdopodobieństwo, że zostały błędnie przypisane do innych form słownikowych.

Anotacja morfosyntaktyczna

Cała anotacja korpusu – podobnie jak lematyzacja – została przeprowadzona w pełni automatycznie. W związku ze specyficzną konstrukcją tagsetu anotacja, w złożeniu morfosyntaktyczna, w pewnych wypadkach zawiera także określone informacje semantyczne lub na temat grafii. Bezpośredniej anotacji składniowej i anotacji znaczeniami słów brak.

Wadą korpusu FidaPLUS jest brak możliwości wyświetlenia znaczników przypisanych danemu segmentowi. O poprawności anotacji możemy więc wnioskować jedynie na podstawie próbnych zapytań o określoną formę gramatyczną. Do próby wybrano trzy równobrzmiące formy należące do lematu FANT 'chłopiec, chłopak': D l.poj., B l.poj. i M l.podw. (*fanta*). Otrzymane rezultaty były znacznie lepsze niż w przypadku próby poprawności lematyzacji. Na 25 losowych poświadczeń dla zapytań o każdą z trzech wymienionych form wyników zgodnych z zapytaniem otrzymano: 20 dla D l.poj., 21 dla B l.poj. i 23 dla M l.podw. Poniżej podano przykłady poprawnej anotacji:

D l.poj. *Sally se boji, da nikoli ne bo našla* ***fanta*** *svojih sanj.*
[HOPLA 0000662]
'Sally się boi, że nigdy nie znajdzie chłopaka swoich marzeń.'

B l.poj. *Policisti domnevajo, da gre za 15-letnega* ***fanta*** *iz Kameruna.*
[VECER 0000095]
'Policjanci przypuszczają, że chodzi o 15-letniego chłopaka z Kamerunu.'

M l.podw. ***Fanta*** *znata svoj posel.* [0002806 0000292]
'Chłopcy są dobrzy w swoim fachu (dosł. umieją swój fach).'

Opcje wyszukiwania i składnia zapytań

Wyszukiwarka korpusu oferuje dwie metody wyszukiwania: podstawową i rozszerzoną. Wyszukiwanie podstawowe umożliwia użycie pełnej składni zapytań

i wszystkich opcji programu z wyjątkiem zawężenia wyszukiwania do korpusu wydzielonego na podstawie kryteriów klasyfikacji tekstów, opisanych w ustępie *Struktura tekstowa*.

Zapytanie o formę wyrazową ma postać tej formy. Zapisując zapytanie małymi literami w wynikach otrzymujemy różne warianty graficzne danej formy, np. zapytanie *fant* daje m.in. następujące wyniki:

Dala sta mu denar in ***fant*** *je odšel do dobavitelja.* [DELO 0000073]
'Dali (l.podw.) mu pieniądze i chłopak poszedł do dostawcy.'

... iz vodnjaka, imenovanega ***Fant*** *z ribo.* [DNEVNIK 0000437]
'... z fontanny zwanej Chłopiec z rybą.'

FANT *RAD DELA.* [EVA 0000618]
'Chłopak lubi pracować.'

Użycie w zapytaniu co najmniej jednej dużej litery daje wyłącznie wyniki zgodne pod względem wielkości liter z zapytaniem.

W zapytaniu o ciąg wyrazów należy połączyć je podkreślnikiem, w przeciwnym wypadku (jeżeli pozostawimy między nimi odstęp) wyszukiwarka wyświetli poświadczenia osobno dla każdego z tych wyrazów w akapitach, w których w dowolnym miejscu występują także pozostałe wpisane wyrazy. Por. wyniki dla zapytań velika_težava i velika težava 'duży problem':

Ljubosumje je ***velika težava*** *v Afriki.* [MLADINA 0000160]
'Zazdrość jest dużym problemem w Afryce.'

Večna ***težava*** *Haloz je še vedno* ***velika*** *razdrobljenost poseljenosti.*
[VECER 0000867]
'Wiecznym problemem regionu Haloze jest wciąż duże rozproszenie osadnictwa.'

Zapytanie o wszystkie formy odmiany przypisane do danej formy hasłowej wymaga użycia nieparzystego (pierwszego, trzeciego lub piątego) kanału wyszukiwania (różnice między nimi wyjaśniono w ustępie *Lematyzacja*). W tym celu zapytanie należy rozpocząć od znaku <#>, następnie bez spacji podać numer kanału i szukany lemat, np. zapytanie #1zgodovina zwraca poświadczenia słowa *zgodovina* 'historia' we wszystkich liczbach i przypadkach, np.:

Tako je treba poznati ***zgodovino*** *fašističnega nasilja nad Slovenci...*
[MLADINA 0000277]
'Trzeba zatem znać historię faszystowskiej przemocy wobec Słoweńców...'

To je tisti del ***zgodovine****, ki ga Hrvatje objokujejo kot križev pot.*
[MLADINA 0001208]
'To jest ta część historii, którą Chorwaci opłakują jak drogę krzyżową.'

W zapytaniu o ciąg lematów należy użyć – podobnie jak w wypadku zapytania o grupę wyrazów – podkreślnika, np. <#1hiter_#1avto> 'szybki samochód':

*Vseeno je, ali svoj **hitri avto** parkirajo tik ob prehodu za pešce.*
[COSMOPOLITAN 0000577]
'Wszystko jedno, czy swój szybki (okr.) samochód parkują tuż przy przejściu dla pieszych.'

*... tisti hudič s **hitrejšim avtom** pa nas že ne bo prehitel.*
[NEDELJSKI 0000951]
'... a ten diabeł z szybszym samochodem już nas nie wyprzedzi.'

We wszystkich zapytaniach istnieje możliwość zastosowania dwóch symboli wieloznacznych: <*> zastępuje dowolną liczbę (także 0) dowolnych znaków, zaś <?> – dokładnie jeden dowolny znak. Przykładowe wyniki dla zapytań:

fant*

*V nasprotju z mano pa si ti videti **fantastična**.* [0022830 0003340]
'W przeciwieństwie do mnie ty wyglądasz fantastycznie.'

hitr*_avt*

*V teku za plenom doseže **hitrost avtomobila** na avtocesti.*
[KMECKI.GLAS 0000134]
'W pogoni za ofiarą osiąga prędkość samochodu na autostradzie.'

Do zapytań o części mowy (klasy słów) i formy gramatyczne służą parzyste kanały wyszukiwania (drugi, czwarty i szósty). Zapytanie np. o czasownik będzie się składało z symbolu <#>, numeru kanału, oznaczenia klasy czasowników <g> oraz symbolu wieloznacznego <*> wskazującego, że wszystkie przypisane w tagsecie klasie czasowników kategorie mają pozostać nieokreślone. Zapytanie #2g* daje m.in. następujące wyniki:

*V bližini Yogye **stojita** veličastna templja.* [NOVI.TEDNIK 0000002]
'W pobliżu Yogye stoją ogromne świątynie (l.podw.).'

*Od zunaj so se **zaslišali** zvonovi.* [ZIV.IN.TEH 0000005]
'Z zewnątrz dobiegł dźwięk dzwonów.'

*Medved je šel **iskat** svojega sinka.* [CICIDO 0000347]
'Niedźwiedź poszedł szukać (supinum) swojego synka.'

Możemy też wyszukiwać wyrazy reprezentujące daną część mowy w określonej formie gramatycznej. W tym celu należy w zapytaniu wpisać szereg symboli odpowiadających poszczególnym kategoriom przypisanym danej klasie słów. Jeżeli danej kategorii nie chcemy zaznaczyć lub jeśli w danym przypadku nie przysługuje ona wyszukiwanej formie, odpowiadający jej symbol zastępujemy znakiem <?>. Kilka kolejnych znaków <?> możemy zastąpić znakiem <*> wyłącznie na końcu ciągu tworzącego zapytanie, ponieważ wewnątrz zapytania liczba kategorii, którym nie przypisano wartości, ma znaczenie dla prawidłowej interpretacji przez program zapytania. Lista symboli części mowy i kategorii im przypisanych znajduje się w dziale *Pomoč* na stronie korpusu. Na przykład zapytanie o czasownik w 2.os. l.mn. tr.rozk. będzie miało postać #2g?v?dm*, wyszukiwarka korpusu zwróci m.in. następujące poświadczenia:

***Napišite**, s kom se je poročil Alejandro.* [HOPLA 0000434]
'Napiszcie, z kim ożenił się Alejandro.'

*Ne **spreglejte** ga!* [NEDELJSKI 0000268]
'Nie przegapcie go!'

Nie jest natomiast możliwe skonstruowanie zapytania o samą formę gramatyczną bez określania części mowy (np. o wszystkie formy wyrazowe rodzaju nijakiego). Można co prawda symbol części mowy zastąpić w zapytaniu symbolem <?>, ale metoda ta w większości wypadków nie daje pożądanych rezultatów, ponieważ na określenie tych samych kategorii w zapytaniach o różne części mowy przeznaczono różne miejsca w ciągu znaków, zaś te same symbole mogą mieć różne znaczenie w zależności od wyszukiwanej klasy słów i pozycji w zapytaniu. Na przykład w ciągu #2??s* symbol „s" na piątym miejscu (jako określenie drugiej w kolejności kategorii) w przypadku rzeczowników będzie oznaczał rodzaj nijaki, zaś w przypadku przymiotników – stopień najwyższy niezależnie od rodzaju. W efekcie dla tego zapytania uzyskujemy m.in. następujące wyniki:

*Kakršnokoli **pranje** denarja je povsem izključeno.* [MLADINA 0000052]
'Jakiekolwiek pranie pieniędzy jest całkowicie wykluczone.'

*En lot pa je **najmanjša** količina, ki jo je možno kupiti.* [MLADINA 0000163]
'Jeden łut jest najmniejszą ilością, którą można kupić'

Możliwe jest jednoczesne użycie w jednym zapytaniu kanału nieparzystego i parzystego. W tym celu do łączenia warunków stosujemy symbole <&> 'i' oraz <&~> 'nie'. Możemy więc sformułować np. zapytanie o poświadczenia lematu FANT w M l.mn.:

#1fant&#2s??mi*

W wynikach otrzymujemy oboczne postaci tej formy:

film ***Fantje*** *ne jočejo* [DOLENJ.LIST 0000114]
'film Chłopaki nie płaczą'

Vsi ***fanti*** *si zaslužijo pohvale.* [MAG 0008000]
'Wszyscy chłopcy zasługują na pochwałę'

Natomiast użycie symbolu <&∼> zwróciłoby poświadczenia tego lematu we wszystkich formach z wyjątkiem formy określonej w drugim warunku zapytania.

Wyszukiwanie kolokacji odbywa się w dwóch etapach. Najpierw należy stworzyć konkordancję dla ośrodka kolokacji, a następnie skorzystać z pierwszej funkcji dostępnej po kliknięciu symbolu dwóch kół zębatych na pasku menu wyszukiwarki (symbol ten pojawia się po zakończeniu lub zatrzymaniu przeszukiwania korpusu). Funkcja ta umożliwia określenie, czy kolokacje z różnymi formami przypisanymi do jednego lematu mają być analizowane osobno, czy też łącznie, oraz ustalenie liczby słów przed lub za ośrodkiem kolokacji branych pod uwagę. Po kliknięciu przycisku *Izvedi* 'Wykonaj' otrzymujemy listę kolokacji oraz szereg danych statystycznych; kolejno są to: frekwencja względna, frekwencja całkowita, informacja wzajemna (MI i MI3) oraz logarytm prawdopodobieństwa. Listę można sortować alfabetycznie lub według każdej z wymienionych wartości statystycznych. Nie jest natomiast możliwe analizowanie kolokacji ze słowami spełniającymi określone warunki, np. kolokacji rzeczownika *Slovenec* 'Słoweniec' z przymiotnikami w stopniu równym.

Kolejna opcja umożliwia sortowanie wyników według lewego lub prawego kontekstu. Trzecią z dostępnych opcji jest filtrowanie wyników. Pozwala ona na wyłączenie z konkordancji kontekstów niezawierających lub zawierających określone segmenty w ustalonej odległości od słowa, o które sformułowano zapytanie. Ostatnia opcja umożliwia losowe ograniczenie liczby wyświetlanych wyników.

Zastosowania

Jak dotąd korpus FidaPLUS (lub jego wcześniejszą wersję FIDA) wykorzystano w pracach nad dwoma słownikami dwujęzycznymi: *Veliki angleško-slovenski slovar* (VASS 2006) i *Mali hrvaško-slovenski in slovensko-hrvaški slovar* (MHS-SHS 2009), przy czym podkreślano duże znaczenie tego korpusu dla badań nad współczesną leksyką (http://slovarji.dzs.si/Slovar/?id=67). Na konferencji poświęconej idei stworzenia nowego obszernego słownika języka słoweńskiego podnoszono konieczność wykorzystania do tego celu korpusu jako jednego ze źródeł (Gorjanc 2009: 45), jednak projekt ten do dziś nie wszedł w fazę realizacji.

Korpus mówionego języka słoweńskiego

Na zakończenie warto wspomnieć o najnowszym słoweńskim korpusie *Korpus govorjene slovenščine* (http://www.korpus-gos.net/). Jest to milionowy korpus zróżnicowanych tekstów mówionych, lematyzowanych i anotowanych morfosyntaktycznie, udostępniający konkordancje nie tylko w formie tekstów transkrybowanych, ale także w postaci nagrań dźwiękowych. Tagset został przeniesiony bez zmian z korpusu FidaPLUS. Co ciekawe, głównym wykonawcą projektu i koordynatorem prac nad korpusem było prywatne przedsiębiorstwo Amebis z Kamnika. Zespołem, w którego skład oprócz pracowników tej firmy weszli także specjaliści ze Słoweńskiej Akademii Nauk i Sztuk, Uniwersytetu w Lublanie i Instytutu Jožef Stefan, kierował Miro Romih. Korpus ten jest szczególnie cenny, ponieważ daje bezpośredni wgląd w najmniej dotąd zbadany obszar języka słoweńskiego, jakim jest język potoczny.

Bibliografia

Arhar Š., Gorjanc V. (2007). *Korpus FidaPLUS. Nova generacija slovenskega referenčnega korpusa*. „Jezik in slovstvo", 2, s. 95-110.

Erjavec T., Gorjanc V., Stabej M. (1998). *Korpus Fida*, http://nl.ijs.si/isjt98/zbornik/sdjt98-Gorjanc.pdf.

Erjavec T., Krek S. (2008). *The JOS Morphosyntactically Tagged Corpus of Slovene*. W: Calzolari N. (red.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, s. 322-326.

Gorjanc V. (2009). *Jezikovnotehnološka podpora slovarskemu delu*. W: Predih A. (red.), *Strokovni posvet o novem slovarju slovenskega jezika*. Ljubljana, s. 45--46.

SrebotRejec T. (1998). *O slovenskih samoglasniških sestavih zadnjih 45 let*. „Slavistična revija", 4, s. 339-346.

Stabej M. (1998). *Besedilnovrstna sestava korpusa FIDA*. „Uporabno jezikoslovje", 6, s. 96-106.

Stabej M., Vitez P. (2000). *KGB (korpus govorjenih besedil) v slovenščini*. W: Erjavec T., Gros J. (red.), *Informacijska družba IS'2000: Jezikovne tehnologije*. Ljubljana, s. 79-81.

Toporišič J. (2000). *Slovenska slovnica*. Maribor.

10. Praktyczny przewodnik po korpusie języka bułgarskiego

Ignacy Doliński
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Informacje ogólne

Bułgarski Korpus Narodowy (BKN, Български национален корпус – БНК, Bulgarian National Corpus – BNC) został utworzony w Instytucie Języka Bułgarskiego (Институт за български език „Проф. Любомир Андрейчин“) przez kolektyw Sekcji Lingwistyki Komputerowej i Sekcji Bułgarskiej Leksykologii i Leksykografii (Секция по компютърна лингвистика, Секция за българска лексикология и лексикография).

Ogólna objętość BKN to ok. 5,4 mld słów, z czego „jądro” korpusu to 1,2 mld słów zawartych w ponad 240 tys. tekstów. Materiały włączone do Korpusu oddają stan języka bułgarskiego – głównie w postaci pisanej – w okresie od połowy XX w. (1945 r.) do dziś.

Obecnie pracami kolektywów kieruje prof. Svetla Koeva. Główna strona internetowa BNK: http://ibl.bas.bg/BGNC_bg.htm. Strona wyszukiwania: http://search.dcl.bas.bg/en/

Kontakt:

1113 София
бул. Шипченски проход № 52, бл. 17
Институт за български език
Български национален корпус
bulnc@dcl.bas.bg

lub:

52 Shipchenski Prohod
Blvd. Block 17
Sofia, 1113
tel/fax: + 359 2 72 23 02
e-mail: ibl@ibl.bas.bg

Dostęp do zasobów korpusu jest bezpłatny, można z nich korzystać przez internet bez rejestracji (wtedy odpowiedź na zapytania jest ograniczona do 30

losowo wybranych poświadczeń). Rejestracja wymaga wypełnienia formularza elektronicznego (dane typu login + password, instytucja, adres itp.), wysłania go pod adres dcl.bas.bg i potwierdzenia drogą mailową. Zarejestrowany użytkownik ma pełen dostęp do: systemu wyszukiwawczego, zapytań dotyczących kolokacji, pobierania wyciągów i subkorpusów z Korpusu, korzystania ze słowników frekwencyjnych tworzonych na bazie Korpusu.

Struktura korpusu

BKN łączy w sobie kilka korpusów elektronicznych, opracowanych w okresie 2001-2009 r.; jest regularnie powiększany o nowe teksty. Składa się z jednojęzycznej części bułgarskiej i 48 korpusów równoległych (tj. w języku bułgarskim i 47 językach obcych) różnej wielkości. Korpusy równoległe budowane są z tekstów posiadających odpowiednik po stronie bułgarskiej. Wielkość poszczególnych korpusów równoległych: od 50 tys. słów (bułgarsko-japoński) do 260 mln (bułgarsko-angielski); korpus bułgarsko-polski liczy 197,8 mln słów (8. pod względem wielkości), natomiast niezwykle ważny dla Bułgarów kulturowo korpus bułgarsko-rosyjski – stosunkowo łatwy do sporządzenia dzięki zbieżności alfabetów – zawiera ledwie ok. 3,3 mln słów (28. miejsce co do wielkości). Być może przyczyną tych dysproporcji jest pozyskiwanie tekstów paralelnych z dokumentacji Unii Europejskiej (20 największych korpusów opiera się na językach UE). Łączna objętość wszystkich korpusów równoległych to prawie prawie 4,2 mln słów zawartych w ok. 1,8 mln tekstów. Korpus bułgarsko-angielski składa się z tekstów: beletrystycznych (21,3%), administracyjnych (68,6%), naukowych (0,4%), potoczno-beletrystycznych (3,5%), publicystycznych (5,4%), popularnonaukowych (0,04%), popularnych (0,5%), nieokreślonych (0,1%).

Z BKN zintegrowane są także następujące korpusy, służące także do badań na próbkach i ćwiczeń (dostęp do nich daje ogólna rejestracja użytkownika):

– Korpus „Brown" Języka Bułgarskiego („Brown" корпус за български език, Structured "Brown" Corpus of Bulgarian, http://dcl.bas.bg/Corpus/home_bg.html) – 500 tekstów po min. 2000 słów w 15 kategoriach 2 typów – artystycznych i informacyjnych; łącznie ok. 1 mln słów tekstu powiązanego („свързан текст");

– Korpus „Brown" Języka Bułgarskiego Tekstów Ciągłych („Brown" корпус за български език с целите текстове, Brown Corpus of Bulgarian with full-length texts, BCB) – ok. 4,8 mln słów;

– Bułgarski Korpus Anotowany Części Mowy (Български POS анотиран корпус, БулПосКор, The Bulgarian Part-of-Speech Annotated Corpus, BulPosCor), http://dcl.bas.bg/poscor/bg/; pojedyncze zapytanie daje możliwość rozpoznania gramatycznego danej formy; 174,7 tys. jednostek leksykalnych z anotacją z Bułgarskiego Słownika Gramatycznego;

– Bułgarski Korpus Anotowany Semantycznie (Български семантично анотиран корпус, БулСемКор, The Bulgarian Sense-Annotated Corpus, BulSemCor), http://dcl.bas.bg/semcor/bg/ – 95,1 tys. jednostek leksykalnych z synonimami, pochodzenie: bułgarski wordnet;

– Bułgarsko-angielski Korpus Równoległy Zdań „Współzależnych" (Българско-английски паралелен корпус със съотнесени изречения и клаузи, The Bulgarian-English Sentence- and Clause-Aligned Corpus, BulEnAC, http://dcl.bas.bg/en/clauseAlignedCorpus_en.html – ok. 260,7 mln tokenów dla języka bułgarskiego i 263,1 mln dla angielskiego;

– Korpus Anotowany Złożonych Jednostek Językowych (Wiki1000+ – анотиран корпус със съставни лексикални единици, Wiki1000+ corpus with annotated MWEs), http://dcl.bas.bg/wikiCorpus.html – 6311 jednostek tekstowych i 13,4 mln słów – wpisów z bułgarskiej Wikipedii (Уикипедия, http://bg.wikipedia.org/) według 25 dyscyplin wiedzy; korpus można pobrać w postaci pliku .zip.

Niemal każda karta (zakładka) bułgarskiego opisu Korpusu ma swój pełny i aktualny (tj. prowadzony bez opóźnień) odpowiednik w języku angielskim.

Struktura tekstowa

Główny zrąb BKN składa się z tekstów w języku bułgarskim (1,2 mld słów w 240 tys. dokumentów). Teksty oryginalne stanowią 37,1% Korpusu, przekłady – 40,5%, co do 22,4% brak jednoznacznego opisu pod tym względem. 97,35% źródeł to teksty pisane, 2,65% – ustne (np. wykłady i wypowiedzi deputowanych w parlamencie). Od autorów i wydawców pochodzi ok. 2,5% materiału językowego. Pozostała część to efekt automatycznej i ręcznej ekscerpcji, dokonanej przez kolektyw BKN.

Proporcje stylistyczne pochodzenia tekstów:

beletrystyczne	41,8%
administracyjne	18,9%
naukowe	3,0%
potoczno-beletrystyczne	1,7%
publicystyczne	30,9%
popularnonaukowe	3,4%
popularne	0,26%
nieokreślone	0,04%

Metody pozyskiwania tekstów to głównie skanowanie tekstów drukowanych, automatyczna i ręczna ekscerpcja internetu oraz dary elektroniczne wydawców i autorów (por. wyżej).

Na temat zawartości korpusów równoległych (na przykładzie korpusu bułgarsko-angielskiego) – por. wyżej.

Wynika stąd, że dobrze zaprezentowane są teksty pisane, bogate leksykalnie o „klasycznym", kontrolowanym doborze słownictwa (poza źródłami internetowymi). Jakiekolwiek badania frekwencyjne z konieczności będą chyba mniej wiarygodne.

Jako podkorpusy prekompilowane i gotowe do pobrania do dyspozycji są (dla użytkowników zarejestrowanych):

– paralelny korpus oficjalnych tekstów administracyjnych UE – w 23 językach, największe to: angielski, niemiecki, rumuński, polski i grecki;

– paralelny korpus tekstów publicystycznych z SETimes.com, w 9 językach bałkańskich (bułgarskim, rumuńskim, macedońskim, serbskim, albańskim, greckim, tureckim, chorwackim, bośniackim) oraz angielskim;

– korpus tekstów bułgarskich z Wikipedii;

– paralelny korpus naukowo-administracyjny z dużą częścią medyczną – w 23 językach.

Chcąc wypreparować samemu podkorpus z BKN, należy ustalić jego zakres, styl, gatunek tekstu, ramy czasowe, język i rozmiar.

Przykładowe adresy gotowych korpusów (o wolnym dostępie):

– bułgarskie
  http://dcl.bas.bg/BulNC-registration/feeds/ADMIN-EMEA.BG.zip
  http://dcl.bas.bg/BulNC-registration/feeds/ADMIN-EUR.BG.zip
  http://dcl.bas.bg/BulNC-registration/feeds/JOURNALISM.BG.zip
  http://dcl.bas.bg/BulNC-registration/feeds/POPULAR_SCIENCE.BG.zip

– angielskie
  http://dcl.bas.bg/BulNC-registration/feeds/ADMIN-EMEA.EN.zip
  http://dcl.bas.bg/BulNC-registration/feeds/ADMIN-EUR.EN.zip
  http://dcl.bas.bg/BulNC-registration/feeds/JOURNALISM.EN.zip

– polskie
  http://dcl.bas.bg/BulNC-registration/feeds/ADMIN-EMEA.PL.zip
  http://dcl.bas.bg/BulNC-registration/feeds/ADMIN-EUR.PL.zip

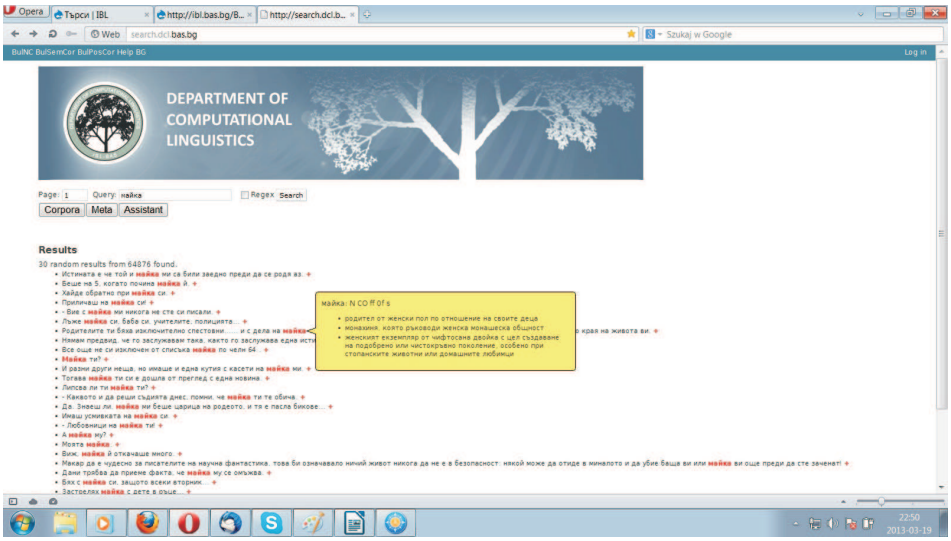
Dla przykładu, korpus http://dcl.bas.bg/BulNC-registration/feeds/POPULAR_SCIENCE.BG.zip zawiera po rozpakowaniu ok. 0,5 GB tekstu, składającego się z pojedynczych haseł (średnio po ok. 2 tys. znaków) pogrupowanych w foldery według dziedzin wiedzy (archeologii, biologii, chemii, fizyki, ekonomii, filozofii, geografii, historii, literatury, medycyny itd.), przypominających zawartością hasła Wikipedii. Służy to jednak do zestawień i badań „wybiórczych" w małych kontekstach. Brak w nich możliwości segregowania według 25 kryteriów („pól"), o których się wspomina w ogólnych instrukcjach posługiwania się Korpusem (filename, path, date_added_to_corpus, author_info, author, translator_info, year_of_creation, source_type itd.). Owe kryteria

można zastosować dopiero wtedy, kiedy się wyszukuje dane standardowym sposobem (czyli pod adresem http://search.dcl.bas.bg/). Tam właśnie można wybrać według własnych upodobań rodzaj tekstów (Corpora: F-InformalFiction, H-Popular, B-Science, C-Che-istry itd.) i na to nałożyć wcześniejsze kryteria. Widok tej operacji – por. rys. 10.3. i 10.4. oraz stosowny komentarz do nich.

Anotacja zewnętrzna

W odpowiedzi na proste zapytanie bez logowania (o słowo *майка* 'matka') otrzymujemy zawsze porcjami po 30 pełnych cytatów-zdań zawierających dane słowo (dla użytkowników zarejestrowanych wszystkie – porcjami, poczynając od zdań najdłuższych; użytkownicy niezarejestrowani otrzymują wynik dobrany losowo). Bardzo wiele słów składających się na cytat po najechaniu kursorem ujawnia anotację gramatyczną i semantyczną (komentarz typu słownika jednojęzycznego – zatem dla każdego przeciętnego użytkownika odpowiedź korpusu stanowi „żywy", choć nie pozbawiony błędów elektroniczny słownik języka bułgarskiego).

Rys. 10.1. Przykładowe zapytanie o leksem *майка* 'matka'.



Po kliknięciu czerwonego krzyżyka dołączonego na końcu cytatu otrzymujemy m.in. metadane:

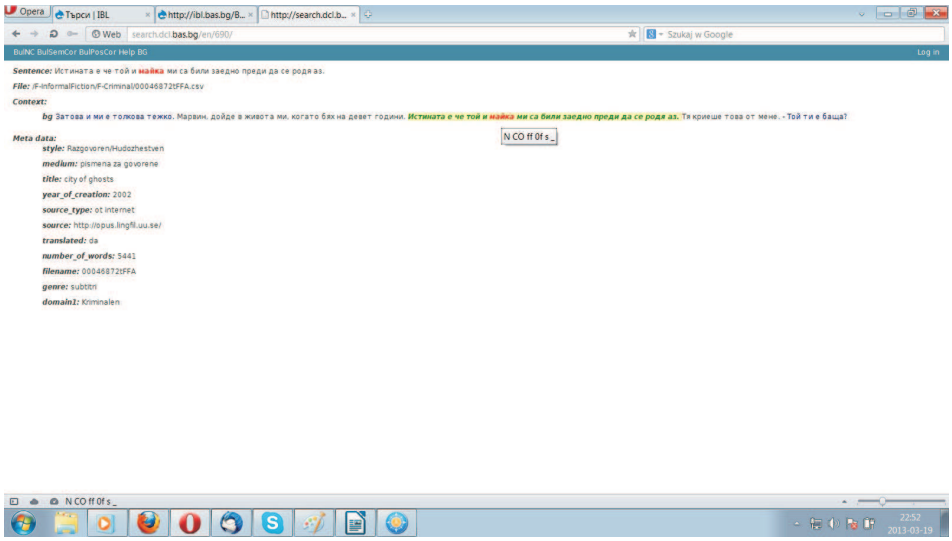
– powtórkę zdania-cytatu (Sentence: *Истината е че той и* ***майка*** *ми са били заедно преди да се родя аз.* ['Prawdą jest, że on i moja **matka** byli razem, zanim się urodziłem.']), przy czym każde słowo cytatu opatrzone jest anotacją gramatyczną, np. *майка* N CO ff 0f s _;

– nazwę dokumentu (File: /F-InformalFiction/F-Criminal/00046872tFFA .csv);

– szerszy kontekst (5-zdaniowy, pełne zdania, bez ograniczeń), (Context: bg *Затова и ми е толкова тежко. Марвин, дойде в живота ми, когато бях на девет години. Истината е че той и* ***майка*** *ми са били заедно преди да се родя аз. Тя криеше това от мене. – Той ти е баща?* [‘Dlatego też jest mi tak ciężko. Marvin pojawił się w moim życiu, kiedy miałem dziewięć lat. Prawdą jest, że on i moja **matka** byli razem, zanim się urodziłem. Ona ukrywała to przede mną. – On jest twoim ojcem?’]);

– metadane po bułgarsku w transliteracji łacińskiej (por. załączony obraz ekranu):

Rys. 10.2.



Komentarze semantyczne, pojawiające się samoczynnie w żółtych polach po najechaniu kursorem na wyraz, odnoszące się do słów autosemantycznych są dość obszerne, do pozostałych – oszczędne. Anotacje gramatyczne wymagają wprawy w deszyfracji lub posiadania indeksu symboli, dlatego np. w pasku menu można by umieścić choćby odsyłacz do adresu: http://dcl.bas.bg/en/BulgarianTagset_en.html, gdzie jest pełny wykaz symboli.

Anotacja wewnętrzna

Kryterium podziału na tokeny jest w BKN tradycyjne, tj. „słowa rozumiane jako ciągi znaków nie będących separatorami słów” (por. NKJP 61) – separatorami są spacje oraz znaki interpunkcyjne (przy czym apostrofu system

bułgarski nie stosuje, zaś dywiz nie może kończyć słowa). Jednostką wyższego rzędu w czasie wyszukiwania jest pełne zdanie (niezależnie od długości).

Niektóre zagadnienia istotne dla konstrukcji i użytkowania np. korpusów polskich są w języku i korpusie bułgarskim nieobecne. Język bułgarski nie zna np. form aglutynowanych; jedyne bliskie łączliwością to morfemy postpozycyjnego określnika, które jednak prawie nigdy (jedynie w tekstach stylizowanych poetycko) się nie odklejają i nie przemieszczają. W szczególności bułgarszczyzna nie zna rozłącznych elementów aglutynowanych (jak polskie *-by*, *-że*, *-li*), zaś formalne elementy analityczne (w zasadzie aglutynowane, jak *по-* i *най-* przy stopniowaniu przymiotników i przysłówków) nie odrywają się od słowa podstawowego. Dostępne narzędzia nie pozwalają badać takich konstrukcji jako połączeń ani o nie odrębnie pytać. Badane w korpusach słowiańskich systemów syntetycznych zasady traktowania skrótów i skrótowców nie przystają do założeń BKN – formacje tego typu są traktowane jako odrębne słowa.

Lematyzacja

W BKN nie przestrzega się opozycji wielkich/małych liter, zaś wyrazy niepoprawne (np. **сушоар* zamiast *сешоар* 'suszarka do włosów') są opisywane tak jak poprawne, zgodnie z ich cechami morfologicznymi. Lematyzacja jest w zasadzie poprawna, lecz nie ma w niej pełnej konsekwencji. Zdarzają się błędy systemowe polegające na uznaniu formy nieokreślonej part praes act za V (verbum), zaś jego postaci określonej – za A (adiectivum), por.: *пишещ* V IM T s y, *пишещият* A mf df s (sic!).

Tagset

Pod adresem: http://dcl.bas.bg/en/BulgarianTagset_en.html znajdujemy informacje nt. tagsetu. Kryterium nadrzędnym podziału słów (kategorią lematu) jest przynależność do części mowy, por.: Noun (tagowane jako N), Verb (jako V), Adjective (A), Pronoun (P), Adverb (D), Numeral (C), Preposition (R), Conjunction (J), Participle (T), Interjection (I).

Z kolei poszczególne części mowy przybierają następujące wartości i tagi:

PoS	**Attribute**	**Value**	**Tag**	**Position**
Noun	Type	Common	C	2nd
Noun	Type	Proper	P	2nd
Noun	Gender	Masculine	M	3rd
Noun	Gender	Feminine	F	3rd
Noun	Gender	Neutral	N	3rd
Noun	Gender	Masculine/Neutral	S	3rd
Noun	Gender	Masculine/Feminine	E	3rd
Noun	Type	Family name	A	3rd
Noun	Number	Singularia tantum	S	4th
Noun	Number	Pluralia tantum	P	4th
Verb	Type	Personal	2	2nd
Verb	Type	Impersonal	M	2nd
Verb	Transitivity	Transitive	T	3rd
Verb	Transitivity	Intransitive	I	3rd
Verb	Aspect	Perfective	P	4th
Verb	Aspect	Imperfective	I	4th
Pronoun	Type	Personal	P	2nd
Pronoun	Type	Possessive	L	2nd
Pronoun	Type	Reflexive	H	2nd
Pronoun	Type	General	B	2nd
Pronoun	Type	Relative	G	2nd
Pronoun	Type	Indefinite	I	2nd
Pronoun	Type	Negative	W	2nd
Pronoun	Type	Indefinite	I	2nd
Pronoun	Type	For persons and objects	B	3rd
Pronoun	Type	For attributes	C	3rd
Pronoun	Type	For quantity	D	3rd
Pronoun	Type	For possession	F	3rd
Numeral	Type	Ordinal masculine	K	2nd
Numeral	Type	Ordinal approximate	O	2nd
Numeral	Type	Ordinal one	Q	2nd
Numeral	Type	Ordinal two	S	2nd
Numeral	Type	Ordinal fraction	U	2nd
Numeral	Type	Cardinal	V	2nd

Kategorie „słowoform" (wyrazów tekstowych, Categories of the Word Form) są następujące:

PoS	**Attribute**	**Value**	**Tag**
Noun, Verb, Adjective, Pronoun, Numeral	Number	Singular	s
— " —	— " —	Plural	p
Noun	— " —	Countable	c
— " —	Case	Vocative	v
Noun, Verb, Adjective, Pronoun, Numeral	Definiteness	Indefinite article	0
— " —	— " —	Definite article	d
— " —	— " —	Short definite article	h
— " —	— " —	Long definite article	l
— " —	Gender	Masculine	m
— " —	— " —	Feminine	f
— " —	— " —	Neuter	n
Verb	Tense	Present	R
— " —	— " —	Aorist	E
— " —	— " —	Imperfect	D
— " —	Mood	Imperative	I
— " —	Participle	Present	Y
— " —	— " —	Past perfective	X
— " —	— " —	Past imperfective	W
— " —	— " —	Verbal adverb	Z
Verb, Pronoun	Person	First	1
— " —	— " —	Second	2
— " —	— " —	Third	3
Numeral	Form	Counting	b
— " —	— " —	Masculine personal	b
— " —	— " —	Approximate	x
Pronoun	Case	Accusative	a
— " —	— " —	Dative	d
— " —	Value Description	Masculine	M
— " —	— " —	Feminine	F
— " —	— " —	Neuter	N
— " —	Possessor's Number	Singular	S
— " —	— " —	Plural	P
— " —	Possessor's Person	First	1
— " —	— " —	Second	2
— " —	— " —	Third	3
— " —	Form	Clitic	C

Jak widać z ostatniego zestawienia, oznaczane (i poddające się badaniu ze strony użytkownika) w BKN są tylko formy proste (np. wyłącznie 3 proste czasy z ogólnej sumy 9; jedynie proste formy zaimków, z pominięciem tzw. podwojonego dopełnienia zaimkowego, składającego się z formy pełnej i krótkiej).

Anotacja morfologiczna, jeszcze niedoskonała systemowo i niekonsekwentna, ujawnia się już po najechaniu kursorem na wyraz w cytacie, mimo że uwzględnia wiele cech morfologicznych (włącznie np. z określonością). *Nota bene*: w zapytaniu można wpisać całą frazę (np. analityczną konstrukcję czasownikową), lecz anotacja odnosi się jedynie do pojedynczych elementów tej frazy. System odrzuca zapytania o formy z łącznikiem (zatem np. o stopień wyższy i najwyższy adi i adv). Przykładowe anotacje (strona zapytań: http://search.dcl.bas.bg/):

– dla sb: *майка* 'matka' N CO ff 0f s, *майката* 'matka (def)' N CO ff df s, *майки* 'matki' N CO ff 0f pf, *майките* 'matki (def)' N CO ff df pf.

Nie ma możliwości, nawet zapytując wprost (np. o formę *Майка*), uzyskania odpowiedzi różnicującej użycie wyrazu jako nomen proprium w opozycji do appellativum.

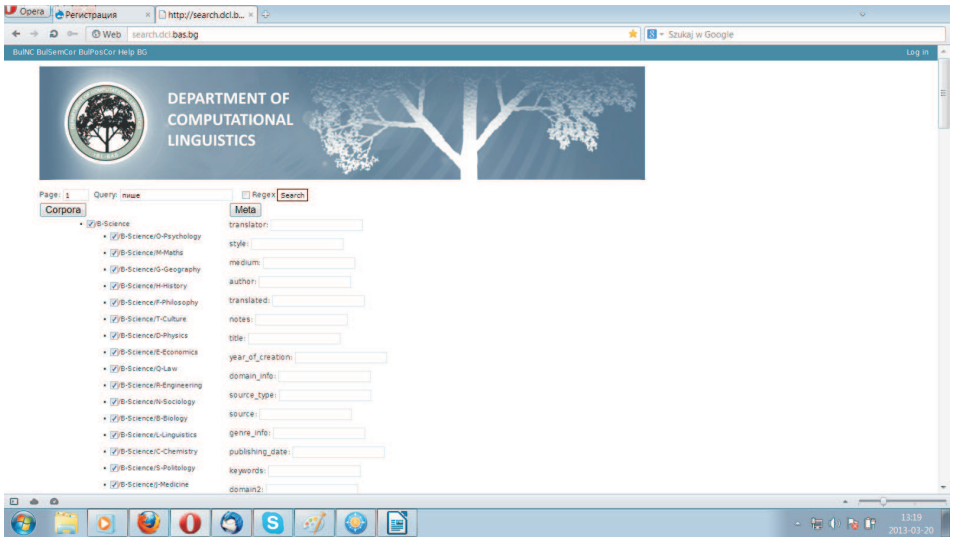
– dla adi: *майчин* 'matczyn' A mf 0f s, *майчина* 'matczyna' A ff 0f s, *майчино* 'matczyne (sg)' A nf 0f s, *майчини* 'matczyne/-i (pl)' A 0f s, *майчиният* 'matczyn (def subiect)' A mf df s, *майчиния* 'matczyn (def obiect)' A mf df s (sic! Poważny defekt opisu – obie ostatnie formy bez różnicy w lematyzacji!), *майчината* 'matczyna (def)' A ff df s, *майчиното* 'matczyne (sg def)' A nf df s, *майчините* 'matczyne/-i (pl def)' A df pf.

– dla verbum (wybrane formy): *пиша* 'piszę' V IM r T s 1, *пишат* 'piszą' V IM r T pf 3, *пишете* (homograf praes//imperat, 'piszecie//piszcie') 1) MISC lub 2) V IM r T pf 2, *пиши* 'pisz' V IM T s 2, *писахте* 'pisaliście (aorist)' V IM e T pf 2, *пишехте* 'pisaliście (imperfectum)' V IM j T pf 2, *писал* 'part praet perfectivum' (i dalsze participia) V IM T s x, *пишел* 'part praet imperfectivum' V IM T s q, *пишещ* 'part praes activum' V IM T s y, *пишещият* 'participium praes activum def' A mf df s (sic! – kolejny błąd systemowy dużej rangi).

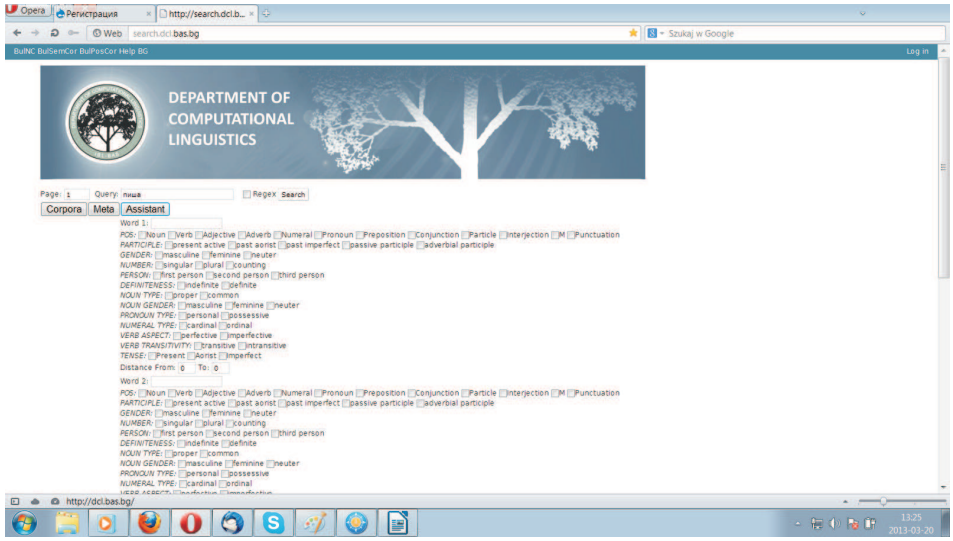
Podstawowe zapytania

Zadając pytanie na stronie http://search.dcl.bas.bg/, można wybrać konkretnie typ dokumentu, sam dokument (opcja Corpora), dokonać selekcji dokumentów według metadanych (opcja Meta: translator, style, medium, author, translated, notes, title: year_of_creation, domain_info, source_type, source, genre_info, publishing_date, genre etc.), dobrać parę szukanych wyrazów (lematów) na podstawie parametrów gramatycznych (na potrzeby badań kolokacji) – opcja Assistent. Można też skorzystać z opcji Regex (wyrażenie regularne) – to narzędzie jednak działa niepewnie. Porównaj obrazy:

Rys. 10.3.



Rys. 10.4.

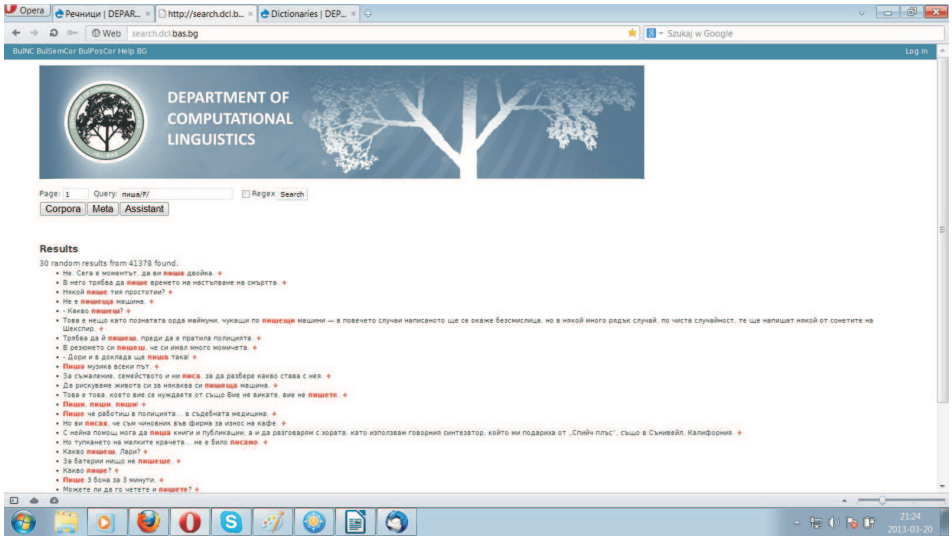


Wyszukiwanie form wyrazowych

Rys. 10.5.

Wyszukiwanie lematu

Rys. 10.6.



Wyszukiwanie form wyrazowych z wariantami ortograficznymi (pisanych małą i wielką literą)

Nie udało się znaleźć możliwości zapytania o zróżnicowane zapisy wielką/małą literą (mimo że istnieje możliwość zapytania o nomen appellativum/proprium) – wyszukiwarka traktuje je identycznie.

Wyszukiwanie grup wyrazów

Brak możliwości wyrażenia takiego zapytania; jedyna szansa to nomen rodzaju męskiego jako podmiot/niepodmiot – niestety, brak takiej opcji w narzędziach. Można pytać o dowolny ciąg wyrazów bez komentarza gramatycznego. (por. niżej).

Wyszukiwanie grup lematów

Rys. 10.7.

Wyszukiwane kategorii gramatycznej

Efektem próby wyszukania sb masc pl defin było zbudowanie przez wyszukiwacz wyrażenia zgodnego z instrukcją wyszukiwania: (<*{D=0f}[0,0]*{POS=NFG=mf N=pf D=df}>), jednak następnie program oddał (jako odpowiedź na własne zapytanie) komunikat Error.

Praca z konkordancją

Kolokator stosuje się tylko do korpusów ćwiczebnych; trudno jest – mimo instrukcji – uzyskać dane co do kolokacji słów z korpusu ogólnego. W ograniczonym zakresie można wykorzystać (dostępne tylko dla użytkowników zarejestrowanych) narzędzie Assistant (http://ibl.bas.bg/BGNC_search_bg.htm). Uzyskane informacje stanowią pełną reprezentację zawartości korpusu. Przykładowe zapytanie o konkretne połączenie słów Word 1: *майчината* 'matczyna' + Word 2: *любов* 'miłość' daje efekt taki jak na rys. 10.7. (62 poświadczenia takiej pary, por. wyżej), zaś analogiczne, jednak dla dowolnego rzeczownika żeńskiego (np. *усмивка* 'uśmiech') jako Word 2 (czyli: <майчината[0,0]*{POS=N}>) – kończy się komunikatem „No results":

Funkcje statystyczne

Brak tak zaawansowanych narzędzi do badań nad ogólną zawartością BKN (jedynie dla towarzyszących korpusów „ćwiczebnych"; por. niżej).

Zaawansowanym zarejestrowanym użytkownikom Korpus Bułgarski i Sekcja Lingwistyki Komputerowej Instytutu Języka Bułgarskiego BAN oferuje narzędzia (o różnych wymaganiach) do samodzielnej obróbki danych typu korpusowego i zwykłych tekstów. Są to programy:

– Chooser – niezależny od platformy, wielofunkcyjny system anotacji, umożliwiający szybkie i intuicyjne anotowanie z szybkim dostępem do szczegółowych informacji semantycznych (http://dcl.bas.bg/Chooser_bg.html); program z podręcznikiem w formacie .pdf do pobrania z internetu;

– Hydra – niezależny od platformy program do tworzenia i ewaluacji sieci leksykalno-semantycznych; (http://dcl.bas.bg/hydra-bg.html); do pobrania z internetu (z podręcznikiem w formacie .pdf);

– MacEst – program do korekty ortografii, pracujący na systemach MacOSX;

– WinEst – analogiczny do poprzedniego program, działający na platformie MS Office (od 2007, wcześniejszych wersji nie obsługuje) w ramach systemów Windows XP 32-bitowych i nowocześniejszych; uwaga, błędny adres pobierania aplikacji (przekierowuje do WebEst);

– WebEst – aplikacje internetowe do redakcji i korekty tekstu bułgarskiego; mamy do dyspozycji dwa niezbyt wiekopomne aspekty pracy: 1) w okienko wpisujemy formę podejrzaną, a program oddaje nam wersję poprawną, 2) wpisujemy dowolną formę wyrazową, program oddaje nam jej formę podstawową (słownikową), informację o przynależności do części mowy i charakterystykę gramatyczną;

– „ItaEst – Taka e!" – system wspomagania korekty (wraz z dzieleniem wyrazów) i redakcji tekstów bułgarskich, działający z „Microsoft Office 2000, XP i 2003" (sic! Nie wiadomo, czy chodzi o pakiety biurowe, czy o systemy

operacyjne i czy odnosi się do bitowości); program podobno jest do ściągnięcia (III 2013), jednak jego strona rodzima odrzuca wszystkie próby wejścia;

– WebEst+ – korekta ortograficzna tekstu on-line (także internetowego, blogów itd.);

– BGDictionary – program pokazujący paradygmaty i charakteryzujący słowa gramatycznie;

– bgMWE – program do pobierania (np. z internetu), rozpoznawania, konwertowania i tagowania złożonych jednostek leksykalnych (ang. MWE); do pobrania z internetu;

– Speechlab – bułgarski syntezator mowy – czyta tekst elektroniczny, strony internetowe, pocztę elektroniczną; bardzo niskie wymagania systemowe (Win 98) i sprzętowe (Pentium 266 MHz, 64 MB RAM); darmowy dla osób z udokumentowaną wadą wzroku;

– System przeszukiwania korpusów (Система за търсене в корпуси);

– Уеб сървис за колокации (Corpus collocation service) – zgodnie z nazwą służący do badania kolokacji (pracuje wyłącznie na korpusach paralelnych);

– BGTokenizer (Bulgarian Sentence Splitter and Tokenizer) – tokenizator, rozpoznaje skróty, nazwy własne, liczby, daty i granice zdań;

– Уеб базирана инфраструктура за лингвистична обработка на данни на български език – A Web-Based Infrastructure for Bulgarian Data Processing – system podstawowej obróbki i anotacji językowej tekstów; przeprowadza podział zdań, tokenuje, charakteryzuje słowa gramatycznie, lematyzuje;

– BgTagger – składnik poprzedniego systemu, służy do tagowania; dokładność rzędu 96,6%; rodzima platforma: Linux/UNIX;

– BGWSD – program probabilistyczny do usuwania wieloznaczności semantycznej.

Zastosowania

Na bazie danych uzyskanych z BKN i korpusów powiązanych można prowadzić typowe badania: częstości występowania poszczególnych leksemów w zależności od epoki, rodzaju oraz stylu tekstów i autorów. Wyniki badań są też wykorzystywane do tworzenia i aktualizacji bułgarskich słowników jedno- i dwujęzycznych. Szczególnie cenne są relatywnie duże korpusy paralelne, stanowiące bardzo cenne źródło obserwacji dla tłumaczy praktyków i dydaktyków translatoryki.

Projekty słownikowe (w dużej mierze oparte na materiałach i obserwacjach korpusowych), w których opracowaniu brali/biorą udział językoznawcy związani z BNK:

– Граматичен речник на българския език – Grammar Dictionary of Bulgarian – 85 tys. jednostek (http://dcl.bas.bg/est/dict.php);

– Синтактичен речник на българския език – Frame Dictionary of Bulgarian;

– Български граматичен речник на съставните лексикални единици – Bulgarian Grammatical Dictionary of MWEs;

– Речникът на съставните лексикални единици в българския език – Bulgarian MWE Dictionary;

– Честотни речници – Frequency Dictionaries;

– Многоезиков речник на 6 езика – Multilingual Dictionary in 6 Languages;

– Граматически речник с обща лексика на български език – General Lexis Grammar Dictionary of Bulgarian;

– Специализирани речници – Specialised Dictionaries:
  – Медицински речник – Dictionary of Medical Terms,
  – Речник с правни и съдебни термини – Dictionary of Legal Terms,
  – Икономически речник – Dictionary of Economic Terms,
  – Политически речник – Dictionary of Political Terms,
  – Военен речник – Dictionary of Military Terms,
  – Спортен речник – Dictionary of Sports Terms,
  – Речник на собствени имена – Dictionary of Personal Names,
  – Речник на съкращения – List of abbreviations.

Bibliografia

Wybrane nowsze publikacje kolektywu twórców korpusu związane z problematyką korpusową:

Koeva S., Dekova R., Stoyanova I., Rizov B., Genov A. (2012). *Bulgarian X-language Parallel Corpus.* W: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* (format pdf)

Koeva S., Rizov B., Tarpomanova E., Dimitrova Ts., Dekova R., Stoyanova I., Leseva S., Kukova H., Genov A. (2012). *Application of Clause Alignment for Statistical Machine Translation.* W: *Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, Jeju, Republic of Korea, 12 July 2012*, s. 102-110. (format pdf)

Koeva S., Rizov B., Tarpomanova E., Dimitrova Ts., Dekova R., Stoyanova I., Leseva S., Kukova H., Genov A. (2012a). *Bulgarian-English Sentence- and Clause--Aligned Corpus.* W: *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon, 29 November 2012.* Lisboa, s. 51-62. (format pdf)

Koeva S., Stoyanova I., Leseva S., Dimitrova Ts., Dekova R., Tarpomanova E. (2012). *The Bulgarian National Corpus: Theory and Practice in Corpus Design.* „Journal of Language Modelling", 2012, No. 1, s. 65-110. (także format pdf)

Коева С., Благоева Д., Колковска С. (2011). *Проектът Български национален корпус – резултати и перспективи.* „Български език", 58 (2011), 3, s. 34-53.

Коева С., Стоянова И., Димитрова Ц., Лесева С. (2012), *Традиции и новаторство в корпуснатa лингвистика: Българският национален корпус*. „Списание на Българската академия на науките", 2012, 3.

11. Praktyczny przewodnik po korpusie języka macedońskiego

Marjan Markovik'
Uniwersytet Cyryla i Metodego, Skopie
Tłum. Magdalena Bogusławska
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Aktualnie narodowy korpus elektroniczny języka macedońskiego jeszcze nie istnieje.

Profesor George Goce Mitrevski z Uniwersytetu w Auburn (USA) przygotował nieanotowany korpus języka macedońskiego. Korpus można przeglądać według kilku kryteriów – zarówno uwzględniając poszczególne formy wyrazowe, jak i z odstępami pomiędzy kilkoma wyrazami. Można wyszukiwać całe wyrazy tekstowe (słowoformy), a także ich części, np. początek, środek czy koniec słowa. Można także wyszukiwać po kilka wyrazów tekstowych, by wskazać elementy znajdujące się pomiędzy nimi oraz sposób ich linearyzacji. Korpus jest dostępny *online* pod adresem: http://www.auburn.edu/∼mitrege/mac-literature.html. Z profesorem Goce Mitrevskim można się kontaktować, korzystając z adresu: mitrevski@auburn.edu.

Równolegle czynione są starania, by został utworzony narodowy anotowany korpus języka macedońskiego. Z inicjatywy Zakładu Lingwistyki i Nauki o Literaturze Macedońskiej Akademii Nauk i Sztuk (MANU; Одделението за лингвистика и литературна наука на МАНУ), w ramach działalności Centrum Badawczego Lingwistyki Arealnej im. Božidara Vidoeskiego (ICAL; Истражувачкиот центар за ареална лингвистика „Божидар Видоески") w kwietniu 2011 r. odbyła się konferencja naukowa *Nauczanie języka macedońskiego – stan obecny i perspektywy* (*Наука за македонскиот јазик – состојби и перспективи*). W konferencji uczestniczyło ponad 40 lingwistów z wielu instytucji oświatowo-naukowych, a wnioski z obrad zostały przyjęte przez władze Macedońskiej Akademii Nauk i przekazane kompetentnym instytucjom.

Formułując wnioski z konferencji, stwierdzono, że najistotniejszym zadaniem, jakie należy wykonać i które wymaga przygotowań niezbędnych dla badań i afirmacji języka macedońskiego na forum międzynarodowym, jest opracowanie *Gramatyki opisowej współczesnego języka macedońskiego* (*Дескриптивна граматика на современиот македонски стандарден јазик*), a równolegle

z przygotowaniem takiego kompendium należy rozpocząć prace nad Elektronicznym korpusem języka macedońskiego (Elektronski korpus na makedonskiot jazik). Pod koniec roku 2012 ministerstwo kultury do narodowej strategii rozwoju kultury w Macedonii włączyło także opracowanie *Gramatyki opisowej współczesnego języka macedońskiego*.

W tym kontekście działające przy Macedońskiej Akademii Nauk Centrum Badawcze Lingwistyki Arealnej oraz Laboratorium Systemów i Sieci Kompleksowych (Лабораторијата за комплексни системи и мрежи), w styczniu 2013 roku rozpoczęły wspólnie przygotowania do realizacji Elektronicznego korpusu współczesnego języka macedońskiego.

Prace nad korpusem znajdują się w fazie wstępnej. Obecnie odbywają się liczne spotkania lingwistów i informatyków, dotyczące ustalenia kierunków i zasad tworzenia korpusu. Informatycy z Laboratorium Systemów Kompleksowych jako podstawę bazowej anotacji wykorzystywać będą system DataSet (http://nl.ijs.si/ME/V4/msd/html/msd-mk.html) z międzynarodowego projektu MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages), zlokalizowanego na: http://nl.ijs.si/ME/. Kolejną fazą będzie zebranie tekstów i ich wstępna anotacja.

12. Praktyczny przewodnik po korpusie języka rosyjskiego

Maksim Duszkin
Polska Akademia Nauk
Instytut Slawistyki

Informacje ogólne

Rosyjski Korpus Narodowy (ros. Национальный корпус русского языка) – dalej RKN – jest dostępny w internecie od 10 lat – od kwietnia 2004 roku (Сичинава 2005: 21). Od początku swojego istnienia do dziś funkcjonuje on na stronie http://www.ruscorpora.ru[1].

RKN jest dziełem wielu osób, wspieranym przez wiele instytucji. Na etapie pierwszych pomysłów i próbnych wdrożeń (w okresie przed oficjalnym uruchomieniem RKN 29 kwietnia 2004 roku) korpus był związany przede wszystkim z kręgiem moskiewskich naukowców z kręgu „Centrum Dokumentacji Lingwistycznej" („Центр лингвистической документации"). Grupa ta działała pod kierownictwem Władimira Plungiana i Michaiła Daniela w Centrum Ustawicznej Edukacji Matematycznej (Центр непрерывного математического образования). Uczeni spotykali się na seminariach poświęconych problematyce lingwistyki stosowanej. W roku 2000 odbyło się pierwsze seminarium na temat lingwistyki korpusowej (pod kierownictwem Siergieja Szarowa), a już w 2001 r. z inicjatywy Plungiana zawiązała się grupa naukowców, która przystąpiła do prac nad przyszłym Rosyjskim Korpusem Narodowym (Сичинава 2005: 22-24).

Pod koniec roku 2001 decyzję o włączeniu się do pracy nad korpusem podjęła Katedra Lingwistyki Teoretycznej i Stosowanej Państwowego Uniwersytetu Moskiewskiego im. Łomonosowa (Кафедра теоретической и прикладной лингвистики МГУ им. М.В. Ломоносова) (Сичинава 2005: 24).

W 2002 roku skonstruowanie korpusu stało się tematem statutowym opracowywanym przez Zakład Badań Lingwistycznych Rosyjskiego Instytutu Informacji Naukowej i Technicznej (Всероссийский институт научной и технической информации РАН). W ramach tego zadania Instytut współpracował

[1] Wcześniej, od roku 2002 mieścił się na niej rosyjski korpus z usuniętą homonimią gramatyczną, który później stał się podstawą RKN (Сичинава 2005: 30).

z Zakładem Zbiorów Maszynowych Instytutu Języka Rosyjskiego im. W. Winogradowa Rosyjskiej Akademii Nauk (Институт русского языка им. В.В. Виноградова РАН) (Сичинава 2005: 25-26).

Oprócz specjalistów z dwóch wyżej wymienionych instytutów, w projekcie uczestniczą naukowcy z ośrodków naukowych Moskwy, Petersburga, Kazania, Woroneża i Saratowa. Są to: Институт языкознания РАН, Институт проблем передачи информации РАН, Институт лингвистических исследований РАН в Санкт-Петербурге (z którym współpracuje Санкт-Петербургский государственный университет [СПбГУ]), Казанский (Приволжский) федеральный университет, Воронежский государственный университет, Саратовский государственный университет (http://www.ruscorpora.ru/corpora-about.html – dostęp: 1 III 2013). Na różnych etapach i w różnym zakresie w opracowywaniu korpusu brało też udział wiele innych osób (zob. np. Сичинава 2005: 26).

W okresie 2003-2012 r. projekt budowy RKN był finansowany przez następujące instytucje:

– Отделение историко-филологических наук Российской академии наук (programy „Филология и информатика", „Русский язык, литература и фольклор в информационном обществе: формирование электронных научных фондов", „Генезис и взаимодействие социальных, культурных и языковых общностей", „Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации" (2003-2011));

– Президиум РАН (program badań podstawowych „Историко-культурное наследие и духовные ценности России" (2009-2012));

– Российский гуманитарный научный фонд (projekty: 03-04-00226а, 06-04-03817в, 06-04-03818в, 08-04-12127в, 09-04-12159в);

– Российский фонд фундаментальных исследований (projekty: 06-06-80133а, 08-06-00371-а);

– Федеральное агентство по образованию, федеральная целевая программа „Русский язык" (umowy: 1028, 890, 608 z 14 XII 2006, 219 z 18 VI 2007, 66 z 11 IV 2008). (http://www.ruscorpora.ru/corpora-about.html – dostęp z 1 III 2013).

Ogromną rolę w projekcie – głównie techniczną, ale nie tylko – odegrała firma Yandex (http://yandex.ru). Jest ona znana przede wszystkim z wyszukiwarki internetowej obsługującej rosyjski segment internetu. Firma Yandex była zainteresowana rozwojem korpusu, widząc w nim narzędzie służące doskonaleniu programów do automatycznej obróbki tekstu (Сичинава 2005: 24).

Yandex finansował pierwsze prace nad przyszłym korpusem (np. usuwanie homonimii w próbnym korpusie tekstów z drugiej połowy XX w. – rok 2001),

a także dostarczył niektórych programów niezbędnych podczas wykonywania tych prac (Сичинава 2005: 25, 26). W 2002 r. firma uruchomiła na swoim serwerze stronę korpusu (ruscorpora.ru) (Сичинава 2005: 29; Плунгян 2005: 15) i do dziś zapewnia jej obsługę techniczną (http://www.ruscorpora.ru/corpora-about.html). Wyszukiwanie w korpusie odbywa się obecnie z wykorzystaniem oprogramowania tej firmy – systemu wyszukiwania Яндекс.Сервер 3.8 Professional (http://www.ruscorpora.ru/corpora-about.html). W pracach nad korpusem bierze udział wielu programistów tego przedsiębiorstwa (http://www.ruscorpora.ru/corpora-team.html).

Warto podkreślić, że RKN to projekt otwarty: stale dodawane są nowe teksty, powstają nowe podkorpusy, stopniowo pojawiają się nowe narzędzia techniczne. Z tego powodu stan projektu czasami wyprzedza opis merytoryczny i techniczny, dostępny na jego stronie internetowej[2].

Charakterystyka RKN. Struktura ogólna projektu

Korpus narodowy w rozumieniu twórców RKN to przede wszystkim korpus reprezentatywny. Korpus reprezentatywny odzwierciedla „wszystkie możliwe typy tekstów istniejące w danym języku w danym okresie" (Плунгян 2005: 8). Język rosyjski, który ma być obiektem prezentacji RKN, z założenia nie jest tożsamy zatem tylko z wzorcową odmianą literacką. Oprócz zgodnych z normą wzorcową tekstów pisanych, należących do różnych gatunków (literatury pięknej, naukowej, publicystyki itd.), w korpusie takim należałoby więc umieścić także teksty mówione, odbiegające od normy wzorcowej danego języka („potoczne"), a nawet teksty różnych odmian regionalnych (teksty dialektalne). Do RKN włączono je, jednak na odmiennych zasadach, organizując w odpowiedni sposób strukturę projektu.

RKN obecnie to zbiór kilku odrębnych korpusów rosyjskich. W jego skład wchodzą mianowicie Korpus Główny oraz kilka korpusów specjalistycznych (http://www.ruscorpora.ru/corpora-structure.html, http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=300&Itemid=127#2, http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=301&Itemid=128)

W Korpusie Głównym są zawarte oryginalne pisane teksty rosyjskie z okresu XVIII–XXI w., przy czym są to utwory wyłącznie prozaiczne (nie poetyckie) – nie tylko teksty literatury pięknej różnych gatunków (choć stanowią one aż 40% zawartości Korpusu Głównego), ale też publicystyka, krytyka literacka, wiadomości prasowe, literatura religijna, teksty naukowe, dydaktyczne itd. W korpusie obowiązują ściśle ustalone proporcje ilościowe pomiędzy reprezentowanymi w nim odmianami tekstów. Jest to zatem korpus zrównoważony.

[2] Dane przedstawione w tym artykule opisują stan projektu w marcu 2013 roku.

Oddzielny korpus stanowi Korpus Publikacji Prasowych (*Газетный корпус* (*корпус современных СМИ*), reprezentujący artykuły z prasy rosyjskiej wydane po roku 2000. Jego zawartość wciąż się powiększa, w związku z czym nie można go dołączyć do Korpusu Głównego; zachwiałoby to bilans ilościowy pomiędzy odmianami tekstów, w tym proporcje tekstów z różnych ram czasowych obejmujących okres od XVIII do XXI w. (http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=301&Itemid=128#23, http://www.ruscorpora.ru/corpora-stat.html). Stale aktualizowany korpus prasowy pozwala natomiast na bieżąco śledzić zmiany zachodzące w języku.

W skład RKN wchodzą także:

– Anotowany morfosyntaktycznie Korpus Składniowy
  Można w nim wyświetlić informację o strukturze składniowej każdego wyekscerpowanego zdania, co nie jest możliwe w tekstach Korpusu Głównego.
– Korpus Tekstów Dialektalnych
  Zawiera zapis tekstów z różnych regionów Rosji, przy czym nie w transkrypcji fonetycznej, lecz w zapisie ortograficznie zbliżonym do oryginału.
– Korpus Tekstów Poetyckich
  Korpus reprezentuje poezję z okresu od XVIII w. do dziś.
– Korpus Tekstów Mówionych
  Zawiera transkrypcje nagrań tekstów mówionych (publicznych i niepublicznych), a także filmów.
– Korpus Akcentologiczny
  Udostępnia informacje o rosyjskim akcencie[3] i pozwala uwzględniać akcent w kryteriach wyszukiwania.
– Korpus Dydaktyczny
  Składa się w dużej mierze z tekstów lektur szkolnych. Zastosowano w nim anotację gramatyczną zgodną z tradycyjną gramatyką szkolną.
– Korpus Multimedialny
  Korpus ten obejmuje fragmenty filmów, począwszy od lat trzydziestych XX w. Użytkownik może obejrzeć wideo, a także przeczytać transkrypcję dialogów i opis gestów. Korpus zawiera między innymi anotację typów gestów (kiwnięcie, klepanie po ramieniu itd.) i prezentowanych aktów językowych (zgoda, ironia itd.).
– Korpus Historyczny
  Obejmuje podkorpus tekstów cerkiewnosłowiańskich oraz udostępniony 17 I 2013 roku podkorpus tekstów XV–XVIII w. („корпус среднерусских текстов")[4].

[3] Akcent można wyświetlić także w wynikach wyszukiwania w Korpusie Poetyckim, a także Dydaktycznym, Multimedialnym, Mówionym i Historycznym oraz w niektórych tekstach Korpusu Głównego. Służy temu specjalna opcja (*версия с ударениями*) (http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=300&Itemid=127#12). W Korpusie Głównym i Dydaktycznym akcent rozstawiony został automatycznie.

[4] W chwili powstania artykułu korpus ten nie był jeszcze dokładnie opisany na stronie projektu. Przytoczyć tu można komunikat, który pojawił się na stronie głównej (dostęp: 1 III 2013): „17 января 2013 года. Открыт новый исторический корпус – корпус сред-

– Korpus Równoległy

Korpus zawiera teksty rosyjskie zestawione z ich przekładami na inne języki lub teksty obcojęzyczne zestawione z tłumaczeniem na rosyjski. Jednym z uwzględnianych w korpusie języków jest polski. (http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=301&Itemid=128#25, http://ruscorpora.ru/corpora-structure.html).

Statystycznie rzecz biorąc, według danych umieszczonych na stronie projektu Korpus Główny – pod względem liczby użytych słów – stanowi obecnie ok. 57% zawartości RKN, Prasowy – 31%, Równoległy – ok. 6%, Mówiony – ok. 3%, Poetycki – ok. 2%, pozostałe – po mniej niż 1% każdy (http://www.ruscorpora.ru/corpora-stat.html).

Niektóre korpusy pod względem zawartości mogą się częściowo dublować, tzn. zawierać te same teksty. Np. Korpus Akcentologiczny zawiera teksty z Korpusu Poezji oraz z Korpusu Mówionego, w Korpusie Równoległym zaś można natrafić na oryginalne teksty rosyjskie, które włączono też w skład Korpusu Głównego[5].

Jeśli chodzi o typy, odmiany stylistyczne i gatunkowe tekstów, to ta sama odmiana tekstów wystąpić może w kilku korpusach. Dla przykładu, artykuły prasowe pojawiają się nie tylko w Korpusie Prasowym, ale też w Korpusie Głównym.

Analizując strukturę RKN, na który się składa właściwie 11 korpusów, należy stwierdzić, że, niestety, nie jest ona jednolita i trudno znaleźć dokładny opis kryteriów podziału. Wydaje się jednak, że odzwierciedla ona założenia autorów co do typów zadań możliwych do zrealizowania przy pomocy korpusu. Chodzi mianowicie o to, że odrębne podkorpusy RKN powstawały z myślą o specyficznych przeznaczeniach. Do badań ogólnych nad tekstami języka pisanego stworzono Korpus Główny. Natomiast pozostałe korpusy projektu RKN zgromadzono w celach specjalistycznych i zaopatrzono w narzędzia do ich realizacji. Przykładowo Korpus Składniowy przeznaczony jest głównie dla naukowców, którym ma dostarczać specjalistycznych informacji o strukturze składniowej zdań i prowadzenia odpowiednich badań. W korpusie tym zastosowano unikalną anotację struktury składniowej zdań, niestosowaną w innych korpusach RKN. Korpus Dydaktyczny natomiast służy do nauczania języka rosyjskiego w szkole i ma odpowiednio dostosowaną anotację morfologiczną. Korpus Ak-

нерусских текстов (XV – начало XVIII века). Объём корпуса – 3 миллиона словоупотреблений: литературные произведения, летописи, жития, деловые грамоты, бытовая переписка. Доступен поиск точных форм (без морфологической разметки), в том числе с использованием символа *, а также задание подкорпуса.".

[5] Informację tę uzyskano w korespondencji z Dmitrijem Siczinawą, jednym z twórców RKN. Natomiast na stronie *Образовательный портал национального корпуса русского языка* podano, że zawartość korpusów jest rozłączna, co nie jest zgodne ze stanem bieżącym (por. http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=300&Itemid=127#2).

centologiczny przeznaczony jest do badań nad akcentem, Dialektalny – nad odmianami gwarowymi języka, Historyczny – nad tekstami dawnymi, Prasowy – nad procesem zmian językowych i specyfiką tekstów prasowych, Poetycki – nad zjawiskami związanymi z wersyfikacją itd. Innymi słowy, zgodnie z przeznaczeniem korpusu dobierano teksty i stosowano anotację na odpowiednim poziomie.

Ze względu na liczbę korpusów specjalistycznych w projekcie RKN (obecnie 10) i zróżnicowanie ich budowy, w niniejszym opisie jesteśmy zmuszeni je pominąć. Skupimy uwagę na Korpusie Głównym.

Struktura tekstowa Korpusu Głównego

Pod względem wyodrębnianych i reprezentowanych odmian tekstów Korpus Główny jest najbardziej zróżnicowanym korpusem wśród korpusów RKN – przynajmniej jeśli chodzi o korpusy zawierające wyłącznie teksty pisane prozą. Tylko on był tworzony jako z założenia reprezentatywny i zrównoważony: reprezentuje wszystkie typy współczesnych tekstów rosyjskich pisanych prozą, przy czym owe typy zostały dobrane w apriorycznie ustalonych, ściśle przestrzeganych proporcjach. Literatura piękna stanowi 44% jego zawartości, pozostałe teksty – resztę, czyli 56% (ogólnej liczby słów w Korpusie Głównym, która wynosi obecnie ponad 209 milionów).

W Korpusie uwzględniono też duże zróżnicowanie gatunkowe w obrębie typów tekstu.

Po pierwsze, ustalono, w jakich proporcjach mają być reprezentowane poszczególne gatunki literatury pięknej. Otóż 61% pochodzących z niej słów to tak zwana proza pozagatunkowa („нежанровая проза"), czyli proza, którą trudno byłoby scharakteryzować za pomocą jakichkolwiek definicji gatunkowych (http://www.ruscorpora.ru/instruction-main.pdf – s. 32). Poza tym w korpusie reprezentowane są: proza dokumentalna[6] (7,7% słów z literatury pięknej), fantastyka (6,1%), proza historyczna (6,0%), literatura detektywistyczna wraz z literaturą akcji (5,8%), literatura dziecięca (4,4%), literatura humorystyczna i satyra (4,1%), literatura przygodowa (2,2%), dramaturgia (1,9%) oraz literatura miłosna (0,9%) (http://www.ruscorpora.ru/corpora-stat.html – dane z 1 III 2013).

Ustalono też, ile miejsca w korpusie zajmą poszczególne odmiany tekstów nienależących do literatury pięknej. Wyodrębniono dziedziny, w których teksty te funkcjonują („сфера функционирования") i ustalono ich udział procentowy. 63,8% słów z tekstów niebędących literaturą piękną to publicystyka. Są poza tym teksty naukowo-dydaktyczne (23%), teksty z dziedziny komunikacji codziennej („бытовая сфера использования") (3,7%), ze sfery komunikacji

[6] Chodzi tu o literaturę pamiętnikarską, która jest traktowana jako gatunek literatury pięknej nieco wbrew rosyjskiej tradycji literaturoznawczej (Плунгян 2005: 11).

urzędowej (3,2%), religii (2,7%), komunikacji internetowej (1,9%), produkcji i techniki (1,1%) oraz z reklamy (0,5%). (http://www.ruscorpora.ru/corpora-stat.html – dane z 1 III 2013).

Udział procentowy poszczególnych typów tekstów wyodrębnionych w Korpusie Głównym RKN z założenia ma odzwierciedlać ich realny udział w języku (http://www.ruscorpora.ru/corpora-structure.html, http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=300&Itemid=127#8)[7].

Warto zauważyć, że w pozostałych – specjalistycznych – korpusach RKN reprezentowane są tylko niektóre odmiany tekstów pisanych prozą. Przykładowo, Korpus Składniowy zawiera tylko współczesną rosyjską literaturę piękną, artykuły popularnonaukowe i polityczne z czasopism z lat 1980-2004 oraz drobne wiadomości publikowane przez agencje prasowe w internecie (http://www.ruscorpora.ru/instruction-syntax.html). Korpus Dydaktyczny reprezentuje głównie literaturę piękną z programu szkolnego (choć z drugiej strony zawiera on także niewielki zasób wypowiedzi mówionych). Zawarte w rosyjskiej części Korpusu Równoległego teksty to literatura piękna, nie tylko rosyjska, ale również tłumaczenia. Teksty podkorpusu cerkiewnosłowiańskiego Korpusu Historycznego są wyraźnie ograniczone pod względem zakresu użycia (teksty liturgiczne, naukowe, prawne etc., ale już nie literatura piękna) itd.

Anotacja zewnętrzna w Korpusie Głównym RKN

Anotacja zewnętrzna czy też – w terminologii autorów – *метаразметка текстов* (http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=300&Itemid=127#9), *параметры текстов* (http://www.ruscorpora.ru/corpora-parameter.html), jest w RKN dość znacznie rozbudowana. Anotacji zewnętrznej podlega każdy dokument (tekst) włączany do korpusu. Podawana jest informacja dotycząca autorstwa tekstu (imię i nazwisko autora, data jego urodzenia), tytułu tekstu oraz daty jego powstania. Natomiast w Korpusie Głównym podaje się także następujące informacje o zawartych w nim tekstach:

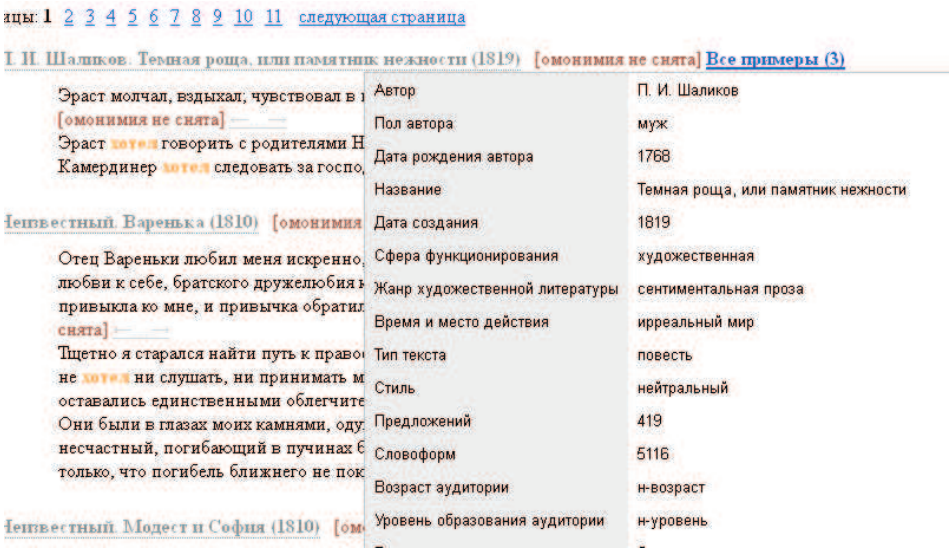
– odmiana tekstu: literatura piękna – teksty nie należące do literatury pięknej;
– w przypadku literatury pięknej określany jest gatunek (literatura dziecięca, fantastyka itd.), typ („autoidentyfikacja" tekstu: powieść, bajka, opowieść, esej itd.) i informacja o czasie i miejscu opisywanych w tekście wydarzeń (np. „Średniowiecze", „Europa: Średniowiecze", „Rosja: 1900-1914" itd. oraz „Świat fikcyjny");

[7] Warto zauważyć, że odzwierciedlana jest raczej liczba/objętość tekstów danego typu (tzn. ile tekstów danego typu napisano), nie zaś ich popularność (ile tekstów danego typu się czyta). Innymi słowy, stosowane jest inne podejście niż w Narodowym Korpusie Języka Polskiego, w którym – jeśli chodzi o teksty pisane – próbuje się „zrekonstruować strukturę czytelnictwa" (NKJP 2012: 28).

– dla innych tekstów określana jest dziedzina funkcjonowania (np. komunikacja codzienna, komunikacja oficjalno-urzędowa itd.), typ („autoidentyfikacja" tekstu, np.: życiorys, umowa, dziennik, przepis, przewodnik, artykuł, podręcznik, instrukcja, sprawozdanie itd.) oraz tematyka (np. prawo, technika, sport, informatyka, psychologia itd.). (http://ruscorpora.ru/mycorpora-main.html).

Anotacja zewnętrzna dokumentów w korpusie pozwala wyświetlić dane o tym, skąd pochodzą znalezione przykłady, jakiego rodzaju jest to tekst, kto jest jego autorem, kiedy powstał, co się w tekście opisuje itd. (zob. rys. 12.1). Ponadto dzięki anotacji zewnętrznej użytkownik może tworzyć własne podkorpusy. Służy do tego opcja *задать подкорпус* w górnej części strony wyszukiwania w Korpusie Głównym (link bezpośredni: http://www.ruscorpora.ru/mycorpora-main.html). Interesujących nas wyrażeń można szukać np. tylko wśród tekstów danego autora (np. Puszkina), bądź tekstów pochodzących np. z okresu 1900-1915, czy tylko w tekstach z literatury pięknej; lub – na odwrót – tylko wśród instrukcji technicznych bądź dokumentów prawnych. Jest to ważna możliwość i warto z niej korzystać.

Rys. 12.1. Anotacja zewnętrzna tekstów (informacja o autorze, tytuł, data powstania itd.).



Anotacja wewnętrzna w Korpusie Głównym RKN

Segmentacja

Zamiast pojęcia „segment" w RKN konsekwentnie stosowane jest pojęcie „użycia słowa" („словоупотребление"). Czy jest to istotne? Otóż okazuje się, że tak.

Zauważmy najpierw, że w Narodowym Korpusie Języka Polskiego segmentami są zarówno leksemy, jak też znaki interpunkcyjne (NKJP 2012: 61)[8]. Znaki interpunkcyjne stanowią w tym korpusie „klasę fleksyjną" („interpunkcja"), w tagsecie zaś przewidziano dla niej odpowiedni tag (NKJP 2012: 63, 66-67). Z punktu widzenia tego korpusu ciąg w rodzaju „*ten, który ...*" stanowi trzy segmenty[9].

W RKN natomiast „segmentami" („словоупотребление") są tylko i wyłącznie słowa, natomiast znaki interpunkcyjne potraktowano tu inaczej. Decyzja taka ma następujące konsekwencje praktyczne:

Po pierwsze, wszystkie dane statystyczne w RKN dotyczą liczby słów (użyć), nie obejmując znaków interpunkcyjnych[10]. Po drugie, znaki interpunkcyjne nie stanowią żadnej kategorii gramatycznej uwzględnianej w tagsecie. Po trzecie, znak interpunkcyjny nie jest traktowany jako słowo, czyli nie jest uwzględniany w zapytaniu. By znaleźć zdania w rodzaju *Я решилась сказать, что никогда не соглашусь быть его женою и идти в монастырь...* (pol. 'Zdecydowałam się powiedzieć, że nigdy się nie zgodzę zostać jego żoną i pójść do klasztoru...') zawierające połączenie *сказать, что* pytamy o występujące tuż przy sobie wyrazy *сказать* oraz *что* (czyli właściwie o ciąg „*сказать что*"). Po czwarte, znaki interpunkcyjne nie są uwzględniane przy sortowaniu wyników według lewego lub prawego kontekstu szukanego wyrazu/połączenia. Np. przy sortowaniu według lewego kontekstu wyników zapytania o połączenie „*твой голос*" zdanie [...] *кажется, твой голос*, [...] będzie występowało tuż przed *Когда твой голос величавый* [...], choć ze względu na przecinek mogłoby zostać wyświetlone znacznie wcześniej (gdyby znaki interpunkcyjne byłyby traktowane tak samo jak słowa).

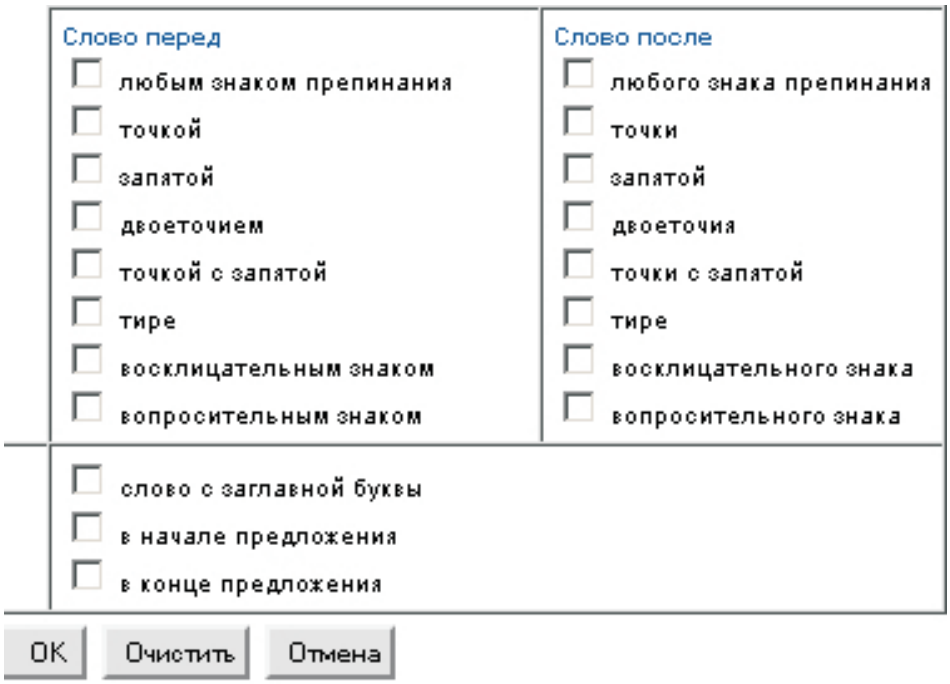
Nie możemy szukać znaków interpunkcyjnych jako odrębnych „pełnowartościowych" segmentów; możemy jednak wskazać, w jakim sąsiedztwie interpunkcyjnym powinien wystąpić wyraz, którego szukamy (np. zapytanie o zaimek *кто* wyłącznie po przecinku). Służą temu odpowiednie opcje we „właściwościach dodatkowych" szukanych wyrazów (opcja *Доп. признаки*). Wyraz wystąpić może przed znakiem interpunkcyjnym lub po nim (zob. rys. 12.2.).

[8] Znaki będące częściami słów, tzn. niektóre dywizy i apostrofy, nie były traktowane jako segmenty, np. dywiz w słowie *ping-pong* czy apostrof w formie *Chomsky'ego* (NKJP 2012: 61).

[9] W tym sensie przecinek i np. rzeczownik są tu traktowane jako jednostki tego samego poziomu i wagi. Segment w NKJP nie odpowiada zatem temu, co w języku potocznym nazwałoby się „wyrazem". Odbija się to między innymi na języku zapytań.

[10] W NKJP natomiast statystyka dotyczy segmentów, czyli także znaków interpunkcyjnych. Warto zdawać sobie sprawę z tej różnicy, choćby wtedy, gdy porównujemy rozmiar tych korpusów (np. 300 mln segmentów w NKJP to nie to samo, co 300 mln użyć słów w RKN).

Rys. 12.2. Zapytanie o wyraz w określonym kontekście znaków interpunkcyjnych.



W uproszczeniu można zatem powiedzieć, że znaki interpunkcyjne nie są samodzielnymi segmentami RKN, a najwyżej parametrami wyrazów (można je uwzględniać jako „sąsiedztwo" wyrazu w zapytaniach).

Jeśli chodzi o wyrażenia w rodzaju *Нью-Йорк*, *какой-нибудь*, *чего-то*, to w RKN dywiz nie rozbija tych słów na dwa; jest to zgodne z intuicją zwykłego nosiciela języka. O takie wyrażenia pytamy tak samo jak o pozostałe (np. *какой-нибудь* jak *какой*, *серый*, *сделать* itd.).

Skrótowce w rodzaju *СССР*, *ВДНХ*, *ФСБ* są przedstawiane jako wyrazy.

Na szczególną uwagę zasługuje segmentacja leksemów wielowyrazowych (ros. неоднословные лексические единицы) w rodzaju *без ведома*. W RKN użycie leksemu tego typu to nie segment, a wiele segmentów[11]. Sprawdźmy to na przykładzie wspomnianego wyżej przyimka *без ведома*. Przy wyszukiwaniu go jako jednego wyrazu (wyszukiwanie według podstawowej formy gramatycznej wyrazu: *Лексико-грамматический поиск*) otrzymujemy wynik zerowy (żadnego użycia). Natomiast jeśli będziemy szukali wystąpień dwóch słów – *без* i *ведомо* – otrzymujemy ponad 500 poświadczenie. W RKN są to zatem dwa wyrazy (segmenty).

[11] W Korpusie Głównym przyjęto zasadę, że anotacja morfologiczna dotyczy każdego wyrazu „ortograficznego" (oddzielanego od innych spacją) (http://ruscorpora.ru/corpora-morph.html). To założenie pociągnęło za sobą odpowiednią segmentację w całym korpusie.

Zasada dzielenia leksemu wielowyrazowego na wyrazy ortograficzne obowiązuje również wtedy, gdy wyrazy te nie stanowią rzeczywiście odrębnych jednostek słownikowych. Np. ciąg literowy *ведома* (który czysto formalnie zinterpretować należy jako dopełniacz lp. rzeczownika *ведомо*) występuje wyłącznie w połączeniach *с ведома* oraz *без ведома*, natomiast leksem **ведомо* nie funkcjonuje samodzielnie we współczesnym języku rosyjskim. Warto zatem pamiętać, że, szukając wystąpień jakiegokolwiek leksemu w RKN, musimy ująć go w zapytaniu zgodnie z zasadami segmentacji przyjętymi w tym korpusie.

Należy też dodać, że na stronie projektu udostępniono listę leksemów wielowyrazowych (wielowyrazowych przyimków, przysłówków itd.), których użycia znaleźć można w RKN (http://www.ruscorpora.ru/obgrams.html). Lista ma charakter pomocniczy.

Rodzaje anotacji wewnętrznej w Korpusie Głównym RKN

Anotacja wewnętrzna dotyczy rozmaitych właściwości wyrazów w dokumentach włączonych do korpusu. W Korpusie Głównym RKN stosowana jest anotacja trzech podstawowych rodzajów: 1. Anotacja gramatyczna, 2. Częściowa anotacja semantyczna, 3. Anotacja tzw. „właściwości dodatkowych"[12].

Anotacja ostatniego typu pozwala na przykład określić, czy dany wyraz znajduje się na końcu, czy na początku zdania, czy po nim albo przed nim występuje jakiś znak interpunkcyjny (przecinek, dwukropek, kropka, nawias itd.), a także pewne inne właściwości (w głównej mierze – rozmaite charakterystyki kontekstu wyrazu).

Anotacja semantyczna dotyczy znaczenia wyrazów. Przeprowadzano ją całkowicie automatycznie, z użyciem programu *Semmarkup* autorstwa Aleksieja Polakowa (http://www.ruscorpora.ru/corpora-sem.html). Anotowano tylko te wyrazy, które są zawarte w słowniku semantycznym, z którego korzysta wspomniany program. Słownik opiera się na klasyfikacji leksyki rosyjskiej, zastosowanej w bazie danych *Лексикограф* (Отдел лингвистических исследований ВИНИТИ РАН под рук. Е.В. Падучевой и Е.В. Рахилиной). Bazę tę jednak poszerzono, klasyfikację udoskonalono, poza tym dodano opis właściwości słowotwórczo-derywacyjnych (http://www.ruscorpora.ru/corpora-sem.html). Stosowany jest tagset bazujący na skrótach angielskojęzycznych, np. t:constr – budynki, pt:part – części itd. Opis tagsetu można znaleźć na stronie: http://www.ruscorpora.ru/corpora-sem.html.

[12] W korpusie dostępne jest też wyszukiwanie uwzględniające budowę słowotwórczą wyrażeń (panel *Словообразование*). Można wpisać np. *дом*, wskazując, że to rdzeń, i w wynikach otrzymamy *домашний*, *надомный*, *домой* itd. Taka opcja sugeruje, że w korpusie pojawiła się anotacja opisująca budowę słowotwórczą wyrazów, obecnie jednak na stronie Korpusu nie ma żadnej wzmianki o tym, nie ma także dokładnego opisu wyszukiwania tego typu.

Właściwości semantyczne, które mogą być przypisane dowolnemu wyrazowi występującemu w tekście korpusu, zależą od części mowy, do której wyraz ten przynależy. Obecnie w Korpusie semantycznie anotowane są tylko rzeczowniki, przymiotniki, przysłówki, zaimki, liczebniki oraz czasowniki. Na pierwszym etapie anotacji semantycznej rzeczownika, przymiotnika, zaimka lub liczebnika określana jest jego kategoria semantyczna. Przykładowo, w korpusie wyodrębnia się 3 kategorie semantyczne rzeczowników:

– r:concr – nazwy obiektów konkretnych (*девочка*, *стол*, *молоко*),

– r:abstr – nazwy abstrakcyjne (*яркость*, *время*),

– r:propn – imiona własne (*Эйнштейн*).

Trzy kategorie wyodrębniono również w przypadku liczebników:

– r:card – liczebniki główne (*два*, *пять*, *десять*),

– r:card:pauc – liczebniki oznaczające liczby małe (*два*, *три*, *четыре*, *оба*, *пол*, *полтора*),

– r:ord – liczebniki porządkowe (*первый*, *второй*, *десятый*).

Kategorii semantycznej nie precyzuje się natomiast przy anotacji czasowników i przysłówków.

Anotowanym semantycznie wyrazom przypisuje się także właściwości w ramach następujących pól znaczeniowych: taksonomia (klasa tematyczna leksemu), mereologia (właściwości związane z relacją 'część – całość', 'element – zbiór'), topologia (status topologiczny oznaczanego obiektu), kauzacja, status czasownika posiłkowego, ocena. Tagi właściwości taksonomicznych przypisuje się jedynie rzeczownikom, przymiotnikom, czasownikom oraz przysłówkom. Por. np.: t:animal – 'zwierzęta' (*корова*, *жираф*, *сорока*), t:food – 'jedzenie i napoje' (*пирог*, *каша*, *молоко*). Właściwości mereologiczne przypisuje się rzeczownikom będącym nazwami obiektów konkretnych i abstrakcyjnych, np.: pt:part – 'części' (*верхушка*, *кончик*, *половина*, a także np. *начало*, *финал*), pt:part & pc:tool:instr – 'części narzędzi' (*топорище*, *лезвие*). Tagi topologiczne dotyczą właściwości semantycznych nazw obiektów konkretnych: top:contain – 'kontenery lub pojemniki' (*кошелек*, *комната*, *озеро*, *ниша*), top:horiz – 'powierzchnie poziome' (*пол*, *площадка*). Stosowane w korpusie tagi kauzacji (ca:caus, ca:noncaus) i tagi statusu posiłkowego czasownika (aux:phase, aux:caus) mogą być przypisane tylko czasownikom. Por. np.: ca:caus – czasowniki kauzatywne (*показать*, *вертеть*), ca:noncaus – czasowniki niekauzatywne (*видеть*, *вертеться*); aux:phase – czasowniki fazowe (*начать*, *продолжать*, *прекратить*), aux:caus – czasowniki posiłkowe kauzatywne (*вызывать*, *привести* (*к*)). W przypadku nazw obiektów konkretnych oraz nazw abstrakcyjnych, przymiotników i przysłówków anotowana jest także ocena. Wiążą się z nią trzy tagi semantyczne: ev – ocena (nieokreślona); ev:posit – ocena pozytywna (*везучий*, *ладный*), ev:neg – ocena negatywna (*продажный*, *сварливый*).

Poza tym, być może nieco nieintuicyjnie, anotacja semantyczna dotyczy też cech słowotwórczych (derywacyjnych). Np. wśród przysłówków wyodrębniane są deminutywy w rodzaju *немножко*, *быстренько* (tag d:dim), przysłówki odrzeczownikowe w rodzaju *вверху*, *дома* (tag der:s), przysłówki odczasownikowe w rodzaju *отродясь*, *стоймя* (tag der:v) i kilka innych typów i odpowiadających im tagów.

Anotacja semantyczna korpusu pozwala, z jednej strony, wyświetlić w wynikach wyszukiwania informacje znaczeniowe o wyrazach, z drugiej zaś, umożliwia określanie klasy semantycznej wyrazu w zapytaniach. Warto dodać, że wyrazy, których znaczenie zinterpretować można wielorako, mają kilka wersji anotacji semantycznej (kilka wariantów właściwości).

Anotacja gramatyczna w korpusie to informacja o każdym wyrazie (ciągu tekstowym zidentyfikowanym jako wyraz rosyjski) zawierająca:

- lemat (podstawowa <słownikowa> forma leksemu, którego użyciem jest dany wyraz)[13];
- część mowy, do której należy reprezentowany przez wyraz leksem;
- klasyfikujące właściwości gramatyczne reprezentowanego przez wyraz leksemu (wartość kategorii, ze względu na którą leksem się nie odmienia, a jednocześnie – tym różnić się może od innych leksemów tej samej części mowy, np. rodzaj rzeczownika: *река* – r.ż., *снег* – r.m. itd.);
- właściwości fleksyjne wyrazu będącego realizacją określonego leksemu w tekście (np. wartość kategorii przypadku dla rzeczownika, wartość kategorii osoby dla czasownika itd.).

Dla przykładu, anotacja gramatyczna wyrazu *звонкими* stanowiłaby informację: „*звонкий*" (lemat), „przymiotnik", „liczba mnoga", „narzędnik", „pełna forma". W przypadku języków fleksyjnych, z których jednym jest język rosyjski, anotacja gramatyczna, a zwłaszcza informacja o lematach, jest szczególnie ważna. Pozwala ona zrealizować możliwości wyszukiwania wszystkich form fleksyjnych danego leksemu na podstawie jednego zapytania (o lemat)[14]. Oprócz tego, dzięki anotacji gramatycznej użytkownik budować może zapytania o wystąpienia określonej klasy gramatycznej, wybranej wartości danej

[13] Warto podkreślić, że lematu nie należy utożsamiać z jednostką słownikową (leksemem w sensie lingwistycznym). Możliwa jest sytuacja, kiedy ten sam lemat mają wyrazy stanowiące różne leksemy. Przykładowo, w tekście może wystąpić wyraz *ужа* (formę dopełniacza rzeczownika *уж*) i *уж*-partykułę, którym w RKN zostaje przypisany ten sam lemat *уж*. Rozróżnienie tych wyrazów w anotacji polega na przypisaniu im różnej wartości części mowy.

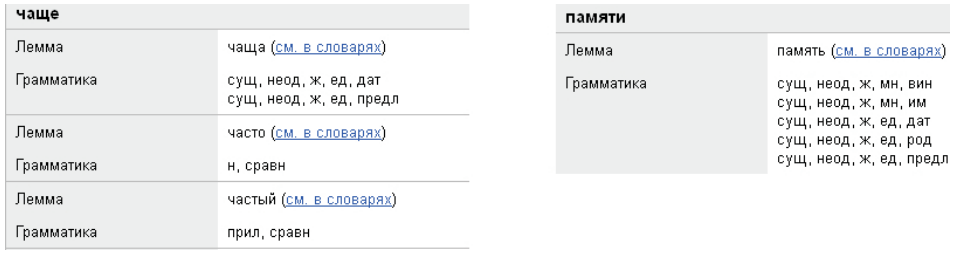
[14] Dla przykładu, pytamy o *текст*, a w wynikach mamy zdania zawierające też inne formy gramatyczne tego leksemu: *текста*, *тексту*, *тексте*, *тексты* itd. W rozumieniu technicznym szuka się wszystkich segmentów o danym lemacie i – ewentualnie – określonych w zapytaniu właściwościach gramatycznych w rodzaju części mowy, pozwalających odróżnić wystąpienia jednego leksemu od wystąpień leksemów homonimicznych, np.: czasownik *знать* i rzeczownik *знать*.

kategorii fleksyjnej lub klasyfikującej, konkretnej wartości kategorii fleksyjnej lub klasyfikującej wśród wyrazów klasy gramatycznej itd. Możemy, dla przykładu, pytać o <dowolny> „rzeczownik", „rzeczownik rodzaju nijakiego", „narzędnik", „czasownik 1 os. liczby mnogiej". Omawiane opcje pozwalają na wyszukiwanie także konstrukcji składających się z wyrazu o pewnych charakterystykach gramatycznych i wyrazu o danym lemacie (np. <dowolny> „przymiotnik" + *домик*).

Przyjęty w RKN standard gramatyczny opiera się na podstawowym słowniku Andrieja Zalizniaka *Грамматический словарь русского языка* (Зализняк 19772003)[15], informującym o odmianie i akcentowaniu wyrazów rosyjskich. Z tego słownika korzysta też program *Mystem* (autor: Ilja Segałowicz, firma Yandex), który był używany podczas automatycznej lematyzacji i anotacji gramatycznej tekstów w Korpusie Głównym[16]. Program ten nie usuwał jednak homonimii gramatycznej, tzn. tworzył wszystkie możliwe interpretacje gramatyczne występujących w tekście form wyrazowych (np. *уж* jako rzeczownik 'żmija', a także jako partykuła) (Ляшевская, Плунгян, Сичинава 2005: 117). Po obróbce tym programem teksty są zatem niedezambiguowane: anotacja uwzględnia wszystkie możliwe interpretacje gramatyczne każdego występującego w nich wyrazu. W terminologii RKN są to teksty z nieusuniętą homonimią gramatyczną.

W RKN z homonimią gramatyczną mamy do czynienia na poziomie nie tylko anotacji lematów wyrazu, ale też przy opisie jego formy fleksyjnej. Por. rys. 12.3 (trzy możliwe lematy wyrazu *чаще* z pewnego tekstu i możliwe interpretacje fleksyjne formy *памяти* odzwierciedlane w anotacji).

Rys. 12.3. Kilka interpretacji segmentu na poziomie anotacji gramatycznej.



чаще	
Лемма	чаща (см. в словарях)
Грамматика	сущ, неод, ж, ед, дат сущ, неод, ж, ед, предл
Лемма	часто (см. в словарях)
Грамматика	н, сравн
Лемма	частый (см. в словарях)
Грамматика	прил, сравн

памяти	
Лемма	память (см. в словарях)
Грамматика	сущ, неод, ж, мн, вин сущ, неод, ж, мн, им сущ, неод, ж, ед, дат сущ, неод, ж, ед, род сущ, неод, ж, ед, предл

[15] W standardzie RKN są też jednak pewne odstępstwa, omawiane w artykule Laszewskiej, Plungiana i Siczinawy *О морфологическом стандарте Национального корпуса русского языка* (Ляшевская, Плунгян, Сичинава 2005: 120-130).

[16] Program ten może nie tylko interpretować znane jego słownikowi wyrażenia, ale też generować hipotezy co do wyrażeń, których w nim nie ma.

Niektóre teksty zostały jednak zdezambiguaowane (ujednoznacznione pod względem interpretacji gramatycznych wyrazów). W takich tekstach poprawną interpretację wybierał lingwista, pozostałe zaś były usuwane[17]. Tekstów takich jest w korpusie znacznie mniej niż tekstów z nieusuniętą homonimią gramatyczną. Wyszukiwanie w Korpusie Głównym odbywa się w tekstach obu typów jednocześnie. Przewidziano jednak możliwość wyszukiwania w tekstach wyłącznie jednego z nich (tylko w tekstach z usuniętą homonimią gramatyczną albo tylko w tekstach, z których jej nie usuwano): opcje *Только тексты со снятой грамматической омонимией*, *Только тексты с неснятой грамматической омонимией* w zakładce *задать подкорпус* (http://ruscorpora.ru/mycorpora-main.html).

Warto zaznaczyć, że teksty z usuniętą homonimią gramatyczną były anotowane także pod względem cech akcentologicznych.

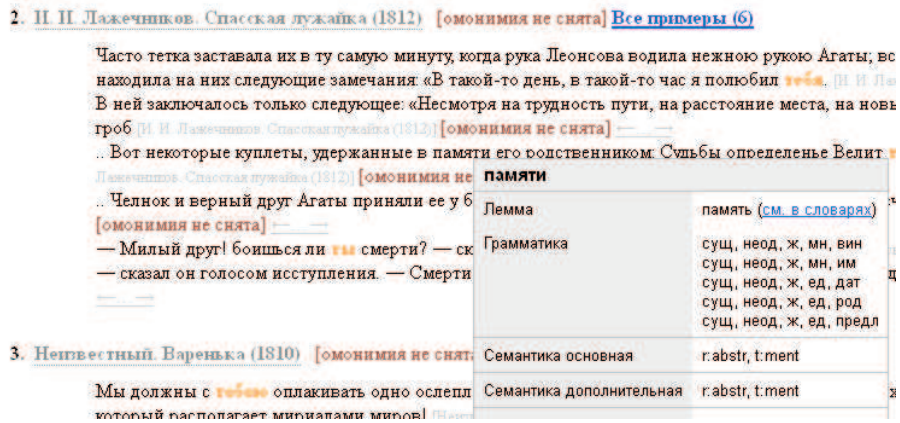
Opis tagsetu gramatycznego i używanych przez niego kategorii znaleźć można na stronie http://ruscorpora.ru/corpora-morph.html. Do swobodnego, świadomego korzystania z narzędzi gramatycznych RKN może on czasami nie wystarczać. Pewne decyzje co do przyjętej struktury gramatycznej są bowiem nieoczywiste, nieintuicyjne, zaś opis nie zawiera bardziej szczegółowych wyjaśnień.

Wśród części mowy wyodrębniono np. predykatywy. Natomiast wielu badaczy rozpatruje predykatywy jako wyrazy z różnych części mowy, występujące w zdaniach pewnego typu w funkcji składniowej orzeczenia (Плотникова 1997). Czy podejście RKN ma jakieś korzyści praktyczne? Wydaje się, że minimalne. Zgodnie z podjętą w RKN decyzją należy wyodrębniać np. *холодно*-przysłówek (np. *Он холодно посмотрел на сестру*) i *холодно*-predykatyw (np. *Мне холодно*). W praktyce zaś każde wystąpienie formy *холодно* w korpusie interpretowane jest na oba sposoby (za wyjątkiem tekstów z usuniętą homonimią gramatyczną): predykatywne użycia otrzymują także interpretację przysłówkową, a przysłówkowe – predykatywną. Wprowadzono zatem dyskusyjne rozróżnienie teoretyczne, z którego korzyść praktyczna jest wątpliwa.

Na stronie RKN brakuje zatem bardziej szczegółowego opisu tagsetu i reprezentowanych w nim kategorii gramatycznych. Wiele wątpliwości można jednak rozwiązać, analizując wyniki wyszukiwań. Na stronie z wynikami można mianowicie kliknąć dowolny wyraz w znalezionych zdaniach (nie tylko ten, którego szukano) i przeczytać informacje o nim w wyskakującym okienku (zob. rys. 12.4).

[17] Korzystano też z dodatkowych programów i skryptów ułatwiających pracę lingwisty, np. z programu *Диалинг*, „częściowo prognozującego poprawne opisy fleksyjne słowoform homonimicznych" (Ляшевская, Плунгян, Сичинава 2005: 117).

Rys. 12.4. Wyświetlanie informacji gramatycznej po kliknięciu wyrazu.



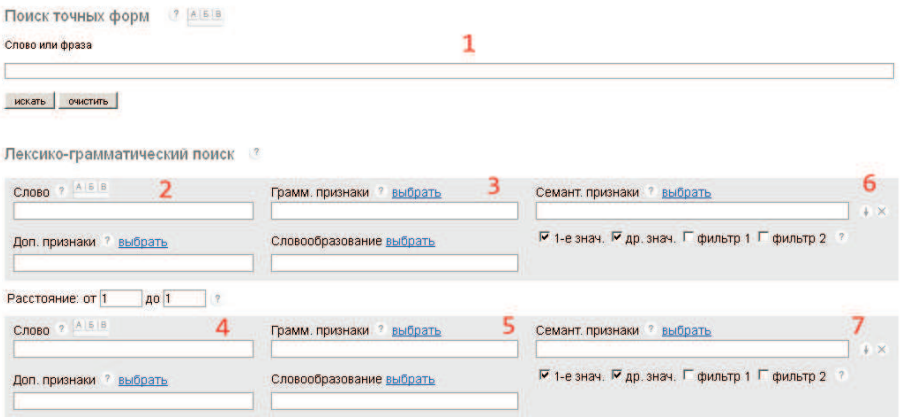
Podstawowe zapytania

Dokładny opis zapytań w RKN znaleźć można w instrukcji użytkownika: http://www.ruscorpora.ru/instruction-main.pdf (rozdział *Поиск в Корпусе*, s. 37-61). Tutaj zaś omówimy jedynie kilka podstawowych przykładów.

Formularz wyszukiwania w Korpusie Głównym RKN umieszczono na stronie http://www.ruscorpora.ru/search-main.html. Podzielono go na kilka pól, których się używa w zależności od rodzaju zapytania.

Wyszukiwania dokładnie sprecyzowanej przez nas formy wyrazowej bądź ciągu kilku takich form można dokonać, wpisując odpowiednie zapytanie w polu *Поиск точных форм* (zob. pole 1 na rys. 12.5). Wpisywać tu można wyrazy w dowolnej formie fleksyjnej, nie tylko kanonicznej. Inne formy wyrazów niż podana nie będą wyszukiwane. Przykładowo, jeśli w polu wpiszemy wyraz *вечера*, wyszukiwarka znajdzie formę *вечера*, a pominie inne formy fleksyjne leksemu *вечер*: *вечере*, *вечером*, *вечерах* itd.

Rys. 12.5. Formularz wyszukiwania w Korpusie Głównym RKN.



Tak samo pytamy o grupę wyrazów, np. *теплым вечером* (po prostu wpisujemy interesujące nas połączenie w polu tekstowym i klikamy przycisk *искать*)[18]. W tym trybie wyszukiwania nie są rozróżniane wielkie i małe litery. Dla przykładu, zapytania *вечером*, *Вечером*, *ВЕЧЕРОМ* czy *ВеЧЕРОМ* będą miały ten sam efekt: wyszukiwana będzie każda możliwa kombinacja wielkich i małych liter w tym słowie (zarówno forma *вечером*, jak i *Вечером* itd.).

Dowolne inne zapytania (np. o kilka wariantów ortograficznych słowa, kategorię gramatyczną czy lemat w określonej kategorii gramatycznej) wprowadzać należy w polach oznaczonych wspólną nazwą *Лексико-грамматический поиск*.

O lemat pytamy, wpisując go w polu *Слово* (pole 2 na rys. 12.5). Powinno to być słowo rosyjskie w postaci kanonicznej (słownikowej). W praktyce wyszukiwane będą wszystkie formy wyrazowe, których lemat pokrywa się z wpisanym przez nas słowem. Jeśli, na przykład, wprowadzimy wyraz *человек*, wyniki obejmą też zdania z formami *человеком*, *люди*, *людям* itd. Musimy pamiętać, że ten sam lemat mają nie tylko formy fleksyjne leksemu, ale też leksemy homonimiczne. Przykładowo, wpisując *течь*, znajdziemy wystąpienia zarówno *течь*-czasownika, jak i *течь*-rzeczownika. Jeśli interesuje nas tylko jeden z homonimów, musimy podać w zapytaniu jego właściwości gramatyczne, semantyczne itd., które pozwolą odróżnić go od innych wyrazów o tym samym lemacie.

[18] Z technicznego punktu widzenia mamy do czynienia z narzędziem wyszukiwania segmentów lub ciągów segmentów. Nie należy myśleć o nim jako o narzędziu wyszukiwania dowolnego ciągu znaków. Dla przykładu, w wyniku wyszukiwania *сегм* – ciągu liter stanowiącego jedynie część istniejącego w języku słowa – otrzymujemy 0 wyników, natomiast po wpisaniu *сегментов* – pełnego wyrazu (segmentu) – uzyskujemy 296 jego wystąpień. (Podobnie: *всех сегм* – 0 wyników, *всех сегментов* – 8 poświadczeń.).

Należy unikać stosowania znaków interpunkcyjnych przy wyszukiwaniu danego typu. Wyszukiwanie w tym trybie nie rozróżnia bowiem liter i niektórych znaków, np. kropki, znaku zapytania, wykrzyknika. Po wpisaniu takiego znaku interpunkcyjnego tuż po wyrazie wyszukiwany jest segment, który składałby się z danego ciągu znaków (łącznie ze znakiem interpunkcyjnym, traktowanym jako litera). Prowadzi to do błędnego wyszukiwania i, w konsekwencji, braku wyników. Przykładowo, wyszukiwarka nie może znaleźć w korpusie żadnego wystąpienia *ужина. вечером* (wyraz kończący zdanie, kropka i wyraz zaczynający kolejne zdanie), korpus natomiast zawiera tekst, który mógłby pasować, gdyby kropkę nie utożsamiano z częścią zapytania o segment: *Шитьё у девок обычно продолжалось до ужина. Вечером же невесту мог навещать жених* [...]. (Możemy się o tym przekonać po zapytaniu o segment *вечером*).

Przecinek przy wyszukiwaniu tego typu jest traktowany jako brak znaku. Wyszukiwarka interpretuje w ten sposób zarówno przecinek wpisany przez użytkownika w zapytaniu, jak też przecinek, który znajdzie ona w tekście korpusu; np. wynikiem wyszukiwania: ‚*ужина* lub *ужина*, (przecinek przed lub po wyrazie w zapytaniu) będą też zdania, zawierające segment *ужина* (bez przecinka obok), a po wyszukiwaniu *ужина вечером* otrzymamy także zdania zawierające przecinek pomiędzy dwoma segmentami ([...] *опять зашагать по Парижу до ужина, вечером мучить жену, затихать к девяти* [...]). Nie wiadomo, co jest powodem takiej osobliwości (usterka oprogramowania czy świadoma decyzja twórców).

Właściwości gramatyczne, semantyczne itd. szukanego wyrazu wpisujemy w odpowiednich polach tekstowych (*Грамм. признаки*, *Семант. признаки*, *Доп. признаки*, *Словообразование*) obok okienka *Слово*. Przykładowo, by znaleźć wystąpienia czasownika *течь*, w opcji *Слово* wpisujemy *течь*, a w polu *Грамм. признаки* wpisujemy V (zob. pola 2 i 3 na rys. 12.5.). Jeśli zaś szukamy rzeczownika o tym samym lemacie, w okienku *Грамм. признаки* zamiast V podajemy S[19]. Nie musimy pamiętać tekstowego oznaczenia kategorii gramatycznych, semantycznych i innych opcji. Można kliknąć link *выбрать* przy każdym z tych pól (właściwości gramatycznych, semantycznych itd.), a następnie w pojawiającym się okienku zaznaczyć potrzebne opcje i kliknąć przycisk *OK* – potrzebny nam element zapytania pojawi się w odpowiedniej pozycji formularza.

Wyszukiwanie zaawansowane *Лексико-грамматический поиск* pozwala też szukać wystąpień określonej kategorii gramatycznej. Zapytać możemy np. o wyraz rodzaju męskiego w bierniku liczby pojedynczej. W polu *Слово* nie podajemy wtedy żadnego lematu, a w okienku *Грамм. признаки* wpisujemy: *m*, *acc*, *sg*. Znalezione zostaną zdania zawierające wyrazy, które spełniają określone przez nas kryteria (w danym przypadku będą to rozmaite rzeczowniki, przymiotniki, liczebniki porządkowe itd., np. *И это происходит несмотря на непомерный рост цен на энергоносители.* [Встречи В.В. Путина на Си-Айленде // «Дипломатический вестник», 2004], *Обновить гардероб? Легко! Сочетания, создающие стиль* [Обновить гардероб? Легко! Сочетания, создающие стиль // «Даша», 2004]).

Możemy też wpisywać pytania dotyczące grupy lematów. Każdy lemat z wyszukiwanego przez nas połączenia wpisujemy w osobne pole *Слово*. Jeśli, na przykład, szukamy połączenia *красный домик* w dowolnych formach przypadkowych, liczbowych itd., to w pierwszym okienku *Слово* wpisujemy *красный*, a w drugim, znajdującym się niżej polu *Слово – домик* (zob. pola 2 i 4 na rys. 12.5).

Szukając połączeń, możemy korzystać z kryteriów gramatycznych. Dość często zdarza się, że musimy znaleźć połączenie dwóch słów, z których jedno jest ustalone, a drugim może być dowolne słowo o określonej charakterystyce gramatycznej. Możemy np. zapytać o połączenie „dowolny przymiotnik w liczbie mnogiej + *домик* w liczbie mnogiej": w pierwszym polu *Слово* nic nie wpisujemy, a w drugim (*Грамм. признаки*) wprowadzamy A,pl (zob. pola 2 i 3 na rys. 12.5); w kolejnym polu *Слово* wprowadzamy *домик*, a w okienku *Грамм. признаки* obok – pl (pola 4, 5 na rys. 12.5). Znalezione zostaną zdania z połączeniami w rodzaju *гостевые домики*, *крохотных домиках*.

[19] Musimy jednak pamiętać, że w pewnej części korpusu, anotowanej automatycznie, nie usunięto homonimii gramatycznej. Dlatego czasami do efektywnego korzystania z opcji wyszukiwania według właściwości gramatycznych może być konieczne zaznaczenie opcji wyszukiwania wyłącznie w tekstach z usuniętą homonimią w zakładce *задать подкорпус*.

Domyślną opcją jest szukanie dwóch lematów – przewidziano to w formularzu zapytania. Można też tworzyć zapytania o większe grupy lematów. Miejsce na nowy lemat dodajemy, klikając przycisk dodania przy lemacie, pod którym chcemy go umieścić w formularzu zapytania (zob. przyciski 6, 7 na rys. 12.5).

Na stronie *Образовательный портал национального корпуса русского языка* podano informację, że pytać też można o kilka lematów w ramach zapytania o jeden (ściśle rzecz biorąc, jest to pytanie o alternatywę), rozdzielając je pionową kreską, np. *ветер|дождь|солнце|град* (http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=73&Itemid=84#7). Obecnie jednak to wyszukiwanie nie działa.

Funkcje i narzędzia wyszukiwarki RKN

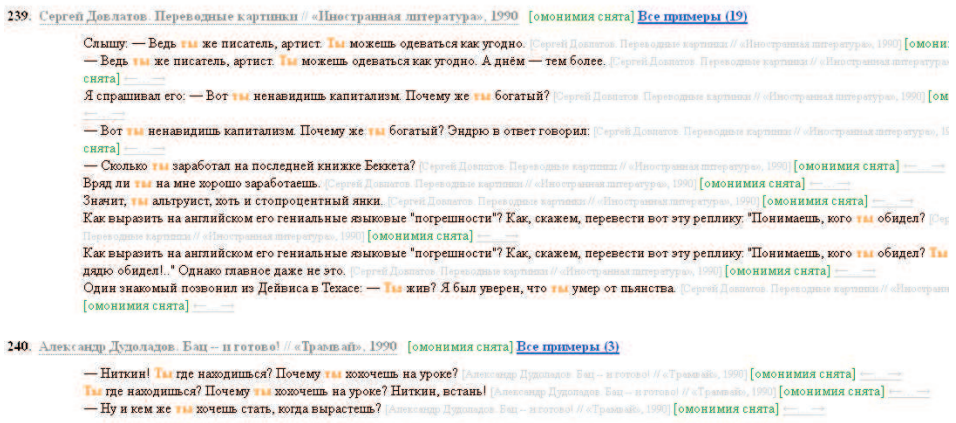
Część opcji wyszukiwarki opisano już wyżej. Użytkownik ustalić może własny podkorpus. Zapytanie o semantyczne, fleksyjne, słowotwórcze i inne właściwości wyrazu (lub wyrazów) ułatwia specjalny interfejs graficzny. Pytający może wyświetlić całą informację o każdym wyrazie znajdującym się w znalezionym fragmencie tekstowym (nie tylko o tym, którego szukano).

Wpisując słowa rosyjskie do wyszukiwarki, możemy skorzystać z cyrylickiej klawiatury wirtualnej. Uruchamiamy ją, kliknięciem przycisku *Virtual keyboard*, znajdującego się obok pól tekstowych. Ta funkcja ułatwia korzystanie z korpusu użytkownikom będącym nosicielami języków zapisywanych alfabetem łacińskim.

W pracy z korpusem ważne są też opcje sortowania wyników i opcje formatu wyświetlania.

Domyślnie wyświetlane są pojedyncze zdania (przykłady), w których jeden raz lub więcej wystąpiło wyszukiwane wyrażenie. Wyrażenie to wyodrębniane jest kolorem pomarańczowym.

Rys. 12.6. Przykłady pochodzące z tego samego tekstu wyświetlane razem.



Przykłady pochodzące z tego samego dokumentu (tekstu) są grupowane razem (zob. rys. 12.6). Każda grupa ma przypisany numer oraz tytuł. W tytule opisane są właściwości tekstu (autor, tytuł itd.), poza tym podana jest informacja, czy usunięto w nim homonimię gramatyczną. Domyślnie wyświetlanych jest maksymalnie 10 przykładów. Jeśli jest ich więcej, można kliknąć link *Все примеры* i obejrzeć pominięte zdania (liczba znalezionych zdań podawana jest w nawiasach w tytule linku). Można też zmienić domyślną wartość liczby przykładów z jednego tekstu w zakładce *Настройки* (na górze w prawym rogu). Jeśli chcemy obejrzeć większy kontekst, w którym wystąpiło poświadczenie, klikamy przycisk „←...→" po prawej. Pozwoli to wyświetlić fragment o maksymalnej długości 5 zdań.

Domyślnie na jednej stronie wyświetlane są przykłady pochodzące z 10 tekstów. Zmiana tych wartości jest dostępna także w zakładce *Настройки*. W tej zakładce można też ustawić sposób sortowania wyników. Domyślne sortowanie odbywa się według daty powstania tekstu w kolejności od najpóźniejszej do najwcześniejszej.

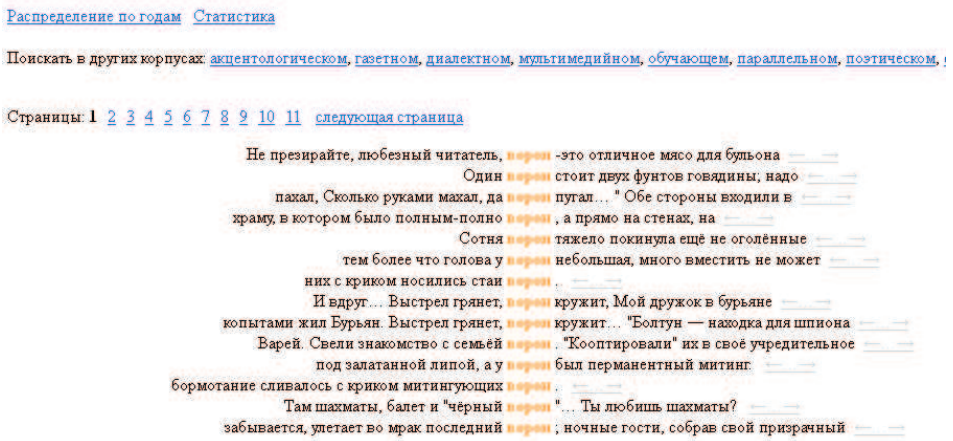
Istotne jest sortowanie według kontekstu – prawego bądź lewego. Zdania są sortowane alfabetycznie według wyrazu, który znajduje się najbliżej wyszukiwanego (po jego prawej bądź lewej stronie). W przypadku zbieżności uwzględniany jest drugi wyraz (po lewej lub prawej) itd.

Poświadczenia można sortować także według form, zwróconych w wyniku zapytania o lemat. Wystąpienia zawierające tę samą formę fleksyjną leksemu można dzięki temu sortowaniu wyświetlać razem. (Przykładowo, jeśli posortujemy w ten sposób wyniki zapytania o lemat *добрый*, to poświadczenia z formą *добр* znajdą się przed takimi, w których znaleziono formę *добрым*). Dalsze sortowanie odbywa się według prawego bądź lewego kontekstu. Warto zauważyć, że przy takich trybach sortowania przykłady pochodzące z tego samego tekstu nie są grupowane razem.

Można też skorzystać z innego trybu wyświetlania: z formatu KWIC. Wyświetlany jest tylko mały fragment zdania[20], zaś wyszukiwany wyraz (leksem, połączenie wyrazów itd.) jest umieszczany w środku (zob. rys 12.7). Ułatwia to analizowanie jego prawego bądź lewego kontekstu.

[20] Jego długość można ustalić w panelu *Настройки*.

Rys. 12.7. Wyniki wyszukiwania wyświetlane w formacie KWIC.



Na stronie wyników wyszukiwania (niezależnie od tego, czy wyświetlane są standardowo, czy też w formacie KWIC) znajdziemy także podstawowe dane statystyczne:

– dane o objętości korpusu bądź podkorpusu, w którym odbywało się wyszukiwanie oraz informacja o liczbie wyników zapytania wyświetlane są na górze strony;
– liczbę lematów, a także form wyrazowych, zwróconych jako wyniki zapytania i wyświetlonych na stronie przeglądanej przez użytkownika, znajdziemy na jej dole;
– poza tym, klikając link *Распределение по годам*, wyświetlić można wykres przedstawiający częstotliwość występowania wyszukiwanego wyrazu (lub połączenia) w tekstach z różnych lat. Ustalić też można okres, który nas interesuje (domyślnie wyświetlane są dane z okresu 1800-2010). Poza tym można też obejrzeć dane o liczbie wystąpień w poszczególnych latach w postaci tabeli (link *Показать таблицы* pod wykresem na dole). Pozwala to śledzić dynamikę zmian użycia wyrazu[21];

[21] Narzędzie dostępne również poprzez stronę http://www.ruscorpora.ru/ngram.html. W polu tekstowym na górze wpisujemy dowolne wyrażenie lub połączenie, którego statystyka w korpusie nas interesuje. Można też porównać ze sobą statystykę kilku wyrazów lub połączeń. Wprowadzamy je wówczas kolejno, oddzielając przecinkami.

Wpisywane wyrazy traktowane są jako formy wyrazowe, nie zaś lematy. Cyfry w tabelach przedstawiają frekwencję całkowitą wystąpień wyrazu w poszczególnych latach, natomiast wykres przedstawia wartości względne *ipm* (względną częstotliwość użycia wyrazu w określonym roku). Jest ona definiowana jako liczba użyć słowa w ciągu roku dzielona przez objętość korpusu z tego roku i pomnożona przez 1 milion (http://www.ruscorpora.ru/help-ngram.html), uwzględnia zatem nie tylko liczbę wystąpień wyrażenia, ale

– korpus pozwala także wyświetlić szczegółową statystykę występowania wyszukiwanego wyrazu (lub grupy wyrazów) u różnych autorów, w tekstach z różnych dziedzin (literatura piękna, publicystyka, teksty dydaktyczno-naukowe itd.), typów (powieść, list itd.), o różnej tematyce (technika, sztuka i kultura itd.) czy gatunku (np. proza historyczna, literatura dziecięca). Służy temu link *Статистика* (obok linku *Распределение по годам*) w górnym lewym rogu strony z wynikami. Korzystając z tego narzędzia, dowiedzieć się można np., że forma wyrazowa *регион* najczęściej występuje w tekstach nienależących do literatury pięknej (ponad 650 razy), a w literaturze pięknej bardzo rzadko (tylko 13 wystąpień), a forma *ворон* najczęściej wystąpiła w tekstach E. Szwarca (64 wystąpienia w jednym tekście) oraz M. Priszwina (w 14 tekstach tego pisarza 43 razy łącznie).

W RKN bardzo brakuje narzędzi do ekscerpcji kolokacji[22] (narzędzi do generowania list kolokacyjnych na podstawie obliczeń T-score, Chi^2 etc.). Pracować możemy jedynie z konkordancjami, opierając się na ogólnych danych statystycznych (liczbie słów i zdań w korpusie oraz zwróconych poświadczeń wyszukiwania).

Zastosowania Rosyjskiego Korpusu Narodowego

Jak już wspominaliśmy, w RKN nie ma narzędzi do automatycznej ekscerpcji kolokacji. Sprawia to, że przeprowadzenie badań lingwistycznych opartych na zaawansowanej analizie statystycznej dużego materiału (metodach statystycznych pochodzących z nauk ścisłych) jest na nim raczej niemożliwe, przynajmniej do chwili, kiedy narzędzia takie zostaną zaimplementowane.

Natomiast zaawansowany system anotacji zewnętrznej i wewnętrznej oraz prostsze narzędzia statystyczne czynią z korpusu poważne źródło danych czysto lingwistycznych i obserwacji ilościowych. Używając korpusu, możemy m.in. stwierdzić, czy dane połączenie wyrazów występuje w języku, czy też nie (np. *еле* czy *едва* w kontekście liczebnika porządkowego), czy wyraz pojawia się w tekstach pewnego gatunku, kiedy wszedł w użycie (np. w tekstach z XVIII i XIX wieku nie ma żadnego użycia, w tekstach z lat 30. – kilka użyć, a w tekstach z drugiej połowy XX w. – kilka tysięcy) itd.

---

też liczbę słów w korpusie z tego roku. Bardziej szczegółową informację można znaleźć na stronie: http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=303&Itemid=130#9.

[22] Z punktu widzenia tych narzędzi kolokacja to takie połączenie szukanego słowa z innym słowem, które w korpusie występuje częściej niż można byłoby prognozować na podstawie danych statystycznych o pojedynczych słowach wchodzących w jego skład. Takie połączenie może się okazać lingwistycznie nieprzypadkowe (tzn. sygnalizować np. frazeologizm, jednostkę wielowyrazową itd.). Narzędzia do automatycznego wyszukiwania kolokacji nie zostały wbudowane do wielu korpusów.

RKN posłużył do wielu badań lingwistycznych dotyczących zjawisk morfologicznych, składniowych, semantycznych, leksykalnych, pragmatycznych i akcentologicznych. Niektóre z tych prac przedstawiono na stronie *Образовательный портал Национального корпуса русского языка*. Są to między innymi artykuły (wykaz ponad 50 tego typu publikacji obejrzeć można w zakładce: http://studiorum.ruscorpora.ru/index.php?option=com_docman&Itemid=67), kilka prac doktorskich (http://studiorum.ruscorpora.ru/index.php?option=com_docman&Itemid=106), wiele referatów konferencyjnych (http://studiorum.ruscorpora.ru/index.php?option=com_docman&Itemid=103). Prace te dotyczą zarówno bardziej ogólnych zjawisk (np. Татевосов 2009), jak też pojedynczych leksemów (np: Оскольская, Холодилова 2009).

Z korpusu korzystano też np. przy opracowaniu podręcznika praktycznej stylistyki rosyjskiej dla uczelni wyższych autorstwa A. Lewinzona oraz E. i N. Dobruszyn (Левинзон, Добрушина, Добрушина 2010), gdzie jeden z rozdziałów zawiera ćwiczenia z wykorzystaniem RKN (http://studiorum.ruscorpora.ru/index.php?option=com_content&view=article&id=198&Itemid=117).

Na podstawie materiałów z RKN opublikowano również szereg eksperymentalnych słowników elektronicznych (http://dict.ruslang.ru/), które powstały z udziałem pracowników Zakładu Lingwistyki Korpusowej i Poetyki Lingwistycznej Instytutu Języka Rosyjskiego im. W. Winogradowa Rosyjskiej Akademii Nauk (Отдел корпусной лингвистики и лингвистической поэтики Института русского языка им. В.В. Виноградова РАН). Są to mianowicie: *Грамматический словарь новых слов русского языка* (Гришина, Ляшевская), *Новый частотный словарь русской лексики* (Ляшевская, Шаров), *Словарь русской идиоматики. Сочетания слов со значением высокой степени* (Кустова) oraz *Словарь глагольной сочетаемости непредметных имен русского языка* (Бирюк, Гусев, Калинина).

W najbliższych latach można się spodziewać przeprowadzenia wielu nowych badań i publikacji prac opracowanych na materiale z RKN.

Bibliografia

NKJP = Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.) (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa. Dostęp w internecie: http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf (1 III 2013).

Бирюк О.Л., Гусев В.Ю., Калинина Е.Ю. *Словарь глагольной сочетаемости непредметных имен русского языка*. Dostęp w internecie: http://dict.ruslang.ru/abstr_noun.php (1 III 2013).

Гришина Е.А., Ляшевская О.Н. *Грамматический словарь новых слов русского языка*. Dostęp w internecie: http://dict.ruslang.ru/gram.php (1 III 2013).

Зализняк А.А. (19772003). *Грамматический словарь русского языка*, Москва (wyd. 1 – 1977, wyd. 4 – 2003).

Кустова Г.И. *Словарь русской идиоматики. Сочетания слов со значением высокой степени.* Dostęp w internecie:
http://dict.ruslang.ru/magn.php (1 III 2013).

Левинзон А.И., Добрушина Е.Р., Добрушина Н.Р. (2010). *Практикум по культуре речи. Учебное пособие для вузов.* (ред.) Левинзон А.И. Москва.

Ляшевская О.Н., Плунгян В.А., Сичинава Д.В. (2005). *О морфологическом стандарте Национального корпуса русского языка.* W: *Национальный корпус русского языка: 2003-2005. Результаты и перспективы.* Москва, с. 111--135. Dostęp w internecie:
http://ruscorpora.ru/sbornik2005/08lashevs.pdf (1 III 2013).

Ляшевская О.Н., Шаров С.А. *Новый частотный словарь русской лексики.* Dostęp w internecie:
http://dict.ruslang.ru/freq.php (1 III 2013).

Оскольская С., Холодилова М. (2009). *Некоторые особенности употребления предлога ЁкромеV.* „Русская филология" 20, Сборник научных работ молодых филологов. Tartu.

Плотникова (Робинсон) В.А. (1997). *Предикативы.* W: *Русский язык. Энциклопедия*, Изд. 2., (ред.) Караулов Ю.Н., Москва, s. 368-369.

Плунгян В.А. (2005). *Зачем нужен Национальный корпус русского языка? Неформальное введение.* W: *Национальный корпус русского языка: 2003-2005.* Москва, s. 6-20. Dostęp w internecie:
http://ruscorpora.ru/sbornik2005/02plu.pdf (1 III 2013).

Сичинава Д.В. (2005). *Национальный корпус русского языка: очерк предыстории.* W: *Национальный корпус русского языка: 2003-2005.* Москва, s. 21-30. Dostęp w internecie:
http://ruscorpora.ru/sbornik2005/03sitch.pdf (1 III 2013).

Татевосов С.Г. (2009). *Множественная префиксация и анатомия русского глагола.* W: *Корпусные исследования по русской грамматике*, (red.) Киселева К.Л., Плунгян В.А., Рахилина Е.В., Татевосов С.Г., Москва, s. 92-156.

13. Praktyczny przewodnik po korpusie języka ukraińskiego

Natalia Kotsyba
Polska Akademia Nauk
Instytut Podstaw Informatyki

Informacje ogólne

Korpusy języka ukraińskiego są obecnie rozwijane w kilku ośrodkach, zarówno na Ukrainie, jak i poza jej granicami. Na szczególną uwagę zasługują dwa z nich: Ukraiński Narodowy Korpus Lingwistyczny (w oryginale „Український Національний Лінгвістичний Корпус", dalej będziemy używać skrótu UNLK) opracowywany głównie w latach 2005-2010 pracowników instytucji „Ukraiński Zbiór Lingwistyczno-Informacyjny" (w oryginale „Український Мовно-Інформаційний Фонд", dalej będziemy korzystać ze skrótu UFLI) przy Narodowej Akademii Nauk Ukrainy w Kijowie pod kierownictwem prof. Wołodymyra Szyrokowa, znajdujący się pod adresem http://lcorp.ulif.org.ua/virt_unlc/, oraz Korpus Języka Ukraińskiego (w oryginale „Корпус Української Мови", dalej będziemy używać skrótu KUM), budowany od 2011 w Laboratorium Lingwistyki Komputerowej na Uniwersytecie Narodowym im. Tarasa Szewczenki w Kijowie pod kierownictwem dr Natalii Darczuk, dostępny w sieci pod adresem http://www.mova.info/corpus.aspx.

Język ukraiński jest reprezentowany również w kilku korpusach równoległych, dostępnych publicznie: równoległy korpus rosyjsko-ukraiński w ramach projektu NKJR[1] (ok. 38 mln słów w obu językach), polsko-ukraiński korpus równoległy PolUKR[2] (3,5 mln słów w obu językach), wielojęzyczny korpus równoległy ParaSol[3] (ok. 1 mln słów w tekstach ukraińskich). Wszystkie trzy korpusy równoległe umożliwiają wyszukiwanie według lematów oraz informacji gramatycznych. Dostępny w internecie w postaci konkordancji jest także wielojęzyczny korpus napisów filmowych, stworzony w Kijowskim Narodowym Uniwersytecie Lingwistycznym[4] (Lebediew: 36-37), zawierający ok. 200 tys.

[1] http://www.ruscorpora.ru/search-para-uk.html.
[2] http://www.domeczek.pl/∼polukr/index.php?option=welcome.
[3] http://parasol.unibe.ch/.
[4] http://complinguide.com.ua/Corpus.aspx.

wyrazów ukraińskich. Nie jest on jednak lematyzowany ani tagowany morfologicznie.

Oprócz istniejących, publicznie dostępnych zasobów, w literaturze przedmiotu można znaleźć także poważne deklaracje prac nad podobnymi zasobami (Demska-Kulczycka 2005, 2011), które jak na razie nie doczekały się realizacji praktycznej.

Miano narodowego korpusu języka ukraińskiego nosi UNLK, jest on także jest kilkakrotnie większy objętościowo od KUM oraz bardziej szczegółowo opisany, ma też dłuższą historię rozwoju. Jednak ze względu na brak bezpośredniego dostępu do niego oraz nieudostępnioną informację gramatyczną będziemy opisywać oba wymienione korpusy jednojęzyczne. Zwłaszcza, że z perspektywy porównawczej lepiej można ocenić mocne i słabe strony oraz określić perspektywicznie cechy rozwoju obu korpusów oddzielnie bądź korpusu połączonego, który mógłby kiedyś powstać.

Charakterystyka korpusów

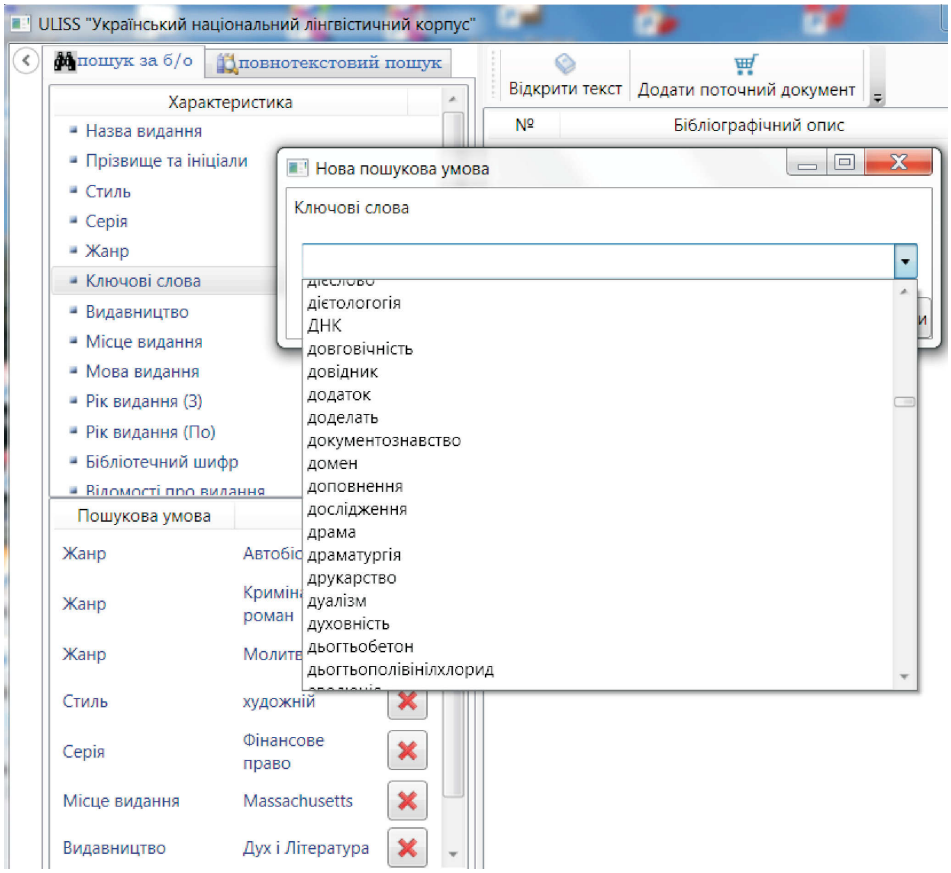
Nie ma możliwości zobaczyć tego bezpośrednio, ale ze starszych (Szyrokow et al. 2005) oraz nowszych artykułów (Szyrokow et al. 2011: 1) wiemy, iż na UNLK składa się korpus ogólny (76 mln słów) oraz korpus języka ustawodawstwa (18 mln słów). Korpus ogólny jest bardzo zróżnicowany pod względem gatunków, stylów, szczegółowo oznakowany następującymi rodzajami metainformacji: nazwa wydania; nazwisko autora, seria; gatunek (ok. 50 pozycji); styl (religijny, naukowo-dydaktyczny; ludowy; naukowy; naukowo-informacyjny; popularnonaukowy; dyplomatyczny; ustawodawczy; poetycki; publicystyczny; konwersacyjny; literatura piękna); wydawnictwo (kilkadziesiąt pozycji); miejsce wydania (ok. 100 pozycji); język wydania[5]; rok wydania (od/do, z możliwością wpisywania konkretnych lat); liczba stron, oraz kilka innych informacji bibliograficznych, mniej istotnych dla użytkownika pod względem lingwistycznym informacji bibliograficznych. Niestety, brakuje w dokumentacji informacji statystycznych, związanych z metainformacją, jednakże, sądząc z opisów w literaturze (Szyrokow et al. 2005) oraz subiektywnej oceny wyników wyszukiwania w UNLK, można go uznać za w miarę zrównoważony i równomiernie przedstawiający różne gatunki językowe. Należy zwrócić uwagę jednak na to, że UNLK zawiera niewielką liczbę tekstów tłumaczonych z innych języków, czego należy być świadomym, pracując z tym korpusem. W zależności od celu badawczego może wyniknąć potrzeba odfiltrowania tekstów tłumaczonych[6].

[5] Ta cecha w danym przypadku jest redundantna, prawdopodobnie pozostała w ramach ujednolicenia ze stosowanymi w UNLK standardami metainformacji UNMARC (Biblioteki Kongresu USA) [Szyrokow et al. 2005: 267-271].

[6] Nie wydaje się, żeby to można było uczynić w prosty sposób, za pomocą jednego kliknięcia, jednak opcja dowolnej aranżacji tekstów do wyszukiwania w UNLK teoretycznie daje

Ciekawą opcją jest wyszukiwanie według słów kluczowych – do wyboru jest kilkaset kręgów tematycznych, np. *awiacja*, *Azarow*[7], *język angielski*, *hematologia*, *sanskryt* itd. Takie bogactwo parametrów wyszukiwawczych pozwala określić użytkownikowi podkorpus dopasowany do jego potrzeb. Może może być szczególnie przydatne przy badaniach terminologii czy różnych aspektów lingwistycznych gatunków tekstów. Rys. 1. pokazuje możliwości definiowania podkorpusu UNLK według różnych metainformacji.

Rys. 13.1. Uściślanie warunków metainformacyjnych dla wyszukiwania w UNLK.



Po lewej stronie u góry widoczny jest fragment listy rodzajów metainformacji do wyboru („Характеристика"), pod nim wybrane przez nas opcje[8]: gatunek 'autobiografia' albo 'kryminał' czy 'modlitwa', styl 'literatura piękna',

taką możliwość, a opcje podkorpusu można zapisać w osobnym pliku i później z nich korzystać ponownie.

[7] Obecny premier Ukrainy.

[8] Przy takiej konfiguracji zostanie zwrócony pusty zbiór tekstów.

seria 'prawo finansowe', miejsce wydania 'Massachusetts', wydawnictwo 'Duch i Litera'[9]. W okienku „нова пошукова умова" (nowy warunek wyszukiwania) widać fragment listy dostępnych słów kluczowych, ułożonych alfabetycznie.

Wyszukiwania w tekstach (wybranych według wskazanych parametrów, np. tekstach publicystycznych) dokonujemy w sposób następujący: otwieramy zakładkę 'wyszukiwanie według bibliografii'; zadajemy warunek wyszukiwania podwójnym kliknięciem kryterium *жанр* 'gatunek', wybieramy interesujący nas gatunek i klikamy *додати умову* 'dodać warunek'. Następnie klikamy przycisk *пошук* 'wyszukiwanie', żeby zobaczyć, ile dokumentów odpowiada zadanemu kryterium gatunku oraz jaki jest ich szczegółowy opis bibliograficzny. Dalej, żeby ograniczyć wyszukiwanie do znalezionych tekstów, możemy dodać wszystkie albo wybrane znalezione teksty do koszyka (klikamy *кошик* 'koszyk', trzeci przycisk u góry, następnie *додати в кошик всі відібрані документи* 'dodaj do koszyka wszystkie wybrane dokumenty' oraz kontynuować wyszukiwanie, powracając do zakładki *повнотекстовий пошук* 'wyszukiwanie pełnotekstowe' z zaznaczeniem opcji *пошук за текстами з поточного кошика* 'wyszukiwanie w tekstach z obecnego koszyka'[10].

KUM zawiera łącznie 13 mln słów[11]. Ma inną budowę niż UNLK. Składa się z czterech podkorpusów; są to: literatura piękna (proza), literatura piękna (poezja), folklor oraz inne. Jednocześnie można korzystać tylko z jednego podkorpusu. Jest to istotna wada, ponieważ zmusza użytkownika, który chce dostać dane z całego korpusu, do zadawania czterech pytań zamiast jednego i dodatkowej obróbki danych Proporcjonalnie rozkład według gatunków w KUM wygląda następująco (Darczuk 2012a: 102-103):

Gatunek	**Liczba słów**
poezja	1 mln
literatura piękna	7 mln
folklor	32 tys.
publicystyka	4 mln
literatura naukowa społeczno-humanistyczna	1 mln
teksty oficjalne	1 mln

W podkorpusach KUM możliwe jest wybieranie płci autora tekstów podczas wyszukiwania, co może być bardzo pomocne dla badań genderowych i socjolingwistycznych. Dostępne są także inne metainformacje, ale dopiero po wyszukaniu, podpięte oddzielnie do każdego poświadczenia. Po kliknięciu linku *джерело* 'źródło' po prawej stronie od poświadczenia, na osobno otwierającej

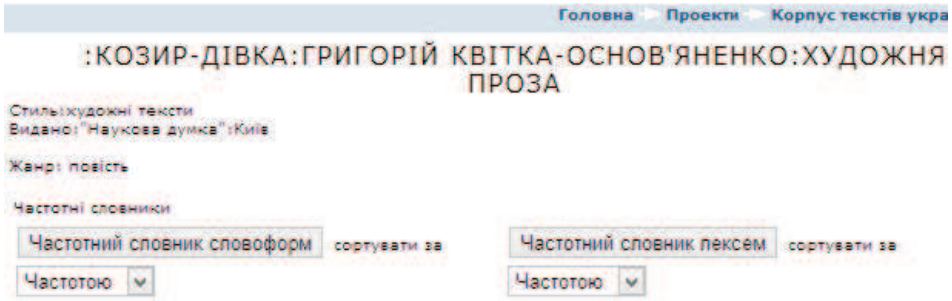
[9] Do metainformacji w UNLK wkradł się błąd – faktyczna nazwa wydawnictwa to „Дух і літера".

[10] Wypowiedź prywatna Nadii Sydorczuk, pracownika UFLI (maj 2011 r.), podana została tutaj ze względu na brak innych dostępnych instrukcji do pracy z UNLK.

[11] Na podstawie strony internetowej KUM: http://www.mova.info/Page.aspx?l1=6. Informacje dotyczące objętości KUM są sprzeczne – Darczuk (2012b) deklaruje 17 mln słów.

się stronie można zobaczyć m.in. nazwę utworu, jego autora, nazwę podkorpusu, styl, wydawnictwo i miejsce wydania, gatunek tekstu. Poniżej także znajdą się linki do słowników frekwencyjnych danego tekstu, o których bardziej szczegółowo mówimy niżej, zob. rys. 13.2.

Rys. 13.2. Metainformacje dotyczące konkretnego poświadczenia w KUM.



Metainformacje – zarówno w UNLK, jak i w KUM – podane są w sposób wyczerpujący i rzetelny. Pewien niedosyt pozostawia jednak brak podsumowań statystycznych według różnych metadanych – pozwalałoby to użytkownikom lepiej zrozumieć architekturę obu korpusów.

Segmentacja tekstu

Segmentacja tekstu w obu korpusach oparta jest głównie na wyrazach, nie ma widocznego dla użytkownika podziału na zdania czy akapity. Natomiast segmentacja wewnątrzwyrazowa jest organizowana w omawianych korpusach w sposób odmienny. W KUM łącznik jest uważany za integralną część wyrazów. Ma to konsekwencje takie, że nie można wyszukiwać według jednej z części wyrazu, zawierającego łącznik; musi zostać wpisany cały wyraz. Dotyczy to także form, dopuszczających duży stopień swobody połączeń, np. kolorów czy języków (*żółto-czerwony*, *polsko-niemiecki*). Samo wpisanie formy *жовто-* 'żółto-' w KUM nie zwraca żadnych wyników. Ta sama forma bez łącznika funkcjonuje jako przysłówek i wpisanie jej do okienka zapytań oznacza zwrot, odpowiednio, jedynego poświadczenia, które jest w korpusie z tą formą przysłówkową. Natomiast wpisanie formy *жовто-червоний* 'żółto-czerwony' zwraca już większą liczbę poświadczeń. Jeżeli forma nie funkcjonuje samodzielnie, to odpowiedź jest pusta.

Lepiej pomyślane jest wyszukiwanie form z łącznikiem w UNLK. Wpisanie formy częściowej *жовто-* zwraca wszystkie formy, zawierające tę kombinację znaków, nawet te, w których stanowi ona część wyrazu trójczłonowego *зелено-жовто-червону* 'zielono-żółto-czerwoną'. Z tego należy wnioskować,

że w UNLK łącznik jest traktowany jako segment łączący wyrazy. Wyszukiwanie według samego łącznika zwraca zbiór pusty. Należy zaznaczyć, że w UNLK trafiają się poświadczenia zawierające nieprawidłową ortografię. Jest to pozostałość skanowania i digitalizacji wersji papierowych tekstów zawierających dywizy rozdzielające części wyrazów przy przenoszeniu ich do następnej linijki. Odpowiednio, przy indeksowaniu łącznik nierozdzielający jest traktowany jako łącznik, dlatego wyszukiwanie według kombinacji liter, od których mogą zaczynać się wyrazy, np. *жов* zwraca nieistniejące lematy, będące w rzeczywistości częściami innych, np. *жов-нір* 'żoł-nierz'.

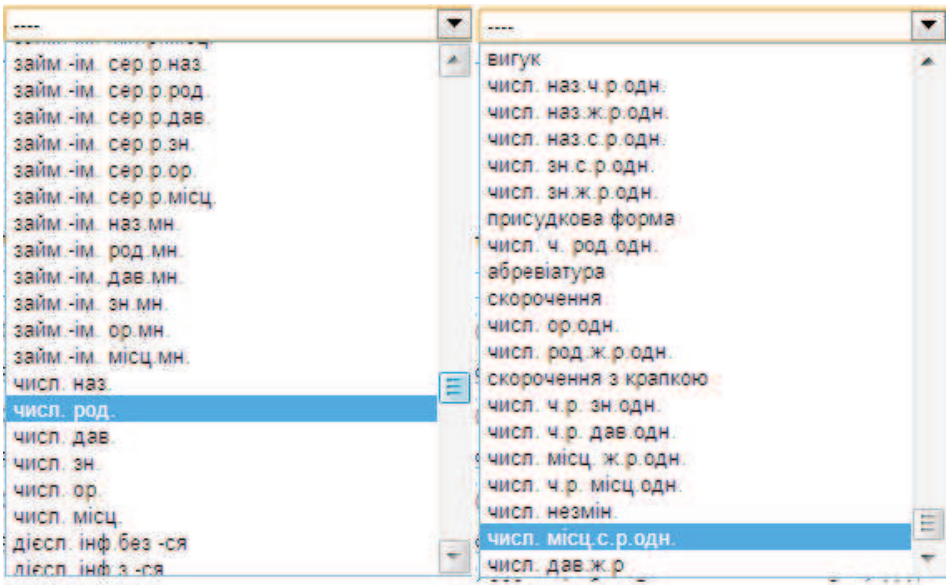
Skrótowce w UNLK można wyszukiwać, podając formę z kropką lub bez niej. Jednakowe wyniki otrzymamy, wpisując *тис* 'tys (= tysiąc)' albo *тис.* 'tys.'[12]. W wyrazach zawierających dywiz sygnalizujący odmianę, np. *30-ма* 'trzydziestoma', *30-ти* 'trzydziestu', *30-х* 'trzydziestu', *30-річний* '30-letni', jest on traktowany tak samo jak łącznik. Zapytanie sformułowane liczbowo, np. *30*, zwraca wyżej wymienione przykłady oraz inne, łącznie pisane poświadczenia nieliterowe: *30%*, *№30*, *с. 28-30.*, *20-30.*

Tagset

Tagset stosowany w UNLK jest dobrze opisany w (Szyrokow et al. 2005: 420--438). Niestety, informacje gramatyczne nie są dostępne w publicznej wersji korpusu. Informacje morfoskładniowe są natomiast dostępne w KUM, lecz tagset KUM ma dokumentację ledwie śladową (Darczuk 2012b). Niemniej jednak, z bardzo podobnego podziału kategorialnego, np. nietypowego traktowania stopniowanych form przymiotników i przysłówków jako odrębnych lematów, wyodrębnienia czasowników z postfiksem *-ся* albo bez niego (pisanego łącznie z formą czasownikową odpowiednika polskiego *się*), dwuliterowego sposobu kodowania oraz samego używania kodów itd. wynika, że opis gramatyczny w obu korpusach wywodzi się z tej samej tradycji. Jedyną widoczną różnicą jest rozbicie kategorii predykatywów w KUM na predykatywne słowa i predykatywne formy, podczas gdy w UNLK jest to jedna kategoria.

[12] Generalnie znaki nieliterowe w oknie wyszukiwania są ignorowane – identyczne wyniki otrzymamy na zapytanie *30* oraz *30%*.

Rys. 13.3. Fragmenty listy rozwijanej z kombinacjami kategorii gramatycznych w KUM.



Warunki gramatyczne, według których można uściślać wyszukiwanie w KUM, są pokazywane jako kategorie bądź ich kombinacje na wygenerowanej liście z możliwością wyboru jednej opcji naraz (np. „zaimek pronominalny rodzaju nijakiego w mianowniku", „zaimek pronominalny rodzaju nijakiego w dopełniaczu" itd., zob. rys. 13.3. Przedstawiamy niżej, w tab. 13.1, zwarty wykaz wszystkich dostępnych kombinacji kategorii z tej listy w porządku, w którym się na niej pojawiają, oraz podsumowujemy statystycznie owe kombinacje.

Tab. 13.1. Kategorie gramatyczne i/lub ich kombinacje, których można użyć jako kryteriów wyszukiwania gramatycznego w KUM.

Część mowy	**Cecha 1**	**Cecha 2**	**Cecha 3**	**Cecha 4**	**Liczba kombinacji**
rzeczownik	przypadek (7)				7
		l. mn.			7
	przypadek (7)		rodzaj (3)		21
		l. mn.	rodzaj (3)		21
	nieodmienny				1
			rodzaj (3)		3
	przypadek (7)		rodzaj (3)	nazwa własna	21
	nieodmienny		rodzaj (3)	nazwa własna	3
	przypadek (7)	pl. tantum			7
	nieodmienny	pl. tantum			1

[ciąg dalszy tabeli ze strony poprzedniej]

Część mowy	**Cecha 1**	**Cecha 2**	**Cecha 3**	**Cecha 4**	**Liczba kombinacji**
	przypadek (7)	pl. tantum		nazwa własna	7
przymiotnik	przypadek (6)		rodzaj (3)		18
		l. mn.			6
			rodzaj (3)	nazwa własna	18
zaimek-przymiotnik	przypadek (6)		rodzaj (3)		18
		l. mn.			6
zaimek-rzeczownik	przypadek (6)		rodzaj (3)		18
		l. mn.			6
liczebnik	przypadek (6)				6
czasownik	bezokolicznik			z/bez *-ся*	2
	osoba (3)	liczba (2)		z/bez *-ся*	12
	czas przeszły		rodzaj (3)	z/bez *-ся*	6
		l. mn.		z/bez *-ся*	2
	czas przeszły			z/bez *-ся*	2
	osoba (3)	liczba (2)	czas przyszły	z/bez *-ся*	12
	osoba 1	l. mn.	tryb rozkazujący	z/bez *-ся*	2
	osoba 2	liczba (2)	tryb rozkazujący	z/bez *-ся*	4
imiesłów	dok/niedok			z/bez *-ся*	4
zaimek dzierżawczy			rodzaj (2:m,f)		2
		l. mn.			1
przyimek	przypadek (5)				5
partykuła					1
przysłówek					1
predykatyw1					1
spójnik	współrzędny/ podrzędny/ oba				3
wykrzyknik					1
liczebnik	mianownik	rodzaj (3)	l. poj.		3
	biernik	rodzaj (3)	l. poj.		3
		r. m.	l. poj.		1
predykatyw2					1
abrewiatura					1
skrót					1
liczebnik	narzędnik		l. poj.		1
	dopełniacz	r. ż.	l. poj.		1

[ciąg dalszy tabeli ze strony poprzedniej]

Część mowy	**Cecha 1**	**Cecha 2**	**Cecha 3**	**Cecha 4**	**Liczba kombinacji**
skrót z kropką					1
liczebnik	celownik	r. m.	l. poj.		1
	miejscownik	rodzaj (3)	l. poj.		3
	nieodmienny				1
	celownik	r. ż.			1
Suma					**275**

Lista jest dość długa (275 pozycji), co czyni ją niewygodną w korzystaniu. Nie wszystkie możliwe kombinacje są dostępne: widzimy różne zestawienia przypadków i rodzaju dla rzeczowników ogólnie oraz dla rzeczowników w liczbie mnogiej, ale nie mamy już możliwości wyszukiwania zawężonego do liczby pojedynczej. Podobnie jest z czasem przyszłym czy teraźniejszym (nie mamy możliwości wyszukiwania form czasu przyszłego albo teraźniejszego bez dalszych ograniczeń) oraz kryterium nazwy własnej. Niektórych kombinacji brakuje, np. zaimka dzierżawczego rodzaju nijakiego, niektóre się powtarzają (np. czasownik 1 os. l.mn. z *-ся* pojawia się na liście dwa razy). W sposób bardzo niekonsekwentny podane są połączenia zaimków i liczebników. Wymaga też objaśnienia niezupełnie oczywista różnica między kategoriami „присудкова форма" i „присудкове слово" (w tabeli nazwane odpowiednio „predykatyw1" i „predykatyw2"), jak i same te kategorie[13].

Lematyzacja

Lematyzacja w obu korpusach prezentuje dość wysoki poziom, ale i tu czyhają na użytkownika pułapki. W KUM problemy pojawiają się przy wyrazach rzadkich, nowych, którym lematy, podobnie jak tagi, są przypisywane automatycznie na podstawie „zgadywania automatycznego". Np. wyraz *єврозоні* 'eurostrefie' został zinterpretowany automatycznie jako rzeczownik rodzaju męskiego, którego lematem hipotetycznie miał być nieistniejący **єврозон*. Tego rodzaju błędy trudno dostrzec, ponieważ informacje o formie podstawowej nie są wyświetlane w wynikach bezpośrednio. Pojawiają się natomiast podczas wyszukiwania według charakterystyk morfoskładniowych, np. rzeczowniki rodzaju męskiego w miejscowniku. Problemem mnożącym błędy lematyzacyjne w UNLK są wspomniane wcześniej łączniki nierozdzielające, stosowane przy

[13] W gramatykach akademickich spotykamy różne traktowanie kategorii predykatywu, zob. też analizę w (Kotsyba, Derzhanski 2008).

podziałach słów, które *de facto* rozdzielają jeden właściwy, istniejący wyraz na dwa niepoprawne, które są później indeksowane jako lematy.

Anotacja morfosyntaktyczna

W wersji publicznej UNLK anotacja morfoskładniowa nie została udostępniana wcale, ze względu na niewystarczająco wysoki poziom jakości dezambiguacji[14]. Tagi morfoskładniowe w KUM są dostępne pośrednio, tzn. tylko w momencie formułowania zapytania. W gotowych wynikach znaczników morfologicznych nie zobaczymy. Tagi w KUM były przypisywane w sposób zautomatyzowany (Darczuk 2012b); formom niejednoznacznym był przypisywany jeden tag na podstawie zadanych przez twórców korpusu reguł kontekstowych. Jakość dezambiguacji morfoskładniowej, niestety, nie jest najwyższa. Mimo że autorzy twierdzą, że poprawność dezambiguacji osiąga poziom 93-95% (Darczuk 2012b: 19), wyniki przeprowadzonych prób wyszukiwania pokazują zupełnie odmienny obraz – jest wiele błędów i niedokładności.

W odpowiedzi na zapytanie o niejednoznaczne formy wyrazu *хатки* 'domki' jako rzeczownika w mianowniku liczby mnogiej znalazła się m.in. forma w bierniku liczby mnogiej:

і, ще інші спорудили на базі балконів барвисті хатки (KUM 12 III 2013)
'i jeszcze zbudowali na bazie balkonów barwne domki'

Zapytanie o tę samą formę w bierniku liczby mnogiej zwróciło 12 zdań (w jednym z nich tego wyrazu użyto dwa razy), z których 6 form jest w mianowniku, 6 w bierniku oraz jedna forma w dopełniaczu liczby pojedynczej. Zapytanie zaś o formę dopełniacza liczby pojedynczej tego samego wyrazu też zwróciło błędne odpowiedzi, których liczba znacznie przekracza dopuszczalny w takich przypadkach błąd statystyczny[15].

Poniższe poświadczenia zostały zwrócone przez korpus na zapytanie o formę *мати* (1. 'matka', 2. 'mieć') jako czasownika; są to formy rzeczownikowe (średnio jest po 3-4 formy rzeczownikowe na każde 50 poświadczeń):

Мати *– в Італії, батько п'є, дитина – на шиї в бабусі, якій 70 років.*
(KUM 22 III 2013)

[14] Wypowiedź nieoficjalna pracowników ULFI.

[15] Biorąc pod uwagę liczbę danych językowych do opracowania oraz stopień złożoności morfologii języków słowiańskich, trudno spodziewać się stuprocentowej poprawności tagowania. Niemniej jednak, w stosunku do korpusów porównywalnych języków – czeskiego, polskiego – podawano poziom poprawnych tagów w zakresie 90-95% całości. W korpusie narodowym języka rosyjskiego stosowana jest ręczna dezambiguacja i możliwe było wyszukiwanie tylko w tym podkorpusie. W pozostałych wypadkach badacze otrzymują informacje o wszystkich możliwych interpretacjach morfoskładniowych, co mogłoby być lepszym niż obecne rozwiązaniem proponowanym dla KUM.

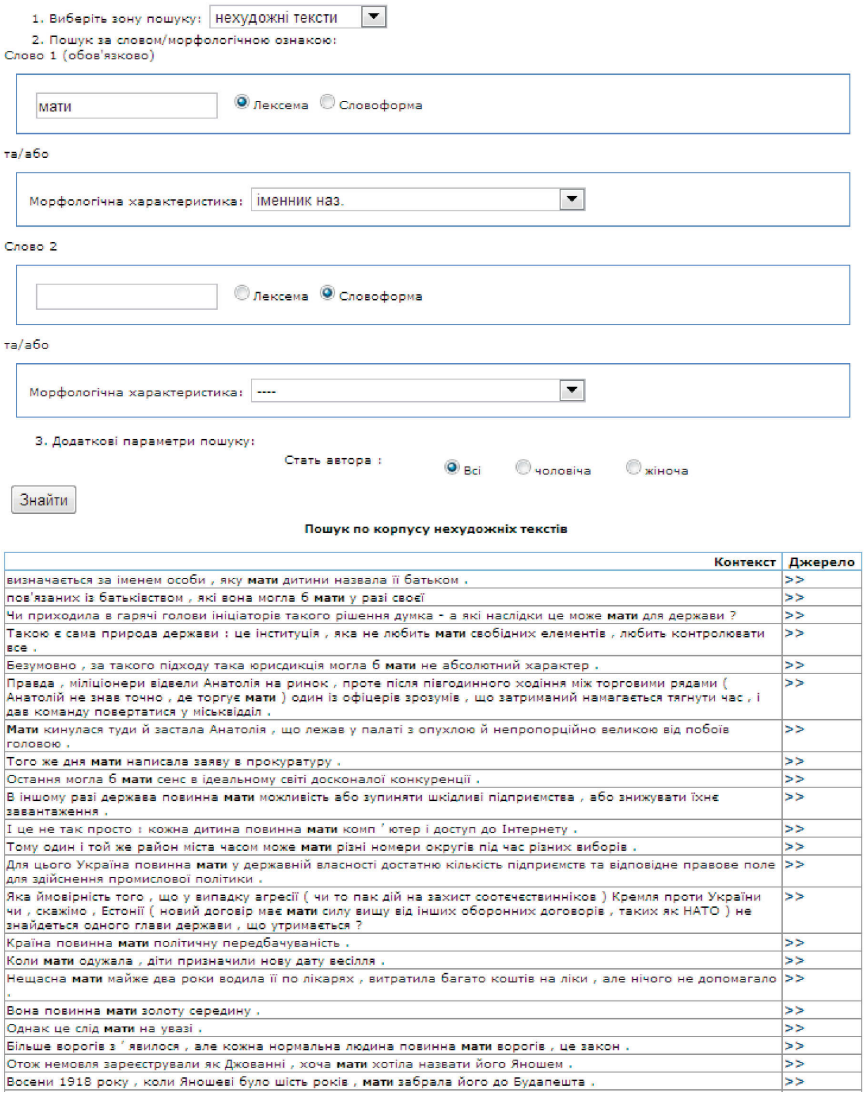
'Matka – we Włoszech, ojciec pije, dziecko – na szyi u babci, która ma 70 lat.'

Вода тяжка й густа пророка колихала, наче ***мати*** *сина* (KUM 22 III 2013)

'Woda ciężka i gęsta proroka kołysała, jak matka syna'

Odwrotne, z zaznaczeniem opcji gramatycznej „rzeczownik w mianowniku" zwraca wyniki, wśród których czasowników już jest 14 (poświadczenia: 2-5, 9-15, 18-20) spośród pierwszych 22 widocznych poświadczeń, tzn. tylko 1/3 wyników jest właściwa, zob. rys. 13.4. poniżej.

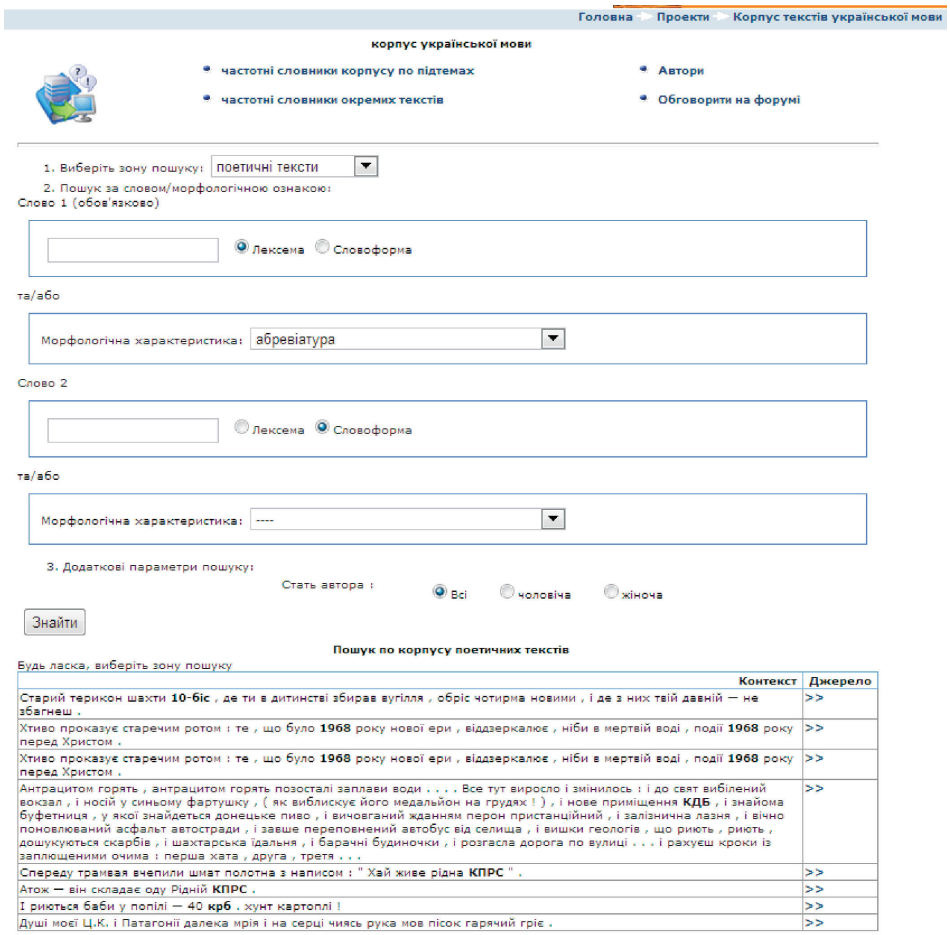
Rys. 13.4. Wyszukiwanie w KUM jednocześnie według zadanego lematu i kryteriów morfologicznych.



Контекст	Джерело
визначається за іменем особи , яку **мати** дитини назвала її батьком .	>>
пов'язаних із батьківством , які вона могла б **мати** у разі своєї	>>
Чи приходила в гарячі голови ініціаторів такого рішення думка - а які наслідки це може **мати** для держави ?	>>
Такою є сама природа держави : це інституція , яка не любить **мати** свобідних елементів , любить контролювати все .	>>
Безумовно , за такого підходу така юрисдикція могла б **мати** не абсолютний характер .	>>
Правда , міліціонери відвели Анатолія на ринок , проте після півгодинного ходіння між торговими рядами (Анатолій не знав точно , де торгує **мати**) один із офіцерів зрозумів , що затриманий намагається тягнути час , і дав команду повертатися у міськвідділ .	>>
Мати кинулася туди й застала Анатолія , що лежав у палаті з опухлою й непропорційно великою від побоїв головою .	>>
Того же дня **мати** написала заяву в прокуратуру .	>>
Остання могла б **мати** сенс в ідеальному світі досконалої конкуренції .	>>
В іншому разі держава повинна **мати** можливість або зупиняти шкідливі підприємства , або знижувати їхнє завантаження .	>>
І це не так просто : кожна дитина повинна **мати** комп ' ютер і доступ до Інтернету .	>>
Тому один і той же район міста часом може **мати** різні номери округів під час різних виборів .	>>
Для цього Україна повинна **мати** у державній власності достатню кількість підприємств та відповідне правове поле для здійснення промислової політики .	>>
Яка ймовірність того , що у випадку агресії (чи то пак дій на захист соотечественників) Кремля проти України чи , скажімо , Естонії (новий договір має **мати** силу вищу від інших оборонних договорів , таких як НАТО) не знайдеться одного глави держави , що утримається ?	>>
Країна повинна **мати** політичну передбачуваність .	>>
Коли **мати** одужала , діти призначили нову дату весілля .	>>
Нещасна **мати** майже два роки водила її по лікарях , витратила багато коштів на ліки , але нічого не допомагало .	>>
Вона повинна **мати** золоту середину .	>>
Однак це слід **мати** на увазі .	>>
Більше ворогів з ' явилося , але кожна нормальна людина повинна **мати** ворогів , це закон .	>>
Отож немовля зареєстрували як Джованні , хоча **мати** хотіла назвати його Яношем .	>>
Восени 1918 року , коли Яношеві було шість років , **мати** забрала його до Будапешта .	>>

W odpowiedzi na zapytanie o skrótowce w tekstach poetyckich znalazło się osiem poświadczeń (jeden z nich powtórzony), do skrótowców zostały zaliczone m.in.: liczby (*1968*), wyrazy, gdzie indziej klasyfikowane jako skróty z kropką: *крб.* 'karbowaniec' (rubel radziecki, była waluta ZSSR), nazwy własne nieskrócone: *10-біс* '10-bis (nazwa kopalni)', zob. rys. 13.5. poniżej.

Rys. 13.5. Wyszukiwanie w KUM: parametry i wyniki.



W kolejnym teście zapytanie o znalezienie skrótów w tekstach literatury publicystycznej (podkorpus „нехудожні тексти") zwraca dziewięć poświadczeń (składających się na pięć osobnych zdań bez powtórzeń), z czego cztery to są *км* 'km' (wszystkie w jednym zdaniu), dwa *млрд.* 'mld' (też w jednym zdaniu), oraz po jednym *рр.* 'lata' i *млн.* 'mln'. Natomiast wyszukiwanie samej tylko formy *км* w tekstach tegoż podkorpusu zwraca sześć stron wyników, każdy po 50 poświadczeń. Niestety, nie ma możliwości sprawdzenia, pod jakim

tagiem (i czy w ogóle tag został formie przypisany) można wyszukiwać dane formy, ponieważ informacja morfoskładniowa w znalezionych tekstach nie jest wyświetlana. Takie formy nie są wyświetlane ani jako abrewiatury, ani jako skróty z kropką.

Podsumowując, należy stwierdzić: jakość informacji gramatycznej KUM znajduje się obecnie w takim stanie, że należy być bardzo ostrożnym, decydując się na jej wykorzystanie w swoich pracach.

Podstawowe zapytania

Oba korpusy umożliwiają wyszukiwanie według form wyrazowych albo lematów poprzez zaznaczenie odpowiedniej opcji w oknie wyszukiwania. Podobnie w obu korpusach warianty ortograficzne (np. małą lub wielką literą) są unifikowane automatycznie, bez możliwości wyszukiwania jednej bądź drugiej opcji osobno. Niemniej jednak takie rozwiązanie raczej ułatwia pracę badacza niż ją utrudnia.

W obu korpusach możliwe jest także wyszukiwanie grupy wyrazów. W KUM wpisujemy każdy z wyrazów oddzielnie, zaznaczając przy tym, czy jest to lemat, czy forma wyrazowa. Istnieje możliwość wpisania maksymalnie dwóch wyrazów/lematów. Natomiast UNLK umożliwia wyszukiwanie dość długich ciągów wyrazów, nawet całych zdań, przy czym można też zadać odległość między szukanymi wyrazami w korpusie.

KUM także łączy warunki gramatyczne, a nawet gramatyczne i leksykalne razem. W odróżnieniu od KUM, w UNLK możemy zaznaczyć opcję lematyzacji tylko jeden raz i będzie ona dotyczyła jednocześnie wszystkich szukanych wyrazów ciągu, co jest pewnym ograniczeniem.

Teoretycznie KUM umożliwia w niewielkim stopniu wyszukiwanie kolokacji dzięki temu, że w grupie szukanych wyrazów możemy wpisać jeden z elementów jako lemat bądź formę wyrazową, a zamiast drugiego podać ograniczenie gramatyczne. Np. moglibyśmy szukać epitetów do wyrazu *листя* ‘liście’, zaznaczając, że wyraz poprzedzający tę formę jest przymiotnikiem[16]. W praktyce zapytania tego rodzaju nie są technicznie możliwe – pojawia się błąd systemu[17]. W odwróconym szyku (kiedy najpierw podany jest lemat/forma wyrazowa, a charakterystyka gramatyczna jest drugim członem zapytania) system zwraca wyniki, z tym, że do adiektywów zaliczane są także formy czasownikowe czasu przeszłego, co stwarza dodatkowe niedogodności przy wyszukiwaniu

[16] Trzeba pamiętać, że przymiotnik jako samodzielna część mowy nie występuje w tagsecie, zob. tab. 13.1. Stosowany natomiast jest bardziej ogólny termin „adiektyw”, który obejmuje także liczebniki porządkowe oraz imiesłowy przymiotnikowe. Dlatego teoretycznie taka lista zawierałaby także wyrazy funkcjonalne.

[17] Wyszukiwaliśmy siedem różnych konfiguracji jednostek. Skutek był za każdym razem ten sam.

kolokacji. Tego problemu można byłoby częściowo się pozbyć, jeżeliby można było wyniki zapisywać w postaci tabeli z możliwością sortowania, ale ta opcja obecnie nie jest zaimplementowana. Jeszcze jedna niedogodność, to niemożliwość wybrania przymiotnika jako kategorii ogólnej. Użytkownik jest zmuszany do wybrania szczegółowej charakterystyki: adiektyw wraz z określonym rodzajem i w określonym przypadku. Wymaga to sformułowania 18 zapytań zamiast teoretycznie możliwego jednego.

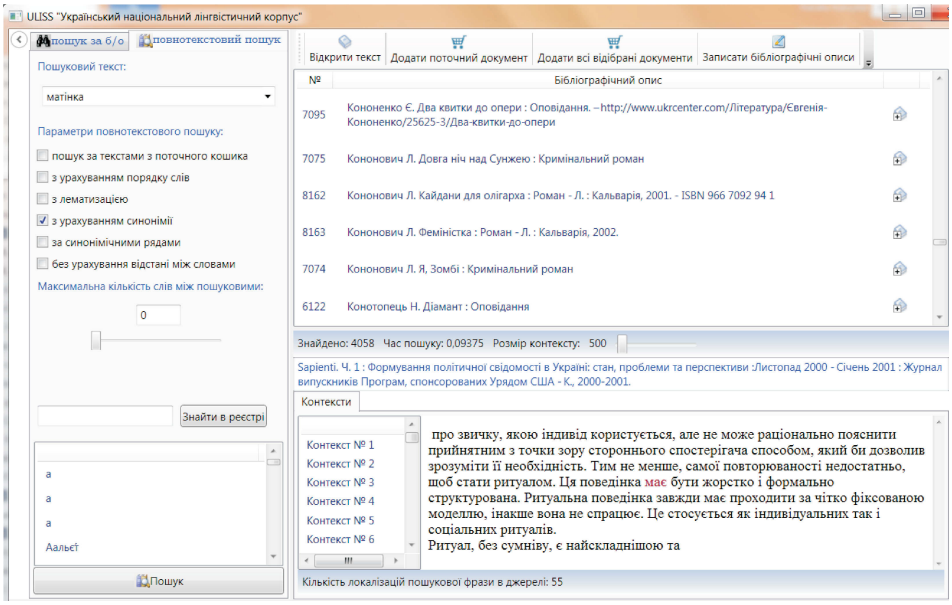
KUM umożliwia sformułowanie zapytania o samą kategorię gramatyczną, np. formę wyrazową rodzaju męskiego w bierniku liczby pojedynczej czy abrewiaturę, zob. rys. 13.5. Możliwe też są zapytania o lemat w określonej kategorii gramatycznej, np. мати '1) matka, 2) mieć' jako rzeczownik w mianowniku, zob. rys. 13.4.

Dostępne funkcje w oprogramowaniu korpusów języka ukraińskiego

Oba korpusy zwracają wyniki wyszukiwania w postaci konkordancji. Żaden z nich nie daje możliwości wyświetlania wyników w postaci konkordancji KWIC ani nie pozwala odfiltrowywać znalezionych danych. KUM nie podaje żadnych informacji statystycznych o znalezionych wynikach. Wyświetla się po 50 wyników na stronę; jeżeli jest to wiele stron, to na dole pojawia się link z napisem *наступна сторінка* 'kolejna strona'), ale nie widać, ile jest stron łącznie. UNLK natomiast wyświetla liczbę poświadczeń oraz liczbę tekstów, w których je znaleziono, a także liczbę poświadczeń w każdym z tekstów. Możliwe jest też dopasowanie odległości między wyrazami w grupie wyrazowej. Na zapytanie жовте листя 'żółte liście' (odstęp 3 wyrazy) UNLK zwrócił 77 wyników (dla porównania przy jednowyrazowym odstępie – 70; przy dwuwyrazowym – 76).

W UNLK istnieje możliwość wyszukiwania od razu w ramach całego rzędu synonimicznego do podanego wyrazu. Może to być funkcja bardzo przydatna w badaniach semantycznych, kognitywistycznych oraz w ogóle naukach humanistycznych, które bardziej interesuje informacja niż konkretne formy wyrazowe. Należy jednak być świadomym możliwych na tym etapie błędów lematyzacji. Na przykład, wyszukiwanie wyrazu матінка (zdrobnienie od 'matka') wraz z synonimami zwraca konteksty z wyrazem 'mieć', którego lemat jest tożsamy z lematem *мати* 'matka'.

Rys. 13.6. Wyszukiwanie w UNLK z użyciem opcji rzędów synonimicznych.



Bardzo pożyteczną funkcją w UNLK jest możliwość zapisywania wyników. Możemy zdecydować, czy chcemy zapisać wyniki z wybranego źródła, czy ze wszystkich, czy ograniczyć liczbę zapisywanych wyników z każdego źródła albo ustalić długość kontekstu (domyślnie – 500 słów). Wyniki są zapisywane w formacie HTML, szukane wyrazy/frazy są wydzielane specjalnym tagiem stylu, co daje możliwość dalszej obróbki automatycznej wyników.

Częściowe dane statystyczne o korpusie w postaci słowników frekwencyjnych są osiągalne w czasie pracy z KUM. Można wygenerować słownik na podstawie lematów albo form wyrazowych do dowolnego tekstu. Bardzo pożyteczną jest funkcja „wykrojenia" części słownika frekwencyjnego dla konkretnej części mowy, na żądanie pokazywane są także niektóre dane typu statystycznego, np. liczba tekstów, średnia częstotliwość, odchylenie standardowe, współczynnik stabilności. Dostępne są także słowniki frekwencyjne morfemów. W przypadku niektórych słowników zaznaczono jednak, że dezambiguacja nie została przeprowadzona całkowicie, dlatego możliwe są odchylenia od stanu faktycznego.

Funkcja, której najbardziej brakuje w obu korpusach, a która może zostać dopracowana względnie niewielkim kosztem, to sortowanie wyników wyszukiwania według lematu wyszukiwania, lewego czy prawego kontekstu. W przypadku KUM bardzo brakuje podstawowych informacji statystycznych jak np. liczba zwróconych poświadczeń, a także możliwości zapisywania wyników do plików tekstowych w celu dalszego ich opracowywania. Warto jeszcze dodać, że

przy KUM działa otwarte forum internetowe, przeznaczone do dyskusji o korpusie[18].

UNLK obecnie działa jednocześnie jako korpus tekstów oraz baza bibliograficzna, co oznacza, że informacje bibliograficzne są bardzo obszerne i dobrze ustrukturowane (istnieje nawet możliwość tworzenia „koszyków" bibliograficznych). Położenie akcentu na dane bibliograficzne oraz połączenie tych funkcji jest związane z konkretnymi potrzebami leksykograficznymi twórców UNLK. Przeciętnemu użytkownikowi korpusu wydaje się ono zbędne.

Metodologia i perspektywy badawcze

Wymienione korpusy przede wszystkim są cenne przy wyszukiwaniu poświadczeń językowych według lematu bądź formy wyrazowej albo ciągu wyrazów oraz tworzeniu konkordancji, co jest bardzo pożyteczne dla celów leksykograficznych. Ze względu na zawarte informacje morfoskładniowe (KUM) nadają się także do badań nad składnią. Szkoda, że ciąg wyszukiwawczy jest ograniczony do dwóch jednostek wyrazowych, co poważnie zmniejsza możliwości badawcze. Można także prowadzić badania corpus-driven, ponieważ korpusy są w miarę duże i zrównoważone. Niestety, nie ma jeszcze możliwości wyciągania informacji statystycznych ani kolokacji z żadnego z korpusów. Należy się spodziewać, że nowe funkcje wyszukiwawcze będą się pojawiać wraz ze wzrostem zainteresowania korpusami i metodami korpusowymi wśród badaczy ukraińskich.

UNLK, po gwałtownym wzroście w latach 2005-2010, nie okazuje ostatnio oznak rozwoju. Dalsza budowa i popularyzacja korpusu nie jest zadaniem priorytetowym ULFI, skupionego głównie na wyzwaniach leksykograficznych, dla których m.in. został stworzony UNLK. W chwili obecnej głównym zastosowaniem korpusu jest dostarczanie cytatów leksykograficznych oraz monitoring znaczeń leksykalnych na potrzeby powstającego 20-tomowego słownika języka ukraińskiego. Całkiem praktyczne podłoże miało także powstanie KUM – jego głównym w danym momencie produktem są zróżnicowane i w miarę jakościowe słowniki frekwencyjne, częściowo udostępniane przez autorów. Ogólnie rzecz ujmując, zarówno KUM, jak i UNLK są zasobami zamkniętymi, nie pozwalającymi na odtworzenie istniejących wyników przez niezależnych badaczy.

Do marca 2013 roku nie udało nam się znaleźć opublikowanych, dostępnych prac o charakterze lingwistycznym innym niż wyżej wspomniane, wykorzystujące istniejące zasoby korpusowe dla języka ukraińskiego, zwłaszcza pisane

[18] Obecnie (marzec 2013 r.) zawiera ono cztery wpisy, z których dwa to informacje o nowych dodanych funkcjach korpusu, kolejne dwa są pytaniami użytkowników, dotyczącymi korpusu bezpośrednio (jedno z listopada 2012, drugie z sierpnia 2011 roku), na które nie dano żadnej odpowiedzi. Zakładanie forum jest niezaprzeczalnie bardzo dobrą inicjatywą twórców korpusu, wymaga jednak ono należytego administrowania.

przez badaczy spoza jednostek, gdzie się je opracowuje. Obecnie uwaga językoznawców skupia się głównie na koncepcjach tworzenia zasobów korpusowych dla języka ukraińskiego (S. Buk, O. Demska, V. Starko, O. Siruk) albo na technicznych aspektach ich tworzenia (W. Szyrokow, N. Darczuk, N. Sydorczuk, N. Kotsyba). Wygłaszane są poglądy (Perebyjnis, Bobkowa: 4), że w dużej mierze jest to związane z sytuacją lingwistyki komputerowej w kraju w ogóle, mającą podłoże historyczne i polityczne. Brak informacji o badaniach lingwistycznych, prowadzonych na materiale korpusów języka ukraińskiego[19], świadczy o tym, że – z jednej strony – niewystarczająca jest świadomość potencjalnych użytkowników owych korpusów, spowodowana zbyt słabą ich popularyzacją, z drugiej zaś – jakość tych korpusów oraz poziom ich dostępności tymczasem nie pozwala na podejmowanie poważnych badań na ich materiale. Z tego powodu twórcy korpusów języka ukraińskiego wciąż stoją przed zadaniem ich jakościowej poprawy, a także dalszego rozbudowywania poziomów znakowania oraz atrakcyjnych metod wyciągania informacji już istniejącej w korpusach. Dobrym rozwiązaniem byłoby połączenie wysiłków różnych ośrodków oraz zaproszenie do tego przedsięwzięcia szerokich kręgów językoznawców Ukrainy w celu stworzenia scentralizowanego, monitorowanego pod względem jakości standardów międzynarodowych narodowego korpusu języka ukraińskiego.

Bibliografia

Derzhanski I., Kotsyba N. (2008). *The Category of Predicatives in the Light of Consistent Morphosyntactic Tagging.* W: „Lexicographic Tools and Techniques", Proceedings of MONDILEX First Open Workshop, Moscow, Russia, 3-4 October. Moskwa, s. 68-79.
http://domeczek.pl/∼natko/papers/ID_NK_tagSlav.pdf

Kotsyba N. (2012). *PolUKR (a Polish-Ukrainian Parallel Corpus) as a Testbed for a Parallel Corpora Toolbox.* „Prace Filologiczne", t. LXIII. Warszawa, s. 181-196.
http://www.domeczek.pl/∼natko/papers/NKotsyba_SlaviCorp2010_paper.pdf

Дарчук Н. (2012a). *Корпус Украинского Языка.* „Prace Filologiczne", t. LXIII. Warszawa, s. 99-108.

Дарчук Н. (2012b). *Морфологічне анотування Корпусу української мови.* „Комп'ютерна лінгвістика: сучасне та майбутнє". Kijów, s. 16-19.

Демська О. (2011). *Текстовий корпус: ідея іншої форми.* Kijów.

Демська-Кульчицька О. (2005). *Основи національного корпусу української мови.* Kijów.

[19] Śledzone i analizowane są konferencje językoznawcze oraz korpusowe krajowe i międzynarodowe oraz strony internetowe większych lingwistycznych jednostek badawczych Ukrainy.

Комп'ютерна лінгвістика: сучасне та майбутнє. Матеріали Міжнародної науково-практичної конференції (2012). Kijów. http://www.mova.info/zbirnyk.pdf

Лебедєв К. (2012). *Створення Багатомовного корпусу паралельних текстів.* W: „Комп'ютерна лінгвістика: сучасне та майбутнє". Kijów, s. 36-37.

Перебийніс В., Бобкова Т. (2012). *Історія лабораторії комп'ютерної лінгвістики КНЛУ. Комп'ютерна лінгвістика: сучасне та майбутнє.* W: „Комп'ютерна лінгвістика: сучасне та майбутнє". Kijów, s. 3-4.

Сидорчук Н.М. (2005). *Архітектурні та системотехнічні підходи до конструювання Українського національного лінгвістичного корпусу.* W: „Бионика интеллекта", 2 (63), s. 107-110.

Широков В.А., Бугаков О.В., Грязнухіна Т.О., Костишин О.М., Кригін М.Ю., Любченко Т.П., Рабулець О.Г., Сидоренко О.О., Сидорчук Н.М., Шевченко І.В., Шипнівська О.О., Якименко К.М. (2005). *Корпусна лінгвістика.* Kijów.

Широков В.А., Сидорчук Н.М., Бугаков О.В., Кригін М.Ю. (2011). *Застосування Українського національного лінгвістичного корпусу в лексикографії та лінгвістичних експертизах.* W: „Українська лексикографія в загальнослов'янському контексті: теорія, практика, типологія". Kijów, s. 285-294. http://ovbugakov.org.ua/articles/Bugakov_Oleg_article2.pdf

14. Praktyczny przewodnik po korpusie języka białoruskiego

Alyaxey Yaskevich
(badacz niezależny)

Białoruskie projekty korpusowe znacznie się różnią od propozycji przekładanych przez twórców korpusów innych języków słowiańskich. Jest to spowodowane powolnym rozwojem białoruskiej lingwistyki korpusowej. Pierwsze ogólnodostępne projekty powstały dopiero na przestrzeni ostatnich paru latach.

Ilustracją nie najlepszego stanu białoruskiej lingwistyki korpusowej może być konferencja *Slavicorp* (22-24 listopada 2010 r., Warszawa), poświęcona korpusom słowiańskim, na której z regionu krajów wschodnio- i zachodniosłowiańskich nie było tylko przedstawicieli projektów korpusowych z Białorusi[1].

Zainteresowanie podejściem korpusowym w badaniu języka na Białorusi zaczęło się w latach 1993-1994 (Совпель 1994, Рычкова 1994). Mówimy tu o korpusach w postaci współczesnej. Początkowo takie zbiory tekstów nazywano *машынны фонд* 'zasób komputerowy'[2]. Owoce tych zainteresowań pojawiły się dopiero po 5 latach, w postaci korpusów dydaktycznych (stwarzanych dla potrzeb nauczania języka), np. projekty Ryczkowej[3] (Рычкова 1997, 1998). Pomimo tego, że korpusy te stanowiły ważne ogniwo w procesie rozwoju lingwistyki korpusowej oraz przyniosły korzyści białoruskiej dydaktyce, z przyczyn gospodarczych i technologicznych projekty korpusowe na Białorusi końca XX wieku nie mogły być w pełnym zakresie realizowane. Sytuacja w społeczeństwie i w nauce była niestabilna, brakowało specjalistów, informatyka we współczesnej postaci zaczynała się dopiero rozwijać.

W latach 2000-2005 Instytut Językoznawstwa[4] Narodowej Akademii Nauk (dalej – NANB) i Narodowe Centrum Naukowo-Edukacyjne im. Francyska

[1] Z krajów południowosłowiańskich nie przyjechali przedstawiciele Macedonii i Bośni.

[2] Biał. *машынны фонд* jest kalką rosyjskiego *машинный фонд* (Марковская, Филимонова 2008: 71), gdzie słowo *фонд* – w podstawowym znaczeniu 'zasób' – użyte jest w znaczeniu 'zbiorów (np. muzealnych)'. Taka metafora wygląda archaicznie w XXI wieku. Użycie tego związku wyrazowego też jest znamienne.

[3] Pol. Ludmiła Ryczkowa, biał. Людміла Рычкова, ros. Людмила Рычкова.

[4] W listopadzie 2007 r. Instytut Językoznawstwa został połączony z Instytutem Literatury, powstał Instytut Języka i Literatury. W 2012 r. wszedł on jako jednostka niesamodzielna (filia) do nowo utworzonego Centrum Badań Białoruskiej Kultury, Języka i Literatury.

Skaryny zajmowały się projektem „Problemy reprezentacji i budowy korpusu języka białoruskiego". Jego celem było ogólne wypracowanie koncepcji, a także zbiór i digitalizacja tekstów. Twórcy korpusu zaplanowali jego wersję demonstracyjną. Założona liczba słów – 500 tys. – miała odpowiadać wymogom reprezentacyjności. Wyniki tego projektu nie zostały jednak opublikowane.

Do roku 2006 został także ukończony projekt pt. „Zasoby komputerowego języka białoruskiego", który był realizowany na zamówienie Ministerstwa Informacji. Na wykonawcę wybrano Laboratorium Systemów Informacyjnych Wydziału Matematyki Stosowanej (Białoruski Uniwersytet Państwowy) pod kierunkiem Igora Sovpela[5]. W ramach projektu powstał korpus tekstów języków rosyjskiego i białoruskiego (10 mln słów) i rosyjsko-białoruski korpus równoległy (2 mln słów) (Рубашко 2006:74). Wyniki tego projektu w postaci zasobów lingwistycznych także nie zostały udostępnione. Rezultaty tego grantu zostały w całości przekazane zleceniodawcy – Ministerstwu Informacji, a ograniczony dostęp został przydzielony tylko niektórym zespołom naukowym. Na przykład na podstawie danych tego korpusu (ale bez bezpośredniego do niego dostępu – poświadczenia uzyskano od twórców korpusu) i tekstów z innych źródeł Aleksander Barkowicz[6] (Барковіч 2008) napisał i obronił rozprawę doktorską na temat nowej leksyki białoruskiej.

Mimo że sytuacja polityczna i socjolingwistyczna nie sprzyja badaniom języka białoruskiego w instytucjach państwowych, podejmowane są indywidualne inicjatywy naukowe. Tak więc, niezależnie od akademickich projektów korpusowych i badań nad przetwarzaniem języka naturalnego, w pierwszej dekadzie bieżącego wieku Fiodar Piskunow[7] pracował nad własną bazą danych, zawierającą dane gramatyczne słów języka białoruskiego. W 2004 r. na podstawie tych zasobów powstał program do sprawdzania ortografii białoruskiej „Litara" (kompatybilny z pakietem Microsoft Office edycji 2000-2003).

Ostatnich pięć lat zaowocowało lepszymi rezultatami. Najnowsze projekty (przynajmniej niektóre) zostały publicznie udostępnione.

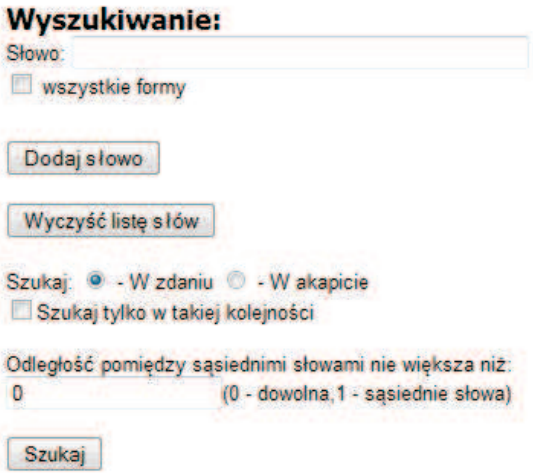
W ramach projektu BalticGRID II na początku 2010 r. powstał Corpus Albaruthenicum. Podstawowe założenia tego projektu (metody obliczeniowe i zastosowane systemy zachowywania danych) pozostawiały wiele do życzenia. Ale dzięki inicjatywie Ihara Mikłaszewicza[8], kierownika Laboratorium dynamiki systemów i mechaniki materiałów Białoruskiego Narodowego Uniwersytetu Technicznego i entuzjasty języka białoruskiego, korpus ten powstał. Autorzy projektu piszą, że „korpus tekstów akademickiego języka białoruskiego jest pierwszą próbą stworzenia korpusu obejmującego jeden zakres przedmiotowy" (http://grid.bntu.by/corpus_albaruthenicum), podkreślając zarazem, że w sy-

[5] Pol. Igor Sowpiel, ros. Игорь Совпель, biał. Ігар Соўпель.
[6] Pol. Aleksander Barkowicz, biał. Аляксандр Барковіч, ros. Александр Баркович.
[7] Pol. Fiodor Piskunow, biał. Фёдар Піскуноў, ros. Фёдор Пискунов.
[8] Pol. Igor/Ihar Mikłaszewicz, biał. Ігар Міклашэвіч, ros. Игорь Миклашевич.

tuacji, kiedy pisany język białoruski jest głównie używany w literaturze pięknej i gazetach, a w nauce tylko w pracach filologicznych, jest to naprawdę ważne. Celem projektu jest „stworzenie publicznego zbioru źródeł akademickiego języka w języku białoruskim, który dotychczas nie był łatwo dostępny dla szerokiej społeczności" (http://grid.bntu.by/corpus_albaruthenicum).

Korpus zawiera 350 tys. słów, 74 teksty naukowo-techniczne (szczegóły jego struktury nie zostały precyzyjnie opisane).

Rys. 14.1. Wygląd strony korpusu (w naszym tłumaczeniu, następne zdjęcia podane są z oryginalnym tekstem interfejsu).



Funkcje dostępne w tym korpusie są typowe dla konkordancera: wyszukiwanie tylko formy lub wyszukiwanie według lematu. Jeśli zapytanie zawiera kilka słów, to można ograniczyć wyniki do wyszukiwania słowa w jednym zdaniu albo w jednym akapicie, w dowolnej kolejności lub w takiej samej kolejności jak w zapytaniu. Można także określić odległość między słowami w tekście.

Na stronie wyników widnieje wyłącznie zdanie (dodatkowo można odtworzyć stronę z akapitem, jeśli wybrane zostało wyszukiwanie słów w ramach akapitu). Interfejs korpusu nie daje dostępu do pełnego tekstu ani opisu bibliograficznego źródła. Wyszukiwanie trwa przeciętnie 1-10 sekund – w zależności od częstości słów i trybu wyszukiwania (wyszukiwanie według lematu trwa dłużej).

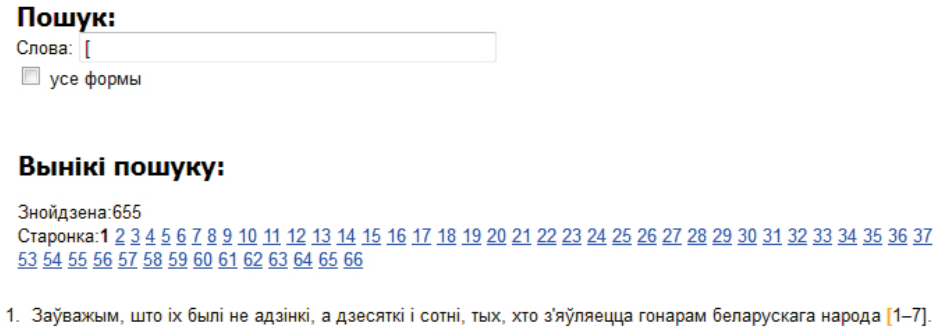
Jeżeli chodzi o strukturę wewnętrzną korpusu, to strona projektu zawiera informacje, że teksty zostały oznakowane za pomocą języka programowania, zbudowanego na bazie XML (http://grid.bntu.by/corpus_albaruthenicum) przez elektroniczną bazę danych języka białoruskiego, stworzoną w Instytucie Języka i Literatury Narodowej Akademii Nauk Białorusi (Капылоў, Кощанка

2010: 16). Wyniki wyszukiwania są wyświetlane po 10 zdań na stronie (nie można zmienić tego parametru). Interfejs korpusu jest bardzo prosty a możliwości opisywanego zasobu – ograniczone. Próbując wyszukać słowo z apostrofem, znaleźliśmy błąd bazy danych MySQL. Dla poprawnego wyszukiwania trzeba dodać symbol \ przed apostrofem. Zatem słowo *аб'ект* 'obiekt' musi być zapisane jako *аб\'ект*. Nie można też zostawiać spacji przed słowem wpisanym do pola wyszukiwania. W korpusie są pseudowyrazy: np. tam, gdzie w tekstach były formuły, zostały one zamienione na słowo FORMULA i w tej postaci teksty zostały umieszczone w korpusie.

Rys. 14.2. Przykład strony wyszukiwania w korpusie z wynikami (*Вынікі пошуку* 'Wyniki wyszukiwania', *Знойдзена* 'Znalezione', *Старонка* 'Strona').

Warto zauważyć, że można tu wyszukiwać nie tylko zwykłe hasła, ale też znaki przestankowe. W ten sposób możemy np. znaleźć zdania zawierające odwołania, por. niżej.

Rys. 14.3.



Ta sama baza gramatyczna została też zastosowana do Wielkiego Korpusu Języka Białoruskiego w Mińskim Lingwistycznym Uniwersytecie Państwowym (dalej – MLUP), który obejmuje właściwy korpus białoruski oraz kilka korpusów równoległych (angielski – Е.В. Таболич 2008), niemiecki – Е.В. Марковская, Т.А. Филимонова 2008, rosyjski) (Капылоў, Кощанка 2010: 15). Według Aleksandra Zubowa[9], kierownika projektów korpusowych MLUP, od 2006 r. MLUP wspólnie z Instytutem Języka i Literatury Narodowej Akademii Nauk Białorusi (dalej – IJL) pracował nad tematem „Stworzenie wielkiego korpusu tekstów białoruskiego języka i jego wykorzystanie do badań języka białoruskiego i jego stosunków z innymi językami Europy". Rezultatem tej współpracy jest białoruski podkorpus – 1 mln słów oraz korpus równoległy rosyjsko-białoruski – 300 tys. słów (Зубов 2010: 516). Format znakowania został wypracowany na podstawie standardu CES (http://www.cs.vassar.edu/CES/). Teksty zostały opracowane w trybie półautomatycznym (Зубов 2010: 516).

Od 2009 r. Zakład Informatyki i Lingwistyki Stosowanej MLUP w ramach tematu naukowego „Strukturalno-semantyczna analiza tekstów z użyciem metod lingwistyki korpusowej" pracuje nad stworzeniem rosyjsko-białoruskiego równoległego korpusu tekstów dydaktycznych (też CES i znakowanie półautomatyczne) (Зубов 2010: 516). Niestety, żaden z korpusów z MLUP nie jest publicznie dostępny. Pomimo tego doświadczenie i niektóre wyniki z tych propozycji były wykorzystane w trakcie pracy nad następnym projektem.

Od 2011 r. Instytut Języka Rosyjskiego Rosyjskiej Akademii Nauk wspólnie z Instytutem Języka i Literatury[10]. NANB pracuje nad Równoległym Korpu-

[9] Pol. Aleksander Zubow / Aliaksandr Zubau, biał. Аляксандр Зубаў, ros. Александр Зубов.

[10] Pol. Igor Kopyłow / Ihar Karyłou, biał. Ігар Капылоў, ros. Игорь Копылов; pol. Władzimir Koszczenko / Uladzimir Koszczanka, biał. Уладзімір Кощанка, ros. Владимир Кощенко; pol. Olga Mickiewicz, biał. Вольга Міцкевіч, ros. Ольга Мицкевич; pol. Inna Glinnik, biał. Інa Гліннік, ros. Инна Глинник.

sem Rosyjsko-Białoruskim oraz Białorusko-Rosyjskim. Według słów koordynatora Zespołu korpusów równoległych Narodowego Korpusu Języka Rosyjskiego (dalej – NKJR) Dmitrija Siczinawy[11], jest to kontynuacja wyżej opisanego projektu MLUP realizowanego pod kierunkiem Zubova (2010: 31). Jak pisze Siczinawa, wszystkie „równoległe korpusy NKJR są formatowane w XML i wyrównane według zdań" (Сичинава, Шведова 2010: 31). Korpusy są przetwarzane za pomocą analizatorów morfologicznych firmy Yandex i, jak poinformowano na stronie NKJR, wyszukiwanie działa w wyspecjalizowanym systemie *Yandex.Server* (http://company.yandex.ru/technology/server/). W marcu 2013 r. został ukończony projekt ze strony białoruskiej (ściślej mówiąc, część sfinansowana z Funduszu badań podstawowych Republiki Białoruś). Równoległe korpusy z NKJR są propozycją znacznie znacznie obszerniejszą, można więc oczekiwać udoskonaleń, chociaż brak informacji o dalszych planach. Na obecnym etapie korpus podkorpusów równoległych zawiera około 38 mln słów, z których rosyjsko-białoruska część wynosi 2,84 mln słow. Chociaż na stronie z informacją o typach korpusów NKJR (http://www.ruscorpora.ru/corpora-structure.html) *korpus rosyjsko-białoruski* i *korpus białorusko-rosyjski* wyraźnie rozróżniono, nie umieszczono o nich żadnych dodatkowych informacji dotyczących ich struktury czy danych ilościowych.

Korpus rosyjsko-białoruski znajduje się na stronie NKJR i faktycznie jest podkorpusem korpusu wielojęzycznego NKJR. Również interfejs wyszukiwania jest ten sam, a korpus rosyjsko-białoruski nazywa się w nim „podkorpusem użytkownika".

Metadane zawierają informacje o imieniu i nazwisku autora (1), tłumacza (7), tytule tekstu (2), roku urodzenia autora (3) i powstania tekstu (4), jego objętości – liczby zdań (5) i słów (6), oraz o języku oryginału (8). Nie każdy tekst ma opis według wszystkich elementów tej listy: obligatoryjne są pola (1), (2), (5), (6), (8).

Tab. 14.1.

Автор (1)	**Светлана Алексиевич**	**А. Бутэвіч**
Название (2)	Чернобыльская молитва	Дні Максіма Багдановіча ў Ялце
Дата рождения автора (3)	1948	
Дата создания (4)	1996	2011
Предложений (5)	19914	86
Словоформ (6)	117139	1531
Переводчик (7)	Мікола Гіль	Л.Ф. Кондухович
Язык (8)	Рус	Бел

[11] Pol. Dmitrij Siczinawa, ros. Дмитрий Сичинава.

Wyszukiwanie według dodatkowych parametrów, np. tytuł, autor, rok urodzenia autora (zakres), rok napisania tekstu (zakres), tłumacz, jest możliwe za pomocą strony umożliwiającej tworzenie własnego podkorpusu z tekstów korpusów równoległych. Ale z powodu niekonsekwentnego i rzadkiego oznakowania ten interfejs nadaje się tylko do tego, żeby sprawdzić, czy korpus zawiera dany tekst. Nie można polegać na wynikach wyszukiwania w korpusie stworzonym przez wybór metadanych. Jednak trzeba wyraźnie podkreślić, że wyszukiwanie w podkorpusie stworzonym na podstawie podstawowych elementów zewnętrznego znakowania działa lepiej (ale nie całkiem dobrze); pozwala bowiem tworzyć nowe podkorpusy, np. korpusy autorów. Tak więc, jeśli wpisać do pola *Автор текста* 'Autor tekstu' *Уладзімір Караткевіч*[12], to wyświetli się strona umożliwiająca tworzenie podkorpusu.

Уладзімір Караткевіч. Зброя (1981) [омонимия не снята]
Уладзімір Караткевіч. Дзікае паляванне караля Стаха (1964) [омонимия не снята]
Уладзімір Караткевіч. Маленькая балерына (1961) [омонимия не снята]
Уладзімір Караткевіч. Сівая легенда (1960) [омонимия не снята]
Уладзімір Караткевіч. Лісты не спазняюцца ніколі (1958) [омонимия не снята]

Niestety, nawet podstawowe znakowanie zawiera błędy. Jeśli wpiszemy tylko nazwisko autora *Караткевіч*, to otrzymamy więcej tekstów:

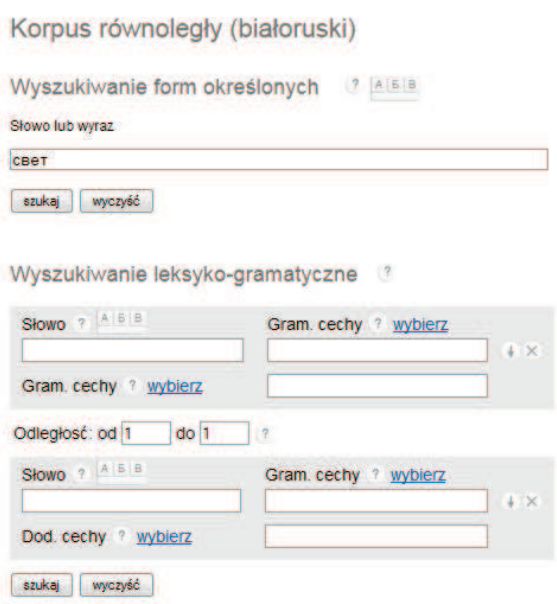
Уладзімір Караткевіч. Зброя (1981) [омонимия не снята]
Уладімір Караткевіч. Ладдзя Роспачы (1968-1975) [омонимия не снята]
Уладзімір Караткевіч. Дзікае паляванне караля Стаха (1964) [омонимия не снята]
Уладімір Караткевіч. Каласы пад сярпом тваім (1962-1964) [омонимия не снята]
Уладзімір Караткевіч. Маленькая балерына (1961) [омонимия не снята]
Уладзімір Караткевіч. Сівая легенда (1960) [омонимия не снята]
Уладзімір Караткевіч. Лісты не спазняюцца ніколі (1958) [омонимия не снята]

Jak widać, w znakowaniu dwóch tekstów imię zostało zapisane błędnie (opuszczona została litera *З*).

Interfejs wyszukiwania, jak zostało wcześniej wspomniane, jest wspólny dla wszystkich korpusów NKJR. Poniżej prezentujemy zdjęcie wyszukiwarki opatrzonej polskimi napisami (tłum. wł.), w istocie narzędzia NKJR są dostępne tylko w języku rosyjskim, chociaż wśród korpusów równoległych znajduje się np. korpus polsko-rosyjski. W polu wyszukiwania wpisane jest słowo *свет* 'świat'.

[12] Pol. Uładzimir Karatkiewicz – pisarz, klasyk literatury białoruskiej.

Rys. 14.4.



Interfejs pozwala na wyszukiwanie niektórych form (przy czym można korzystać z wyrażeń regularnych, np. * dla dowolnego znaku). Ten tryb działa po naciśnięciu pierwszego od góry przycisku *искать* 'szukaj'.

Można także wyszukiwać leksemy we wszelkich formach albo leksemy o określonych cechach morfologicznych, albo też formy leksemów posiadające pewne właściwości. Dla wyszukiwania w tym trybie trzeba wypełnić pola pod linijką *Лексико-грамматический поиск* 'Wyszukiwanie leksykalno-gramatyczne' i nacisnąć dolny przycisk *искать* 'szukaj'. Interfejs z dwoma przyciskami obsługującymi jedną funkcję wydaje się nie najlepszym rozwiązaniem.

Oprócz tego w okienku wyboru kategorii gramatycznych znajdują się tylko wartości charakterystyczne dla języka rosyjskiego, chociaż morfologia białoruska różni się rosyjskiej. Np. czasownik rosyjski w 1 osobie trybu rozkazującego ma tylko formę pobudzającą[13] typu *пойдёмте*, natomiast w białoruskim obok niej (*хадземце*) używa się jeszcze formy typu *ходзьма* 'idźmy', analogicznej do polskiego *chodźmy*. Wołacza w białoruskim w zasadzie już nie ma, choć bywa używany w kontekstach familiarnych (*сынку* 'synku', *браце* 'bracie'), a w rosyjskim powstaje nowy wołacz (*мам* 'mamo', prócz tradycyjnych *Боже* i *Господи*). W wynikach na zapytanie *сынку* i *браце*, pierwsza forma została wyświetlona jako celownik/miejscownik, a druga jako miejscownik.

Ponadto teksty białoruskie zostały oznakowane za pomocą uboższego zestawu tagów niż rosyjskie. Z tego powodu niektóre klasy gramatyczne, które

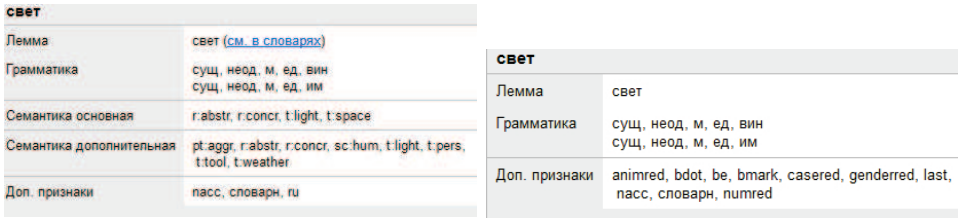
[13] W interfejsie korpusu nazywana jest 'rozkazujący II'.

możemy wybrać w interfejsie, zostaną odnalezione tylko w tekstach rosyjskich, nawet jeśli w białoruskim istnieją ich identyczne odpowiedniki (np. *повелительное наклонение 2* 'tryb rozkazujący II' z znaczeniem wspólnego działania).

Pocieszające jest natomiast, że specyfika tekstów białoruskich została w pewnym stopniu wzięta pod uwagę; teksty zostały ujednolicone ortograficznie – na zapytanie *свет* zostaną także zwrócone poświadczenia zawierające segment *съвет*. W przeciwnym kierunku wyszukiwarka nie działa; nie znajdzie niczego na zapytanie o postać *съвет*.

Do każdego słowa w wynikach dodana jest informacja, którą można wyświetlić od razu. Podana jest ona w postaci skrótów, jasnych dla twórców korpusu, ale nie dla jego użytkowników. Na rysunkach podane są okienka z informacją o słowie *свет* w tekście rosyjskim (z lewej strony) oraz w białoruskim (z prawej strony)[14].

Rys. 14.5.



свет	
Лемма	свет (см. в словарях)
Грамматика	сущ, неод, м, ед, вин сущ, неод, м, ед, им
Семантика основная	r:abstr, r:concr, t:light, t:space
Семантика дополнительная	pt:aggr, r:abstr, r:concr, sc:hum, t:light, t:pers, t:tool, t:weather
Доп. признаки	nacc, словарн, ru

свет	
Лемма	свет
Грамматика	сущ, неод, м, ед, вин сущ, неод, м, ед, им
Доп. признаки	animred, bdot, be, bmark, casered, genderred, last, nacc, словарн, numred

Największą wadą opisywanego korpusu jest nierozróżnianie języków. Oznacza to, że ciąg liter cyrylickich *свет* będzie wyszukiwany we wszystkich tekstach podkorpusu, niezależnie od języka. Tymczasem w języku białoruskim znaczy on 'świat', a w rosyjskim – 'światło'[15]. Zmniejsza to zaufanie badacza do statystyk, bo nie wiadomo, co konkretnie zostało znalezione, a wyniki trzeba sprawdzać ręcznie.

Nie jest też jasne, jak mogło dojść do tego, że przynajmniej jedna para tekstów została błędnie oznakowana. Chodzi o powieść *W kraju ptaku rajskiego* Janki Maura[16]: tekst białoruski został oznakowany jako rosyjski, a jej tłumaczenie na rosyjski – jako tekst białoruski.

Homonimię w korpusach równoległych NKJR usuwa się ręcznie, ale na podstawie doświadczenia naszej pracy z korpusem, możemy stwierdzić, że w białoruskich tekstach homonimia została zachowana w całości.

[14] *Лемма* 'lemma', *грамматика* 'gramatyka', *доп. признаки* 'cechy dodatkowe', *сущ.* 'rzeczownik', *неод.* 'nieżywotny', *м.* 'rodzaj męski', *ед.* 'liczba pojedyncza', *вин.* 'biernik', *им.* 'mianownik'.

[15] Znaczenie 'świat' też istnieje, ale jest archaiczne i spotyka się w związkach frazeologicznych i podobnych do nich kontekstach.

[16] Janka Maur – białoruski pisarz radziecki, autor książek dla dzieci i sztuk teatralnych.

Objętość białoruskiej części szacuje się na 1,5 mln słów. Nie jest to zatem wystarczająca liczba danych do przeprowadzania reprezentatywnych badań językowych.

Jednak mimo wad korpus na pewno może być stosowany do badania jednego lub obu języków, co swoim przykładem pokazuje jeden z twórców projektu, Dmitrij Siczinawa (Сичинава 2012) (w jego pracy zostały wykorzystane też dane z innych źródeł, ponieważ materiał korpusowy okazał się niewystarczający).

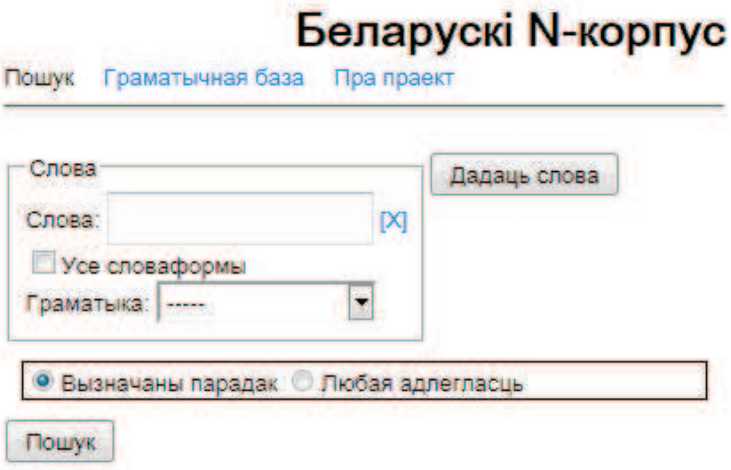
Prace nad zasobami lingwistycznymi, stosowanych do tworzenia niektórych z wyżej opisanych korpusów, były prowadzone w ostatniej dekadzie. Na razie wobec IJL stoi zadanie ułożenia nowego słownika języka białoruskiego (oficjalne zadanie to „Wypracowanie podstaw teoretycznych, bazy leksykograficznej i stworzenie automatycznego analizatora tekstów dla nowego fundamentalnego objaśniającego słownika języka białoruskiego i połączonego słownika białoruskich gwar ludowych") (Капылоў, Кощанка 2010: 12). Korpusy wchodzą w zakres działalności przede wszystkim Działu Leksykografii i Leksykologii, którym kieruje Ihar Kapyłou, prace korpusowe koordynuje Uładzimir Koszczanka.

Korpus, który jest w trakcie opracowywania, jeszcze nie ma ustalonej nazwy, umownie można go nazwać Akademickim Korpusem dla Słownika Objaśniającego. Jego twórcy (Kapyłou i Koszczanka) sądzą, że nowy korpus może powstać na podstawie korpusów MLUP i Corpus Albaruthenicum (Капылоў, Кощанка 2010: 15). W 2010 r. już istniał zbiór tekstów (obejmujący obszary tematyczne: media, nauka, literatura piękna) o objętości 20 mln słów. Autorzy projektu zamierzają go uzupełnić, dodając teksty z innych kanałów, np. z internetu, radia i telewizji, także teksty ustne (Капылоў, Кощанка 2010: 16). Trudno ocenić, w jakim stopniu te zamierzenia są realistyczne, jednak nawet dzisiejsza liczba – 20 mln – już jest przesunięciem górnej granicy. Z drugiej strony, nie można pominąć faktu, że korpus interesuje autorów projektu nie jako zasób samodzielny, ale jako baza danych będąca podstawą słownika.

W maju 2013 r. pojawiła się strona Białoruski N-korpus (http://bnkorpus.info/) – interfejs do wyszukiwania w bazie tekstów z 1945 r. o objętości 15 mln form (http://bnkorpus.info/about.html). Korpus składa się z utworów literatury pięknej, jest morfologicznie oznakowany (homonimia nie została usunięta). Teksty zawierają metadane (autor, tytuł, rok), z których w wynikach wyszukiwania pokazywane są tylko nazwisko i imię autora oraz tytuł tekstu. Autorzy informują, że planowana objętość korpusu to 100 mln słów. Strona nie zawiera danych na temat twórców korpusu ani o procesie powstania projektu. Interfejs jest wyposażony w podstawowe narzędzia (podobnie jak w interfejsach korpusów wyżej opisanych): bezpośrednie wyszukiwanie słowa, wyszukiwanie słowa we wszystkich formach, wyszukiwanie kilku słów, także nie sąsiadujących bezpośrednio ze sobą. Projekt jest jeszcze w trakcie realizacji, więc nie wszyst-

kie narzędzia działają w nim poprawnie, ale już teraz zasób ten jest przydatny do wybierania przykładów[17].

Rys. 14.6.



Oprócz właściwych projektów korpusowych istnieją inne projekty z zakresu przetwarzania języka i zawierające zadania stworzenia językowych zasobów komputerowych.

Wymieniony wcześniej twórca komputerowej bazy gramatycznej języka białoruskiego Fiodar Piskunow wiosną 2012 r. na podstawie owej bazy wydał *Wielki słownik języka białoruskiego* (Піскуноў 2012) o objętości 223 tys. haseł. W Laboratorium Rozpoznawania i Syntezy Mowy Połączonego Instytutu Problemów Informatyki NANB pod kierunkiem Jurasia Hecewicza[18] został stworzony szereg zasobów lingwistycznych dla specyficznych zadań laboratorium, które mogą być też stosowane do przetwarzania tekstów w celach lingwistycznych (Hetsevich, Hetsevich 2012).

W podsumowaniu trzeba powiedzieć, że stan lingwistyki korpusowej (jak i szerzej – lingwistyki i w nauki w ogóle) oraz sytuacja polityczno-socjolingwistyczna w kraju nie wpływa dobrze na rozwój projektów. Jednym z głównych problemów jest brak młodych specjalistów. Jeśli spojrzeć na publikacje z białoruskiej lingwistyki korpusowej, to można zauważyć, że w 2013 r. lista nazwisk jest – z małymi wyjątkami – ta sama, co w roku w czasach powstania tej

[17] Elementy interfejsu wyszukiwania: *Слова* 'Słowo', *Дадаць слова* 'Dodaj słowo', *Усе словаформы* 'Wszystkie formy słowa', *Граматыка* 'Gramatyka' (chodzi o część mowy), *Вызначаны парадак* 'Szyk określony', *Любая адлегласць* 'Dowolna odległość', *Пошук* 'Wyszukiwanie'.

[18] Pol. Juraś Hecewicz, biał. Юрась Гецэвіч, ros. Юрий Гецевич.

dziedziny na Białorusi. Z drugiej strony, pewne prace już zostały wykonane i można się na nich opierać. Ponadto niektóre projekty korpusowe i zbliżone zostały całkowicie lub częściowo zrealizowane przez entuzjastów, osoby, które zawodowo „nie muszą" tym się zajmować. Czasem też białoruscy studenci i doktoranci wykazują zainteresowanie pracą tworzeniem zasobów korpusowych (О.А. Волчек, В.В. Порицкий), choć w żadnej instytucji edukacyjnej Białorusi w programach nauczania kierunków lingwistycznych nie ma praktycznych kursów z zakresu przetwarzania języka naturalnego.

Ogólnie rzecz biorąc, korpusy na Białorusi nie reprezentują wysokiego poziomu w porównywaniu z korpusami budowanymi w innych krajach.

Bibliografia

BalticGrid II. Corpus Albaruthenicum; http://grid.bntu.by/projects

BalticGrid-II Web Site; http://www.balticgrid.org/

Corpus Albaruthenicum; http://grid.bntu.by/corpus/

Corpus Encoding Standard; http://www.cs.vassar.edu/CES/

Belarusian National Technical University – Corpus Albaruthenicum; http://grid.bntu.by/corpus_albaruthenicum

Hetsevich Y., Hetsevich S. (2012). *Overview of Belarusian and Russian dictionaries and their adaptation for NooJ // Selected Papers from the NooJ 2011 International Conference "Automatic Processing of Various Levels of Linguistic Phenomena", Dubrovnik, Croatia.* Newcastle, s. 29-40.

Барковіч А.А. (2008). *Лексічны патэнцыял беларускай мовы ў кантэксце корпуснай лінгвістыкі (на матэрыяле новай дзеяслоўнай лексікі). Аўтарэферат дысертацыі на атрыманне вучонай ступені кандыдата филалагічных навук.* Мінск.

Беларускі N-корпус: пошук па корпусе, http://bnkorpus.info/

Беларускі N-корпус: Пра праект, http://bnkorpus.info/about.html

Волчек О.А., Порицкий В.В. (2012). *Проект корпуса белорусскоязычной периодики и художественной прозы* W: *Компьютерная лингвистика: научное направление и учебная дисциплина : сборник научных статей.* Вып. 2. В.И. (red.) Коваль В.И. [и др.]. – Гомель, s. 16-19.

Зубов А.В. (2010). *Лингво-методические возможности русско-белорусского параллельного корпуса текстов / А.В. Зубов* W: *Русский язык: Исторические судьбы и современность. ИВ Международный конгресс исследователей русского языка. Труды и материалы.* Москва, s. 516-517.

Капылоў І.Л., Кощанка У.А. (2010). *Корпус тэкстаў як аснова аптымізацыі падрыхтоўкі фундаментальнага «Тлумачальнага слоўніка беларускай мовы»* W: *Компьютерная лингвистика: научное направление и учебная дисциплина: сборник научных статей.* Вып. 1. (red.) Коваль В.И. [и др.]. Гомель, s. 12-17.

Капылоў І., Кощанка У., Міклашэвіч І. (2010). *Corpus Albaruthenicum як частка міжнароднага праекта «BalticGrid-II»*. „В мире науки", №10 (92), http://innosfera.org/node/822.

Марковская Е.В., Филимонова Т.А. (2008). *О возможностях использования в учебном процессе немецко-белорусского параллельного корпуса тэггированных текстов* W: *Актуальные проблемы прикладной лингвистики. Сборник научных статей. В двух частях.* (red.) Зубов А.В. Минск, ч. 2, s. 35-37.

Национальный корпус русского языка. Состав и структура, http://www.ruscorpora.ru/corpora-structure.html

Піскуноў Ф.А. (2012). *Вялікі слоўнік беларускай мовы : арфаграфія, акцэнтуацыя, парадыгматыка (каля 223 000 слоў).* W: Ф.А. Піскуноў. Мінск: Тэхналогія, XVI + 1208 s.

Рубашко Н.К. [и др.] (2006). *Компьютерный фонд белорусского языка и его приложения* W: *Информационные системы и технологии (IST'2006) : Материалы* III *Международной конференции. В двух частях.* Минск, ч. 2. s. 71-76.

Рычкова Л.В. (1994). *К вопросу о машинном фонде белорусского языка.* W: *Шануючы спадчыну Я. Карскага: Чацвёртыя навуковыя чытанні. В двух частях.* ч. 2. Гродна, s. 132-134.

Рычкова Л.В. (1998). *Использование специальных полнотекстовых баз данных в исследовании номинативных единиц различных уровней.* W: *Словообразование и номинативная деривация в славянских языках: материалы VI международной научной конференции. Гродно, 28-29 мая 1998 г. В двух частях.* ч. 2. Гродно, s. 107-115.

Рычкова Л.В. (1999). *Камп'ютэрная версія драматургічных твораў Янкі Купалы: прынцыпы стварэння, напрамкі выкарыстання.* W: *Міжнародныя купалаўскія чытанні: матэрыялы навуковай канферэнцыі, Гродна, 25-27 лістапада 1997 г.* Гродна, s. 257-261.

Сичинава Д.В. (2012). *Русская конструкция с было и белорусский плюсквамперфект: сопоставительный анализ на материале белорусско-русского и русско-белорусского параллельных корпусов* W: *Компьютерная лингвистика: научное направление и учебная дисциплина : сборник научных статей.* Вып. 2. (red.) Коваль В.И. [и др.]. Гомель, s. 95-104.

Сичинава Д.В., Шведова М.А. (2010). *Параллельные корпуса в составе национального корпуса русского языка: технологии и решаемые задачи* W: *Компьютерная лингвистика: научное направление и учебная дисциплина: сборник научных статей.* Вып. 1. (red.) Коваль В.И. [и др.]. Гомель, s. 30--34.

Совпель И.В. (1994). *Машинный фонд белорусского языка и его приложения (проект).* W: *Современные информационные системы и технологии.* Минск, s. 11-13.

Таболич Е.В. [и др.] (2008). *Об англо-белорусском параллельном тэггированном корпусе учебных текстов.* W: *Актуальные проблемы прикладной лингвистики. Сборник научных статей. В двух частях.* (red.) Зубов А.В. ч. 2. Минск, s. 28-29.

Яндекс. Поисковые технологии; http://company.yandex.ru/technology/server/

15. Praktyczny przewodnik po korpusach równoległych. Wiadomości wstępne. Korpus ParaSol i Korpus polsko-rosyjski UW

Marek Łaziński
Uniwersytet Warszawski
Instytut Języka Polskiego

Korpus równoległy, dwu- lub wielojęzyczny umożliwia znalezienie zdania z wyszukiwanym słowem lub konstrukcją w języku wyjściowym oraz odpowiedników tego zdania w języku docelowym. Od korpusów równoległych (ang. *parallel*) należy odróżnić korpusy porównywalne (*comparable*), grupujące teksty oryginalne teksty w dwóch lub wielu językach z tej samej dziedziny, np. teksty medyczne, prawne czy wiadomości prasowe z tego samego okresu. Korpusy równoległe są znacznie bardziej popularne i często stosowane zarówno praktyczne w nauce języków i w pracy tłumacza, jak też teoretycznie w badaniach lingwistycznych.

Równoległe publikowanie tekstów w różnych językach ma długą przedkorpusową tradycję, związaną – jak mechanizm konkordancji – z biblistyką i chrześcijaństwem. Najstarszym znanym tekstem równoległym jest dekret staroegipski z II w. p.n.e. zapisany hieroglifami oraz po grecku na kamieniu z Rozetty; bez niego nie udałoby się odczytać hieroglifów.

Od epoki renesansu znane są dwujęzyczne wydania Biblii i jej fragmentów. W roku 1511 Erazm z Rotterdamu wydał *Novum Instrumentarium* – równoległy tekst greckiego Nowego Testamentu oraz własnego przekładu łacińskiego (poprawionej Wulgaty). W roku 1519 wydano w Madrycie *Poliglotę Kompluteńską* – równoległe przekłady hebrajskie, łacińskie oaz greckie z łacińskimi glosami.

W roku 1555 Konrad Gessner wydał *Mithridates de differentis linguis* z modlitwą *Ojcze nasz* w 22 językach, a w roku 1806 u progu narodzin nowożytnego językoznawstwa Johann Ch. Adelung opublikował aż 500 przekładów Modlitwy Pańskiej: *Mithridates oder allgemeine Sprachenkunde, mit dem Vater--Unser als Sprachprobe in beinahe fünfhundert Sprachen und Mundarten* (imię starożytnego poligloty Mithrydatesa długo było w Europie alegorią wielojęzyczności). Traktat Adelunga można uznać za początek porównawczej analizy języka,

w której porównywanie przekładów służy nie tylko praktycznemu zrozumieniu tekstu.

Dzisiejsze korpusy równoległe także łączą cele teoretyczne z praktycznymi, są bowiem doskonałym narzędziem dla tłumaczy. Pierwszy korpus równoległy dostępny w sieci – Hansard – powstał ze stenogramów debat parlamentu kanadyjskiego, które i tak były archiwizowane w języku angielskim i francuskim (http://www.isi.edu/natural-language/download/hansard). Hansard to nazwisko XVII-wiecznego autora systemu transkrypcji debat parlamentu brytyjskiego.

Obecnie w sieci dostępne są liczne korpusy dwu- i wielojęzyczne, w tym korpus debat Parlamentu Europejskiego – http://www.statmt.org/europarl i europejskich aktów prawnych. O udostępnianie na wspólnej platformie różnych zasobów korpusowych języków europejskich dbają lingwiści w programach META-NET (META – Multilingual Europe Technology Alliance) i CLARIN (Common Language Resources and Technology Infrastructure). Dostępne są m.in. akty prawne UE w korpusie JRC-Acquis – http://www.clarin.eu/acquis-communautaire, doniesienia prasowe w językach UE w korpusie Europress – http://www.presseurop.eu. Komentarze polityczne publicystów z różnych krajów w kilku wersjach językowych (choć nie po polsku) gromadzi praski projekt Syndicate – http://www.project-syndicate.org.

Jeśli chodzi o analizę porównawczą języków słowiańskich, trzeba wspomnieć o kilku ważnych projektach, z których dwa dostępne *online* zostaną tu omówione dokładniej. Najstarszym słowiańskim korpusem równoległym jest Amsterdam Slavic Parallel Aligned Corpus – ASPAC. Zawiera on kilkadziesiąt tekstów literackich wydanych po rosyjsku oraz w innych językach słowiańskich i zachodnioeuropejskich. Tekstów rosyjskich jest 69 (3 mln słów), ale odpowiedników w innych językach – znacznie mniej. Tekstów polskich jest 58, w tym 3 oryginalne. Korpus jest dostępny jedynie lokalnie na Uniwersytecie Amsterdamskim, informacje na stronie: http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/b/a/a.a.barentsen/a.a.barentsen.html.

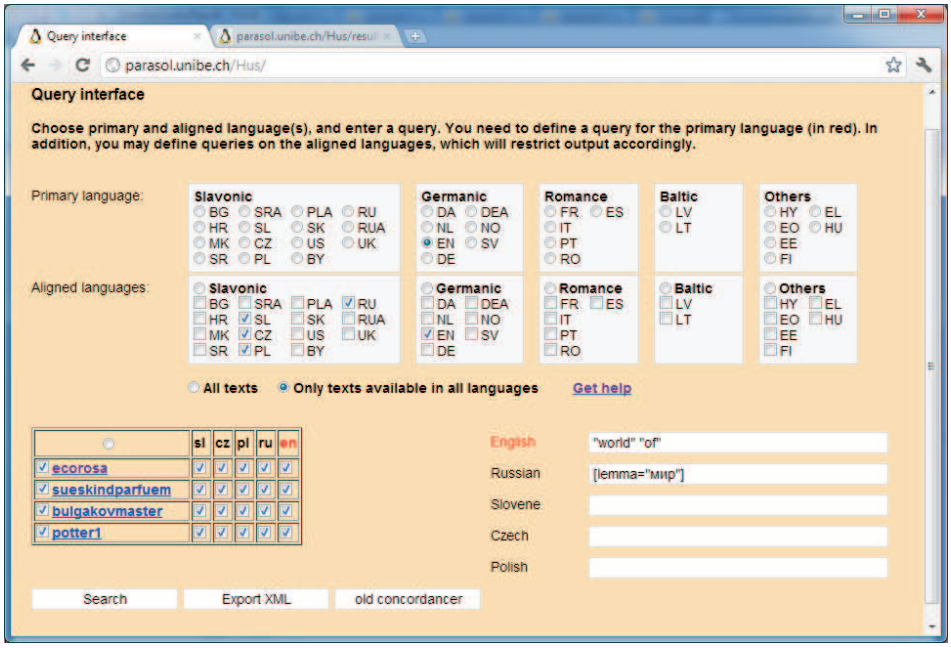
Inne korpusy równoległe zawierają tekst rosyjski, np. na uniwersytecie w Oslo dostępny jest lokalnie korpus wielkości 8 mln słów w tekstach rosyjskich, norweskich i angielskich: http://www.hf.uio.no/ilos/english/research/projects/run/corpus. Duży, 30-milionowy korpus polsko-rosyjski powstał na Uniwersytecie Warszawskim. Dwa największe i najważniejsze słowiańskie korpusy wielojęzyczne to: InterCorp (zob. rozdział *Praktyczny przewodnik po korpusie równoległym InterCorp*) ramach Czeskiego Korpusu Narodowego oraz ParaSol tworzony we współpracy uniwersytetów w Bernie i Ratyzbonie. W tym miejscu przedstawimy korpus ParaSol oraz Korpus Polsko-Rosyjski UW.

ParaSol (Parallel aligned corpus of Slavic and Other Languages) zawiera teksty literackie we wszystkich językach słowiańskich oraz kilkunastu innych europejskich. Korpus liczy 25 mln słów w 31 tekstach przełożonych na 32 języki,

z czego 5 tekstów jest dostępnych w ponad 12 językach. W odróżnieniu od znacznie większego korpusu InterCorp w ParaSolu większą wagę przykłada się do reprezentacji i zrównoważenia jak największej liczby języków i żaden z tych języków nie jest podstawowy, tak jak podstawowy jest czeski w korpusie InterCorp, gdzie każdy tekst musi mieć czeską wersję.

W każdym wyszukiwaniu należy wybrać jeden język źródłowy oraz jeden lub więcej języków docelowych. Po ograniczeniu wyszukiwania do przekładów dostępnych jednocześnie po angielsku, polsku, czesku, słowacku i rosyjsku liczba tekstów spada z 31 do 4: *Imię róży* Eco, *Pachnidło* Suskinda, *Mistrz i Małgorzata* Bułhakowa oraz *Harry Potter* Jane Rowling.

Rys. 15.1. Strona wyszukiwania ParaSol.



Najprostsze wyszukiwanie obejmuje słowa ortograficzne. Służy do tego operator *world*. Aby wyszukać polskie słowo *świata*, należy wpisać [word="świata"][1] lub po prostu "świata". Wyszukiwanie słów zachowuje kasztowość, a zatem aby znaleźć słowa zaczynające się małą lub wielką literą, należy zastosować wyrażenie regularne [word="[Śś]wiata"]. Na ilustracji przedstawiono wyszukiwanie angielskiego ciągu *world of*, a w przekładzie rosyjskim – leksem *mir*. Możliwość ograniczania wyszukiwania do obu badanych języków bardzo się przydaje przy analizie przekładu, jeśli chcemy np. sprawdzić, kiedy angielskie *word* tłumaczone jest na *mir*, a kiedy nie.

[1] W wyszukiwarce NKJP zamiast operatora *word* mamy *orth*.

W wyszukiwaniu ciągów znaków można stosować wyrażenia regularne. Jeśli np. chcemy znaleźć słowo, które zawiera co najmniej 3 litery, wpisujemy [word=".*a.*a.*a.*"], a żeby znaleźć słowo, które zawiera dokładnie 3 litery *a* – [orth=".*a.*a.*a.*" & orth!=".*a.*a.*a.*a.*"] lub ".*a.*a.*a.*[^a].*"

Wyszukiwaniu leksemów służy operator *lemma*[2], np. [lemma="świat"]. Wyszukiwarka leksemów nie reaguje na kasztowość, znajdzie więc formy *Świata*, *świata*, *świecie* itd. Na ilustracji przedstawiono wyszukiwanie rosyjskiego leksemu *мир* odpowiadającego angielskiemu ciągowi *world of*.

Nie we wszystkich językach teksty są lematyzowane i tagowane gramatycznie, ale w polskim, rosyjskim, czeskim, słoweńskim i słowackim – tak. Projekt ParaSol wykorzystuje tagsety poszczególnych korpusów, a więc formułowanie zapytań wymaga ich znajomości (podobnie jest w korpusie Intercorp). Aby wyszukać część mowy lub wartość kategorii gramatycznej, należy odpowiedni symbol umieścić w polu wartości operatora [tag]. Wpisaną wartość trzeba poprzedzić symbolem wielokrotności dowolnego znaku .* i taki sam symbol postawić po niej.

Oto pytania o bezokolicznik w trzech różnych językach i tagsetach:
polski: [tag=".*inf.*"]
rosyjski: [tag=".*Vmn.*"]
czeski: [tag=".*Vf.*"].
Stosowne tagsety są dostępne odpowiednio na stronach NKJP i Czeskiego Korpusu Narodowego (www.korpus.cz), a dla języka rosyjskiego na stronie slawistyki Uniwersytetu w Leeds: http://corpus.leeds.ac.uk/mocky/msd-ru.html. Tagsety dla rosyjskiego oraz dla języków południowosłowiańskich są zgodne z systemem MultexEast (http://nl.ijs.si/ME). Warto zwrócić uwagę na różnice między tagami i składnią zapytań dla różnych języków. W zapytaniach rosyjskich i czeskich ważne są nie tylko odpowiednie symbole części mowy i kategorii, ale też ich kolejność. W tagsecie polskim, zgodnym z NKJP, kolejność symboli nie odgrywa roli; za to klasy gramatyczne są węższe niż w tradycyjnej gramatyce szkolnej. Bezokolicznik jest w tagsecie rosyjskim i czeski reprezentowany jako uszczegółowienie czasownika, odpowiednio VMn (m – czasownik pełnoznaczny, n – bezokolicznik) i Vf (f – bezokolicznik), a w tagsecie polskim jest oddzielnym fleksemem (inf) odrębnym od reszty form czasownikowych, ponieważ ma inne właściwości fleksyjne i inną łączliwość.

W korpusach równoległych prawdopodobnie częściej niż struktury gramatyczne porównujemy leksemy, ich znaczenia i konteksty. Dlatego zakończmy prezentację korpusu ParaSol porównaniem kolokacji nazw gorących napojów oraz różnej roli kulturowej herbaty i kawy. Wybierzmy trzy języki słowiańskie: polski, rosyjski i czeski i wyszukajmy [lemma="herbata"].

[2] W wyszukiwarce NKJP zamiast operatora *lemma* mamy *base*.

Nie jest zaskoczeniem, że *herbata* ma stosunkowo wyższą frekwencję w korpusie rosyjskim niż w czeskim, a *kawa* – wyższą w czeskim niż w rosyjskim, nie dziwi nas też stosunkowo wysoka frekwencja *herbaty* w polszczyźnie, bliższa frekwencji rosyjskiej niż czeskiej. Wyszukajmy jednak słowo *herbata* w tekstach polskich i spójrzmy dokładniej na odpowiedniki rosyjski i czeskie. W tekstach rosyjskich polskiej *herbacie* zawsze odpowiada *čaj* i odwrotnie. Jednak w czeskich przekładach sytuacja jest bardziej skomplikowana:

Rys. 15.2. *Herbata* i jej czeskie odpowiedniki w korpusie ParaSol.

Otóż w czeskim przekładzie *Pamiętnika znalezionego w wannie* Lema kulturową rolę herbaty jako napoju dla gościa przejmuje kawa:

Poprosił mnie do gabinetu mieszczącego się po przeciwnej stronie korytarza, poczęstował **herbatą**. . .
Pozval mě do protější pracovny, počastoval **kávou**. . .
Затем пригласил меня в кабинет , находившийся напротив по коридору , угостил меня ***чаем***. . .

Na 37 wystąpień słowa *herbata* w tej powieści w czeskim przekładzie aż 22 razy mamy *kawę*. Taka adaptacja nie mogłaby się zdarzyć w przekładzie tekstu silnie osadzonego w tradycji kraju, gdzie pije się dużo herbaty, jak choćby w przekładzie *Mistrza i Małgorzaty* czy *Harry'ego Pottera* (oba teksty dostępne w porównaniu). Ale akcja *Pamiętnika* dzieje się w dalekiej przyszłości w nieznanym kraju, a więc każdy przekład kształtuje zwyczaje tego kraju na obraz

własnego języka i kultury. Częstowanie gościa herbatą wydało się czeskiemu tłumaczowi dziwaczne, dokonał więc swobodnej lokalizacji przekładu.

Przykład zastosowania korpusu wielojęzycznego w badaniach aspektologicznych słowiańskiego opisuje twórca ParaSola, Ruprecht von Waldenfels (2010). Także herbata i kawa przyciąga uwagę autora – Waldenfels (2012).

Oprócz ParaSola, InterCorpu i innych dużych projektów korpusów wielojęzycznych na wielu uniwersytetach i w innych ośrodkach buduje się korpusy dwujęzyczne, często na potrzeby lokalne. Poniżej przedstawiamy korpus polsko-rosyjski tworzony w latach 2010-2013 na Uniwersytecie Warszawskim w ramach grantu Narodowego Centrum Nauki (NN104056638). Historię projektu opisuje artykuł Łaziński et al. (2012).

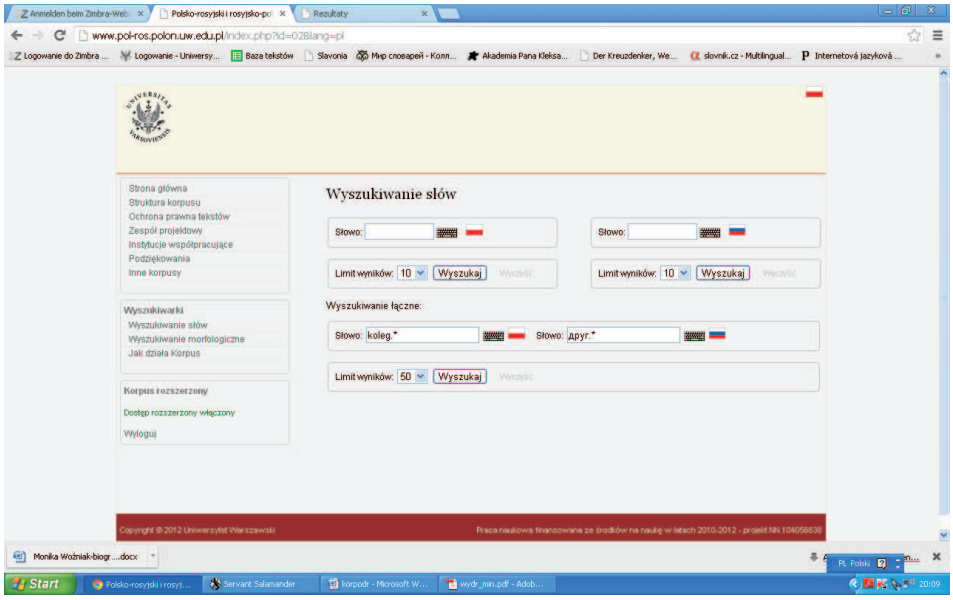
Projekt ten nie wyróżnia się na tle ParaSola, InterCorpu i innych korpusów równoległych poziomem anotacji czy możliwościami wyszukiwania, tylko wielkością, zróżnicowaniem tekstów oraz prostotą interfejsu. Korpus Polsko-Rosyjski UW zawiera 30 milionów słów (po 15 mln słów polskich i rosyjskich) z przekładami w ponad 200 tekstach literackich, stu kilkudziesięciu tekstach prasowych, prawnych i religijnych. Jest to więc zasób znacznie większy niż udział tekstów w wersji polskiej i rosyjskiej w korpusie ParaSol (28 tekstów) czy InterCorp (6 tekstów). Wartością korpusu odróżniającą go od projektów zagranicznych jest także obecność tekstów prasowych interesujących dla obu sąsiednich krajów. Teksty te pochodzą z tygodnika *Forum* (przekłady polskie tekstów rosyjskich) oraz z rosyjskiego serwisu przekładów prasowych inosmi.ru (warto wspomnieć, że wspomniany wcześniej wielojęzyczny serwis prasowy PressEurope zawiera teksty polskie, ale nie rosyjskie, a serwis komentarzy politycznych Syndicate ma wersję rosyjską, ale nie polską).

Korpus działa na zasadzie bazy danych. Każde zdanie jest przechowywane w bazie w dwóch wersjach, a każde słowo w bazie słów, od której jest odnośnik do bazy zdań. Teksty są tagowane programami wykorzystywanymi przez NKJP i Narodowy Korpus Języka Rosyjskiego. Polski tager daje wyniki zdezambiguowane, rosyjski nie, dlatego wyszukiwanie w tekście rosyjskim daje mnóstwo form homonimicznych.

Najważniejszą cechą wyróżniającą korpus warszawski nie jest jednak zestaw tekstów, lecz okienkowy interfejs konstruowania zapytań, znacznie prostszy niż w pozostałych korpusach wielo- i dwujęzycznych i w większości korpusów narodowych. Interfejs jest wzorowany na Narodowym Korpusie Języka Rosyjskiego, który wyróżnia się wśród korpusów narodowych łatwością obsługi. Mamy do dyspozycji moduł wyszukiwania form wyrazowych lub ich ciągów oraz moduł wyszukiwania leksemów bądź wartości kategorii gramatycznych. W module pierwszym należy wpisać w okno wyszukiwania dokładny ciąg znaków, ewentualnie z Boole'owskimi zastępnikami. Zapytanie lubi.. zwróci nam wyniki *lubisz* i *lubimy*, ale nie *lubię* czy *lubiliśmy*, a zapytanie lubi.* – wszystkie wyżej wymie-

nione formy. W tym module możliwe jest wyszukiwanie łączne. Wyszukajmy np. wszystkie polskie zdania zawierające słowo *kolega* i ich rosyjski odpowiedniki ze słowem *drug*:

Rys. 15.3.

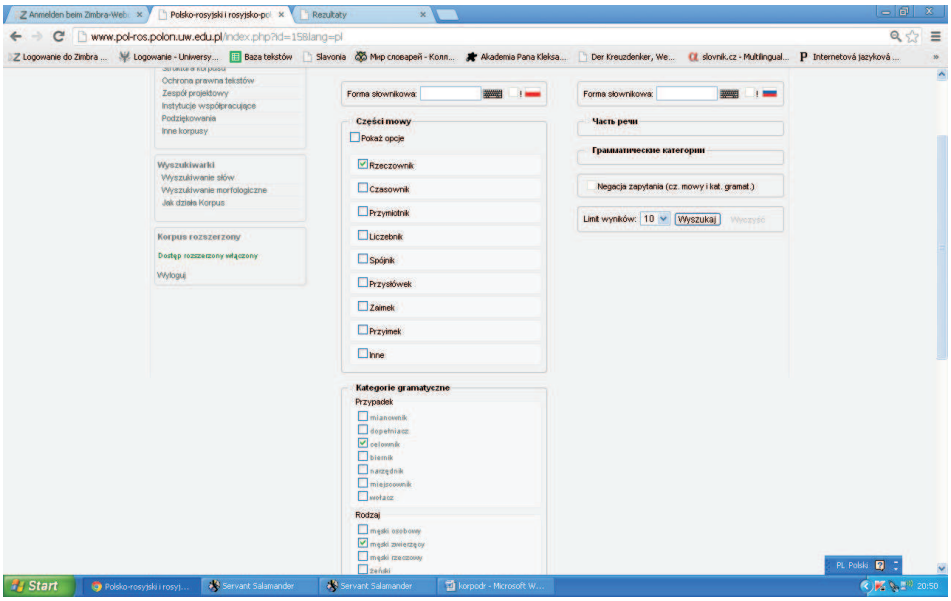


Właśnie słowo *kolega*, a nie *przyjaciel*, jest najczęstszym odpowiednikiem rosyjskiego *drug* czy angielskiego *friend* w kontekście wspólnoty działania lub losu. W korpusie znajdziemy jednak i bardziej skomplikowane przykłady, kiedy to po rosyjsku trzeba jakoś oddać polską różnicę między *kolegą* a *przyjacielem*:

. . . poznajcie się, mój ***kolega****, no,* ***przyjaciel****, powiedzmy.*
(Czeszko: Pokolenie)
. . . poznakomtes', eto moj ***odnoklassnik****, možno daže skazat'* ***drug****.*

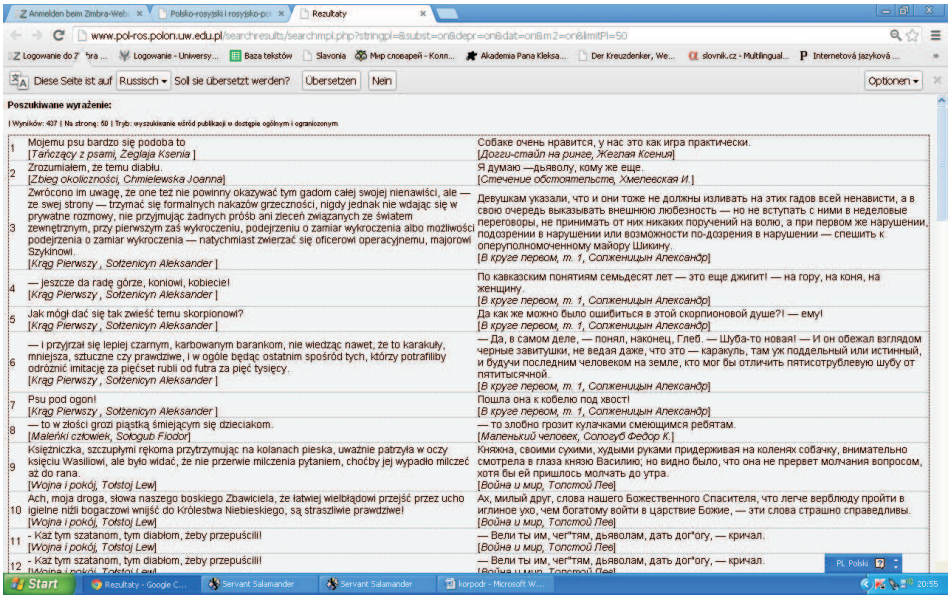
W module wyszukiwania morfologicznego mamy w obu językach rozwijaną listę wyboru części mowy i kategorii gramatycznych. Poniżej wyszukiwanie rzeczowników męsko zwierzęcych w celowniku (zob. Rys. 15.4.).

Rys. 15.4.



Oto wyniki:

Rys. 15.5.



Ponieważ klasy i kategorie gramatyczne zapisane w polskich tekstach według tagsetu NKJP, musieliśmy dokonać uproszczeń i połączyć np. wszystkie klasy czasownikowe w jedną. Najtrudniej było uzyskać wartość czasu przeszłego. Aby odszukać formę *zrobiłem*, wyszukiwarka musi reagować na tagowany oddzielnie ciąg segmenty pseudoimiesłowowe *zrobił* i aglutynanty osobowe *em*. Jeśli jednak chcemy zachować rozróżnienia tagsetu NKJP, np. odrębne klasy dla form finitywnych, bezokoliczników i trybu rozkazującego, może to zrobić, zaznaczając pole *Pokaż opcje* w strefie *Części mowy*).

Ze względu na prawa autorskie do tekstów, nasz korpus dostępny jest obecnie w dwóch wersjach: jednej ogólnej zawierającej teksty, na które uzyskano zgodę lub niewymagające zgody oraz w wersji pełnej, zawierającej także teksty, w których sprawie trwa korespondencja z posiadaczami praw. Teksty te wykorzystywane są wyłącznie lokalnie na serwerze Wydziału Polonistyki UW, a dostęp do wyszukiwania w nich jest chroniony hasłem.

Bibliografia

Łaziński M., Kuratczyk M., Orekhov B., Slobodyan E. (2012). *The Polish-Russian Parallel Corpus and Its Application in the Linguistic Analysis*. „Prace Filologiczne" LXIII, s. 209-218.

Von Waldenfels R. (2012a). *Polish* tea *is Czech* coffee*: advantages and pitfalls in using a parallel corpus in linguistic research*. W: *Methods in Contemporary Linguistics*, (eds.) A. Ender, A. Leemann & B. Wälchli, Berlin, s. 294-301.

Von Waldenfels R. (2012b). *ParaSol: Introduction to a Slavic Parallel Corpus*. „Prace Filologiczne" LXIII, s. 293-302.

16. Praktyczny przewodnik po korpusie równoległym InterCorp

Elżbieta Kaczmarska
Uniwersytet Warszawski
Instytut Slawistyki Zachodniej i Południowej

Alexandr Rosen
Univerzita Karlova
Ústav teoretické a komputační lingvistiky

Informacje ogólne

InterCorp to projekt, który powstał w Pradze na Uniwersytecie Karola; jego celem jest wybudowanie obszernego równoległego korpusu synchronicznego, który obejmowałby jak najwięcej języków. InterCorp jest jednocześnie nazwą ciągle rozrastającego się dużego synchronicznego korpusu paralelnego, obejmującego 32 języki (stan na styczeń 2014).

W tworzeniu InterCorpu uczestniczyli pedagodzy i studenci Wydziału Filozoficznego Uniwersytetu Karola (FF UK), osoby związane z Czeskim Korpusem Narodowym, a także współpracownicy zewnętrzni.

Korpus równoległy ma służyć, między innymi, jako źródło danych do badań teoretycznych, analiz gramatycznych i leksykograficznych, prac translatorskich, projektów dotyczących nauki języków obcych, a także do poszukiwań studenckich.

Źródło i dostęp: http://www.korpus.cz/intercorp/.

Charakterystyka korpusu

Trzonem korpusu InterCorp są teksty ręcznie wiązane[1], przeważnie beletrystyczne. Oprócz nich korpus obejmuje również automatycznie opracowane teksty; są one w interfejsie nazywane kolekcjami. Obecnie udostępnione są (w ramach tych kolekcji) artykuły publicystyczne oraz sprawozdania ze stron http//www.project-syndicate.org i http://www.presseurop.eu/cs, akty prawne Unii

[1] Termin „wiązanie tekstów" (alignment) za: Lewandowska-Tomaszczyk 2005.

Europejskiej z korpusu Acquis Communautaire oraz zapisy posiedzeń Parlamentu Europejskiego z lat 2007-2011 (z korpusu Europarl). W następnej wersji InterCorpu (czyli 7) twórcy zapowiadają udostępnienie bazy napisów dialogowych z serwera Open Subtitles. Teksty w kolekcjach są wiązane automatycznie; w konkordancjach może być więc więcej zdań, które sobie nawzajem nie odpowiadają. Niektóre teksty z korpusów Acquis Communautaire i Europarl były częściowo poprawione i posortowane, dlatego mogą się różnić od oryginału formą i objętością.

Wszystkie teksty w InterCorpie mają zawsze czeski odpowiednik (jako oryginał lub tłumaczenie). Język czeski jest dla InterCorpu językiem kluczowym, a czeska wersja jest wiązana z wersjami obcojęzycznymi. Całkowita objętość dostępnej części korpusu w wersji 6 (z kwietnia 2013) to 138 779 000 słów obcojęzycznych w tekstach w czeskim trzonie oraz 728 508 000 słów w obcojęzycznych tekstach w tzw. kolekcjach.

Twórcy InterCorpu planują również wprowadzenie do korpusu tekstów, które nie mają czeskiej wersji. Wówczas InterCorp straciłby swoją jednoznaczną orientację na język czeski.

Rozkład tekstów w ramach trzonu korpusu oraz w poszczególnych kolekcjach jest uwidoczniony poniżej w tabelkach:

Tab. 16.1. Wielkość korpusu w tysiącach słów – wersja 6.

Skrót	**Język**	**Trzon**	**Syndicate**	**Press-europ**	**Acquis**	**Europarl**	**W sumie**
Ar	arabski	29	0	0	0	0	29
Be	białoruski	1 308	0	0	0	0	1 308
Bg	bułgarski	3 979	0	0	13 816	9 083	26 879
Ca	kataloński	1 758	0	0	0	0	1 758
Da	duński	190	0	0	21 680	13 916	35 785
De	niemiecki	17 256	3 050	1 715	21 724	13 089	56 835
El	grecki	210	0	0	25 070	15 404	40 683
En	angielski	10 019	3 083	1 863	24 208	15 580	54 753
Es	hiszpański	14 552	3 479	1 948	27 001	15 885	62 865
Et	estoński	0	0	0	15 963	10 900	26 862
Fi	fiński	2 131	0	0	16 667	10 241	29 040
Fr	francuski	3 816	3 535	2 054	27 352	17 178	53 936
Hi	hindi	155	0	0	0	0	155
Hr	chorwacki	12 625	0	0	0	0	12 625
Hu	węgierski	2 511	0	0	19 168	12 307	33 985
It	włoski	4 081	247	1 893	24 850	15 489	46 560
Lt	litewski	358	0	0	18 433	11 020	29 811
Lv	łotewski	1 337	0	0	18 745	11 689	31 770
Mk	macedoński	2 664	0	0	0	0	2 664
Mt	maltański	0	0	0	14 133	0	14 133
Nl	holenderski	9 426	0	2 082	24 746	15 563	51 817
No	norweski	2 301	0	0	0	0	2 301
Pl	polski	12 710	0	1 660	20 464	12 805	47 640

[ciąg dalszy tabeli ze strony poprzedniej]

Skrót	Język	Trzon	Syndi-cate	Press-europ	Acquis	Euro-parl	W sumie
Pt	portugalski	2 318	0	2 103	28 599	16 481	49 502
Ro	rumuński	2 433	0	1 917	8 200	9 446	21 995
Ru	rosyjski	4 937	2 651	0	0	0	7 588
Sk	słowacki	8 152	0	0	19 222	12 734	40 108
Sl	słoweński	1 855	0	0	19 646	12 241	33 741
Sr	serbski	6 972	0	0	0	0	6 972
Sv	szwedzki	7 205	0	0	20 615	13 874	41 694
Uk	ukraiński	1 493	0	0	0	0	1 493
W sumie		**138 779**	**16 044**	**17 237**	**430 300**	**264 926**	**867 287**
Cs	czeski	61 962	2 741	1 639	20 285	12 920	99 547

W przypadku tekstów czeskich słowa liczone są tylko jeden raz, bez względu na liczbę wersji obcojęzycznych.

Polsko-czeska część InterCorpu zawiera (w porównaniu z parami innych języków) wysoki odsetek literatury pięknej (180 tytułów i 12,6 mln słów – w tej kategorii jest trzecia po języku niemieckim i hiszpańskim), przy czym należy zauważyć, iż proporcje pomiędzy oryginałami a przekładami są w przypadku tych dwóch języków wyrównane. Większą część beletrystyki z czesko-polskiej części korpusu równoległego stanowią przekłady z innych języków. Jeżeli chodzi o kolekcje, język polski nie jest reprezentowany jedynie w tekstach z serwera Project Syndicate (według danych z wersji 6). Orientacyjne, przykładowe dane przedstawia poniższa tabelka.

Tab. 16.2. Zestawienie ilościowe tekstów pod względem języka źródłowego.

	Czeskie oryginały	Polskie oryginały	Obce oryginały	Razem
W języku czeskim	1 899	2 331	8 402	12 632
W języku polskim	1 987	2 256	8 379	12 622

Anotacja zewnętrzna

W korpusie InterCorp anotowane są zewnętrznie według następujących informacji (poniżej tylko niektóre z nich).

Tab. 16.3. Wybrane symbole i ich znaczenia dostępne w anotacji zewnętrznej.

Atrybut	**Znaczenie**	**Wartość**
doc.lang	język tekstu	bg | ca | cs | da | de | el | en | es | et | fi | fr | ga | hr | hu | it | lt | lv | mt | nl | no | pl | pt | ro | ru | sk | sl | sv | un
div.id	identyfikacja tekstu	*nazwisko_ autora-skrócona_ nazwa_ tekstu* | _ACQUIS | _EUROPARL | _PRESSEUROP | _SYNDICATE
div.group	podział na	jádro (*trzon*)/kolekce (*kolekcja*)
div.version	wersja tekstu	do tej pory wszędzie 00;
div.wordcount	liczba słów tekstu	*Liczba*
div.author	autor tekstu	*nazwisko, imię*
div.title	tytuł tekstu	*Tytuł*
div.publisher	wydawca	
div.pubplace	miejsce wydania	
div.pubyear	rok wydania	*rok*
div.txtype	typ tekstu	drama (*dramat*) | literatura faktu | poezie (*poezja*) | právní texty (*teksty prawnicze*) | próza (*proza*) | publicistika – komentáře (*publicystyka – komentarze*) | publicistika – zprávy (*publicystyka – informacje*) | různé (*różne*) | zápis debaty (*zapis dyskusji*)
div.original	czy tekst jest oryginałem	ANO (*tak*)/NE (*nie*)
div.srclang	język źródła (oryginału)	bg | ca | cs | da | de | el | en | es | et | fi | fr | ga | hr | hu | it | lt | lv | mt | nl | no | pl | pt | ro | ru | sk | sl | sv | un
div.translator	tłumacz tekstu	*nazwisko, imię*
div.transsex	płeć tłumacza	F/M
div.authsex	płeć autora	F/M

Oznaczenie *doc.version* może być wykorzystane, gdy mamy do czynienia z kilkoma wersjami tego samego tekstu w danym języku (np. różne przekłady jednej książki). Dotychczas użycie tego oznaczenia nie jest możliwe, dlatego wszędzie pojawia się znak „00".

Pozostałe oznaczenia mogą mieć znaczenie w zależności od typu przeprowadzanego badania. Istotna może być np. płeć autora lub tłumacza, jeśli dla analizy relewantne jest, czy tekst został napisany przez kobietę, czy mężczyznę. Ważny może być także np. język oryginału, kiedy przedmiotem badania jest materiał językowy, który powstał w danym języku.

Anotacja wewnętrzna

Jeżeli chodzi o anotację wewnętrzną czy segmentację tekstu, dla każdego języka może być inna. Na stronach internetowych InterCorpu znajdują się odnośniki do opisów danych języków. Morfologiczną anotacją opatrzone są teksty w językach: angielskim, bułgarskim, czeskim, estońskim, francuskim, hiszpańskim, holenderskim, litewskim, niemieckim, norweskim, polskim, portugalskim, rosyjskim, słowackim, słoweńskim, węgierskim, włoskim. Wkrótce otagowane mają zostać teksty w językach: chorwackim, fińskim, rumuńskim, serbskim i szwedzkim.

W przypadku tekstów z tagami (a także lematami) zapytania formułowane wprost o zrosty mogą pozostać bez odpowiedzi. Dotyczy to np. angielskich form *can't* albo *I'm*, które tager dzieli na dwa słowa (*ca* + *n't* i *I* + *'m*) z odpowiednimi lematami i znacznikami. Podobnie sytuacja wygląda z polskimi formami typu *byłam* czy *gdybyś* (podzielone na *była* + *m* i *gdyby* + *ś*).

Należy jednak liczyć się z tym, że wyrazy mogą być błędnie dzielone, np. *gdzie ś za Wisłą*[2] (zaimek nieokreślony podzielony tak, jakby był połączeniem zaimka względnego i morfemu czasownikowego, np. *Gdzieś się podziewał?*). Jeśli więc segmentujemy *gdzieś* na cząstki: *gdzie ś*, to według tagera TaKIPI[3] jest to:

gdzie/gdzie/qub ś/być/aglt:sg:sec:imperf:nwok

Jest to błędne znakowanie; zaimek nieokreślony *gdzieś* (np. ze zdania *Położyłam to gdzieś i zapomniałam gdzie.*) powinien być znakowany w poniższy sposób:

gdzieś/gdzieś/qub

Zapytanie o cały zrost trzeba zadać jako połączenie wyrazowe (*Slovní spojení*/*Phrase*), a części zrostu należy oddzielić spacją. Lemat i znacznik mają tylko części zrostu.

Morfologiczne tagi zawierające znaki, które w regularnych wyrażeniach mają specjalne znaczenie (np. „$" w angielskim znaku „wp$"), w zapytaniach należy wpisywać za wstecznym ukośnikiem, np. tag="wp\$".

Na stronach korpusu znajdują się dokładniejsze informacje o znacznikach morfosyntaktycznych. Segmentacja i znakowanie mogą się różnić w zależności od poszczególnych języków:

> Wielosłowne wyrażenia w języku hiszpańskim, np. *a lo largo de* segmentuje się i znakuje jako:
> <w lemma="a~lo~largo~de"

[2] Por. *Jest gdzie*/gdzie/qub *ś*/być/aglt:sg:sec:imperf:nwok *na północy Anglii.*.
[3] Piasecki 2007: 151-167.

tag="PREP">a lo largo de</w>,

Podobnie: <w lemma="por∼el∼momento" tag="ADV">Por el momento</w>

Przykłady z innych języków poniżej (forma, lemat, tag):
czeski: *udělals*: udělals/udělat/VpYS---2R-AA---
polski: *zrobiłeś*: zrobił/zrobić/praet:sg:m1:perf + es/byc/aglt:sg:sec:imperf:wok
francuski: *dit-il*: dit/dire/VER:pres + -il/il/PRO:PER
angielski: *children's*: children/child/NNS + 's/'s/POS
angielski: *parents'*: parents/parent/NNS + '/'/POS
angielski: *I'm*: I/I/PP + 'm/be/VBP
angielski: *can't*: ca/ca/MD + n't/n't/RB

Tryb przypuszczający w języku czeskim i polskim:
czeski: bych/být/Vc-S---1-------
polski: by/by/qub + m/byc/aglt:sg:pri:imperf:nwok
czeski: abychom/aby/J,-P---1-------
polski: że/że/conj + by/by/qub + śmy/być/aglt:pl:pri:imperf:nwok

Dostęp do tekstów w korpusie

InterCorp można przeszukiwać, używając trzech interfejsów – Park (http://korpus.cz/Park), NoSketch Engine (https://korpus.cz/corpora) i KonText (https://kontext.korpus.cz). NoSketch Engine (dostępny od maja 2013 roku) różni się od interfejsu Park nie tylko wyglądem, ale przede wszystkim funkcjonalnością. Natomiast KonText dostępny od stycznia 2014) różni się od NoSketch Engine właściwie tylko wyglądem. Kiedy otworzymy stronę główną korpusu http://www.korpus.cz/, widoczny jest dostęp jedynie do dwóch wspomnianych wyżej interfejsów (KonText i Park). Bezpośredni dostęp do wszystkich trzech oferuje strona http://www.korpus.cz/intercorp. Pełne teksty (książki, dokumenty, publikacje) z pomieszaną kolejnością zdań uzyskać można po podpisaniu umowy licencyjnej.

Korpus InterCorp można badać, posługując się danymi dostępowymi, które wykorzystywane są również do pozostałych korpusów ČNK. Dostęp do korpusów ČNK jest darmowy po uprzedniej rejestracji na stronie *Prohlášení uživatele korpusů ÚČNK / Statement of a User of the ICNC Corpora*.

Interfejs KonText

Przez interfejs KonText można przeszukiwać zarówno korpusy jednojęzyczne, jak i równoległe. Po wpisaniu nazwy użytkownika i hasła otworzy się strona z domyślnie ustawionym korpusem (np. SYN2010). Po kliknięciu przycisku z nazwą tego korpusu pojawi się propozycja dostępnych typów korpusów. Jeżeli klikniemy strzałkę przy pozycji *Paralelní korpus InterCorp*, pojawi się lista części korpusu według języków.

Wybór języków

Kliknięciem któregoś z języków przy pozycji *Paralelní korpus InterCorp* wybieramy pierwszy badany język. Jest to język, w którym będziemy zadawać pytanie. Okienko z zapytaniem w tym języku nie może pozostać puste. Oprócz tego kolejność języków jest ważna jeszcze przy tworzeniu subkorpusów (zob. poniżej). Zakres tekstów, które mają tworzyć subkorpus, można ograniczyć tylko na podstawie tekstów pierwszego języka. W pozostałych przypadkach kolejność języków nie odgrywa żadnej roli.

Po wybraniu pierwszego języka na górze po prawej stronie w niebieskiej ramce pojawi się krótki opis wybranej części korpusu (po kliknięciu na nazwę korpusu w tej ramce wyświetli się więcej informacji: jego wielkość, liczba pozycji, wyrazy tekstowe oraz znaki interpunkcyjne). Chcąc dodać kolejny język, należy go najpierw w ramce *Zarovnané korpusy/Aligned Corpora* wybrać, a potem kliknąć *Přidat/Add*. W dodanym języku można (ale nie trzeba) dodać uzupełnienie pytania. Jeżeli chcemy, aby w odpowiedzi na zapytanie pojawiały się również konkordancje, które w danym języku nie mają ekwiwalentu, należy kliknąć *zobrazit prázdné řádky/include empty lines*. W ten sam sposób dodawać można kolejne języki. Możemy wyszukiwać także tylko w jednej części korpusu równoległego, czyli tylko w jednym języku. W tym przypadku kolejnych języków nie dodajemy, ale wybieramy od razu typ zapytania, a odpowiednie zapytanie wpisujemy.

Zapytania

Można wybierać z kilku opcji *Typ dotazu/Query Type* (zob. poniżej). W przypadku wszystkich typów zapytań oprócz *Základní/Basic* ważna jest wielkość liter. Można używać także wyrazów regularnych. W przypadku zapytania o *Slovní tvar/Word Form* domyślne ustawienie zakłada, że na wielkość liter nie ma znaczenia, ale funkcję *Shoda velikosti/Match case* trzeba włączyć.

Konstruując zapytanie, mamy do dyspozycji następujące opcje:

– *Základní/Basic* – wyszuka określone wyrażenie jako formę wyrazową bez względu na wielkość liter; jeżeli chodzi jednocześnie o podstawową formę słownikową (lemat), wyszukają się także wszystkie jej formy;

– *Lemma* – wyszuka wszystkie formy wytworzone od wpisanego lematu;
– *Fráze/Phrase* – wyszuka sekwencję wpisanych form wyrazowych;
– *Slovní tvar/Word Form* – wyszuka wpisaną formę wyrazową;
– *Podřetězec/Character* (Ciąg znaków) – wyszuka formy wyrazowe zawierające wpisaną sekwencję znaków;
– *CQL* – wyszuka jedną albo więcej form wyrazowych zgodnie z wpisanym wyrażeniem języka zapytań CQL. Wpisanie morfologicznego tagu dla języka czeskiego może ułatwić funkcja *vložit tag/insert tag*, która umożliwia wstawienie na odpowiednią pozycję kodu cechy z listy atrybutów i właściwych wartości. W przypadku wszystkich języków można dodatkowo wykorzystać funkcję *vložit "within"/insert "within"*, która pozwala na zawężenie wyników wyszukiwania w oparciu o metadane. Wykaz atrybutów i ich wartości znajduje się w tabelce dotyczącej anotacji zewnętrznej. Parę atrybut="wartość" należy łączyć operatorem & (spójnik logiczny). Cały warunek "within" musi być wpisany dopiero na końcu zapytania za specyfikacją jednej albo wielu pozycji (w nawiasach kwadratowych). W jednym zapytaniu może być więcej warunków "within". Oboma poniższymi sposobami uzyskamy ten sam wynik, a mianowicie wszystkie formy rzeczowników w wołaczu w oryginałach czeskich dramatów:

```
[tag="N...5.*"] within <div txtype="drama" & rclang="cs" />
```

```
[tag="N...5.*"] within <div txtype="drama" /> within <div srclang="cs" />
```

Wynik otrzymamy po kliknięciu przycisku *Vytvořit konkordanci/Make Concordance*.

Przy drugim i kolejnym języku można dodatkowo ustalić, czy wyszukiwane konkordancje mają zawierać wyrażenia wpisane do okienka zapytania w tymże drugim i kolejnym języku.

Jeżeli np. chcemy w języku polskim znaleźć ekwiwalenty czeskiego czasownika *milovat* inne niż *kochać* i jego derywaty, zakreślamy *NEobsahuje/Does NOT contain* i wpisujemy:

```
[word=".*[Kk]och.*"]
```

Przy takim uzupełnieniu musimy zmienić typ zapytania na CQL. Otrzymujemy wyniki, w których po polskiej stronie nie ma ani *kochać*, ani jego derywatów.

Jeżeli nastawimy polecenie *Obsahuje/Contains* oraz wpiszemy zapytanie:

```
[word!=".*[Kk]och.*"]
```

negacja odnosi się do pierwszej (liczonej od lewej) pozycji danego segmentu, w której nie ma cząstki 'koch'. Otrzymamy wyniki, w których pojawi się w segmencie chociaż jedno słowo, w którym nie będzie 'koch'; łącznik przed słowem *kochać* traktowany jest jako pozycja, jak w poniższym przykładzie:

< <s> - Kocham cię, Vernon. </s>

Płynie stąd wniosek dla użytkownika, że te dwa zapytania po włączeniu opcji *Contains/Does NOT contain* dają inne wyniki:

Contains + [word!=".*[Kk]och.*"]

oraz

Does NOT Contain + [word=".*[Kk]och.*"]

NEobsahuje/Does NOT contain dotyczy zdania, a wykrzyknik w zapytaniu – tylko pozycji.

Uszczegółowienie kontekstu

Jeżeli chcemy wyszukać wyłącznie poświadczenia, w których słowo kluczowe znalazłoby się w jakimś konkretnym otoczeniu, wybieramy odpowiednie parametry w momencie wpisywania zapytania. Określamy, czy dany element ma wystąpić w odpowiednio szerokim kontekście lub nie może się w nim pojawić. Możemy również dobrać szerokość kontekstu.

Zawężenie objętości przeszukiwanych tekstów w zależności od metadanych

Objętość przeszukiwanych tekstów możemy zawęzić poprzez wskazanie metadanych, które miałyby te teksty charakteryzować. Precyzujemy to również w momencie wpisywania zapytania.

Wyniki zapytania

Odpowiedzią na zapytanie jest pojawienie się strony z konkordancjami. Za pomocą menu umieszczonego na górze strony można zmieniać sposób wyświetlania, a także zapisywać, klasyfikować, filtrować, wyszukiwać kolokacje, a wyniki szacować statystycznie. Funkcje z menu dotyczą języka przedstawionego na jasnoniebieskim tle; parametry odpowiedniej części korpusu zostały zaprezentowane w nagłówku (por. rysunek poniżej):

Rys. 16.2.

Kliknięcie paska z nazwą konkordancji zawierającą poświadczenia danego języka umożliwia zmianę tego języka. Jednakże, by to było możliwe, konkordancje w obu językach muszą zawierać wyszukiwane wyrażenie (dzieje się tak po zadaniu równoległych zapytań w obu językach, w innych przypadkach zaś użytkownik jest proszony o wpisanie wyszukiwanego wyrażenia z użyciem opcji filtru pozytywnego). Chcąc wpisać nowe zapytanie, klikamy *Nový dotaz/New query* z menu *Dotaz/Query*.

Opis funkcji znajdujących się w menu

- *Dotaz/Query*
  - *Nový dotaz/New Query* – przejście na stronę wpisywania nowego zapytania;
  - *Předchozí dotazy/Recent queries* – przeniesienie na stronę ukazującą historię zapytań;
  - *Seznam slov/Word list* – przejście na stronę wpisywania listy form wyrazowych lematów, znaczników i metadanych w danym języku, poklasyfikowanych malejąco według frekwencji;
- *Subkorpusy/Subcorpora*
  - *Vytvořit nový.../Create new...* – sposobowi tworzenia subkorpusów poświęciliśmy odrębny podrozdział (zob. poniżej);
  - *Mé subkorpusy.../My subcorpora...* – przejście na stronę z własnymi subkorpusami użytkownika;
- *Uložit/Save* – eksport konkordancji w tekstowym formacie (TXT) lub w formatach CSV i XML. Format tekstowy używa tabulatorów jako elementu oddzielającego kolumny. Parametry poszczególnych formatów można ustawić, wybierając z menu *Vlastní/Custom*.

- *Konkordance / Concordance*
  - *Aktuální konkordance / Current concordance* – widok wyniku aktualnego wyszukiwania;
  - *Třídění / Sort* – umożliwienie sortowania wyników zapytania według różnych atrybutów (forma, lemat, znacznik morfologiczny) dotyczących słowa kluczowego (lub słów) w lewym lub prawym kontekście; można sortować również *a tergo* (*Zpětně / Backward*) i na trzech kolejnych poziomach;
  - *Levý kontext / Left context* | *Pravý kontext / Right context* – sortowanie według lewego lub prawego kontekstu;
  - *KWIC* – sortowanie według słowa kluczowego;
  - *Promíchat / Shuffle* – sortowanie poświadczeń w przypadkowej kolejności;
  - *Vzorek / Sample* – redukcja konkordancji do wybranej liczby przypadkowych poświadczeń;
  - *Popis dotazu / ConcDesc (Concordance description)* – zbiór parametrów aktualnego zapytania z liczbami znalezionych poświadczeń;
  - *Zpět / Undo* – powrót do wyników ostatnio przeprowadzonego działania;
- *Filtr / Filter* – umożliwia redukcję wyświetlonych wyników według wpisanych kryteriów[4]:
  - Filtr pozytywny likwiduje wszystkie poświadczenia, które nie spełniają wpisanego kryterium (zapytania).
  - Filtr negatywny usuwa wszystkie poświadczenia spełniające wpisane kryterium (zapytanie).
- *Frekvenční distribuce / Frequency* – umożliwia sporządzanie (na podstawie wyników zapytania) list pozycji posortowanych według frekwencji, z informacją o częstości całkowitej lub względnej (% albo liczba poświadczeń w przeliczeniu na milion słów). Kolejne opcje generują:
  - *Značky / Node tags* – listy znaczników morfologicznych;
  - *Slovní tvary / Node forms* – listy form wyrazowych;
  - *Dokumenty / Doc IDs* – listy nazw tekstów;
  - *Typy textu / Text Types* – listy według głównych pozycji metadanych;
  - *Vlastní / Custom* – (poprzez wpisanie kryteriów) trzypoziomową listę form, lematów lub znaczników posortowanych według frekwencji lub podobną listę pozycji metadanych.

[4] Jeśli wyszukiwane wyrażenie składa się z większej liczby słów, wówczas wygodniej jest określić wyjściową pozycję jako *Vybraný token / Selected token první / first* albo *poslední / last*. *Rozsah hledání / Search Span* – określa, w jak szerokim kontekście będzie realizowane wpisane kryterium (zapytanie), wyjściową pozycję oznacza się jako 0, ujemne wartości wyrażają liczbę tokenów na lewo od pozycji wyjściowej, pozytywne wartości – na prawo od pozycji wyjściowej. Wymienione funkcje dostępne są zarówno dla filtra pozytywnego, jak i negatywnego.

- *Kolokace/Collocations* – umożliwia sporządzenie listy kolokacji ze słowem kluczowym na podstawie różnych miar statystycznych:
  - *Vlastní/Custom* – umożliwia odpowiednie dobranie parametrów.
  - *Atribut/Attribute* – określa, czy w wynikach jako kolokat mają pojawić się formy wyrazowe, lematy, morfologiczne znaczniki.
  - *V rozsahu od/In the range from* – określa, w jak szerokim kontekście będzie przebiegać wyszukiwanie; wyjściową pozycję określa się jako 0, ujemne wartości oznaczają liczbę tokenów na lewo od wyjściowej pozycji, pozytywne – na prawo od tej pozycji.
  - *Minimální frekvence v korpusu/Minimum frequency in corpus* – ogranicza wyniki według frekwencji kolokacji.
  - *Zobrazit funkce/Show functions* – wyświetla wyniki według odpowiednich funkcji statystycznych.
  - *Setřídit dle/Sort by* – sortuje rezultaty wyszukiwania według danego parametru.
- *Možnosti zobrazení/View options* – parametry wyświetlania wyników:
  - *KWIC* – wyświetlenie ze słowem kluczowym pośrodku kontekstu;
  - *Věta/Sentence* – wyświetlenie całego zdania; ma znaczenie dla drugiego języka i kolejnych – jeżeli jednemu zdaniu w pierwszym języku odpowiada więcej zdań w innym, wyświetli się tylko pierwsze zdanie;
  - *Zarovnání/Alignment* – wyświetlenie wiązanych segmentów; ma znaczenie tylko dla drugiego języka i kolejnych – jeżeli jednemu zdaniu w języku pierwszym odpowiada więcej zdań w innym języku, wówczas wyświetlą się wszystkie zdania;
  - *Atributy, struktury a metainformace.../Attributes, structures and references...* – wybór własności, jakie mają się wyświetlić na stronie: atrybuty form wyrazowych (lemat, morfologiczny znacznik), znaczniki strukturowe (dla dokumentu, sekcji, akapitu, zdania), metadane (autor, tytuł itd.);
  - *Obecné volby zobrazení konkordance.../General concordance view options...* – określenie liczby konkordancji na stronie i szerokości kontekstu liczonego w liczbie pozycji od słowa kluczowego, które mają się wyświetlić na stronie;
- *Nápověda/Help*
  - *Uživatelská příručka.../User manual...* – podręcznik dla użytkownika wzbogacony specjalnymi lekcjami;
  - *Lingvistické pojmy.../Linguistic terms...* – definicje terminów lingwistyki korpusowej;
  - *Dostupné korpusy.../Available corpora...* – korpusy dostępne w Czeskim Korpusie Narodowym.

Subkorpusy – zmniejszenie objętości przeszukiwanych tekstów

Zakres przeszukiwanych tekstów jest domyślnie ustawiony na wszystkie teksty wybranych języków. Zakres ten można ograniczyć przed realizacją zapytania poprzez kliknięcie pozycji *Subkorpusy/Subcorpora* w menu. Nowy subkorpus należy w ramce *Jméno nového subkorpusu/New subcorpus name* nazwać.

Teksty do subkorpusu można wybrać dwoma sposobami (*Specifikovat subkorpus pomocí/Specify subcorpus using*). Wybór *Vlastní 'within' podmínky/Custom 'within' condition* oferuje o wiele większy wachlarz możliwości. Natomiast opcja *Seznam atributů/Attribute list* jest bardziej przystępna dla użytkownika.

Kliknięciem *Seznam atributů/Attribute list* wybieramy spośród poszczególnych atrybutów. Jeżeli w kolumnie danego atrybutu nie zaznaczymy żadnego wyboru, będzie się to równało zaznaczeniu wszystkich możliwości; można to też uczynić, klikając przycisk *Vybrat vše/Select all*.

Po kliknięciu na *Vlastní 'within' podmínka/Custom 'within' condition* możemy określić precyzyjniej subkorpus za pomocą wszystkich atrybutów dostępnych w korpusie. Więcej atrybutów mamy do dyspozycji, wybierając warunek na poziomie części dokumentu (*div*). Dostępne w konkretnej grupie atrybuty ukażą się po kliknięciu na znak zapytania pomiędzy ramkami. Spis atrybutów i ich wartości znajdziemy też w tabelce dotyczącej anotacji zewnętrznej (wyżej w tym rozdziale). Na przykład, jeżeli chcemy do subkorpusu wybrać oryginalne czeskie dramaty, po wybraniu opcji *div*, w ramce do wpisywania warunku umieszczamy sekwencję znaków:

```
txtype="drama" & srclang="cs"
```

Ten sam rezultat otrzymalibyśmy poprzez zaznaczenie pozycji *drama* i *cs* w odpowiednich indeksach atrybutów.

Wybór tekstów do subkorpusu przeprowadzimy po kliknięciu *Vytvořit subkorpus/Create Subcorpus*. Do tego subkorpusu możemy wracać przy wpisywaniu zapytania poprzez wybór *Dostupný korpus/Available subcorpora* w pozycji *Subkorpus*.

Interfejs NoSketch Engine

Przez interfejs NoSketch Engine można przeszukiwać zarówno korpusy jednojęzyczne, jak i równoległe. Po wpisaniu nazwy użytkownika i hasła otworzy się strona z domyślnie ustawionym korpusem (np. SYN2010). Po kliknięciu przycisku z nazwą tego korpusu pojawi się wykaz dostępnych typów korpusów. Jeżeli klikniemy strzałkę przy pozycji *Paralelní korpus InterCorp*, pojawi się lista części korpusu według języków.

Wybór języków

Kliknięciem nazwy któregoś z języków w ramach opcji *Paralelní korpus InterCorp* wybieramy pierwszy badany język. W tym języku będziemy zadawać pytanie. Okienko z zapytaniem nie może pozostać zatem puste. Oprócz tego kolejność języków jest także istotna przy tworzeniu subkorpusów (zob. poniżej). Zakres tekstów subkorpusu można ograniczyć tylko na bazie tekstów pierwszego języka. W innych przypadkach kolejność języków nie odgrywa żadnej roli.

Po wybraniu pierwszego języka na górze strony się pojawi krótki opis wybranej części korpusu, jego wielkość, liczba pozycji (słownych form włącznie z interpunkcją). Chcąc dodać kolejny język, należy go najpierw wybrać w ramce *Zarovnané korpusy/Aligned Corpora*, potem kliknąć *Přidat/Add*. W dodanym języku można (choć nie jest to konieczne) zadać kolejne pytanie. Jeżeli chcemy, by w odpowiedzi na zapytanie pojawiły się również poświadczenia, które w danym języku nie mają ekwiwalentu, powinniśmy kliknąć *zobrazit prázdné řádky/include empty lines*. W ten sam sposób dodawać można kolejne języki. Możemy wyszukiwać także tylko w jednej części korpusu równoległego, czyli tylko w jednym języku. W tym przypadku kolejnych języków nie dodajemy, ale wybieramy od razu typ zapytania, a odpowiednie zapytanie wpisujemy.

Zapytania

Można wybierać z sześciu typów zapytań (*Typy dotazu/Query Type*). We wszystkich opcjach zapytań oprócz wyboru *Základní/Basic* istotna jest wielkość liter. Można używać także wyrazów regularnych. W przypadku zapytania o *Slovní tvar/Word Form* domyślne ustawienie zakłada, że wielkość liter nie ma znaczenia, a funkcję *Shoda velikosti/Match case* należy wtedy włączyć. Konstruując zapytanie, mamy do dyspozycji następujące opcje wyszukiwania:

- *Základní/Basic* – określonych wyrażeń jako form wyrazowych bez względu na wielkość liter; jeżeli chodzi jednocześnie o podstawową formę słownikową (lemat), wyszukamy w ten sposób także wszystkie jej formy;
- *Lemma* – wszystkich form wytworzonych od wpisanego lematu;
- *Fráze/Phrase* – sekwencji wpisanych form wyrazowych;
- *Slovní tvar/Word Form* – wpisanych pojedynczych form wyrazowych;
- *Podřetězec/Character* (ciąg znaków) – form wyrazowych zawierających wpisaną sekwencję znaków;
- *CQL* – jednej lub więcej form wyrazowych zgodnie z wpisanym wyrażeniem języka zapytań CQL. Wpisanie morfologicznego tagu dla języka czeskiego może ułatwić funkcja *vložit tag/insert tag*, która umożliwia wstawienie na odpowiednią pozycję kodu cechy z listy atrybutów i właściwych

wartości. W przypadku wszystkich języków można dodatkowo wykorzystać funkcję *vložit "within"/insert "within"*, która umożliwia zawężenie wyników wyszukiwania w oparciu o metadane tj. bibliograficzne i inne dane dotyczące danego tekstu. Wykaz atrybutów i ich wartości znajduje się w tabelce dotyczącej anotacji zewnętrznej. Parę atrybut=„wartość" należy łączyć operatorem & (spójnik logiczny). Cały warunek „within" musi być wpisany dopiero na końcu zapytania za specyfikacją jednej albo wielu pozycji (w nawiasach kwadratowych). W jednym zapytaniu może być więcej warunków „within". Na przykład ten sam wynik, a mianowicie wszystkie formy rzeczowników w wołaczu w tekstach dramatów czeskich, uzyskamy dwoma sposobami:

[tag="N...5.*"] within <div txtype="drama" & srclang="cs" />

[tag="N...5.*"] within within

Przy drugim i kolejnym języku można dodatkowo ustalić, czy wyszukiwane konkordancje mają zawierać wyrażenia wpisane do okienka zapytania w tymże drugim i kolejnym języku.

Jeżeli np. chcemy w języku polskim znaleźć ekwiwalenty czeskiego czasownika *milovat* inne niż *kochać* i jego derywaty, zakreślamy *NEobsahuje/Does NOT contain* i wpisujemy:

[word=".*[Kk]och.*"]

Otrzymujemy wyniki, w których po polskiej stronie nie ma ani *kochać*, ani jego derywatów.

Jeżeli nastawimy polecenie *Obsahuje/Contains* oraz wpiszemy zapytanie:

[word!=".*[Kk]och.*"]

negacja odnosi się do pierwszej (liczonej od lewej) pozycji danego segmentu, w której nie ma cząstki 'koch'. Otrzymamy wyniki, w których pojawi się w segmencie chociaż jedno słowo, w którym nie będzie 'koch'; łącznik przed słowem *kochać* traktowany jest jako pozycja, jak w poniższych przykładach:

<s> - Kocham cię . </s>

<s> - Kocham cię - szepnął em . </s>

Wniosek dla użytkownika i czytelnika poradnika jest taki, że te dwa zapytania po włączeniu *Contains/Does NOT contain* dają inne wyniki:

Contains + [word!=".*[Kk]och.*"]

oraz

Does NOT Contain + [word=".*[Kk]och.*"]

NEobsahuje/Does NOT contain dotyczy zdania, a wykrzyknik w zapytaniu – dotyczy tylko pozycji.

Wynik otrzymamy po kliknięciu przycisku *Vytvořit konkordanci/Make Concordance*.

Wyniki zapytania

Odpowiedzią na zapytanie jest pojawienie się strony z konkordancjami. W menu umieszczonym w pierwszej kolumnie można zmieniać sposób wyświetlenia, zapisywać, klasyfikować, filtrować konkordancje według kontekstu, wyszukiwać kolokacje, a wyniki oszacować statystycznie. Funkcje z menu dotyczą języka z jasnoniebieskim tłem; parametry odpowiedniej części korpusu są przedstawione w nagłówku (jak na rysunku poniżej).

Rys. 16.3.

Kliknięcie nazwy kolumny konkordancji z danego języka umożliwia jego zmianę, jednakże, by było to możliwe, konkordancje w obu językach muszą zawierać wyszukiwane wyrażenie (dzieje się tak po równoległym zapytaniu w obu językach, w innych przypadkach po zmianie języka użytkownik jest proszony o wpisanie kluczowego wyrażenia przy użyciu filtru pozytywnego). Kolumnę z menu można przesunąć nad konkordancje kliknięciem przycisku *Změnit umístění menu/Switch menu position*. Wprowadzanie nowego zapytania dokonujemy kliknięciem przycisku *Konkordance/Concordance*.

Wyniki skryte za prawym krańcem okna odkryjemy poprzez jego zwiększenie lub za pomocą poziomego suwaka, który znajdziemy po zrolowaniu w dół do końca strony z konkordancjami.

Opis funkcji znajdujących się w lewej kolumnie

- *Konkordance/Concordance* – przejście na stronę wpisywania zapytania.
- *Seznam slov/Word list* – przejście na stronę wpisywania listy form wyrazowych, lematów, znaczników i metadanych w danym języku, poklasyfikowanych malejąco według frekwencji.
- *Uložit/Save* – eksport konkordancji w tekstowym formacie lub w formacie XML. Format tekstowy używa tabulatorów jako elementu oddzielającego kolumny, można go przeformatować do postaci tabelki (nie jest to jednak tak łatwe i oczywiste dla użytkowników programy Windows jak eksport z interfejsu Park do formatu Excel).
- *Zobrazení/View* – parametry wyświetlania wyników:
  - *Vlastní/Custom* – oprócz określenia liczby poświadczeń na stronie i szerokości kontekstu można także wybrać atrybuty form wyrazowych (lemat, tag), znaczniki strukturalne (dokument, sekcja, akapit, zdanie), metadane (autor, tytuł itd.);
  - *KWIC* – wyświetlenie ze słowem kluczowym pośrodku kontekstu;
  - *Věta/Sentence* – wyświetlenie całego zdania; jeżeli jednemu zdaniu w pierwszym języku odpowiada więcej zdań w innym, wyświetli się tylko pierwsze zdanie;
  - *Zarovnání/Alignment* – wyświetlenie wiązanych segmentów; ma znaczenie tylko dla drugiego języka i kolejnych – jeżeli jednemu zdaniu w języku pierwszym odpowiada więcej zdań w innym języku, wówczas wyświetlą się wszystkie zdania.
- *Třídění/Sort* – umożliwienie sortowania wyników zapytania według różnych kryteriów:
  - *Vlastní/Custom* – form, lematów, znaczników morfologicznych odnoszących się do słowa kluczowego (lub słów) w lewym lub prawym kontekście; można sortować również (*a tergo*, *Zpětně/Backward*) i na trzech kolejnych poziomach;
  - *Vlevo/Left* | *Vpravo/Right* – lewego lub prawego kontekstu;
  - *KWIC* – słowa kluczowego słowa;
  - *Reference/References* – nazwy tekstu;
  - *Promíchat/Shuffle* – w przypadkowej kolejności.

- *Filtr/Filter* – umożliwia redukcję wyświetlonych wyników według wpisanych kryteriów:
  - Filtr pozytywny usuwa z wyników wszystkie poświadczenia niespełniające wpisanego kryterium (zapytania);
  - Filtr negatywny likwiduje wszystkie poświadczenia, które wpisane kryterium (zapytanie) spełniają;
  - Jeśli wyrażenie kluczowe składa się z większej liczby słów, wówczas wygodniej jest określić wyjściową pozycję jako *Vybraný token/Selected token první/first* lub *poslední/last*;
  - *Rozsah hledání/Search Span* – określamy, jak szeroki kontekst będzie realizował wpisane kryterium (zapytanie); wyjściową pozycję oznaczamy jako 0, ujemne wartości wyrażają liczbę tokenów na lewo od pozycji wyjściowej, pozytywne zaś – na prawo od pozycji wyjściowej.
- *Frekvenční distribuce/Frequency* – umożliwia sporządzanie (na podstawie wyników zapytania) list pozycji posortowanych według frekwencji, z informacją o częstości całkowitej lub względnej (% albo liczba poświadczeń w przeliczeniu na milion słów). Kolejne opcje generują:
  - *Vlastní/Custom* – trzypoziomową listę form, lematów lub znaczników posortowanych według frekwencji lub podobną listę pozycji metadanych;
  - *Značky/Node tags* – listę znaczników morfologicznych;
  - *Slovní tvary/Node forms* – listę form wyrazowych;
  - *Dokumenty/Doc IDs* – listę nazw tekstów;
  - *Typy textu/Text Types* – listy według głównych pozycji metadanych.
- *Kolokace/Collocations* – umożliwia sporządzenie listy kolokacji ze słowem kluczowym na podstawie różnych miar statystycznych:
  - *Atribut/Attribute* – określa, czy w wynikach jako kolokat mają się pojawić formy wyrazowe, lematy lub znaczniki morfologiczne;
  - *V rozsahu od/In the range from* – określa, w jak szerokim kontekście będą wyszukiwane kolokaty; wyjściową pozycję określa się jako 0, ujemne wartości oznaczają liczbę tokenów na lewo od pozycji wyjściowej, dodatnie – na prawo od niej;
  - *Minimální frekvence v korpusu/Minimum frequency in corpus* – ogranicza wyniki na podstawie częstości kolokacji;
  - *Zobrazit funkce/Show functions* – wyświetla wyniki stosownie do odpowiednich funkcji statystycznych;
  - *Setřídit dle/Sort by* – sortuje rezultaty wyszukiwania według danego parametru.
- *– Popis dotazu/ConcDesc (Concordance description)* – zbiór parametrów aktualnego zapytania wraz z liczbą znalezionych poświadczeń.

Subkorpusy – zmniejszenie objętości przeszukiwanych tekstów

Zakres przeszukiwanych tekstów jest domyślnie ustawiony na wszystkie teksty wybranych języków. Można go ograniczyć przed realizacją zapytania poprzez kliknięcie pozycji *Subkorpus* w menu. Ograniczenie dotyczy pierwszego języka. W ramce *Subkorpus* można potem wybierać spośród poszczególnych atrybutów. Brak zaznaczenia jakiejkolwiek opcji oznacza wybranie wszystkich możliwości; można to też osiągnąć, klikając przycisk *Vybrat vše/Select all.*

Wyboru tekstów do subkorpusu dokonujemy kliknięciem opcji *vytvořit nový/create new* w ramce *Subkorpus*, następnie określamy atrybut w nowym okienku. Możemy nie tylko zaznaczyć odpowiednie pozycje przy kryteriach wyboru, ale także określić precyzyjniej subkorpus za pomocą funkcji *Vlastní 'within' podmínky/Custom 'within' condition.* Najpierw musimy zdecydować, czy warunek przypisujemy dla atrybutów na poziomie całego dokumentu (doc) czy tylko jego części (div). Na tym etapie wyszukiwania nie można łączyć ze sobą atrybutów z obu grup. Spis atrybutów i ich wartości znajdziemy w tabelce odnoszącej się do anotacji zewnętrznej. Na przykład, jeżeli chcemy do subkorpusu wybrać oryginalne czeskie dramaty, po wybraniu opcji *div* do ramki służącej do wpisywania warunku wprowadzamy sekwencję:

```
txtype="drama" & srclang="cs"
```

Ten sam rezultat otrzymalibyśmy poprzez zaznaczenie pozycji *drama* i *cs* w odpowiednich indeksach atrybutów. Nowy subkorpus należy w ramce *Jméno nového subkorpusu/New subcorpus name* odpowiednio nazwać; po kliknięciu przycisku *Vytvořit subkorpus/Create subcorpus* powstanie żądany podkorpus. Możemy do niego wracać, korzystając z opcji *Dostupný korpus/Available subcorpora* w pozycji *Subkorpus.*

Park

Park jest najstarszą wersją interfejsu korpusu InterCorp.

Wpisanie zapytania

Po wpisaniu nazwy użytkownika i hasła otworzy się strona z odnośnikiem do nowego zapytania oraz opcja przełączenia z aktualnej wersji na wersję poprzednią. Jeśli później wrócimy na tę stronę (poprzez kliknięcie przycisku *Domů* 'Początek'), znajdziemy tam listę dotąd aktywnych zapytań, które można wyświetlić ponownie.

Po kliknięciu opcji *Nový Dotaz* 'Nowe zapytanie' program przenosi nas na stronę z listą języków i tekstów, które są aktualnie do dyspozycji. Najpierw należy zaznaczyć przynajmniej dwa języki.

Jeżeli chcemy prowadzić ekscerpcję we wszystkich tekstach w trzonie korpusu i nie interesują nas kolekcje automatycznie opracowanych tekstów ani ich konkretny wybór, możemy przejść do wpisywania zapytania.

Jeżeli natomiast chcemy oprócz trzonu ręcznie skontrolowanych tekstów korpusu przeszukiwać także kolekcje tekstów opracowanych automatycznie, powinniśmy przy odpowiedniej kolekcji kliknąć opcję *Zařadit/Include*. Następnie przechodzimy bezpośrednio do konstruowania zapytania.

Jeżeli chcemy pracować z tekstami spełniającymi określone kryteria, to należy skorzystać z funkcji filtru. Przy jego ustawianiu postępujemy w następujący sposób:

– Najpierw upewniamy się, czy prawidłowo określiliśmy języki, które chcemy analizować;

– Wybieramy żądany język (można to osiągnąć, klikając lewym przyciskiem myszy nazwę języka – powinien przybrać ciemniejszy kolor). Jeśli wybierzemy np. język angielski, na liście tekstów pokaże się informacja o ich angielskich wersjach, latach wydania itp.;

– Możemy także w polu filtrów zaznaczyć bądź odznaczyć typy tekstów, zgodnie z wybranymi danymi bibliograficznymi;

– Filtr możemy również wykorzystać do działań z kolekcją tekstów opracowanych automatycznie (wybór: *Podle filtru/Apply filter*). Po zaznaczeniu opcji *Zařadit/Include* wybrane zostaną teksty z trzonu korpusu według ustawień filtra, lecz dana kolekcja będzie przeszukiwana w całości – bez względu na filtr. Jeśli natomiast zaznaczymy przycisk *Nezařazovat/Exclude*, dana kolekcja nie zostanie przeszukana, niezależnie od ustawień filtra;

– Istnieje także możliwość ręcznego doprecyzowania filtrów. Po wybraniu opcji *Ruční výběr textů/Manual text selection* wyświetli się lista konkretnych tekstów;

– Zaznaczając opcję *Filtruj texty/Filter texts*, uzyskamy wyróżnienie tylko przy tych tekstach, które odpowiadają wcześniej ustawionemu filtrowi. Te parametry możemy jeszcze ręcznie poprawić stosownie do naszych potrzeb.

Należy też pamiętać, że trzeba przestrzegać odpowiedniego porządku operacji, ponieważ dokonanie zmian we wcześniejszych „krokach" powoduje wyzerowanie wszystkich ustawień.

Możliwości zapytania

Użytkownik programu Park dysponuje następującymi możliwościami wyszukiwania:

– formy wyrazowej (*Slovní tvar*/*Word Form*);
– ciągu wyrazów (*Slovní spojení*/*Phrase*);
– wyrażeń języka CQL (analogicznie jak w czeskiej części ČNK);
– lematu (formy podstawowej – *Lemma*) – dla niektórych języków;
– tagu (morfosyntaktická značka); zapytanie trzeba wpisać w formacie CQL – dostępne dla niektórych języków.

Program pozwala prowadzić wyszukiwanie w jednym języku lub w kilku jednocześnie, używając przy wpisywaniu zapytania klawiatury wirtualnej.

Prezentacja konkordancji równoległych

InterCorp dysponuje także możliwością zaprezentowania szczegółowych danych dotyczących konkordancji równoległych. Poniżej przedstawiamy listę dostępnych informacji wraz ze ścieżką dostępu do nich:

– znaczniki strukturalne (*Konkordance*/*Concordance*, *Zobraz možnosti*/*Show options*, *Struktury*/*Structures*);
– dane bibliograficzne i identyfikacja konkordancji (*Konkordance*/*Concordance*, *Zobraz možnosti*/*Show options*, *Reference*/*References*);
– lematy oraz/albo znaczniki morfosyntaktyczne dla słowa kluczowego lub wszystkich wyświetlonych słów (*Konkordance*/*Concordance*, *Zobraz možnosti*/*Show options*, *Atributy*/*Attributes*) – dla niektórych języków;
– „filtrowanie" rezultatów na podstawie obecności lub nieobecności wpisanych wyrażeń w kontekście (*Konkordance*/*Concordance*, *Zobraz filtr*/*Show filter*);
– całe zdania lub wiersze z szukanym wyrażeniem w środku (KWIC);
– większa liczba języków w kolumnach obok siebie lub w wierszach pod sobą (*Pohled: vertikální*/*horizontální*/*View: vertical*/*horizontal*) – przycisk w prawym dolnym rogu ekranu;
– szerszy kontekst (*zobraz kontext*/*show context*);
– eksport konkordancji do tabelki w formacie Excel (Export: xls1, xls2).

Na górze w pasku nawigacji znajdziemy przycisk *Domů/Home/Początek*. Jeśli klikniemy na niego, powrócimy do zapytania.

Na stronach www z opisem korpusu znajdziemy dokładniejsze informacje o znacznikach morfosyntaktycznych.

Porównanie interfejsów

Interfejs KonText jest pod względem funkcjonalności porównywalny z NoSketch Engine, ale jest bardziej intuitywny. Ponieważ ich menu i funkcje są niemal identyczne, poniżej przedstawiamy ich cechy w zestawieniu z interfejsem Park:

- Oba narzędzia służą do wyszukiwania w korpusach jednojęzycznych i równoległych.
- Cechują się szybszą reakcją i mniejszą podatnością na błędy niż Park.
- Dysponują większą liczbą funkcji umożliwiających opracowywanie rezultatów zapytania (klasyfikacje, rozkład częstości, kolokacje) niż Park.
- Ukazują (na życzenie) także poświadczenia, niezawierające ekwiwalentu (funkcja *zobrazit prázdné řádky/include empty lines*).
- Prezentują wielkość korpusu poprzez liczbę pozycji (słów włącznie z interpunkcją), a nie tylko samych słów.
- Wskazują liczbę wyników w postaci liczby pozycji, podczas gdy w Parku chodzi o liczbę słów.
- Oferują zawężenie objętości przeszukiwanych tekstów poprzez sporządzenie subkorpusu.
- Ukrywają kryteria ograniczające objętość przeszukiwanych tekstów.
- Nie dają możliwości eksportu wyników zapytania dla poszczególnych języków bezpośrednio do formatu arkusza kalkulacyjnego (.xls); w Parku jest to możliwe, natomiast w interfejsie KonText jest dostępny eksport do formatu .csv.
- Umożliwiają przeformatowanie do postaci tabelki konkordancji wygenerowanych za pomocą funkcji *Uložit/Save* w formacie *Text* (kodowanie znaków to UTF-8), kolumny zaś są oddzielane tabulatorem.
- Nie pozwalają wyświetlić rezultatów zapytania dla poszczególnych języków kolejno po sobie.
- Udostępniają tylko aktualną wersję korpusu.
- Wyświetlają – podczas przeszukiwania trzech i więcej języków – równoległe konkordancje związane jedynie z językiem prymarnym, a nie ze wszystkimi wyświetlonymi językami.

Wady i zalety korpusu InterCorpu

Niezwykle ważną dla użytkownika opisywaną wyżej funkcją jest możliwość eksportu wyników wyszukiwania do formatu tabelki. Jest to komfortowe narzędzie przy ręcznym wtórnym otagowywaniu wyszukanych wyrażeń.

Żaden z trzech opisywanych wcześniej interfejsów nie jest w tym momencie idealny; Park nie dysponuje funkcjami statystycznymi, a NoSketch Engine/KonText nie umożliwia wygodnego filtrowania tekstów. KonText się jednak ciągle rozwija.

Dla poszczególnych języków różny jest zarówno skład tekstów, jak i ich objętość.

Utrudnieniem w badaniu mogą być także niedokładne metadane (np. oznaczenia języka oryginału); przeszkadza to przy filtrowaniu, tworzeniu subkorpusów. Używając interfejsu Park, można rezultat filtrowania skontrolować i poprawić. Interfejs NoSketch Engine/KonText takiej możliwości nie daje.

W korpusie InterCorp brakuje informacji o typie i dokładności wiązania tekstów (1:1/2:1/1:2, automatycznie/ręcznie kontrolowane); nie ma też możliwości wyszukania tej informacji. Inwentarze znaczników (tagów) dla różnych języków bardzo się różnią, podobnie jak zasady tokenizacji (dotyczy to też języka czeskiego i polskiego); w przypadku niektórych języków w ogóle brakuje anotacji.

W kolekcjach (szczególnie w ACQUIS) dużo jest tekstów zawierających fragmenty w innym języku niż deklarowany. Nie zostało w nich także skontrolowane wiązanie tekstów.

Bibliografia

Čermák P. (2013). *Las posibilidades de estudio ofrecidas por los corpus paralelos: el caso del prefijo español re-*. „Philologica 2/2013 – Romanistica Pragensia" vol. XIX (Les langues romanes à la lumière des corpus linguistiques), Acta Universitatis Carolinae. Praha, s. 123-136.

Čermák F. (red.) (2011). *Korpusová lingvistika Praha 2011: 1 – InterCorp*, vol. 14 of *Studie z korpusové lingvistiky*. Praha.

Čermák F., Corness P., Klégr A., red. (2010). *InterCorp: Exploring a Multilingual Corpus*. Praha.

Čermák F., Kocek J., red. (2010). *Mnohojazyčný korpus InterCorp: Možnosti studia*. Praha.

Čermák F., Rosen A. (2012). *The case of InterCorp, a multilingual parallel corpus*. „International Journal of Corpus Linguistics", 13(3), s. 411-427.

Jindrová J. (2013). *Algumas observações sobre a conorrência dos adjetivos "preto" e "ne- gro" na língua portuguesa*. „Philologica 2 – Romanistica Pragensia", Acta Universitatis Carolinae, vol. XIX, s. 219-229.

Kaczmarska E. (2012). *Czeski czasownik* zdát se *w przekładzie na język polski (na podstawie badań z wykorzystaniem czesko-polskiego korpusu równoległego InterCorp).* „Studia z Filologii Polskiej i Słowiańskiej" (47). Warszawa, s. 247-261.

Kaczmarska E., Rosen A. (2013). *Między znaczeniem leksykalnym a walencją – próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego.* „Studia z Filologii Polskiej i Słowiańskiej" (48). Warszawa, s. 103-121.

Káňa T. (2012). *Wortbildung: Umriss der Theorie mit Aufgaben und Übungen.* Brno.

Káňa T. (2013). *Česká substantivní deminutiva ve světle korpusových dat.* W: *Gramatika & Korpus 2012: 4. mezinárodní konference*, Hradec Králové.

Křen M., Rosen A., Štourač M., Vavřín M., Vondřička P. (2011). *Paralelní korpus InterCorp po sedmi letech [The parallel corpus InterCorp after seven years].* W: Čermák F. (red.) *Korpusová lingvistika Praha 2011: 2 – Výzkum a výstavba korpusů. Studie z korpusové lingvistiky* (15). Praha, s. 105-115.

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

Malá M. (2013). *Translation counterparts as markers of meaning. The case of copular verbs in a parallel English-Czech corpus.* Languages in Contrast 13:2, s. 170-192.

Malá M., Šaldová P. (2012). *Complex condensation – Vilém Mathesius, Josef Vachek, Jiří Nosek, and beyond.* W: Malá M. and Šaldová P. (eds.) *A Centenary of English Studies at Charles University: form Mathesius to Present-day Linguistics.* Praha, s. 135-158.

Nádvorníková O. (2013). *Paralelní korpus InterCorp – co může nabídnout překladatelům a translatologům?.* „PLAV – měsíčník pro světovou literaturu". Praha, s. 47-51. http://www.svetovka.cz/archiv/2013/01-2013-aktualita.htm.

Nádvorníková O. (2013a). *– Paul se rase en chantant, dit-il en bafouillant : Quels types de manière pour le gérondif en français ?* Acta Universitatis Carolinae Philologica 2 (Romanistica Pragensia XIX), s. 31-44.

Peloušková H. (2013). *Český a německý volný dativ ve srovnání.* Brno.

Peloušková H. (2013a). *Konstrukce s formálním objektem v němčině a jejich protějšky v češtině.* W: *Gramatika a korpus 2012: 4. mezinárodní konference.* Hradec Králové.

Piasecki M. (2007). *Polish Tagger TaKIPI: Rule Based Construction and Optimisation.* „TASK Quarterly. Scientific Bulletin of the Academic Computer Centre in Gdansk", vol. 11.

Rosen A. (2010). *Morphological tags in parallel corpora.* W: Čermák F., Klégr A., Corness P. (red.) *InterCorp: Exploring a Multilingual Corpus.* Studie z korpusové lingvistiky (13), s. 205-234. Praha.

Rosen A., Vavřín M. (2012). *Building a multilingual parallel corpus for human users.* W: N. Calzolari et al. (red.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, s. 2447-2452.

Vavřín M., Rosen A. (2008). *InterCorp: A Multilingual Parallel Corpus Project.* W: *Proceedings of the International Conference Corpus Linguistics – 2008.* St. Petersburg, s. 97-104.

Vondřička P. (2010). *TCA2 – nástroj pro zpracovávání překladových korpusů [TCA2 – a tool for processing translation corpora].* W: Čermák F., Kocek J. (red.) *Mnohojazyčný korpus InterCorp: Možnosti studia [Multilingual Corpus InterCorp: Research Options].* Praha, s. 225-231. Praha.