MAGDALENA TURSKA
(bez afiliacji)
NATALIA KOTSYBA
Instytut Slawistyki PAN

Polsko – ukraiński korpus równoległy (PolUKR)[1]

1. Korpusy i korpusy równoległe.

Obecnie można zaobserwować stale i szybko rosnące zainteresowanie metodami empirycznymi w inżynierii językowej, co wymaga tworzenia dużych korpusów tekstów. Wysiłki w tym kierunku są podejmowane zarówno na świecie jak i w Polsce, ich celem jest umożliwienie szerokiego dostępu do mono- i wielojęzycznych zasobów językowych o odpowiedniej objętości i różnorodności. Przykładowe projekty, to angielski British National Corpus (BNC), a w Polsce działalność IPI PAN, gdzie powstał najobszerniejszy korpus języka polskiego (300 mln segmentów) oraz Uniwersytetu Łódzkiego, gdzie prowadzone są prace nad korpusem języka polskiego (tzw. „Polski Korpus Narodowy") oraz równoległym korpusem polsko-angielskim. Pomimo tego, zwłaszcza dla mniej popularnych języków wciąż daje się odczuć brak dostępnych materiałów korpusowych.

Elektroniczne, wstępnie przetworzone zasoby językowe znajdują zastosowanie nie tylko w badaniach naukowych, lecz także wśród szerokiego grona użytkowników słowników i encyklopedii internetowych oraz mono- i wielojęzycznych korpusów tekstów. Zwłaszcza dla uczących się języka dwujęzyczne, równoległe korpusy stanowią w ich studiach przydatne narzędzie. Często w roli korpusu, z braku innych materiałów wykorzystywane są zasoby Internetu, dostępne dzięki sieciowym wyszukiwarkom.

Pojęciem korpusu[2] zazwyczaj opisuje się zbiór danych lingwistycznych, zarówno pisanych i mówionych w jednym bądź wielu językach. Czasem nakłada się nań dodatkowe ograniczenia związane z jego reprezentatywnością i zrównoważeniem por. definicję EAGLES

[1] Za cel przyjęłyśmy możliwie popularną prezentację korpusów tekstów i ich zastosowań. Czytelników zaznajomionych z tym tematem zachęcamy do przejścia od razu do rozdziału 5. Projekt struktury korpusu polsko-ukraińskiego.

[2] Definicje w tym rozdziale za: Corpus Typology
http://webdeptos.uma.es/filifa/personal/amoreno/teaching/cl/corpus-typ.html

(Expert Advisory Group on Language Engineering Standards) „*korpus jest to zbiór fragmentów tekstów wybranych i uporządkowanych w oparciu o jawne kryteria lingwistyczne w celu jego użycia jako próbki języka”*. W naszym artykule będziemy stosować nazwę *korpus* do każdego zbioru tekstów, niezależnie od kryteriów przyświecających jego tworzeniu. Korpus może zatem zawierać prozę, artykuły prasowe, poezję, ale również transkrypcje wypowiedzi mówionych, instrukcje, dokumenty prawne i inne.

Na potrzeby korpusów komputerowych przyjmuje się następujące rozszerzenie definicji: korpus komputerowy to taki, który jest zapisany w standaryzowanej, homogenicznej postaci umożliwiającej automatyczne pozyskiwanie informacji.

Korpus równoległy (parallel) natomiast jest zbiorem tekstów, z których każdy jest przetłumaczony na jeden lub więcej języków. W najprostszym przypadku rozważamy jedynie parę języków. Tym niemniej korpusy równoległe mogą istnieć dla ich większej liczby, ponadto zmianie może ulegać kierunek tłumaczenia: niektóre dokumenty stworzone w języku A są tłumaczone na B (i kolejne języki), a inne napisane w B, tłumaczone na A.

Szczególnie interesujące w przypadku korpusów równoległych jest możliwość zestawienia oryginału i tłumaczeń, co pozwala na formułowanie i testowanie hipotez związanych z procesem translacji. Na takich korpusach można też sprawdzać i trenować systemy wspomagające automatyczne tłumaczenie.

Problemem w przypadku korpusów równoległych są trudności związane z pozyskaniem materiałów do nich. Duże ilości dwu- lub wielojęzycznych dokumentów powstają w państwach i organizacjach posługujących się różnymi językami, jak Unia Europejska czy Kanada. Niestety, dla mniej popularnych par języków liczba tłumaczeń dostępnych dla twórców korpusu równoległego jest zazwyczaj niewielka. W takich sytuacjach zapewnienie odpowiedniej ich liczby jest związane z dużym wysiłkiem organizacyjnym.

Często spotykane jest również pojęcie korpusu porównywalnego (comparable), rozumiane jako taki korpus, który gromadzi podobne teksty z dwu lub większej liczby języków. Co do natury tego podobieństwa nie ma jak dotąd szczegółowych standardów. Zastosowanie korpusów porównywalnych obejmuje analizę porównawczą różnych języków w podobnych sytuacjach komunikacyjnych pozwalając na uniknięcie zniekształcenia wprowadzonego przez tłumaczenie w korpusach równoległych.

2. Znakowanie.

Materiały dołączane do korpusów w większości zostały utworzone dla innych potrzeb, takich jak na przykład publikacja w gazetach czy magazynach. Takie surowe dane (primary data) nie są w żaden sposób oznakowane (anotowane) lingwistycznie, tzn. nie są pierwotnie wzbogacone o wyniki jakiegoś rodzaju analizy lingwistycznej.

Anotacja polega na konsekwentnym oznaczeniu w surowych danych informacji przydatnych do dalszych badań jak np.: obszernych jednostek strukturalnych (rozdziały, akapity wraz z ich tytułami, przypisami etc.), elementów mniejszych od akapitu (zdania, cytaty, nazwy własne, daty, skróty, etc) a także informacji morfosyntaktycznej, danych dotyczących uzgodnienia tekstów równoległych (alignment), prozodii, transkrypcji fonetycznej itd.

Znakowanie może i powinno być przeprowadzone na różnych poziomach (np. segmentacja na zdania i wyrazy, znakowanie morfosyntaktyczne i uzgodnienie tekstów równoległych), jednak rezultaty różnych poziomów anotacji najczęściej są przechowywane niezależnie od danych źródłowych i niezależnie od siebie.

Analiza i anotacja morfologiczna[3] polega na określeniu dla każdego występującego w tekście słowa wszystkich form wszystkich jednostek leksykalnych, których może być ono wykładnikiem. Wyniki tego etapu są zapisywane w formie znaczników morfosyntaktycznych. Dalszym etapem analizy jest ujednoznacznienie na podstawie kontekstu, którą z form faktycznie reprezentuje dane wystąpienie słowa.

Uzgodnienie równoległe to jeden z możliwych poziomów anotacji korpusu, pozwalający na zapisanie informacji o odpowiadających sobie jednostkach strukturalnych dokumentów równoległych. Jednostkami tymi mogą być całe akapity, poszczególne zdania a także pojedyncze wyrazy. Takie znakowanie pozwala później w czytelny sposób przedstawić oraz badać odpowiadające sobie fragmenty.

Anotowany morfosyntaktycznie i uzgodniony równolegle korpus wyposażony w program do tworzenia konkordancji równoległych może na wielu polach zastąpić szereg tradycyjnych słowników[4], mając nad nimi przewagę aktualności (ze względu na możliwość ograniczenia

[3] Opis za [Woliński 2003].
[4] Bliżej opisałyśmy to w [Kotsyba, Turska 2006].

wyszukiwania jedynie do tekstów współczesnych) i licznej bazy przykładów użycia. Pozwala to poradzić sobie z neologizmami, czy nowymi znaczeniami słów, których próżno szukać w słownikach.

Właściwie, przeglądając listę konkordancji, możemy dokonać analizy semantycznej całego hasła, przy czym dawną kartotekę hasła, która jest poprzedniczką konkordancji automatycznej, robi nam program komputerowy w nieporównywalnie krótszym czasie, z wysokim prawdopodobieństwem, że wszystkie możliwe znaczenia zostaną w niej ujęte. Równoległe konkordancje umożliwiają nam tworzenie listy wyrażeń będących tłumaczeniami hasła wejściowego w języku docelowym.

Konkordancje równoległe mają także „korzyści uboczne" w postaci twórczych wynalazków językowych tłumaczy – neologizmów, tworzonych i wprowadzanych do tekstów z konieczności (terminologia, frazeologizmy) lub trafnych opisów nowych pojęć (np. w przypadku nazw kulinarnych oraz innych realiów kulturowych). Częste są też sytuacje, kiedy w tłumaczeniach jest stosowany nie dokładny odpowiednik tłumaczeniowy, lecz jego hiperonim lub hiponim, np. ze względu na różnice kulturowe. Wymienione zamiany często pozytywnie wpływają na jakość tłumaczenia, ale nie odnajdziemy ich przykładów w słownikach.

3. Standard CES[5].

Ze względu na wymianę informacji i łączenie zasobów powstałych w ramach różnych projektów oraz możliwość stosowania ogólnodostępnych narzędzi komputerowych w prowadzeniu badań niezbędna jest standaryzacja korpusów. Wydaje się, że na dzień dzisiejszy standardem obowiązującym jest CES (i jego ciągle rozwijająca się wersja XCES).

CES określa minimalny poziom znakowania, który musi spełniać korpus żeby można go było uważać za standaryzowany. Standard ten pozwala m.in. na znakowanie informacji strukturalnych (podział na rozdziały, akapity, zdania, słowa), typograficznych, morfologicznych, informacji o uzgodnieniu językowym dokumentów równoległych (alignment). Możliwe jest też zapisanie w sformalizowanej postaci licznych metadanych dotyczących tekstu (autor, edytor, data powstania, zmiany w dokumencie). CES dostarcza sposobów na anotację lingwistyczną,

[5] Corpus Encoding Standard http://www.cs.vassar.edu/CES/CES1-0.html

dostosowaną do potrzeb konkretnych projektów. Różne poziomy znakowania są zapisywane niezależnie, co pozwala na dostosowanie poziomu anotacji do możliwości i potrzeb organizacji.

CES może być zastosowany do korpusów zarówno mono- jak i wielojęzycznych. Standard CES (Corpus Encoding Standard) jest częścią wytycznych EAGLES Guidelines stworzonych przez Expert Advisory Group on Language Engineering Standards (EAGLES). CES jest instancją języka SGML zgodną ze specyfikacją wytycznych Text Encoding Initiative *TEI Guidelines for Electronic Text Encoding and Interchange.*

4. Polsko-ukraińskie zasoby językowe.

Dla pary języków: polskiego i ukraińskiego dostępność zasobów lingwistycznych jest bardzo niska[6], zwłaszcza w przypadku źródeł elektronicznych. Jak dotąd nie istnieje żaden równoległy korpus polsko-ukraiński, natomiast powszechnie dostępne są wspomniane na wstępie korpusy języka polskiego, zwłaszcza korpus IPI PAN (www.korpus.pl). Trwają zaawansowane prace nad korpusem ukraińskim w Kijowie, w Ukraińskiej Akademii Nauk, jednak nie jest on jeszcze udostępniony szerszemu gronu odbiorców.

W tej sytuacji dotkliwie odczuwalny jest brak polsko-ukraińskich zasobów lingwistycznych, zwłaszcza w kwestii słowników i korpusu równoległego, które posłużyłyby lingwistom, tłumaczom i osobom uczącym się języka.

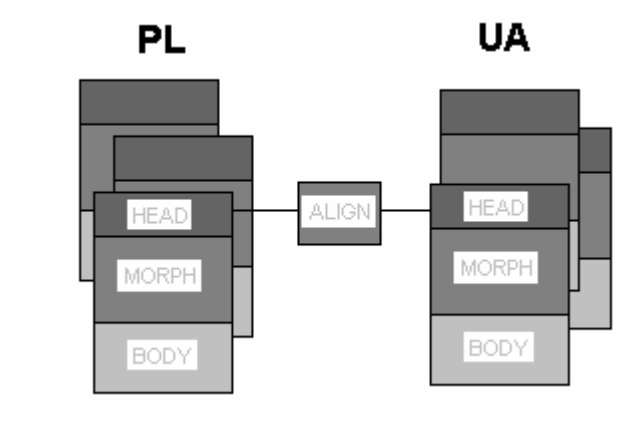
Jak się wydaje odpowiednio obszerny korpus równoległy mógłby znaleźć natychmiastowe zastosowanie jako narzędzie nauki dla studentów i pracy dla tłumaczy, zwłaszcza poprzez możliwość analizy równoległych konkordancji, a także stać się materiałem wyjściowym do analiz lingwistycznych po rozbudowaniu go o narzędzia do anotacji lingwistycznej i zaawansowanego przeszukiwania. Sądzimy także, że w oparciu o korpus równoległy możliwe jest skonstruowanie bardzo wysokiej jakości słowników odzwierciedlających bieżące zmiany w języku[7].

5. Projekt struktury korpusu.

[6] Szczegółowo opisałyśmy tę kwestię w [Kotsyba, Turska 2006].
[7] ibid.

Dokumenty składające się na korpus przechowywane są w postaci plików w formacie XCES. Każdy dokument jest złożony z części nagłówkowej (header) zawierającej metadane, faktycznej treści (body) oraz dodatkowych plików, w których zapisane jest znakowanie morfosyntaktyczne.

Dla każdego dokumentu korpusu równoległego istnieje odpowiadający mu dokument równoległy (co-text) będący jego tłumaczeniem lub oryginałem. Dodatkowo, na każdą parę tekstów równoległych (bitexts) przypada plik z informacją o uzgodnieniu.



Rysunek 1. Struktura korpusu równoległego

Każdemu z tekstów korpusu przypisany jest zawarty w pliku nagłówkowym zbiór metadanych obejmujących m.in. takie informacje jak dane autora, tłumacza, wydawcy tekstu. Dane te mogą zostać później użyte do zawężenia zakresu poszukiwań, np. w celu znalezienia wyłącznie tekstów utworzonych w pewnym okresie. Informacje te są istotne dla ewentualnych badań na materiale korpusowym, choć ich uzyskanie w wielu przypadkach jest trudne bądź nawet niemożliwe (np. dla krótkich notatek prasowych lub dokumentów prawnych).

Nagłówek

Metadane dotyczące autorów i tłumaczy obejmują: datę urodzenia, kraj i region pochodzenia, język ojczysty, ewentualnie inne języki, wykształcenie, miejsce stałego pobytu. Dane związane z samym dokumentem zawierają m.in.: datę utworzenia, pierwszej publikacji, medium, wydawcę, *genre*, docelowego czytelnika oraz informacje o autorze i tłumaczu.

Uzgodnienie równoległe (alignment)

Możliwość wyświetlania dokumentów równoległych w postaci wyrównanej (*aligned*) jest niezbędna dla pełnego wykorzystania korpusu równoległego.

AIESEC (międzynarodowe stowarzyszenie studentów ekonomii i zarządzania) to niepolityczna, naukowa i niekomercyjna organizacja, w całości zarządzana przez studentów. Działamy w ponad 950 szkołach wyższych. Do naszej organizacji należy 60.000 studentów z 84 krajów świata. Naszym celem jest rozwój jednostek i krajów pod względem gospodarczym, kulturalnym i moralnym. Szczególną uwagę przywiązujemy przy tym do międzynarodowego wzajemnego zrozumienia, jedności i współpracy.	AIESEC (міжнародна асоціація студентів, які займаються економією й управлінням) - неполітична, освітня і некомерційна організація, повністю керована студентами. Діємо в понад 950 вищих учбових закладах і об'єднуємо 60 тис. студентів з 84 країн світу. Наша мета - розвиток людей і країн в економічному, культурному і моральному аспектах. Зокрема, особливу увагу ми приділяємо міжнародному взаємозрозумінню, єдності і співробітництву.

Rysunek 2. Dokument uzgodniony na poziomie akapitów

Wyrównanie na poziomie akapitów jest wykorzystywane stosunkowo rzadko, głównie ze względu na zazwyczaj dużą objętość paragrafu. Najbardziej przydatne dla tłumaczy i studentów jest uzgodnienie na poziomie zdań. Należy zwrócić uwagę, że jedno zdanie oryginału może zostać zamienione na dwa lub więcej podczas tłumaczenia i odwrotnie (jak w drugim akapicie naszego przykładu).

AIESEC (międzynarodowe stowarzyszenie studentów ekonomii i zarządzania) to niepolityczna, naukowa i niekomercyjna organizacja, w całości zarządzana przez studentów.	AIESEC (міжнародна асоціація студентів, які займаються економією й управлінням) - неполітична, освітня і некомерційна організація, повністю керована студентами.

Działamy w ponad 950 szkołach wyższych. Do naszej organizacji należy 60.000 studentów z 84 krajów świata. Naszym celem jest rozwój jednostek i krajów pod względem gospodarczym, kulturalnym i moralnym.	Діємо в понад 950 вищих учбових закладах і об'єднуємо 60 тис. студентів з 84 країн світу. Наша мета - розвиток людей і країн в економічному, культурному і моральному аспектах.

Rysunek 3. Ten sam dokument uzgodniony na poziomie zdań.

```
<cesAlign>
  <cesHeader>
...
...
    <translations xml:base="http://corpus.domeczek.pl/corpus">
      <translation trans.loc="exampleAna.ua.xml" lang="ua" xml:lang="ua" n="1" />
      <translation trans.loc="exampleAna.pl.xml" lang="pl" xml:lang="pl" n="2" />
    </translations>
    </profileDesc>
  </cesHeader>

  <linkList>
    <linkGrp id="p1" targType="s">
      <link>
        <align xlink:href="#p1s1" />
        <align xlink:href="#p1s1" />
      </link>
      <link>
        <align xlink:href="#p1s2" />
        <align xlink:href="#p1s2" />
      </link>
    </linkGrp>
    <linkGrp id="p2" targType="s">
      <link>
        <align xlink:href="#xpointer(id('p2s1')/range-to(id('p2s2')))" />
        <align xlink:href="#p2s1" />
```

```
    </link>
    <link>
      <align xlink:href="#p2s3" />
      <align xlink:href="#p2s2" />
    </link>
  </linkGrp>
</linkList>
</cesAlign>
```

Rysunek 4. Fragment pliku z informacją o uzgodnieniu (zdania 1 i 2 z drugiego akapitu są przetłumaczone jako jedno zdanie w dokumencie równoległym).

Anotacja morfologiczna

Drugim warunkiem koniecznym dla sprawnego funkcjonowania korpusu równoległego jest jego anotacja morfologiczna, opisana krótko w punkcie 2, która umożliwia wyszukiwanie według złożonych kryteriów.

Nasz sposób anotacji morfologicznej oparłyśmy na istniejącym korpusie języka polskiego IPI PAN[8]. Niezbędne jednak było rozszerzenie systemu znaczników o informację morfologiczną charakterystyczną dla języka ukraińskiego, zgodnie z rozwiązaniami przyjętymi w Ukraińskiej Akademii Nauk[9].

Znaczniki morfosyntaktyczne są zapisywane jako ciągi wartości: klasy fleksemu i kategorii gramatycznych adekwatnych w przypadku danej formy, np. dla *jechać* będzie to *fin:pl:sec:imperf*. Jeżeli dla danego segmentu występuje niejednoznaczność podawane jest kilka znaczników.

```
<chunk type="p" xlink:href="#p5">
<chunk type="s">
   <tok>
     <orth>dokąd</orth>
     <lex disamb="1">
      <base>dokąd</base>
```

[8] Zob. [Woliński 2003].

[9] www.ulif.org.ua

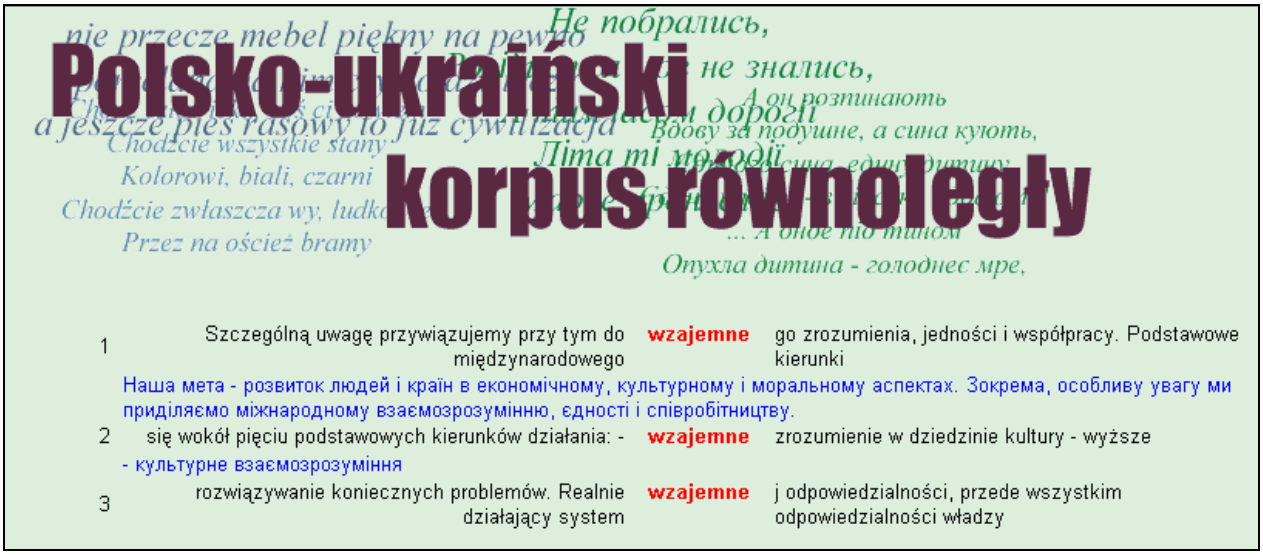
```
      <ctag>qub</ctag>
    </lex>
  </tok>
  <tok>
    <orth>jedziecie</orth>
    <lex disamb="1">
      <base>jechać</base>
      <ctag>fin:pl:sec:imperf</ctag>
    </lex>
  </tok>
  <tok>
    <orth>?</orth>
    <lex disamb="1">
      <base>?</base>
      <ctag>interp</ctag>
    </lex>
  </tok>
</chunk>
</chunk>
```

Rysunek 5. Anotacja morfologiczna fragmentu tekstu.

6. Strategia rozwoju korpusu.

Obecnie z korpusu polsko – ukraińskiego można korzystać poprzez stronę internetową http://corpus.domeczek.pl. Zawiera on ok. 50 dokumentów równoległych zapisanych w standardzie XCES, jednak nie są jeszcze wydzielone segmenty zdaniowe, brak także oznakowania morfologicznego tych tekstów oraz ich uzgodnienia równoległego.

Nowe dokumenty są dołączane do korpusu w miarę ich pozyskiwania, możliwe jest ich przeszukiwanie w prosty sposób oparty na mechanizmie wyrażeń regularnych. Wyniki prezentowane są w postaci konkordancji równoległych uzgodnionych na poziomie akapitów.



Rysunek 6. Wynik wyszukiwania w korpusie w postaci równoległych konkordancji

Prace nad korpusem można podzielić na dwie ścieżki: organizacyjną, której celem jest przede wszystkim zapewnienie współpracy organizacji i osób, będących w posiadaniu tekstów równoległych, które chciałybyśmy włączyć do korpusu i ścieżkę techniczną, której zadaniem jest zapewnienie segmentacji i alignmentu dokumentów na poziomie zdań oraz anotacja morfologiczna korpusu. Dalszy etap prac na ścieżce technicznej to przystosowanie lub stworzenie od podstaw narzędzi do pracy z korpusem równoległym.

Polską część korpusu zamierzamy znakować morfologicznie za pomocą analizatora autorstwa Marcina Wolińskiego [Woliński, 2003], część ukraińską za pomocą analizatora morfologicznego opracowanego przez zespół W. Szyrokowa w Ukraińskim Funduszu Lingwistyczno-Informacyjnym Ukraińskiej Akademii Nauk.

Mamy też nadzieję zorganizować forum dyskusyjne, które może z czasem stać się punktem wymiany idei dla tłumaczy, studentów i badaczy. Liczymy na to, że efektem ubocznym tej działalności będzie stały przyrost nowych materiałów dla korpusu.

Uważamy, że korpus równoległy z wydajną wyszukiwarką, elastycznym językiem zapytań oraz przyjaznym dla użytkownika, czytelnym konkordanserem równoległym będzie jednym z najbardziej pożytecznych zasobów dla wszystkich polskich i ukraińskich badaczy, tłumaczy i studentów.

Bibliografia

Corpus Encoding Standard: http://www.cs.vassar.edu/CES/CES1-0.html

Corpus Typology: http://webdeptos.uma.es/filifa/personal/amoreno/teaching/cl/corpus-typ.html

EAGLES Guidelines: http://www.ilc.pi.cnr.it/EAGLES96/browse.html

Korpus IPI PAN: http://korpus.pl

Korpus PWN: korpus.pwn.pl/

Korpus referencyjny języka polskiego PELCRA: http://korpus.ia.uni.lodz.pl/

Kotsyba, Natalia i Turska, Magdalena (2006): Leksykografia polsko-ukraińska – stan obecny i perspektywy. - [w:] Semantyka i konfrontacja językowa, t. III. Warszawa, SOW.

PolUKR – Polsko-Ukraiński Korpus Równoległy: http://corpus.domeczek.pl

Przepiórkowski, Adam (2004): Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version: http://dach.ipipan.waw.pl/~adamp/Papers/2004-corpus/

TEI Guidelines for Electronic Text Encoding and Interchange: http://etext.virginia.edu/TEI.html

Ukraiński Lingwistyczny Portal: www.ulif.org.ua

Woliński, Marcin (2003): System znaczników morfosyntaktycznych w korpusie IPI PAN. - [w:] Polonica XXII-XXIII, s. 39-55

Широков В.А (1998): Інформаційна теорія лексикографічних систем. - К.: Довіра.

Широков В.А et al. (2004): Технічна документація на програмні продукти та бази даних (dokument wewnętrzny ULIF).