Językoznawcze studia doktoranckie IJP PAN 2015/2016

Leksykografia elektroniczna

Praktyki, metody, narzędzia

Krzysztof Nowak

Pracownia Łaciny Średniowiecznej

Instytut Języka Polskiego PAN

Plan wykładu

- **korpus w pracy leksykografia**
  - po co korpus?
  - jaki korpus?
- **narzędzia korpusowe**
  - prezentacja
  - przeszukiwanie
- **teoria i metody**
- **laboratorium 1**
- **narzędzia leksykograficzne**
  - od surowego korpusu do hasła (SketchEngine etc.)
  - systemy redakcji słowników
- **laboratorium 2**

Pytanie

Co zmieniły korpusy w opisie językoznawczym *tout court*?

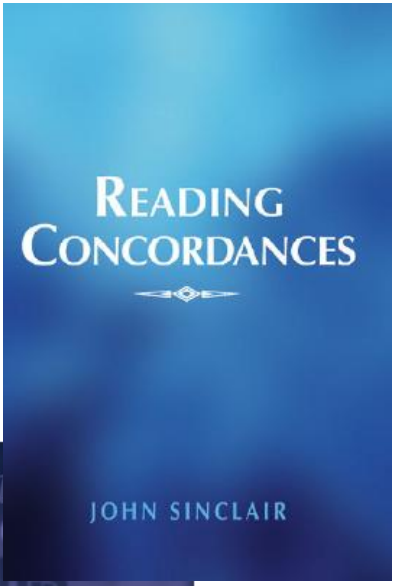
Po co korpus?

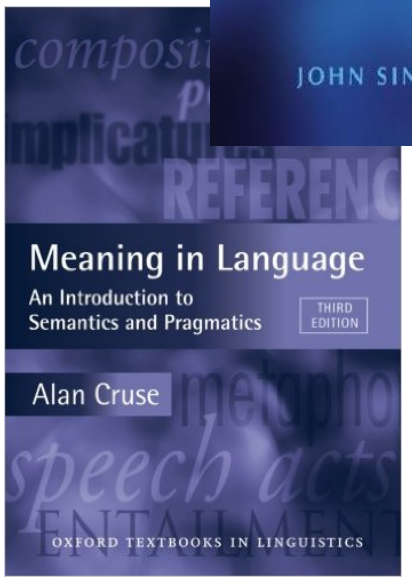
Zalety danych korpusowych

- obiektywizacja obserwacji (frekwencja wyrazów, znaczeń, zjawisk)
- śledzenie wzorców składniowych, pragmatycznych
- lepsze ujęcie frazeologii

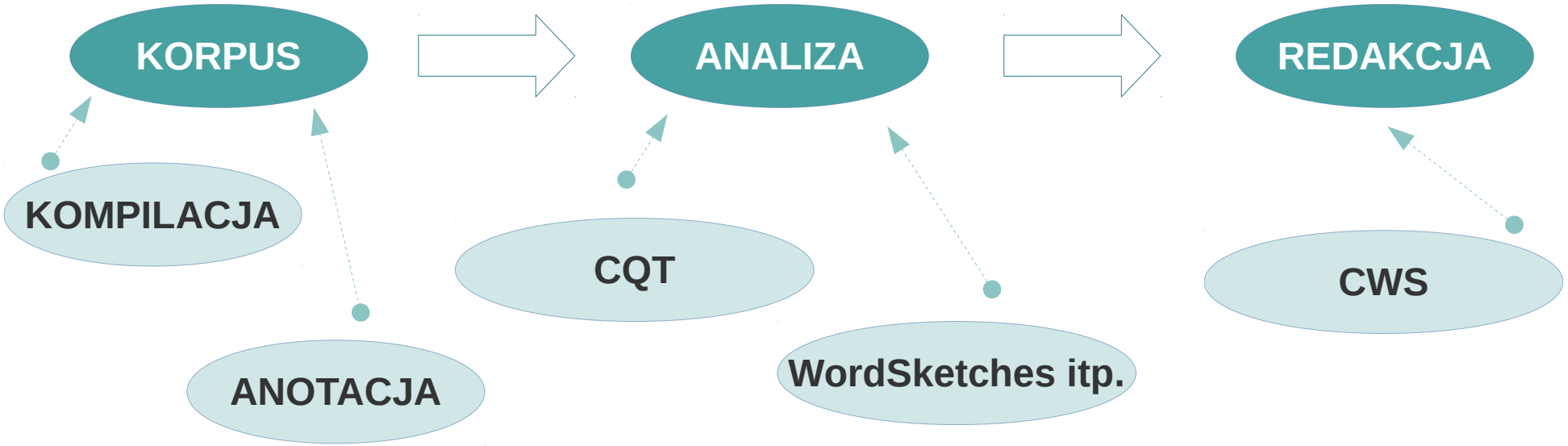
Jaka teoria?

- rola kontekstu w opisie znaczenia (Kilgariff: „I don't believe in word senses")
  - Patrick **Hanks**, projekt *Corpus Pattern Analysis*: wyrazy aktywują potencjał znaczeniowy w kontekście
  - John **Sinclair**, „Reading Concordances"
  - Alan **Cruse**





Procedura


KORPUS
ANALIZA
REDAKCJA
KOMPILACJA
ANOTACJA
CQT
WordSketches itp.
CWS

Jaki korpus dla leksykografii? (1)

jaki korpus dla leksykografii?

nie istnieje korpus do wszystkiego (inny dla języka ogólnego, inny dla słownika dokumentacyjnego)

reprezentatywność

słownik ogólny

język

zróżnicowanie regionalne i dialektalne

czas

synchronia ≠ diachronia

język mówiony czy pisany?

kanał

Sieć ≠ gazeta ≠ książka ≠ czat

domena

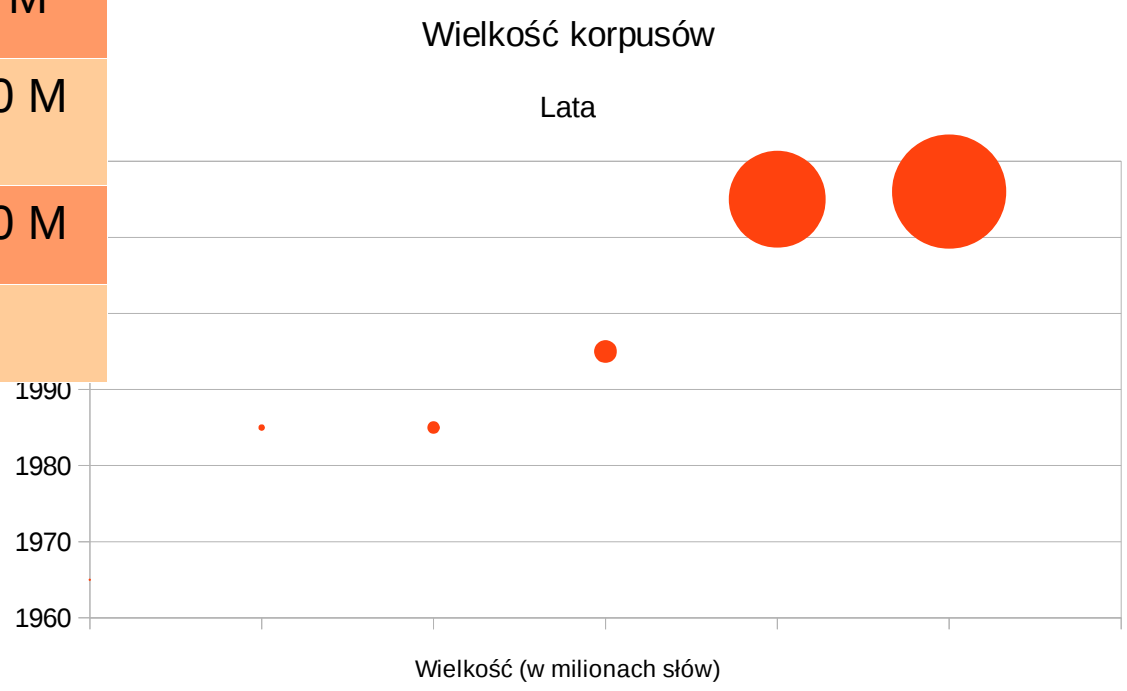
słownictwo specjalistyczne?

korpus równoległy?

leksykografia dwujęzyczna

Jaki korpus dla leksykografii? (2)

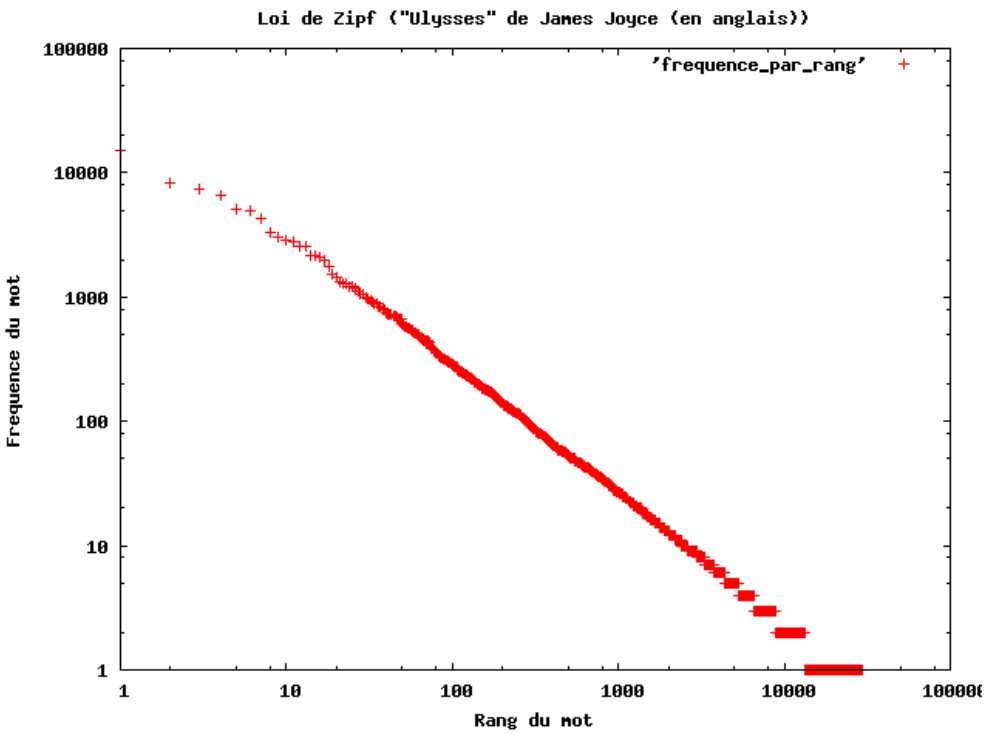
Nazwa korpusu	Wielkość
Brown / LOB Corpus	1 M
Birmingham	8 M
British National Corpus	100 M
Narodowy Korpus Języka Polskiego	1 800 M
Oxford English Corpus	2 500 M
Web as Corpus	?



Jaki korpus dla leksykografii? (3)

Prawo Zipfa…

$$frekwencja \approx \frac{constans}{ranking}$$



… i jego konsekwencje

- mały korpus daje względne pojęcie o słowach o dużej frekwencji
- opis np. 100 k najpopularniejszych słów wymaga już bardzo dużego korpusu
- podobnie jak wyrazy (nawet o dużej frekwencji) zachowują się ich rzadsze znaczenia i użycia

Przygotowanie korpusu

kodowanie znaków

ąęćłźżóśń

metadane

<tytul>Washington Post</tytul>
<rok>1960</rok>
<kanal>prasa<kanal>

typografia

Lorem ipsum dolor sit amet.

tokenizacja

Litwo|ojczyzno|moja|!|

lemmatyzacja

Litwa|ojczyzna|mój|!|

PoS

PROP|NOUN|PRON|INT|

inne

Narzędzia korpusowe

Klasyfikacja (Kilgariff & Kosem)

- lokalne ≠ sieciowe
- związane z określonym korpusem ≠ niezależne
- gotowy korpus ≠ Web as corpus
- proste ≠ zaawansowane

Wybrane korpusy wyposażone w interfejs sieciowy

- NKJP
- Cobuild's Bank of English
- korpusy Institut fur Deutsche Sprache (Mannheim)

Podstawowe funkcje narzędzi korpusowych

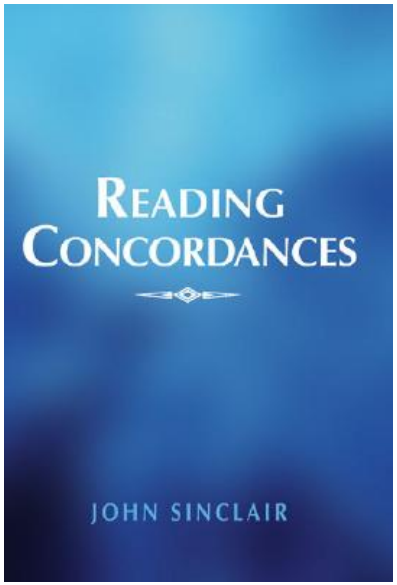
- możliwość ładowania własnego korpusu
- możliwość współdzielenia
- indeksacja
- obsługa anotacji
- obsługa kodowania

Podstawowe *modi*

- listy frekwencyjne
  - kandydaci na wyrazy hasłowe
  - filtrowanie PoS
  - filtrowanie wg metadanych (np. domen)
- listy kolokacji (wg miar asocjacji):
  - znaczenie (Firth)
  - stabilne połączenia
- konkordancja
  - KWIC
  - wyszukiwanie *wildcards*
  - dopasowanie wyświetlania
  - sortowanie wg kontekstu, typu tekstu, czasu powstania itd.
  - randomizacja

TXM / CQPWeb

Jak pracować z konkordancją?



1. **Sprawdzić kotekst: … 2L, 1L, 1R, 2R …**
2. **Zredukować pojedyncze wystąpienie do powtarzających się wzorców**
3. **Postawić hipotezę**
4. **Analizować kolejne wystąpienia**
5. **Sprawdzić, co wnoszą do ujęcia**

leksykogramatyka

znaczenie lokalne

dobra/zła prozodia

klasyfikacja znaczeniowa dopełnień

przenośne - dosłowne

Laboratorium 1: szkic hasła słownikowego na podstawie NKJP

1. Wybór słowa

2. Opracowanie szkicu hasła na podstawie wyszukiwarki PELCRA:

a) kolokator,

b) konkordancje.

3. Zwrócić uwagę na:

a) częste połączenia i inne elementy kontekstu,

b) gatunki tekstu,

c) nacechowanie.

4. Skonfrontować ze znaczeniem odnotowanym w wybranym słowniku: czy analiza korpusowa wniosła cokolwiek?

5. Czy hasło powinno znaleźć się np. w słowniku ogólnym? Słowniku kieszonkowym?

6. Jakie są wady poszczególnych narzędzi analizy korpusu?

SketchEngine

Sumaryzacja danych: WordSketches

- kolokacje + gramatyka
- „jednostronicowe podsumowanie gramatycznego i kolokacyjnego zachowania wyrazu" (Kosem & Kilgariff 2012)

goal *(noun)* ukWaC freq = 168,184 (107.82 per million)

object_of	59,154	3.20	subject_of	25,630	2.00	adj_subject_of	2,159	1.40	modifie
score	8,555	11.03	score	1,029	8.39	galore	27	7.30	ultimat
achieve	9,504	9.69	disallow	270	8.28	achievable	43	6.95	winning
concede	1,418	9.31	concede	223	7.52	attainable	15	6.77	long-te
accomplish	588	7.88	gape	76	6.48	unrealistic	14	5.59	league
reach	1,937	7.42	kick	93	5.45	intact	19	5.04	primar
net	351	7.42	orientate	37	5.12	worthy	36	4.82	second

Sumaryzacja danych: SketchDiffs

- porównanie kolokacji między wyrazami hasłowymi

clever/intelligent ukWaC freqs = 20,593 | 26,128

clever 6.0 | 4.0 | 2.0 | 0 | -2.0 | -4.0 | -6.0 intelligent

and/or	4,831	9,801	2.10	3.40
funny	231	103	6.9	5.7
inventive	22	24	5.9	5.5
witty	133	166	8.3	8.2
resourceful	12	29	5.8	6.3
insightful	11	30	5.2	6.1
affectionate	6	32	4.5	6.2
thoughtful	14	121	5.0	7.7
compassionate	0	27	--	5.9
literate	0	26	--	5.9
well-informed	0	25	--	6.0
adaptive	0	39	--	6.1
informed	0	59	--	6.1

modifier	4,933	3,168	0.90	0.50
fiendishly	45	0	8.1	--
devilishly	17	0	6.8	--
awfully	15	0	6.1	--
terribly	25	0	6.1	--
dead	14	0	5.9	--
diabolically	9	0	5.9	--
amazingly	17	7	5.9	5.0
exceptionally	28	25	5.9	5.9
averagely	0	7	--	6.1
supposedly	0	28	--	6.2
ferociously	0	8	--	6.2
highly	0	572	--	6.9

modifies	11,415	16,772
ploy	72	0
clog	51	0
trick	166	0
twist	94	0
eh	40	0
chap	48	0
wordplay	21	0
ruse	17	0
Huh	18	0
lyric	83	80
creature	11	137
agent	9	465

Systemy redakcji słowników

Po co?

- **kontrola formy hasła**
- **kontrola struktury hasła**
  - następstwo i występowanie części
  - skróty
- **kontrola języka opisu i definicji**
- **połączenie z narzędziami korpusowymi (*TickBox Lexicography*) i leksykalnymi bazami danych**
- **wyeliminowanie niespójności**

Komponenty

- **interfejs edycji**

- **baza danych**

- **zarządzanie użytkownikami**

Następne zajęcia

- **data**
  - piątek, 15 IV 2016, godz. 18
- **temat**
  - Tworzenie i zastosowania słowników komputerowych (Wordnet, Framenet, Babelnet i spółka)
- **wymagania**
  - założenie konta trial na:

    https://www.sketchengine.co.uk

    https://cqpweb.lancs.ac.uk/

    http://ucts.uniba.sk/aranea_about/