Praktyczny przewodnik po korpusach języka ukraińskiego

Natalia Kotsyba

Uniwersytet Warszawski

Informacje ogólne

Korpusy tekstów języka ukraińskiego obecnie są aktywnie rozwijane w kilku ośrodkach, zarówno w 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 **UNKL**) opracowywany głównie w latach 2005-2010 przez pracowników Ukraińskiego funduszu lingwistyczno-informacyjnego (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**), rozwijany od 2011 w Laboratorium Lingwistyki Komputerowej na Uniwersytecie Narodowym im. Tarasa Szewczenki w Kijowie pod kierownictwem dr Natalii Darczuk, dostępny publicznie w sieci pod adresem http://www.mova.info/corpus.aspx.

Język ukraiński jest reprezentowany również w kilku korpusach równoległych, dostępnych do wyszukiwania publicznego: równoległy korpus rosyjsko-ukraiński w ramach projektu NKJR[2] (ok. 38 mln. słów w obu językach), polsko-ukraiński korpus równoległy PolUKR[3] (3,5 mln. słów w obu językach), wielojęzyczny korpus równoległy PARASOL[4] (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 podpisów filmowych, stworzony w Kijowskim Narodowym Uniwersytecie Lingwistycznym[5] [Lebediew:36-37], zawierający ok. 200 tys. wyrazów ukraińskich, nie mniej, bez lematyzacji czy znakowania gramatycznego.

[1] Korpus ten nie jest dostępny bezpośrednio przez stronę internetową, lecz wymaga nieskomplikowanej technicznie instalacji programu do wyszukiwania, dostępnego pod podanym adresem. Ponadto, korzystanie z korpusu wymaga rejestracji w celu otrzymania nazwy użytkownika i hasła oraz logowania się przy każdej sesji.
[2] http://www.ruscorpora.ru/search-para-uk.html.
[3] http://www.domeczek.pl/~polukr/index.php?option=welcome.
[4] http://parasol.unibe.ch/.
[5] http://complinguide.com.ua/Corpus.aspx.

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 UNKL, on także jest kilkakrotnie większy objętościowo od KUMu oraz bardziej szczegółowo i dogłębnie opisany, ma dłuższą historię rozwoju. Jednak ze względu na brak bezpośredniego dostępu do niego oraz nieudostępnioną informację gramatyczną będziemy używać w opisie obu wymienionych korpusów jednojęzycznych. Zwłaszcza, że z perspektywy porównawczej lepiej można ocenić mocne i słabe strony oraz określić perspektywistyczne cechy rozwoju dla obu z nich bądź korpusu połączonego, który mógłby kiedyś powstać.

Charakterystyka korpusów

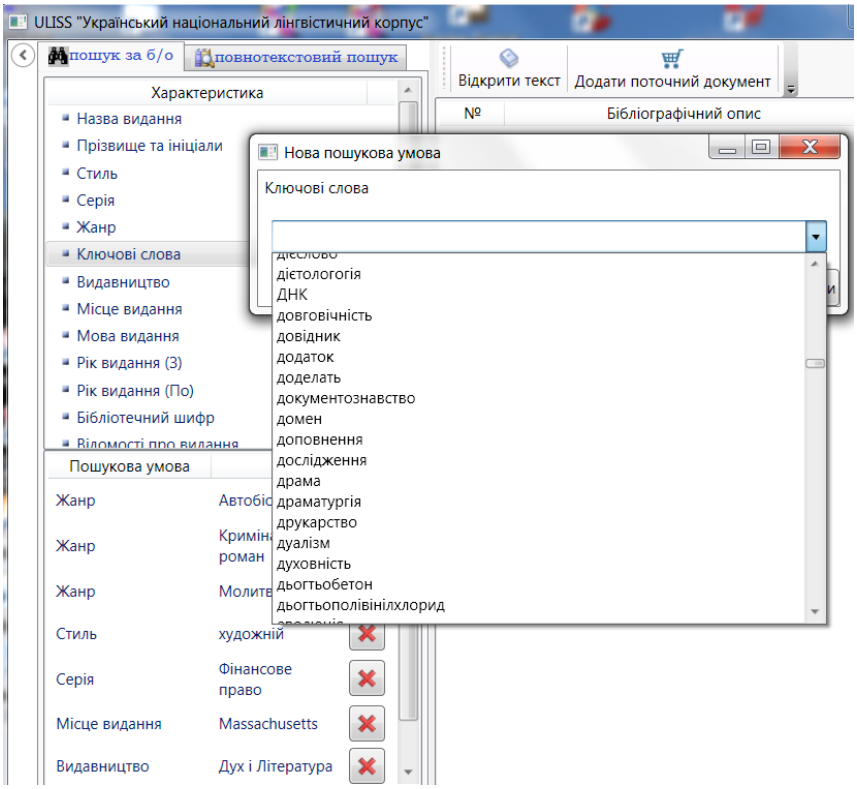
Ze starszych [Szyrokow et al. 2005] oraz nowszych relacji [Szyrokow et al. 2011:1] wiemy, iż UNKL składa się z korpusu ogólnego (76 mln słów) oraz korpusu języka ustawodawstwa (18 mln słów). Oprócz tego korpus ogólny jest bardzo zróżnicowany pod względem gatunków, stylów, szczegółowo oznakowany następującymi rodzajami metainformacji: nazwa wydania; nazwisko autora, seria; gatunek (ok. 50 pozycji); styl (konfesyjny, naukowo-dydaktyczny; ludowy; naukowy; naukowo-informacyjny; naukowo-popularny; dyplomatyczny; ustawodawczy; poetycki; publicystyczny; rozmowny; literatura piękna); wydawnictwo (kilkadziesiąt pozycji); miejsce wydania (ok. 100 pozycji); język wydania[6]; rok wydania (od/do, z możliwością wpisywania okresu); liczba stron, oraz kilka innych, mniej istotnych dla użytkownika pod względem lingwistycznym, informacji bibliograficznych. Niestety brakuje w dokumentacji informacji statystycznych, związanych z metainformacją, ale, sądząc z opisów w literaturze [Szyrokow et al. 2005] oraz subiektywnej oceny wyników wyszukiwania w UNKL, 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

[6] Ta cecha w danym przypadku jest nadmiarowa, prawdopodobnie pozostała w ramach ujednolicenia ze stosowanymi w UNKL standardami metainformacji UNMARC (Biblioteki Kongresu USA) [Szyrokow et al. 2005:267-271].

być świadomym, pracując z tym korpusem. W zależności od celu badawczego może wyniknąć potrzeba odfiltrowania tekstów tłumaczonych[7].

Ciekawą opcją jest wyszukiwanie według słów kluczowych – do wyboru jest kilkaset kręgów tematycznych, np. *awiacja*. *Azarow*[8], *język angielski*, *hematologia*, *sanskryt*, itd. Takie bogactwo parametrów wyszukiwawczych pozwala określić użytkownikowi w dowolny sposób dopasowany do swoich potrzeb podkorpus. Szczególnie to może być przydatne przy badaniach terminologii czy różnych aspektów lingwistycznych gatunków tekstów. Rys. 1 niżej pokazuje możliwości definiowania podkorpusu UNKL według różnych metainformacji.

Rysunek 1 Uściślanie warunków metainformacyjnych dla wyszukiwania w UNKL



Po lewej stronie u góry widoczny jest fragment listy rodzajów metainformacji do wyboru („Характеристика”), pod nim widoczne są wybrane przez nas opcje[9]: gatunek ‘autobiografia’ albo ‘kryminał’ albo ‘modlitwa’, styl ‘literatura piękna’, seria ‘prawo finansowe’, miejsce wydania ‘Massachusetts’, wydawnictwo ‘Duch i Litera’[10]. W okienku

[7] Nie wydaje się, żeby to można było uczynić w zbyt prosty sposób, za pomocą jednego kliknięcia, jednak, opcja dowolnej aranżacji tekstów do wyszukiwania w UNKL teoretycznie daje taką możliwość, a opcje podkorpusu można zapisać w osobnym pliku i później z nich korzystać ponownie.
[8] Obecny Premier Ukrainy.
[9] Przy takiej konfiguracji zostanie zwrócony pusty zbiór tekstów.
[10] Do metainformacji w UNKL wkradł się błąd – faktyczna nazwa wydawnictwa to „Дух і літера”.

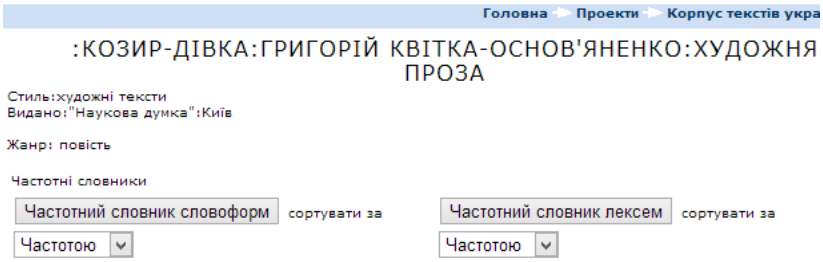
„нова пошукова умова” (nowy warunek wyszukiwania) widać fragment listy dostępnych słów kluczowych, ułożonych alfabetycznie.

Wyszukiwania w tekstach publicystycznych (albo według innych parametrów) dokonujemy w sposób następujący: otwieramy wkładkę ‘wyszukiwanie według bibliografii’; zadajemy warunek wyszukiwania podwójnym kliknięciem na kryterium „жанр” (gatunek), wybieramy interesujący nas gatunek i klikamy „додати умову” (dodaj warunek). Następnie klikamy przycisk „пошук” (wyszukiwanie), żeby zobaczyć, ile dokumentów odpowiada zadanemu kryterium gatunku oraz 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 wkładki „повнотекстовий пошук” (wyszukiwanie pełnotekstowe) z zaznaczeniem opcji „пошук за текстами з поточного кошика” (wyszukiwanie w tekstach z obecnego koszyka).[11]

KUM zawiera łącznie 13 mln słów[12]. Ma on także bardzo odmienną od UNKL aranżację wewnętrzną. Składa się z czterech podkorpusów: literatura piękna (proza), literatura piękna (poezja), folklor oraz inne, przy czym jednocześnie można korzystać tylko z jednego rodzaju podkorpusu, co jest istotną wadą, ponieważ zmusza użytkownika do zadawania czterech pytań zamiast jednego i dodatkowej obróbki danych po tych zapytaniach. Proporcjonalnie rozkład według gatunków w KUM wygląda następująco: poezja (1 mln wyrazów); literatura piękna (7 mln), folklor (32 tys.); publicystyka (4 mln); literatura naukowa społeczno-humanistyczna (1 mln); teksty oficjalne (1 mln) [Darczuk 2012a:102-103]. W korpusach KUM możliwe jest wybieranie płci autora tekstów przy wyszukiwaniu, co może być bardzo pomocne dla badań genderowych i socjolingwistycznych w ogóle. Dostępne są także inne metainformacje do wglądu, ale dopiero po fakcie wyszukiwania, podlinkowane do każdego osobnego poświadczenia. Po kliknięciu na link „джерело” (źródło) po prawej stronie od poświadczenia, na osobnie otwieranej 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 na str. 14, zob. rys. 2.

[11] Relacja osobista Nadiji Sydorczuk, pracownika UFLI (maj 2011 r.), podana została tutaj ze względu na brak innych dostępnych instrukcji do pracy z UNKL.
[12] 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.

Rysunek 2 Metainfromacje dotyczące konkretnego poświadczenia w KUM



Metainformacje zarówno w UNKL, jak i w KUM, podane są w sposób wyczerpujący i rzetelny. Pewien niedosyt pozostaje jednak ze względu na brak podsumowań statystycznych według różnych metadanych – to dawałoby użytkownikom lepsze rozumienie architektury obu korpusów.

Segmentacja tekstu jest oparta w obu korpusach głównie na wyrazach, nie ma widocznego dla użytkownika podziału na zdania czy akapity. Natomiast segmentacja wewnątrzwyrazowa jest organizowana w korpusach w sposób odmienny. W KUM łącznik jest uważany za integralną część wyrazów. Ma to konsekwencje takie, że nie można robić wyszukiwania według jednej z części wyrazu, zawierającego łącznik, musi zostać wpisany cały wyraz. Dotyczy to także form, dopuszczających duży stopień swobody połączeń, np. kolorów albo 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 zwracano, odpowiednio, jedyne poświadczenie, które jest w korpusie z tą formą przysłówkową. Natomiast forma *жовто-червоний* ‘żółto-czerwony’ już zwraca większą liczbę poświadczeń. Jeżeli forma nie funkcjonuje samodzielnie, to odpowiedź jest pusta.

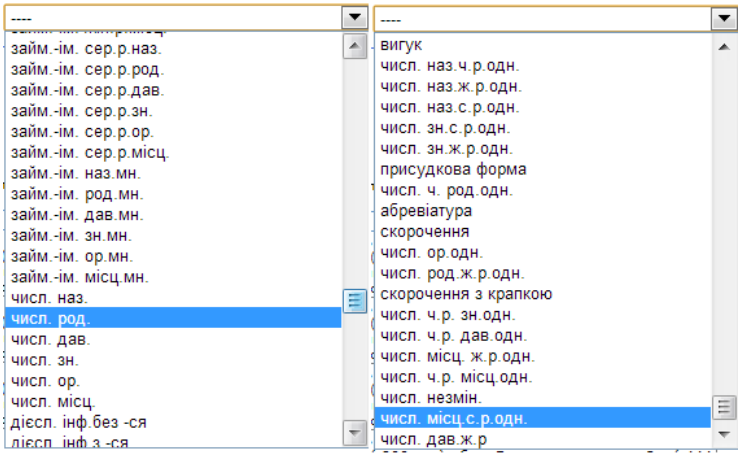
Bardziej pomyślnie jest organizowane wyszukiwanie form z łącznikiem w UNKL. Wpisanie formy częściowej *жовто-* zwraca wszystkie formy, zawierające tę kombinację znaków, nawet te, gdzie ona jest w środku wyrazu trzyczłonowego *зелено-жовто-червону* ‘zielono-żółto-czerwoną’. Z tego należy wnioskować, że w UNKL łącznik jest traktowany jako segment, łączący wyrazy. Wyszukiwanie według samego łącznika zwraca zbiór pusty. Tu jeszcze należy zaznaczyć, że w UNKL trafiają się wypadki nieprawidłowej ortografii, wynikającej z pozostania po skanowaniu i digitalizacji wersji papierowych tekstów łączników, rozdzielających wyrazy przy przeniesieniu ich na następną linijkę. Odpowiednio, przy indeksowaniu łącznik nierozdzielający jest traktowany jako myślnik, dlatego wyszukiwanie według kombinacji liter, z których mogą zaczynać się wyrazy, np. *жов* zwraca nieistniejące lematy, będące w rzeczywistości częściami innych, np. *жов-нір* ’żoł-nierz’.

Skrótowce w UNKL można wyszukiwać, podając formę bez kropki albo z nią. Jednakowe wyniki otrzymamy, wpisując *muc* ‘tys’ albo *muc.* ‘tys.’.[13] W wyrazach, zawierających łącznik sygnalizujący odmianę, np. *30-ма* ‘30-ma’, *30-му* ‘30-stu’, *30-х* ‘30-ch’, *30-річний* ’30-letni’, łącznik jest traktowany tak samo jak myślnik. Zapytanie, sformułowane liczbowo, np. *30*, zwraca wyżej wymienione przykłady oraz inne, łącznie pisane poświadczenia nieliterowe: *30%*, *№30*, *c. 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 śladową [Darczuk 2012b]. Nie mniej, z bardzo podobnego podziału kategorialnego, np. nietypowe traktowanie form stopni porównania przymiotników i przysłówków jako odrębne lematy, wyodrębnienie czasowników z i bez postfiksu *-ся* (pisany łącznie z formą czasownikową odpowiednik polskiego *się*), dwuliterowy sposób kodowania oraz same używane kody, itd., widać, 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, pod czas gdy w UNKL jest to jedna kategoria.

Rysunek 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 liście wypadającej z możliwością wyboru

[13] Generalnie znaki nieliterowe w oknie wyszukiwania są ignorowane – identyczne wyniki otrzymamy na zapytanie *30* oraz *30%*.

jednej opcji na raz (zob. rys. 3). Dla wygody przedstawiamy niżej, w tab. 1, kompaktowy wykaz wszystkich dostępnych kombinacji kategorii z tej listy w porządku, w którym one się na niej pojawiają, oraz robimy podsumowanie statystyczne owych kombinacji.

Tab. 1. Kompaktowe przedstawienie kategorii gramatycznych i/albo ich kombinacji, które można użyć jako kryteria wyszukiwania gramatycznego w KUM

Część mowy	Cecha 1	Cecha 2	Cecha 3	Cecha 4	Liczba kombinacji
rzeczownik	przypadek (7)				7
		liczba mnoga			7
	przypadek (7)		rodzaj (3)		21
		liczba mnoga	rodzaj (3)		21
	niezmienny				1
			rodzaj (3)		3
	przypadek (7)		rodzaj (3)	nazwa własna	21
	niezmienny		rodzaj (3)	nazwa własna	3
	przypadek (7)	pl. Tantum			7
	niezmienny	pl. Tantum			1
	przypadek (7)	pl. Tantum		nazwa własna	7
adiektyw (przymiotnik)	przypadek (6)		rodzaj (3)		18
		l. mnoga			6
			rodzaj (3)	nazwa własna	18
zaimek-przymiotnik	przypadek (6)		rodzaj (3)		18
		l. mnoga			6
zaimek-rzeczownik	przypadek (6)		rodzaj (3)		18
		l. mnoga			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
		liczba mnoga		z/bez *-ся*	2
	czas przeszły			z/bez *-ся*	2
	osoba (3)	liczba (2)	czas przyszły	z/bez *-ся*	12
	osoba 1	liczba mnoga	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
		liczba mnoga			1
przyimek	przypadek (5)				5
partykuła					1
przysłówek					1
predykatyw1					1
spójnik	współrzędny/podrzędny/ oba				3
wykrzyknik					1
liczebnik	mianownik	rodzaj (3)	liczba pojed.		3
	biernik	rodzaj (3)	liczba pojed.		3
		rodzaj męski	liczba pojed.		1
predykatyw2					1

abrewiatura					1
skrót					1
liczebnik	narzędnik		liczba pojed.		1
	dopełniacz	rodzaj żeński	liczba pojed.		1
skrót z kropką					1
liczebnik	celownik	rodzaj męski	liczba pojed.		1
	miejscownik	rodzaj (3)	liczba pojed.		3
	niezmienny				1
	celownik	rodzaj żeński			1
Suma					275

Jak widać, lista jest dość długa – 275 pozycji, co czyni ją niewygodną w korzystaniu. Nie wszystkie możliwe kombinacje są dostępne: widzimy różne kombinacje przypadków i rodzaju dla rzeczowników w ogóle oraz dla rzeczowników w liczbie mnogiej, ale nie ma możliwości wyszukiwania tylko w liczbie pojedynczej. Podobnie jest z czasem przyszłym czy teraźniejszym, kryterium nazwy własnej, itd. – zbiory nie są ortogonalne. Są też braki – nie ma zaimka dzierżawczego rodzaju nijakiego, oraz powtórzenia – czasownik 1 os. l.mn. z -*ся* pojawia się na liście dwa razy. W sposób bardzo niekonsekwentny podane są kombinacje zaimków i liczebników. Wymaga też objaśnienia nie zupełnie oczywista dla przeciętnego użytkownika różnica między kategoriami „присудкова форма" i „присудкове слово" (w tabeli nazwane odpowiednio „predykatyw1" i „predykatyw2"), jak i same te kategorie.[14]

Lematyzacja

Lematyzacja w obu korpusach jest prezentuje dość wysoki poziom, ale należy być świadomym kilku pułapek. W KUM problemy pojawiają się przy wyrazach rzadkich, nowych, którym lematy, jak i tagi, są przypisywane automatycznie na podstawie „zgadywania". 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, składającym się na błędy lematyzacyjne w UNKL, są wspomniane wcześniej łączniki nierozdzielające, stosowane przy podziałach słów, które *de facto* rozdzielają jeden właściwy, istniejący wyraz na dwa niepoprawne, które są później indeksowane jako lematy.

[14] W gramatykach akademickich spotykamy różne traktowanie kategorii predykatywu, zob. też analizę w [Kotsyba, Derzhanski 2008].

Anotacja morfosyntaktyczna

W wersji publicznej UNKL anotacja morfoskładniowa nie jest udostępniana wcale, ze względu na niewystarczająco wysoki poziom jakości dezambiguacji[15]. 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 automatyzowany [Darczuk 2012b], formom niejednoznacznym był przypisywany jeden tag na podstawie zadanych przez twórców korpusu reguł kontekstowych. Jakość dezambiguacji morfoskładniowej niestety nie dostarcza optymizmu. Mimo, że autorzy stwierdzają, że poprawność dezambiguacji osiąga poziomu 93-95% [Darczuk 2012b:19], przeprowadzone eksperymenty pokazują zupełnie inny obraz – pozostało jeszcze wiele błędów i niedokładności.

Zapytanie o znalezienie niejednoznacznej formy wyrazu *хатки* ‘domki’ jako rzeczownika w mianowniku liczby mnogiej m.in. zwróciło formę 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 są w mianowniku, 6 w bierniku oraz jedna forma w dopełniaczu liczby pojedynczej. Zapytanie o formę dopełniacza liczby pojedynczej tego wyrazu też zwróciło błędne odpowiedzi, których liczba znacznie przekracza dopuszczalny w takich przypadkach błąd statystyczny.[16]

Poniższe poświadczenia zostały zwrócone na zapytanie o formę *мати* ‘matka/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]

[15] Relacja osobista pracowników Ukraińskiego Funduszu Lingwistyczno-Informacyjnego.

[16] Biorąc pod uwagę ilość danych językowych do opracowania oraz stopień złożoności morfologii języków słowiańskich, trudno spodziewać się stuprocentowej poprawności tagowania. Nie mniej jednak, w stosunku do korpusów porównywalnych języków – czeskiego, polskiego – podawano poziom poprawnych tagów w zakresie 90-95% od całości. W korpusie narodowym języka rosyjskiego stosowana jest ręczna dezambiguacja i umożliwiano wyszukiwanie tylko w tym podkorpusie. W pozostałych wypadkach badacze otrzymują informacje o wszystkich możliwych interpretacjach morfoskładniowych, co mogłoby być lepszym, niż obecne, rozwiązaniem w przypadku KUM.

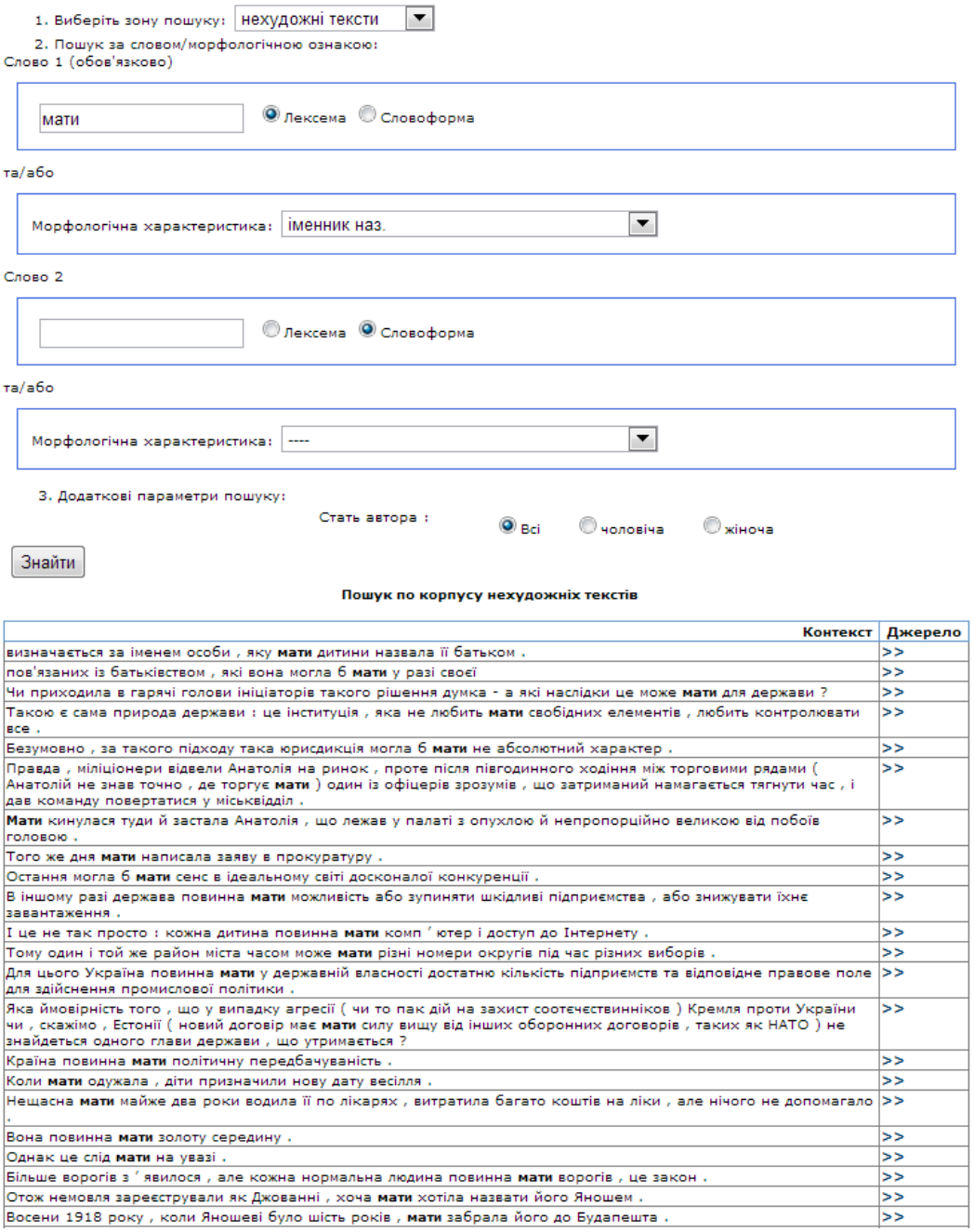
<*Matka – we Włoszech, ojciec pije, dziecko – na szyi u babci, która ma 70 lat.*>

Вода тяжка й густа пророка колихала , наче ***мати*** *сина* [KUM 22.III.2013]

<*Woda ciężka i gęsta proroka kołysała, jak matka syna*>

Odwrotne zapytanie, z zaznaczeniem opcji gramatycznej „rzeczownik w mianowniku" zwraca wyniki, gdzie czasowników już jest 14 (poświadczenia: 2,3,4,5,9,10,11,12,13,14,15,18,19,20) spośród pierwszych 22 widocznych poświadczeń, tzn. tylko 1/3 wyników jest trafiona, zob. rys. 4 niżej.

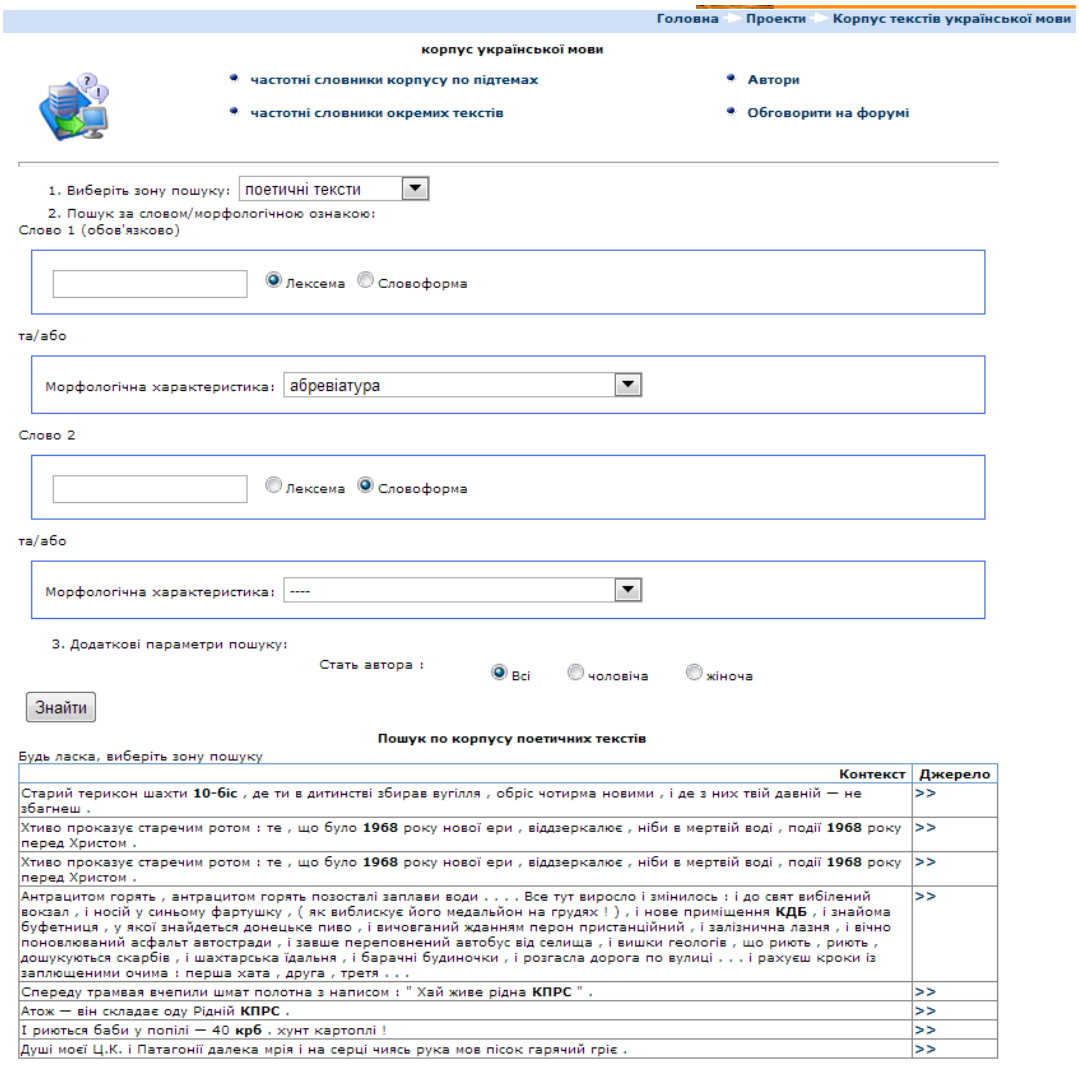
Rysunek 4 Wyszukiwanie w KUM jednocześnie według zadanego lematu i ograniczeń morfologicznych



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 соотечественников) 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 .	>>

Zapytanie o znalezienie skrótowców w tekstach poetyckich zwróciło osiem poświadczeń (jeden z nich powtórzony), gdzie do skrótowców zostały zaliczone m.in. lata: *1968*, segmenty, gdzie indziej klasyfikowane jako skróty z kropką: *крб.* ‘karbowaniec’ (była waluta ZSSR), nazwy własne nieskrócone:*10-біс* ’10-bis’ (nazwa kopalni), zob. rys. 5 niżej.

Rysunek 5 Wyszukiwanie w KUM: parametry i wyniki



Jeszcze jeden test wykazał, że daleko nie wszystkie słowoformy mają przypisany w korpusie właściwy tag morfoskładniowy. Dla przykładu, 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 jednemu *рр*. 'lata' i *млн*. 'mln'. Natomiast, wyszukiwanie samej tylko formy *км* w tekstach tegoż podkorpusu zwraca sześć stron wyników, każda po 50 poświadczeń. Niestety, nie ma możliwości sprawdzenia, pod jakim tagiem (i czy w ogóle tag został formie przypisany) można wyszukiwać danych form, ponieważ informacja morfoskładniowa w znalezionych tekstach nie jest wyświetlana. Nie są te formy wyświetlane ani jako abrewiatury, ani jako skróty z kropką.

Podsumowując, jakość informacji gramatycznej KUM znajduje się obecnie w takim stanie, że należy być bardzo ostrożnym, decydując się na jej wykorzystanie w swoich pracach. Stopień jej pożyteczności bardzo ściśle jest związany z rodzajem prowadzonych badań.

Podstawowe zapytania

Oba korpusy umożliwiają wyszukiwanie według form wyrazowych albo lematów poprzez zaznaczenie odpowiedniej opcji w oknie wyszukiwania. Tak samo w obu korpusach warianty ortograficzne (pisanie wyrazu z małej lub dużej litery) są unifikowane automatycznie, bez możliwości wyszukiwania jednej bądź drugiej opcji osobno. Nie mniej, takie rozwiązanie wydaje się bardziej ułatwiać pracę badacza, niż ją utrudniać.

W obu korpusach możliwe jest także wyszukiwanie grupy wyrazów. W KUMie wpisujemy każdy z wyrazów oddzielnie, zaznaczając przy tym, czy jest to lemat czy forma wyrazowa. Istnieje możliwość wpisania maksymalnie dwóch wyrazów/lematów. Natomiast UNKL umożliwia wyszukiwanie dość długich ciągów wyrazów, nawet całych zdań, przy czym można też zadać możliwy dystans między szukanymi wyrazami w korpusie.

W KUM można także wybrać połączenie warunków gramatycznych i leksykalnych albo tylko warunek gramatyczny. W odróżnieniu od KUMu, w UNKL 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. do wyrazu *листя* ‘liście’ moglibyśmy szukać epitetów, zaznaczając pierwszy wyraz jako przymiotnik[17]. W praktyce zapytania tego rodzaju nie są technicznie możliwe – zwracany jest błąd systemu.[18] W odwróconym szyku (kiedy najpierw podany jest lemat/forma wyrazowa, a charakterystyka gramatyczna jest zaś drugim członem zapytania) system zwraca wyniki, z tym, że do adiektywów zaliczane są także formy czasownikowe czasu przeszłego, co stwarza dodatkowe niedogodności przy wyszukiwaniu kolokacji. Tego problemu można byłoby częściowo się

[17] Przymiotnik jako samodzielna część mowy nie występuje w tagsecie, zob. przedstawienie na str. 7. Używany natomiast jest bardziej ogólny termin „adiektyw”, który obejmuje także liczebniki porządkowe oraz imiesłowy przymiotnikowe.
[18] Wyszukiwaliśmy siedem różnych konfiguracji jednostek. Skutek był za każdym razem ten sam.

pozbyć, jeżeli wyniki można byłoby zapisywać w postaci tabeli z możliwością sortowania, ale ta opcja obecnie nie jest zaimplementowana. Jeszcze jedna niedogodność, która pozostaje, to niemożliwość wybrania przymiotnika jako kategorii ogólnej. Użytkownik jest zmuszany do wybrania szczegółowej charakterystyki: adiektyw wraz z określonym rodzajem i w określonym przypadku. To wymaga formułowania 18 zapytań zamiast teoretycznie możliwego jednego.

KUM umożliwia zapytania o samą kategorię gramatyczną, np. wyraz rodzaju męskiego w bierniku liczby pojedynczej czy abrewiatury, zob. rys. 5. Możliwe też są zapytania o lemat w określonej kategorii gramatycznej, np. *мати* ‘matka/mieć’ jako rzeczownik w mianowniku, zob. rys. 4.

Dostępne funkcje w oprogramowaniu oprócz wyżej wymienionych

Oba korpusy zwracają wyniki wyszukiwania w postaci konkordancji. Żaden z nich nie daje możliwości wyświetlania wyników w postaci tabeli KWIC, ani pozwala odfiltrowywać znalezionych danych. KUM nie podaje żadnych informacji statystycznych o znalezionych wynikach. Wyświetla się po 50 wyników na stronę, jeżeli są kolejne strony, to pojawia się na dole link z napisem „наступна сторінка” (kolejna strona), ale nie widać, ile jest stron w ogóle. UNKL natomiast wyświetla liczbę poświadczeń, jak i liczbę tekstów, w których je znaleziono, a także liczbę poświadczeń w każdym z tekstów. Możliwe jest także dopasowanie odległości między wyrazami w grupie wyrazowej. Na zapytanie *жовте листя* ‘żółte liście’ z możliwą odległością 3 wyrazy UNKL zwrócił 77 wyników (dla porównania 1 – 70; 2 – 76).

W UNLK istnieje możliwość wyszukiwania od razu w ramach całego rzędu synonimicznego do podanego wyrazu. Może to być bardzo przydatna funkcja dla badań semantycznych, kognitywistycznych oraz w ogóle dla nauk 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’, który w formie podstawowej jest tożsamy z formą lematem *мати* ‘matka’.

Rysunek 6 Wyszukiwanie w UNKL z użyciem opcji rzędów synonimicznych

Bardzo pożyteczną funkcją w UNKL jest możliwość zapisywania wyników. Możemy zdefiniować, czy chcemy zapisać wyniki z wybranego źródła czy ze wszystkich, ograniczyć liczbę zapisywanych wyników z każdego źródła, ustalić długość kontekstu (domyślny jest 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 dostępne przy KUMie. Można wygenerować słownik na podstawie lematów albo form wyrazowych do dowolnego tekstu. Bardzo pożyteczną jest funkcja „wykrojenia" części słownika frekwencyjnego dla konkretnej części mowy, pokazywane są na żądanie także niektóre miary statystyczne, np. liczba tekstów, średnia częstotliwość, odchylenie standardowe, współczynnik stabilności. Dostępne są także słowniki częstotliwości morfemów (prefiksy, sufiksy, interfiksy, korzenie). Przy niektórych słownikach zaznaczono jednak, że dezambiguacja nie jest zrobiona całkowicie, dlatego możliwe są odchylenia od stanu faktycznego.

Funkcja, której najbardziej brakuje w obu korpusach, a która może być dopracowana względnie niewielkim kosztem, to sortowanie wyników wyszukiwania według lematu wyszukiwania, lewego czy prawego kontekstów. Bardzo brakuje w przypadku KUM

podstawowych informacji statystycznych, takich np., jak 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. [19]

UNKL obecnie działa jednocześnie jako baza bibliograficzna oraz korpus tekstów, co oznacza, że informacje bibliograficzne są bardzo obszerne i dobrze ustrukturowane (istnieje możliwość tworzenia „koszyków" bibliograficznych). Akcent na danych bibliograficznych oraz połączenie tych funkcji w ogóle zostało zakrojone pod konkretne potrzeby leksykograficzne twórców UNKL, dla przeciętnego użytkownika korpusu wydaje się to jednak zbędne.

Metodologia i perspektywy badawcze

Wymienione korpusy przede wszystkim są cenne dla wyszukiwania poświadczeń językowych według lematu bądź formy wyrazowej albo ciągu wyrazów oraz tworzenia konkordancji, co jest bardzo pożyteczne dla celów leksykograficznych. Ze względu na zawarte informacje morfoskładniowe (KUM) nadają się także do badań nad składnią. Co prawda, ciąg wyszukiwawczy jest ograniczony do dwóch jednostek wyrazowych, co poważnie zmniejsza możliwości badawcze. Ogólnie rozumiane badania *corpus-driven* też są możliwe, ponieważ korpusy są w miarę duże i w miarę zrównoważone. Niestety, nie ma jeszcze możliwości wyciągania informacji statystycznych albo kolokacji w żadnym z korpusów. Należy spodziewać się, że nowe funkcje wyszukiwacze będą pojawiać się wraz z wzrostem zainteresowania korpusami i metodami korpusowymi przez badaczy ukraińskich.

UNKL, który zawiera miano narodowego, po szybkim wzroście w latach 2005-2010, nie okazuje ostatnio znaków rozwoju. Dalszy rozwój i popularyzacja korpusu nie jest zadaniem priorytetowym ULFI, skupionego głównie na wyzwaniach leksykograficznych, dla których m.in. i został stworzony UNKL. W chwili obecnej jego głównym zastosowaniem jest dostarczanie cytatów leksykograficznych oraz monitoring znaczeń leksykalnych do powstającego 20-tomowego słownika objaśniającego języka ukraińskiego. Całkiem praktyczne podłoże miało także powstanie KUM – jego głównym na dany czas produktem są

[19] Obecnie (marzec 2013 r.) zawiera ono cztery wpisy, z których dwa to informacje o nowych dodanych funkcjach korpusu, i 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 intencją ze strony twórców korpusu, wymaga jednak ono należnego administrowania.

zróżnicowane i w miarę jakościowe słowniki frekwencyjne, częściowo udostępnione przez autorów. Ogólnie rzecz ujmując, zarówno KUM jak i UNKL są danymi 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 pisanych przez badaczy spoza jednostek, gdzie one są opracowywane. Obecnie uwaga językoznawców skupia się głównie albo 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). Istnieją poglądy [Perebyjnis, Bobkowa: 4], że w dużej mierze jest to związane z sytuacją z lingwistyką komputerową w kraju w ogóle, mającą podłoże historyczne i polityczne. Brak informacji o badaniach lingwistycznych, przeprowadzonych na materiale korpusów języka ukraińskiego[20], świadczy ogólnie o dwóch istotnych rzeczach: z jednej strony, niewystarczająca jest świadomość potencjalnych użytkowników owych korpusów, a to za sprawą zbyt słabej ich popularyzacji, z innej zaś, jakość tych korpusów oraz poziom ich dostępności na razie nie pozwala na podejmowanie się poważnych badań na ich materiale. Dlatego twórcy korpusów języka ukraińskiego wciąż stoją przed wyzwaniem przede wszystkim ich poprawy jakościowej, ale także dalszego rozwoju pod względem bogactwa poziomów znakowania oraz atrakcyjnych metod wyciągania już zawartej informacji. Dobrym rozwiązaniem byłoby połączenie wysiłków różnych ośrodków oraz załączenia szerokich kręgów językoznawców Ukrainy w celu stworzenia centralizowanego, monitorowanego pod względem jakości i odpowiedniości standardom międzynarodowym, korpusu narodowego 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, 2008, s. 68–79. Moskwa. [http://domeczek.pl/~natko/papers/ID_NK_tagSlav.pdf]

[20] Śledzone i sprawdzane są konferencje językoznawcze oraz korpusowe krajowe i międzynarodowe, strony internetowe większych lingwistycznych jednostek badawczych Ukrainy.

Kotsyba N. (2012). *PolUKR (a Polish-Ukrainian Parallel Corpus) as a Testbed for a Parallel Corpora Toolbox*. W: „Prace Filologiczne", t. LXIII, s. 181–196. Warszawa. [http://www.domeczek.pl/~natko/papers/NKotsyba_SlaviCorp2010_paper.pdf]

Дарчук Н. (2012a). *Корпус Украинского Языка*. W: „Prace Filologiczne", t. LXIII, s. 99–108. Warszawa.

Дарчук Н. (2012b). *Морфологічне анотування Корпусу української мови*. W: "Комп'ютерна лінгвістика: сучасне та майбутнє", s.16–19. Kijów.

Демська-Кульчицька О. (2005). *Основи національного корпусу української мови*. Kijów.

Демська О. (2011) *Текстовий корпус: ідея іншої форми*. 282 s. Kijów.

Комп'ютерна лінгвістика: сучасне та майбутнє. Матеріали Міжнародної науково-практичної конференції (2012), 52 s. Kijów. [http://www.mova.info/zbirnyk.pdf]

Лебедєв К. (2012). *Створення Багатомовного корпусу паралельних текстів*. W: "Комп'ютерна лінгвістика: сучасне та майбутнє", s. 36–37. Kijów.

Перебийніс В., Бобкова Т. (2012). *Історія лабораторії комп'ютерної лінгвістики КНЛУ. Комп'ютерна лінгвістика: сучасне та майбутнє*. W: "Комп'ютерна лінгвістика: сучасне та майбутнє", s. 3–4. Kijów.

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,, 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", s. 285–294. Kijów, Instytut Języka Ukraińskiego. [http://ovbugakov.org.ua/articles/Bugakov_Oleg_article2.pdf]

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, „Dowira", 471 s.