

Korpusy językowe – podstawowa terminologia i metody tworzenia

Natalia Kotsyba

IBI AL Uniwersytet Warszawski

12 i 26 stycznia 2011 r.

Czym jest **korpus**?

- Zbiór tekstów albo zapisanych wypowiedzi, wykorzystywany dla celów analizy lingwistycznej

(The American Heritage® Dictionary of the English Language)

„ciało”

Teraz mówimy o korpusach wyłącznie
elektronicznych

Czy można mówić o **lingwistyce** **korpusowej** tak, jak mówimy o konfrontatywnej, kognitywnej itd.?

- Ma dwie strony: tworzenie korpusów i ich użytkowanie
- Jako użytkowanie jest raczej metodologią, a nie osobną branżą lingwistyki
- Korpusy pomagają językoznawcom obiektywnie popatrzeć na dane lingwistyczne, ale:
- tylko pod warunkiem, że korpus jest wystarczająco duży i reprezentatywny – w pozostałych przypadkach musimy liczyć się z konsekwencjami ograniczeń (świeży przykład: SUM-20)
- Kiedy piszemy prace oparte na własnych korpusach, a nie np. narodowych, konieczny jest opis takiego korpusu
- LK w obu postaciach dopiero się rozwija, więc często brak standardów i gotowych rozwiązań
- Rażąco duża różnica w rozwoju dla różnych języków
- Podobna skala różnicy w ilości dostępnych kwalifikowalnych do korpusów tekstów w różnych językach
- Dużym problemem dla lingwistów są błędy znakowania w korpusach – o tym też warto wspominać w pracach

Rodzaje korpusów

- małe (do 10 mln. wyrazów) – średnie (10 do 50 mln.) – duże (> 50 mln.) klasyfikacja NK 😊
- „surowe/łyste” – adnotowane/znakowane
- jednojęzyczne - dwu-/wielojęzyczne
- równoległe - porównawcze
- korpus referencyjny (*benchmark korpus*): BNC, NKJP, COCA
- korpus zrównoważony (wciąż brak ustalonych norm)
- specjalistyczny
- liczące się z międzynarodowymi standardami, praktykami i rekomendacjami – „odjechane” (granice z przyczyn obiektywnych są obecnie wciąż rozmyte)

Jak tworzymy korpusy?

- Dobieramy obszerny zbiór tekstów reprezentujących różne style językowe

(np. artykuł, notatka prasowa, korespondencja, dokumentacja techniczna, literatura piękna)

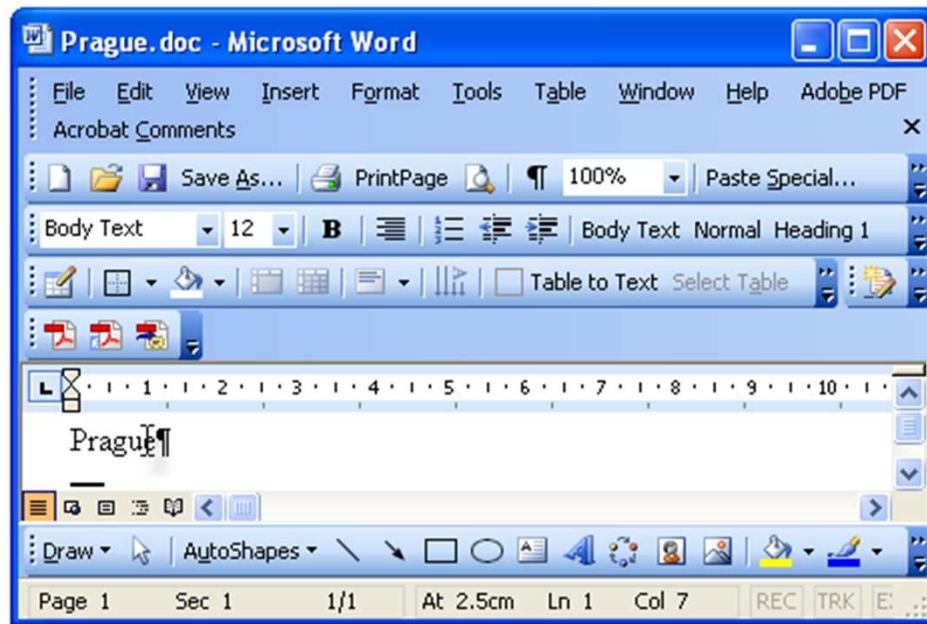
- Znakujemy „surowiec” za pomocą odpowiednich narzędzi

(lematyzator, analizator morfologiczny, aligner)

Wstępne opracowanie tekstów

- zbieranie tekstów
- jeżeli są w postaci papierowej, to:
- skanowanie, czyszczenie skanów od obrazków i czarnych pasków, numeracji stron;
- OCRowanie, redakcja poOCRowa;
- konwertowanie do tekstowego formatu (FineReader --> Word); wyrównywanie kolumn, wyczyszczenie od znaków podziałów stron, przypisów dolnych, znaków przeniesienia wyrazów;
- jeszcze raz wyczytywanie błędów (FineReader wyświetla tylko miejsca, które są "podejrzane" dla niego samego);
- sprawdzenie struktury (podział na akapity, zdania);
- konwertowanie czcionek do wspólnego formatu (przyjęliśmy Unicode/UTF-8);
- konwertowanie z MSWord .doc na .txt
- dodawanie metadanych
- konwertowanie z .txt do .xml
- dodawanie do serwera

Kodowanie w różnych formatach



Prague.doc – 26,064 bytes

Prague.xml = 8,534 bytes

Prague.rtf = 7,893 bytes

Prague.htm = 7,272 bytes

Prague.txt = 8 bytes

Źródło: Mike Scott, „Corpus Linguistic Processing Problems”,
www.lexically.net/downloads/corpus_linguistics

Narzędzia do tworzenia i przeszukiwania korpusów (**menadżerzy** korpusowe)

- Multilingual Corpus Tool kit (Scott Piao)
- Xaira 1.23 (Beta wersja 1.24), BNC, Oxford
- AntConc
- Program „Kolokacje” (A. Buczyński i in.)
- WordSmith
- i wiele innych...

Te programy działają także na korpusach
nieznakowanych

Znakowanie – terminologia

- Jest to wzbogacanie tekstów o informacje, np. lingwistyczne
- Znacznik (tag, ang. ogonek) – zestaw tagów (tagset) – znakować (to tag, poznać) – znakowanie (tagging) – program do znakowania (tagger)
- Także: anotacja, adnotacja, tagowanie, ros. разметка, ale тэг i тэгсэт

Znakowanie – technologia

- Standardem jest znakowanie korpusów w języku XML
- Specjalne moduły tego języka, napisane z myślą o znakowanie tekstów, nie tylko korpusowych, to TEI (Text Encoding Initiative) – zawiera on zdefiniowane znaczniki przewidziane dla różnych potrzeb
- Specjalnie dla korpusów także jest zdefiniowany format XCES, także w ramach XML
- Standard nie jest obowiązkowy, ale wykorzystanie zaleceń umożliwia wzajemne „porozumiewanie się” pomiędzy korpusami

BNC i wcześniejsze korpusy

- Przykład formatu pionowego:

<s>

każdego każdy KOD

wyrazu wyraz KOD

...

</s>

Poziomy znakowania

- struktura: tekst → rozdział → akapit → zdanie → ostatnio też fraza (*chunk*) → wyraz → *segmenty/morfemy;
- osobno – metadane (tytuł, autor, rok napisania, rok i miejsce wydania, gatunek literacki, itd. – zgodnie z zaleceniami TEI);
- właściwości poszczególnych jednostek struktury, dla lingwistów jest to przede wszystkim wyraz (*token*)

Typy znakowania dla wyrazów i fraz

- Lemat (od gr. *λήμμα* „cegiełka”, forma podstawowa, hasłowa)
- Charakterystyka gramatyczna
 - morfologiczna: część mowy (POS) oraz atrybuty i ich wartości
 - składniowa: składnia powierzchniowa albo głęboka; chunking
 - morfoskładniowa: termin umowny ale praktyczny
- Semantyczna
- Referencje (zaimki – rzeczowniki)
- Nazwy własne (named entities)
- itd. w zależności od specyficznych zadań badawczych

Polski tekst nietagowany

- W dzisiejszym posiedzeniu komisji uczestniczy ekspert komisji pan profesor Jan Gajewski.

morphSh.xml - XML Copy Editor

File Edit View Insert XML Tools Help

ulif2.dtd x ulif1.dtd x morphSh.xml x

1 <?xml version="1.0" encoding="UTF-8" ?>

2 <!DOCTYPE cesAna SYSTEM "cesAnaIPI.dtd">

3 <cesAna xmlns:xlink="http://www.w3.org/1999/xlink" type="pre_morph" version="IPI-1.2">

4 <chunkList xml:base="text.xml">

5 <chunk type="p" xlink:href="#dv1hd1">

6 <chunk type="s">

7 <group id="a1" rule="Uncertain: Dobrze PrepNG (z postmodyfikatorem dopełniaczowym) na koncu zdania lub nawiasu, lub przed czasownikiem" synh="a1" semh="a1">

8 <tok id="a1">

9 <orth>W</orth>

10 <lex disamb_sh="0"><base>w</base><ctag>prep:acc:nwok</ctag></lex>

11 <lex disamb="1"><base>w</base><ctag>prep:loc:nwok</ctag></lex>

12 </tok>

13 <tok id="a2">

14 <orth>dzisiejszym</orth>

15 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:m1:pos</ctag></lex>

16 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:m2:pos</ctag></lex>

17 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:m3:pos</ctag></lex>

18 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:n:pos</ctag></lex>

19 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:loc:m1:pos</ctag></lex>

20 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:loc:m2:pos</ctag></lex>

21 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:loc:m3:pos</ctag></lex>

22 <lex disamb="1"><base>dzisiejszy</base><ctag>adj:sg:loc:n:pos</ctag></lex>

23 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:m1:pos</ctag></lex>

24 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:m2:pos</ctag></lex>

25 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:m3:pos</ctag></lex>

26 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:f:pos</ctag></lex>

27 <lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:n:pos</ctag></lex>

28 </tok>

29 <tok id="a3">

30 <orth>posiedzeniu</orth>

31 <lex disamb_sh="0"><base>posiedzenie</base><ctag>subst:sg:dat:n</ctag></lex>

32 <lex disamb="1"><base>posiedzenie</base><ctag>subst:sg:loc:n</ctag></lex>

33 <lex disamb_sh="0"><base>posiedzieć</base><ctag>ger:sg:dat:n:perf:aff</ctag></lex>

34 <lex disamb="1"><base>posiedzieć</base><ctag>ger:sg:loc:n:perf:aff</ctag></lex>

35 </tok>

36 <tok id="a4">

37 <orth>komisji</orth>

38 <lex disamb="1"><base>komisja</base><ctag>subst:sg:gen:f</ctag></lex>

39 <lex disamb_sh="0"><base>komisja</base><ctag>subst:sg:dat:f</ctag></lex>

40 <lex disamb_sh="0"><base>komisja</base><ctag>subst:sg:loc:f</ctag></lex>

41 <lex><base>komisja</base><ctag>subst:pl:gen:f</ctag></lex>

42 </tok>

Current Element

Insert Element

Insert Sibling

Insert Entity

amp

apos

gt

lt

quot

Ukraiński tekst nietagowany

- Львів розташований на етнічних українських землях і є одним з головних нервових вузлів українського народу, найважливішим клапаном його серця, вічним збудником честолюбства, гордості й потягу до волі.

Ukraiński: tagi wejściowe z lematami

Львів<**JDAJJK**><Львів 0|Львів 0|Лев 1|Лев 1> **розташований**<**BDBAV?**><розташований
0|розташований 0|розташувати 0> **на**<**NONOZOPF**><на 4|на 3|на 2|на 1>
етнічних<**AVATAH**><етнічний 0|етнічний 0|етнічний 0>
українських<**AVATAXJIGIKGKJMGM**><український 0|український 0|український
0|Український 0|Український 0|Український 0> **землях**<**FM**><земля 2> **i**<**SSSCNOZO**><i
1|i 3|i 2> **є**<**UPUOUNUKUMUL**><бути 0|бути 0|бути 0|бути 0|бути 0|бути 0>
одним<**HUNQHERQRERU**><один 0|один 0|один 0|один 0|один 0|один 0> **з**<**PE**><z 0>
головних<**AVATAH**><головний 0|головний 0|головний 0> **нервових**<**AVATAH**><нервовий
0|нервовий 0|нервовий 0> **вузлів**<**MIMI**><вузол 2|вузол 1>
українського<**ANADABJDJBKB**><український 0|український 0|український 0|Український
0|Український 0|Українське 0> **народу**<**MBMCMBMS**><народ 0|народ 0|нарід 0|нарід
0>, **найважливішим**<**AQAEAU**><найважливіший 0|найважливіший 0|найважливіший
0> **клапаном**<**ME**><клапан 0> **його**<**FGODOBODOB**><йога 0|воно 0|воно 0|він 0|він
0> **серця**<**NKNHNBN**><серце 0|серце 0|серце 0|серце 0>, **вічним**<**AQAEAU**><вічний
0|вічний 0|вічний 0> **збудником**<**MEME**><збудник 1|збудник 2>
честолюбства<**NB**><честолюбство 0>, **гордості**<**FCFBFF**><гордість 0|гордість 0|гордість
0> **й**<**SSSCZO**><y 1|y 2> **потягу**<**MFMCMGMBMCMFMGFDDGD**><потяг 2|потяг 2|потяг
2|потяг 1|потяг 1|потяг 1|потяг 1|потяга 0|Потяга 0>
до<**NGNFNENDNCNBNNHHNNINJNKLNMMNNPB**><до 2|до 2|до 2|до 2|до 2|до 2|до 2|до
2|до 2|до 2|до 2|до 2|до 2|до 1>
волі<**UOFCEBFEGCBGEFGCGBGHGNNFMF**><воліти 0|воля 0|воля 0|воля 0|Воля
2|Воля 2|Воля 2|Воля 1|Воля 1|Воля 1|Воля 1|Воля 1|воло 0|віл 0>

Przykłady kodów gramatycznych ULIF (384)

Znaczenie gramatyczne	Kod (tag)	Przykład
Czasownik, bezokolicznik, aspekt dokonany, strona czynna	VA	<i>прочитати</i>
Imiesłów przymiotnikowy, rodzaj męski, liczba pojedyncza, mianownik, aspekt dokonany, czas przeszły, czynny	BA	<i>зрослий</i>
Przymiotnik niezmienny	AZ	<i>ультра</i>
Rzeczownik, ogólny, rodzaj żeński, liczba pojedyncza, celownik	FC	<i>квітці</i>
Predykatyw (słowo orzeczeniowe)	X0	<i>слід</i>



10 Cyrilica B I U

1 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

```
.0" encoding="UTF-8"?>
SYSTEM "xcesAnaIPI.dtd">
href="display.css" type="text/css"?>
```

```
іковані</orth>
іковані</base><ctag>adj:pl:nom</ctag></lex>
іковані</base><ctag>adj:pl:acc</ctag></lex>
```

```
кінці</orth>
кінці</base><ctag>adv</ctag></lex>
```

```
оро</orth>
оро</base><ctag>adj:sg:gen:masc</ctag></lex>
оро</base><ctag>adj:sg:acc:masc</ctag></lex>
оро</base><ctag>adj:sg:gen:n</ctag></lex>
```

```
</orth>
</base><ctag>gnoun:sg:gen:masc</ctag></lex>
```

```
th>
se<ctag>int</ctag></lex>
se<ctag>prep</ctag></lex>
se<ctag>prep</ctag></lex>
se<ctag>prep</ctag></lex>
```

```
cial</orth>
cial</base><ctag>ign</ctag></lex>
```

```
/orth>
/base><ctag>ign</ctag></lex>
```

```
вачення</orth>
вачення</base><ctag>gnoun:sg:nom:n</ctag></lex>
вачення</base><ctag>gnoun:sg:gen:n</ctag></lex>
вачення</base><ctag>gnoun:sg:acc:n</ctag></lex>
вачення</base><ctag>gnoun:sg:voc:n</ctag></lex>
вачення</base><ctag>gnoun:pl:nom:n</ctag></lex>
вачення</base><ctag>gnoun:pl:acc:n</ctag></lex>
```

УКРАЇНСЬКИЙ ЛІНГВІСТИЧНИЙ ПОРТАЛ

НОВИНИ

ПРО НАС

ПРОЕКТИ

КНИЖКОВА ПОЛИЦЯ

ON-LINE РЕСУРСИ

КАРТА САЙТУ

МОВНІ ОРГАНІЗАЦІЇ

“Словники України” on-line

Словозміна Синонімія Фразеологія

бути

пошук

Реєстр

[бутербродик](#)
[бутербродний](#)
[Бутефліка](#)
[Бути](#)
[бути](#)
[Бутівля](#)
[бутил](#)
[бутилацетат](#)
[бутилен](#)
[бутилені](#)
[бутилкаучук](#)
[бутиловий](#)
[бутиль](#)
[Бутин](#)
[бутина](#)
[Бутини](#)
[бутинський](#)
[Бутир](#)
[бутират](#)
[бутирометр](#)
[бутирофенон](#)
[бутири](#)
[бутифос](#)
[Бутівка](#)
[Бутівське](#)

бути – дієслово недоконаного виду

Інфінітив	бути	
	однина	множина
Наказовий спосіб		
1 особа		бу́дьмо
2 особа	будь	бу́дьте
МАЙБУТНІЙ ЧАС		
1 особа	бу́ду	бу́демо, бу́де́м
2 особа	бу́деш	бу́дете
3 особа	бу́де	бу́дуть
ТЕПЕРІШНІЙ ЧАС		
1 особа	є	є
2 особа	є, еси́	є
3 особа	є	суть, є
Активний дієприкметник		
Дієприслівник		
бу́дучи		
МИНУЛИЙ ЧАС		
чол. р.	був	були́
жін. р.	була́	
сер. р.	було́	
Активний дієприкметник		
Пасивний дієприкметник		
Безособова форма		
Дієприслівник		
бу́вши		

Lematyzator

- przyporządkowuje formy podstawowe każdej innej w tekście
każdej -> każdy
- zawiera informacje o formach podstawowych wyrazów, o ile posiada słownik;
- nie zawiera opisu gramatycznego form wyrazów, ale może dokonywać częściowej analizy morfologicznej w oparciu o heurystyki (rozumowania)
- jest przydatny w przypadku braku analizatora morfologicznego, wyniki wymagają weryfikacji

Narzędzia do znakowania morfoskładniowego – dane

- Analizator morfologiczny (słownik gramatyczny)

wyznacza formę lub formy (w przypadku homonimii)
podstawowe słowa oraz jego charakterystyki gramatyczne na
podstawie bazy słownikowej oraz reguł gramatycznych

- **miał**

1. „mieć”, czasownik, 3os lp, cz. przeszły, tr. orzekający...
2. „miał”, rzeczownik, mianownik lp. r.męski...

Istniejące analizatory j. polskiego

- MORFEUSZ, Marcin Woliński et al., IPI PAN
(ciągle rozwijany, tryb wsadowy)
- MORFOLOGIK, Marcin Miłkowski, IF PAN
- Krzysztof Szafran, MIM UW
- M. Gajęcki et al. , AGH

Zastosowanie dla automatycznego poprawiania
błędów, tzw. *spell-checkers*

Finalizacja znakowania – ujednoznacznianie

- Po przypisaniu za pomocą słowników wszystkich możliwych wartości do danego wyrazu musimy wybrać wartość właściwą – ta czynność nazywa się **ujednoznacznianiem** albo **dezambiguacją** (*disambiguation*, por. *ambiguous* – niejednoznaczny)
- Ujednoznacznienie może być ręczne albo automatyczne
- Automatyczne ujednoznaczniacze są pisane na podstawie reguł, ale częściej są trenowane na ręcznie ujednoznacznionych fragmentach korpusów, np. 1 mln.; czasami metody są łączone
- Żadna z metod nie daje absolutnie dobrego rezultatu – najwyższy osiągnięty poziom poprawności dla języków słowiańskich to ok. 95 %
- Poziom błędów i strategie dezambiguacji zależą od zdefiniowanego tagsetu – im bardziej szczegółowy, tym więcej potencjalnych błędów
- Ujednoznaczniać trzeba także lematy (typowy błąd TAKIPI to „miał” jako rzeczownik w przypadku czasownika) oraz
- znaczenia (trudny obszar, ale też bardzo uzależniony od słownika referencyjnego)
- a także segmentację (por. gdzieś)

Narzędzia do znakowania morfoskładniowego – programy

- TAKIPI : lematyzator, morfologiczny analizator, ujednoznaczniaacz (dezambiguator), chunker
IPI PAN + Politechnika Wrocławska, ok. 200 tys. słów (Marcin Woliński, Adam Radziszewski i in.)
- SPEJD (IPI PAN): częściowa składnia powierzchniowa (*shallow parsing*)
- **Web Services:** obecnie funkcjonuje nowa wersja TAKIPI PWr; w perspektywie narzędzia IPI PAN (Maciej Ogrodniczuk)

Gdzie można znaleźć informację o korpusach i ich zastosowaniu?

- korpusowa lista mailingowa przy Uniwersytecie w Oslo (corpora mailing list)

<http://mailman.uib.no/listinfo/corpora>

- opis korpusowych resursów w świecie

[http://aclweb.org/aclwiki/index.php?title=List of resources by language](http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language)

- polska lingwistyczna lista mailingowa

<http://bach.ipipan.waw.pl/mailman/listinfo/ling>

Demonstracja

- TEI
- MORFEUSZ
- TAKIPI
- MULTEXT-East