

# Modelowanie tematyczne (topic modeling) treści pozyskanych z internetu

Maciej Eder

Instytut Języka Polskiego PAN  
&  
Uniwersytet Pedagogiczny w Krakowie

część 1

## Porównanie dwóch podkorpusów: słowa kluczowe

# Najpierw jednak słowo o użytych danych

- korpus 100 powieści polskich (1851-1940) [9 mln wyrazów]
- korpus 47 540 artykułów z portalu Onet.pl (komplet dokumentów opublikowanych we wrześniu 2015 r.) [12 mln wyrazów]
- korpus 2533 artykułów z czasopisma „Teksty Drugie” (1990-2014) [10 mln wyrazów]

# Powieść realistyczna vs. publicystyka tabloidowa

Słowa typowe dla literatury:

i pani ja pan się mu rzekł nie jej tak go mnie a mi oczy ją ty jakby tu co ku cóż pana niech panie on rzekła panna znowu nią ani lecz nic ona ale no zaraz zawałań niej sobie chwili ręce twarz nagle nim tam głowę zaś rękę człowiek ...

Słowa istotne dla rejestru publicystyczno-popularnego:

I oraz meczu polski m in zł tys mecz wcześniej trener Polsce ponad sezonie roku podczas r proc Europy czyli reprezentacji września klubu mln sierpnia sytuacji ok euro obecnie również Polska dzięki ligi według zespół osób drużyny gry zespołu piłkarz zdaniem przypadku kolejne spotkania zawodników związku został ostatnio latach spotkaniu ...

# Powieść realistyczna vs. publicystyka tabloidowa

Słowa typowe dla literatury:

i pani ja pan się mu rzekł nie jej tak go mnie a mi oczy ją ty jakby tu co ku cóż pana niech panie on rzekła panna znowu nią ani lecz nic ona ale no zaraz zawałań niej sobie chwili ręce twarz nagle nim tam głowę zaś rękę człowiek ...

Słowa istotne dla rejestru publicystyczno-popularnego:

I oraz meczu polski m in zł tys mecz wcześniej trener Polsce ponad sezonie roku podczas r proc Europy czyli reprezentacji września klubu mln sierpnia sytuacji ok euro obecnie również Polska dzięki ligi według zespół osób drużyny gry zespołu piłkarz zdaniem przypadku kolejne spotkania zawodników związku został ostatnio latach spotkaniu ...

# Powieść realistyczna vs. publicystyka tabloidowa

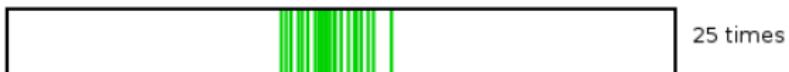
Słowa typowe dla literatury:

i pani ja pan się mu rzekł nie jej tak go mnie a mi oczy ją ty jakby tu co ku cóż pana niech panie on rzekła panna znowu nią ani lecz nic ona ale no zaraz zawałań niej sobie chwili ręce twarz nagle nim tam głowę zaś rękę człowiek ...

Słowa istotne dla rejestru publicystyczno-popularnego:

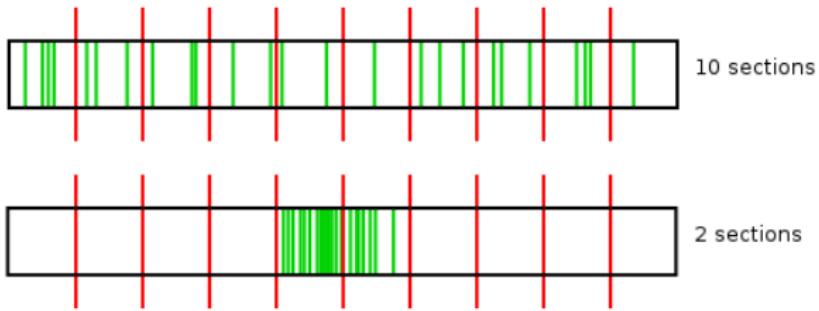
I oraz meczu polski m in zł tys mecz wcześniej trener Polsce ponad sezonie roku podczas r proc Europy czyli reprezentacji września klubu mln sierpnia sytuacji ok euro obecnie również Polska dzięki ligi według zespół osób drużyny gry zespołu piłkarz zdaniem przypadku kolejne spotkania zawodników związku został ostatnio latach spotkaniu ...

# Metoda Zeta (wprowadzona przez Burrowsa)

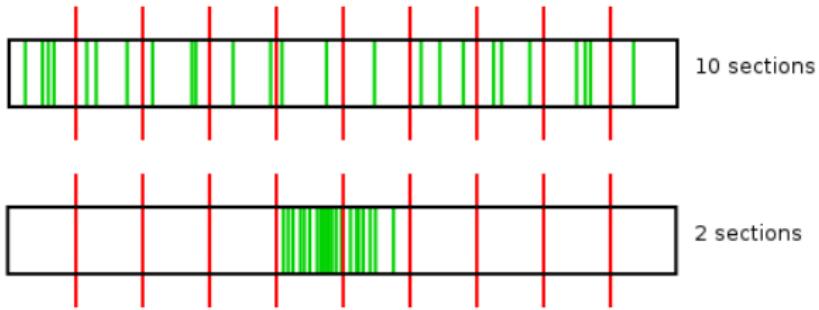


sama frekwencja wyrazu nie mówi nic o jego rozkładzie

# Szukanie słów o różnym rozkładzie



# Szukanie słów o różnym rozkładzie



$$\zeta_{(a,b)} = \left( \frac{f_{(a)} - f_{(b)}}{100} \right) + 1 \quad \zeta_{(a,b)} = \frac{f_{(a)} - f_{(b)}}{f_{(a)} + f_{(b)}}$$

# Panie vs. panowie w literaturze XVIII-XIX wieku

Women:

- A. Brontë (*Agnes Grey, The Tenant of Wildfell Hall*)
- C. Brontë (*Jane Eyre, The Professor, Villette*)
- E. Brontë (*Wuthering Heights*)
- Austen (*Emma, Pride and Prejudice, Sense and Sensibility*)
- Eliot (*Adam Bede, Middlemarch, The Mill on the Floss*)

Men:

- Dickens (*Bleak House, David Copperfield, Hard Times*)
- Fielding (*Joseph Andrews, Tom Jones*)
- Richardson (*Clarissa, Pamela*)
- Thackeray (*Barry Lyndon, The History of Pandennis, Vanity Fair*)
- Trollope (*Barchester Towers, Phineas Finn, The Prime Minister*)

# Kobiety są z Wenus... (?)

Słowa kobiece:

feelings glance effort paused feeling surprise noticed pause  
consciousness dared enjoyment tone listen exclaimed features  
seated continually anxiety solitude inward apparently painful  
entrance respectable relief closed watching feel bent peculiar rain  
suddenly cheerful clear trees aspect watched plan slight doubtless  
reached smile brow vague quiet mere movement gathered suffering  
entered listened observed warm exertion minutes change ...

Słowa męskie:

story although lord bosom honour honest duke parliament city  
score enemy coach coat inn thousand breast bill dozen lordship  
guilty court ain legs bottle captain fight pen battle sum  
nevertheless reader virtue order innocent condition infinite castle  
widow england accident readers laws fellows hundred service king  
stories persons ladyship fly street dearest honours member fortune  
government wig drank papers wretch described honourable pocket

# Kobiety są z Wenus... (?)

Słowa kobiece:

feelings glance effort paused feeling surprise noticed pause  
consciousness dared enjoyment tone listen exclaimed features  
seated continually anxiety solitude inward apparently painful  
entrance respectable relief closed watching feel bent peculiar rain  
suddenly cheerful clear trees aspect watched plan slight doubtless  
reached smile brow vague quiet mere movement gathered suffering  
entered listened observed warm exertion minutes change ...

Słowa męskie:

story although lord bosom honour honest duke parliament city  
score enemy coach coat inn thousand breast bill dozen lordship  
guilty court ain legs bottle captain fight pen battle sum  
nevertheless reader virtue order innocent condition infinite castle  
widow england accident readers laws fellows hundred service king  
stories persons ladyship fly street dearest honours member fortune  
government wig drank papers wretch described honourable pocket

# Kobiety są z Wenus... (?)

Słowa kobiece:

feelings glance effort paused feeling surprise noticed pause  
consciousness dared enjoyment tone listen exclaimed features  
seated continually anxiety solitude inward apparently painful  
entrance respectable relief closed watching feel bent peculiar rain  
suddenly cheerful clear trees aspect watched plan slight doubtless  
reached smile brow vague quiet mere movement gathered suffering  
entered listened observed warm exertion minutes change ...

Słowa męskie:

story although lord bosom honour honest duke parliament city  
score enemy coach coat inn thousand breast bill dozen lordship  
guilty court ain legs bottle captain fight pen battle sum  
nevertheless reader virtue order innocent condition infinite castle  
widow england accident readers laws fellows hundred service king  
stories persons ladyship fly street dearest honours member fortune  
government wig drank papers wretch described honourable pocket

# Powieść realistyczna vs. publicystyka tabloidowa

Słowa typowe dla literatury:

rzekł jakby cóż ku ty niech ręce głowę twarz rękę głosem znowu zaraz nagle pana serce pan człowiek twarzy zawała panie swego drzwi duszy Bóg głos myśl myśl oczach rzekła cicho niby usta nią ja pani zdawało głową chwili panu przecie prędko ach widział swej Boże stary mój ziemi ...

Słowa istotne dla rejestru publicystyczno-popularnego:

I oraz meczu polski m in zł tys mecz wcześniej trener Polsce ponad sezonie roku podczas r proc Europy czyli reprezentacji września klubu mln sierpnia sytuacji ok euro obecnie również Polska dzięki ligi według zespół osób drużyny gry zespołu piłkarz zdaniem przypadku kolejne spotkania zawodników związku został ostatnio latach spotkaniu ...

# Powieść realistyczna vs. publicystyka tabloidowa

Słowa typowe dla literatury:

rzekł jakby cóż ku ty niech **ręce głowę twarz rękę** głosem znowu zaraz nagle pana **serce** pan człowiek **twarz** zawałał panie swego drzwi **duszy** Bóg głos myśl myśli oczach rzekła cicho niby usta nią ja pani zdawała **głową** chwili panu przecie prędko ach widział swej Boże stary mój ziemi ...

Słowa istotne dla rejestru publicystyczno-popularnego:

I oraz meczu polski m in zł tys mecz wcześniej trener Polsce ponad sezonie roku podczas r proc Europy czyli reprezentacji września klubu mln sierpnia sytuacji ok euro obecnie również Polska dzięki ligi według zespół osób drużyny gry zespołu piłkarz zdaniem przypadku kolejne spotkania zawodników związku został ostatnio latach spotkaniu ...

# Powieść realistyczna vs. publicystyka tabloidowa

Słowa typowe dla literatury:

rzekł jakby cóż ku ty niech ręce głowę twarz rękę głosem znowu zaraz nagle pana serce pan człowiek twarzy zawałał panie swego drzwi duszy Bóg głos myśl myśl oczach rzekła cicho niby usta nią ja pani zdawało głową chwili panu przecie prędko ach widział swej Boże stary mój ziemi ...

Słowa istotne dla rejestru publicystyczno-popularnego:

I oraz meczu polski m in zł tys mecz wcześniej trener Polsce ponad sezonie roku podczas r proc Europy czyli reprezentacji września klubu mln sierpnia sytuacji ok euro obecnie również Polska dzięki ligi według zespół osób drużyny gry zespołu piłkarz zdaniem przypadku kolejne spotkania zawodników związku został ostatnio latach spotkaniu ...

część 2

## Modelowanie tematyczne (topic modeling)

# Kolokacje w językoznawstwie korpusowym

- Prawdopodobieństwo wystąpienia słowa A w korpusie:  $P(A)$
- Prawdopodobieństwo wystąpienia słowa B w korpusie:  $P(B)$
- Prawdopodobieństwo ich współwystąpienia:  $P(A) \times P(B)$
- Na przykład:

$$P(A) = 0.01 \quad P(B) = 0.02 \quad P(A) \times P(B) = 0.0002$$

# Kolokacje w językoznawstwie korpusowym

- Niektóre słowa 'lubią się' bardziej niż wynika z teorii:

*strong tea*

*\*powerful tea*

*powerful computer*

*\*strong computer*

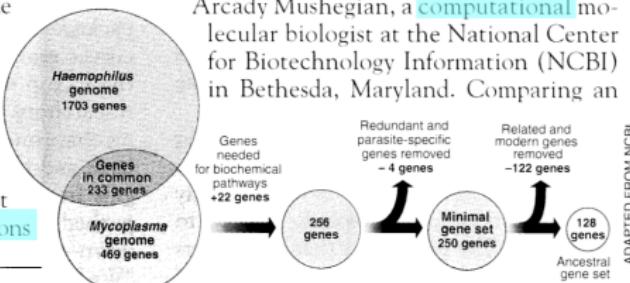
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

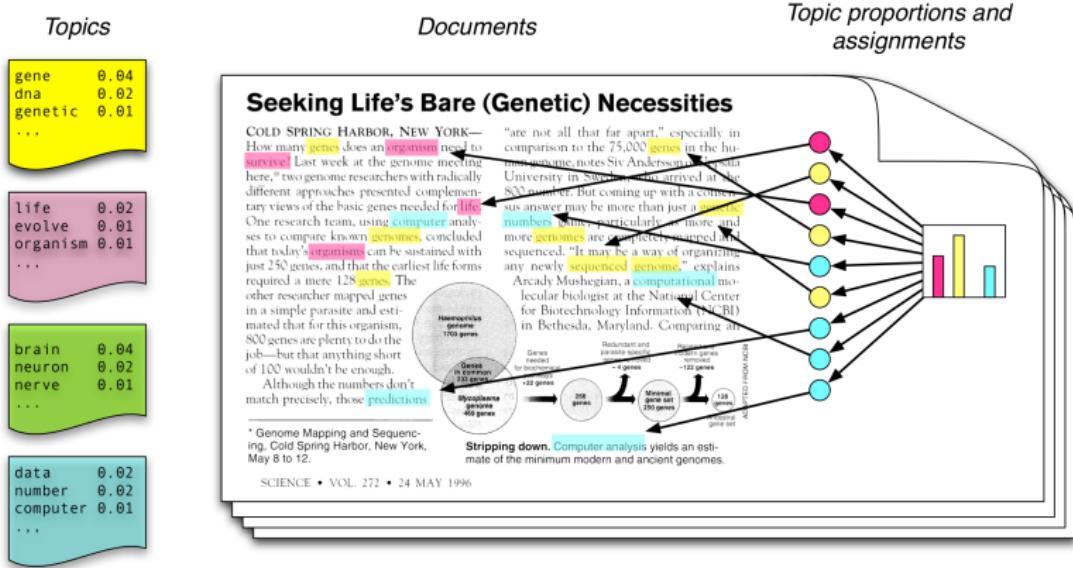
Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

# Słoważbiory nierówno reprezentowane w tekstach



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

# Słowożbiory w 100 powieściach polskich XIX-XX w.

# Słowoziór 1: ‘wsi spokojna, wsi wesoła’

chleb gęsi chłopów  
polach wiatr  
parę słonice ganek  
ganku wieś koni dworze  
tyl lasu WSi ziemia  
dziedzica lesje  
roboty WOZ mieli  
drogi dwór mleka  
dni polu chłop  
konia roboty pola drzewa  
śniegu chłopi  
spod dzień ziemi ludźmi  
las drodze śnieg  
dworu pole dziedzic  
zboża stajni wrócił  
człowiek

# Słowożbiór 28: ‘mydło i powidło (???)’

gdzieś przecie  
kobiet niczym  
niebawem  
bodaj spod wreszcie rycerz  
kobiety czym ulicy  
samego wciąż tymczasem całkiem  
mnich wraz tedyrynkū  
ciała nagle Oto między owo  
piersi ni mistrznazbyt  
onych brat Chwili ducha grodu  
świata wszystkie ledwie niżli  
zgola śród życia znów serca  
wszystkich chyba daremnie  
mnicha snadź kościoła  
powiada

# Słowożbiór 35: ‘w kościołnej kruchcie’

chrystus ojciec  
kaplicy księdzem  
księdu księże matka  
krzyż matki proboszcz  
**szymon** kościół święty  
kościele bogboga śmierci  
piotr ołtarza świętych  
św mszy księży  
modlitw świętej  
świętym ks wiary  
boże rzekłmury  
klasztor modlitwy  
boskiej klasztoru świętego paweł  
dzieci bogu proboszcza  
klasztorze modlić chwila  
nabożeństwo święte

# Słowożbiór 20: 'gimnazjum'

koledzy  
pewnego pewnej  
oczyma  
dopiero  
siódmej itd kolegów  
ażeby  
uczył przecie **lekacji** naukimarcin  
uczyć **gimnazjum** uczeń toteż  
dyrektor klasie między  
jedrekdzieci kolega  
książki lekcje  
ogółe **Szkoleszkoły** uczniów  
wreszcie  
chłopcy dwu profesora  
ławki sali profesor klasa stefan  
tedy **nauczyciel** szkoła  
całej pierwszej czytać  
wszyscy nauczyciela  
podczas miejsca

# Słowoziór 26: ‘przyroda’

skrzydła ciszy  
gałęzie ptaki pieśń ziemię  
razem wiatr powietrzu  
nocy niby kwiatów cały  
pół czasem złote  
mgły drzew niebo gwiazdy  
słońcu słonice wody  
powietrze liście  
czarne białe słońca niebie  
chmur dzień świat drzewa  
pieśni Całe ziemi liści  
chmury kwiaty wysoko  
gałęzi nieba cisza  
księżyc morze czerwone  
cicho głosy

# Słowoziór 8: 'teatr'

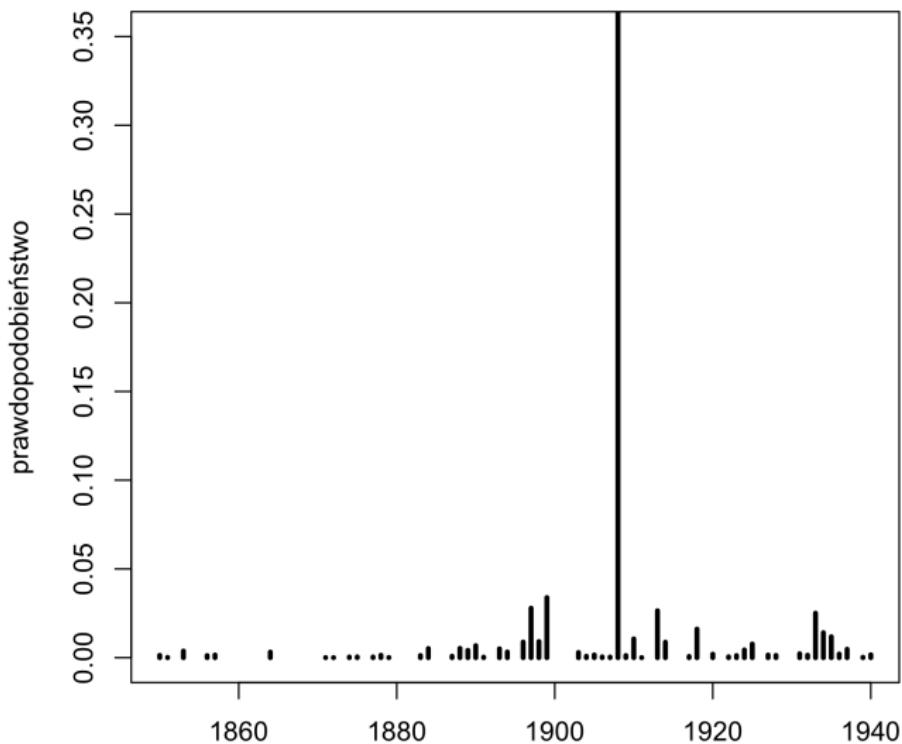
przedstawienie  
towarzystwo  
talentu wszystkich  
zaraz publiczność  
kolega mecenas kobiety  
prawie sztuki gra kieliszek  
ależ teatr teatrze cha wprost  
jakiś la niby brawo roli  
sceny janka sztuka talent  
publiczności teatru role kawiarni  
jakąś teatru loży muzyka  
wina scenie wołał ktoś  
grał scenę aktor  
panowie grac dyrektor muzyki  
władek sztukę dosyć  
towarzystwa sztukę przecież

# Słowożbior 44: 'stylizacja gwarowa'

wnet<sup>ano</sup>  
calkiem nieco  
mateusz wszystkie  
abo antek któryen tyla  
naród zaraz  
świat kaj  
cięgiem juści ino  
trza se znnowu drugie  
stary jał oczy  
wsı wieś kowal  
chatupy ledwie  
dyć niby  
kobiety la całą ludzie  
izbie cicho naraz  
chalupiedopiero czas prawie  
choćby jesus świecie  
dlugo wolna

# Proporcje słwozbioru 44: ‘stylizacja gwarowa’

**temat 44**

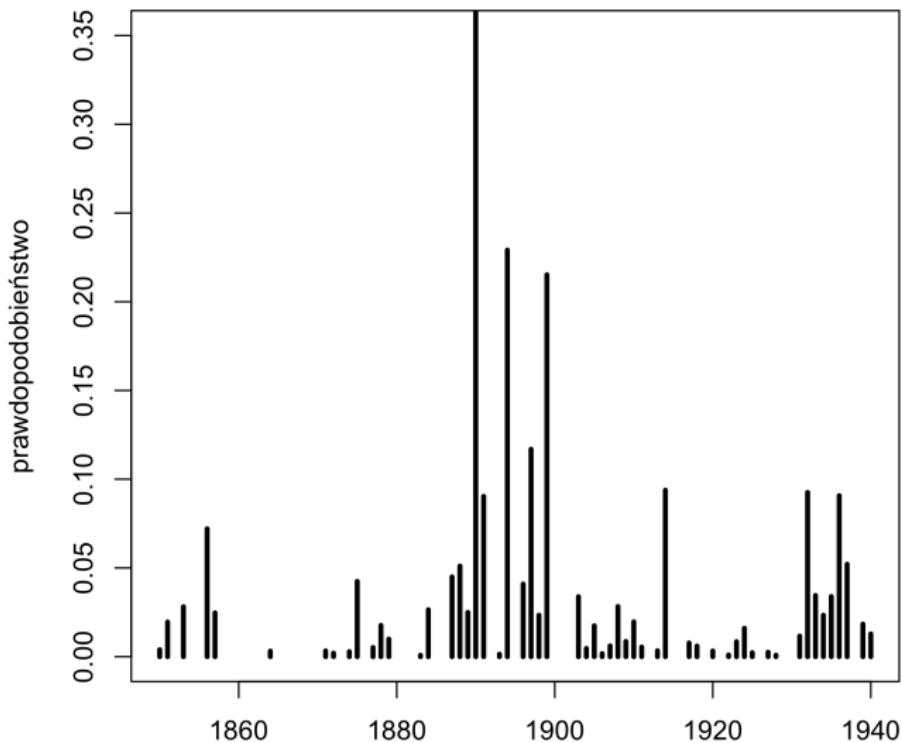


# Słowożbior 41: ‘pieniądz i kapitał’

pięćdziesiąt  
trzydziestu tysiące odparł  
dopiero kupić dwadzieścia stary  
tyle interes niech  
zaczął dom grosza  
sklepu panie sklep  
dzieci zyd majątku  
ojciec rok zrobił  
procent sto rubli ażeby  
stach pieć tysiąc  
**pieniądze** trzy płacić  
kochany **pieniędzy** majątek  
sumę ignacy sklepie  
zresztą sześć sumy  
baronowa dziesięć kieszeni  
złotych cztery

# Proporcje słwozbioru 41: ‘ pieniądze i kapitał’

## temat 41



# Przegląd chronologiczny: „Teksty Drugie” z lat 1990-2014

# Słowożbiór 4: ‘filologia (tradycyjna)’

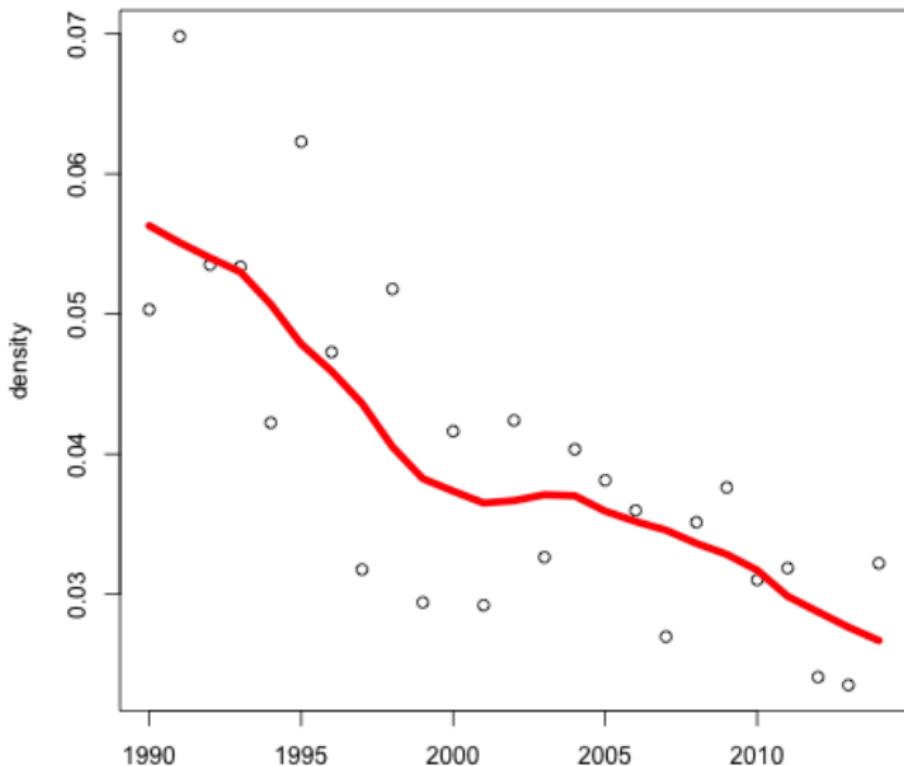
teoretyczny strukturalizm  
markowski interpretator  
kulturowy hermeneutyka  
dekonstrukcja dzieło teza „wiedza  
autor sławiński

# interpretacja

czytanie kontekst literacki  
tekst rorty paradymat  
mówić lektura naukaczytelnik  
koncepcja sens jakis prawda  
filozofia Interpretacyjny znaczenie  
pojęcie kultura metoda  
etyczny derrida odczytać  
założenie literatura fish  
wspólnota intencja  
interpretować język  
humanistyka rozumienie granica stanowisko  
literaturoznawczy

## Słowożbór 4: 'filologia (tradycyjna)'

topic 4

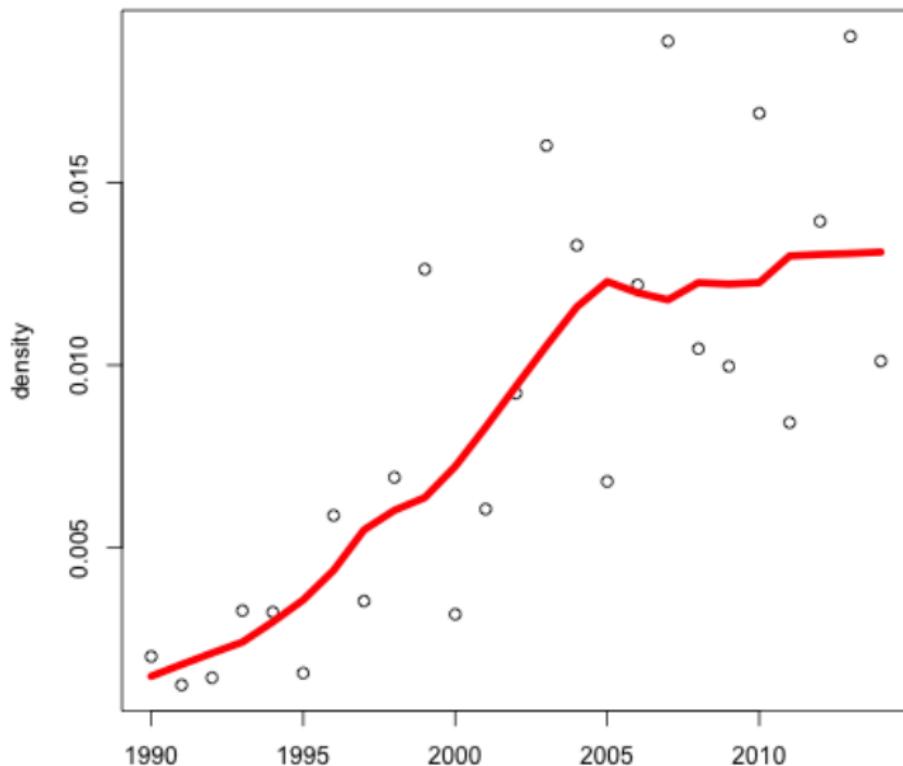


# Słowoziór 14: 'filologia (nowoczesna)'

program tekstowy  
gracz komputerowy  
wizualny medium  
użytkownik informacja między  
internetowy książka różny komputer  
ekran internet sieć elektroniczny  
film znak hipertekst  
tworzący poziom  
kultura  
przykład media  
przekaz e tekstu dostęp  
wirtualny nowa narzędzie  
świat strona rzeczywistość  
czytanie cyfrowy gra forma  
technologia reklama odbiorca  
system medialny wydawnictwo  
komunikacja com okładka

# Słowożbór 14: 'filologia (nowoczesna)'

**topic 14**



# Słowożbior 4 i 14: ‘dwie filologie’

interpretacja  
tekst  
teoria  
czytanie  
kontekst  
tekstura  
mówić  
filozofia  
koncepcja  
Sens  
filozofia  
pojęcie  
interpretacyjny  
kultura  
etyczny  
derrida  
literatura  
zalożenie  
wspólnota  
interpretować  
humanistyka  
literaturoznanieczyzny  
literaturoznanieczyzny  
interpretator  
hermeneuta  
dzieło teza, wiedza  
autoryslawiński  
teoretyczny strukturalizm  
markowski kulturowy  
dekonstrukcja

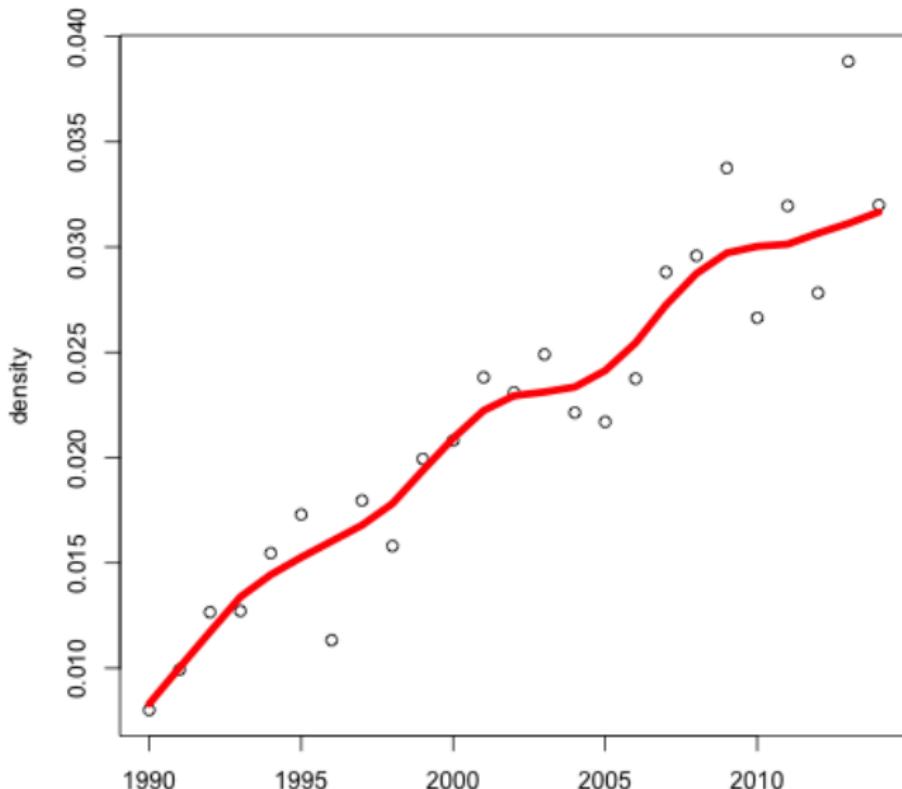
program tekstowy  
gracz komputerowy  
wizualny medium  
użytkownik informacja między  
internetowy książka różny komputer  
ekran internet sieć elektroniczny  
film znak hipertekst  
tworzący kultura poziom  
przykład tekstu dostęp  
przekaz nowa narzędzie  
wirtualny gra rzeczywistość  
świat strona odbiorca  
czytanie cyfrowy reklama  
technologia wydawnictwo  
system medialny com  
okładka komunikacja

# Słowoziór 21: ‘opowieść/pamięć’

zapis dzieje teraźniejszość  
kapuściński autobiografia  
współczesny doświadczanie  
**historyczny pamięć** osobisty historyk pisarz  
opowiadanie fikcja wspomnienie  
relacja autor osoba  
powieść fikcyjny  
wydarzenie historia biografia  
opowiadać narracja życie podróż  
pisanie mówić fakt pisać  
podmiot postać fragment  
czasowy bohater narrator narracyjny  
opowiedzieć przeszłość autobiograficzny  
wiedza prawda  
opowiedzieć zdarzenie forma  
rzeczywistość tożsamość czytelnik

# Słowożbior 21: 'opowieść/pamięć'

topic 21



# Słowoziór 31: 'literatura rosyjska'

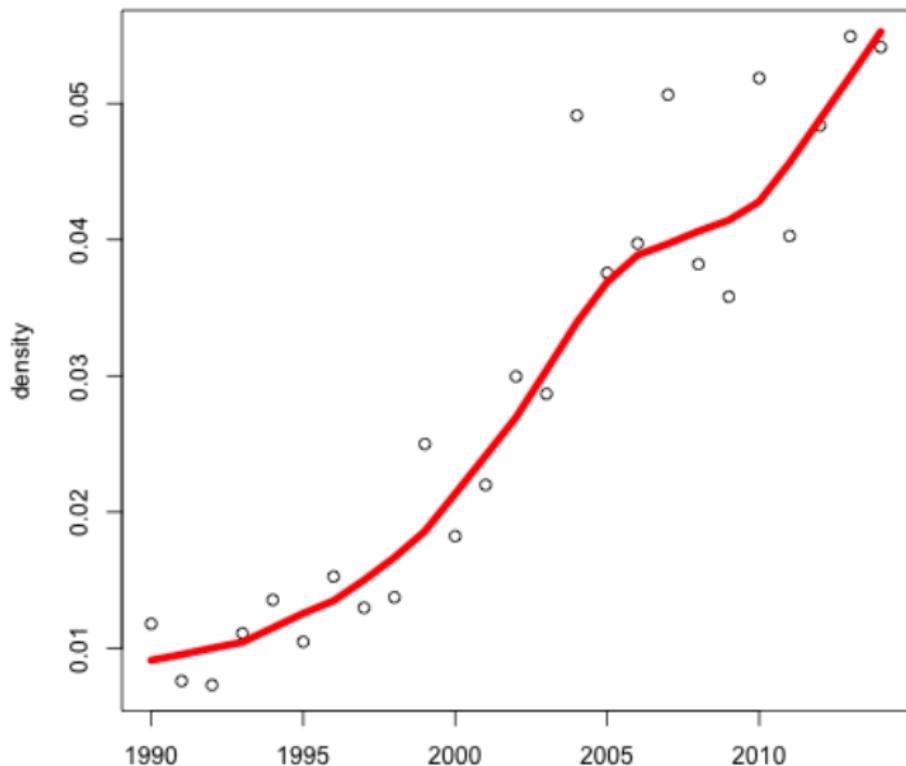
rewolucja życie ruch  
jurij marinetti achmatowa  
konceptja poeta dostojewski  
gogol sekunda lukl rosja włoski  
mandelsztam moskwa russian teoria  
utwór

# rosyjski

csv futuryzm tolstoj zamenhof bachtin m idea bajt  
puszkint język flos zdob pojście  
majakowski iwanow marr  
formalista słowo pieśń  
futurysta bachtinowski

# Słowożbór 31: 'literatura rosyjska'

topic 31



# Słowoziór 17: 'gender/queer studies'

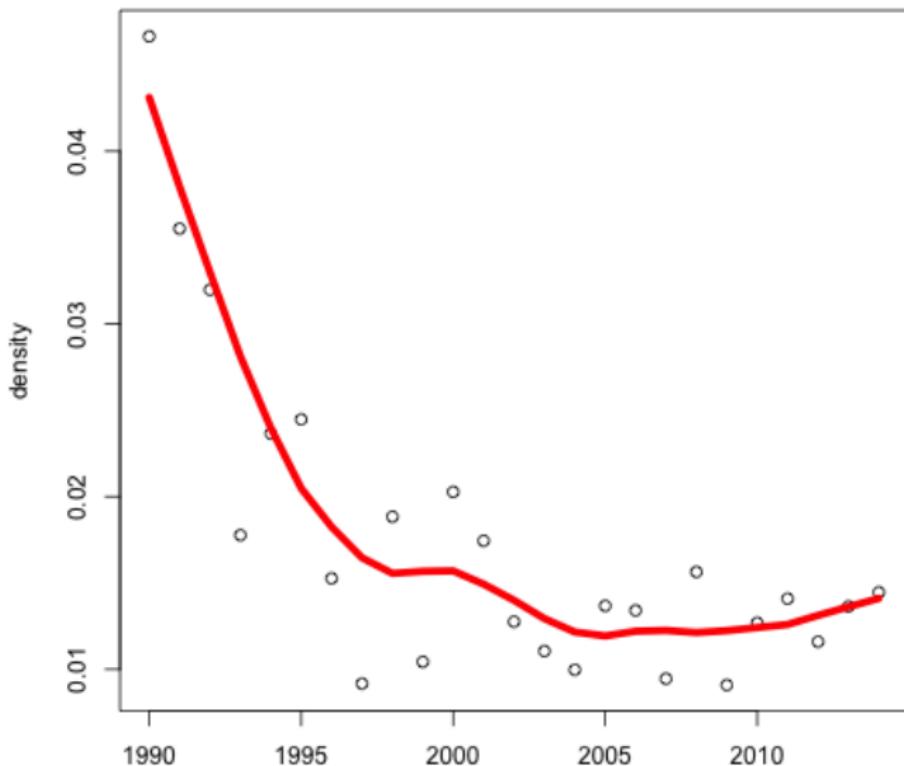
dominujący  
erotyczny heteroseksualny  
kanon odmieniec seksualność  
kobieta rola mężczyzna gender  
camp homoseksualizm parodia  
sontag ritz klara tożsamość film  
dyskurs lechon płeć queer  
lato seksualny queer

# homoseksualny

poprzez męski kamp kampu lesbijka  
znak mit płciowy  
kultura sport iwaszkiewicz gej gay  
związek pożądanie drag gejowski  
homoseksualista sublimacja  
przykład proust tekst kobiecy  
homoseksualność homoerotyczny

# Słowoziór 17: 'gender/queer studies'

topic 17



część 3

## Podsumowanie

# Podsumowanie

- Wyszukiwanie informacji
- Weryfikacja hipotez historycznoliterackich
- Śledzenie chronologii zmian językowych
- Semantyka dystrybucyjna