



Gradacyjna analiza danych korpusowych

Łukasz Dębowski

Emilia Jarochowska

Marek Wiech

Instytut Podstaw Informatyki PAN





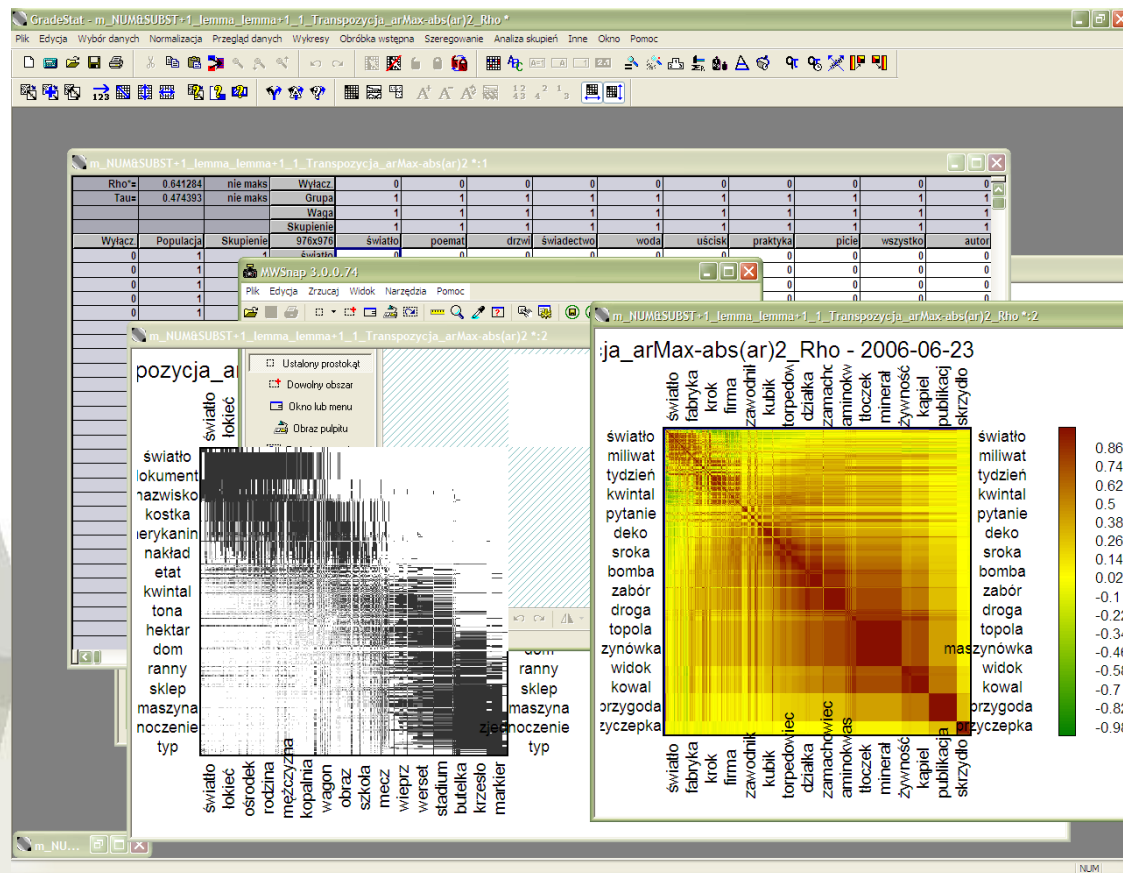
- Narzędzie i dane:
GradeStat i Korpus SFPW
- Mapy nadreprezentacji
- Analiza odpowiedniości (GCA)
- Trzy przykłady
 - klasyfikacja słów nieodmiennych
 - współwystępowanie rzeczowników i liczebników
 - deklinacja rzeczowników





GradeStat wersja 2.6

- implementacja m.in. gradacyjnej analizy danych
- główny autor: dr inż. Olaf Matyja
- wersja demonstracyjna do pobrania pod adresem:
<http://gradeostat.ipipan.waw.pl>





Słownik frekwencyjny polszczyzny współczesnej

I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak
Instytut Języka Polskiego PAN, Kraków, 1990

Korpus SFPW pochodzi z lat 60. XX w.

Zawiera 10 tys. próbek tekstów po około 50 słów.

Ogółem **500 tys słów**, po 100 tys. słów z:

- tekstów popularnonaukowych,
- drobnych wiadomości prasowych,
- tekstów publicystycznych,
- prozy artystycznej
- dramatu artystycznego.



Korpus SFPW jest anotowany

form lemma POS number case gender person degree aspect negation accommodability accentability post-prepositionality agglutination vocalicity punctuation

Sztuka sztuka subst sg nom f - - - - -
utraciła utracić praet sg - f - - perf - - - - nagl - -
swoją swój adj sg acc f - pos - - - - -
moc moc subst sg acc f - - - - -
pobudzającą pobudzający adj sg acc f - pos - - - - -
: : interp - - - - -
przykrym przykry adj sg inst n - pos - - - - -
widowiskiem widowisko subst sg inst n - - - - -
staje stawać fin sg - - ter - imperf - - - - -
się się qub - - - - -
koncert koncert subst sg nom m3 - - - - -
wybitnej wybitny adj sg gen f - pos - - - - -
niegdyś niegdyś qub - - - - -
śpiewaczki śpiewaczka subst sg gen f - - - - -
i i conj - - - - -
nie nie qub - - - - -
uświetnią uświetnić fin pl - - ter - perf - - - - -
go on ppron3 sg gen m3 ter - - - - nakc npraep - - -
nawet nawet qub - - - - -
nigdy nigdy qub - - - - -
dotąd dotąd qub - - - - -
nie nie qub - - - - -
wykonywane wykonywać ppas pl nom m3 - - imperf aff - - - - -
utwory utwór subst pl nom m3 - - - - -
genialnego genialny adj sg gen m1 - pos - - - - -
kompozytora kompozytor subst sg gen m1 - - - - -



- Narzędzie i dane:
GradeStat i Korpus SFPW
- **Mapy nadreprezentacji**
- Analiza odpowiedniości (GCA)
- Trzy przykłady
 - klasyfikacja słów nieodmiennych
 - współwystępowanie rzeczowników i liczebników
 - deklinacja rzeczowników

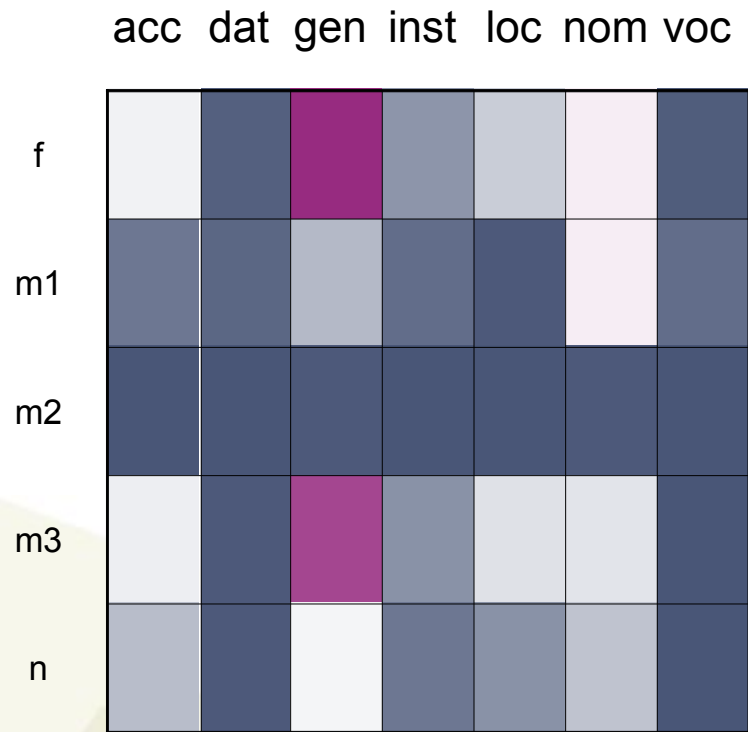




Mapy nadreprezentacji

częstości rodzajów i przypadków dla rzeczowników

dane surowe

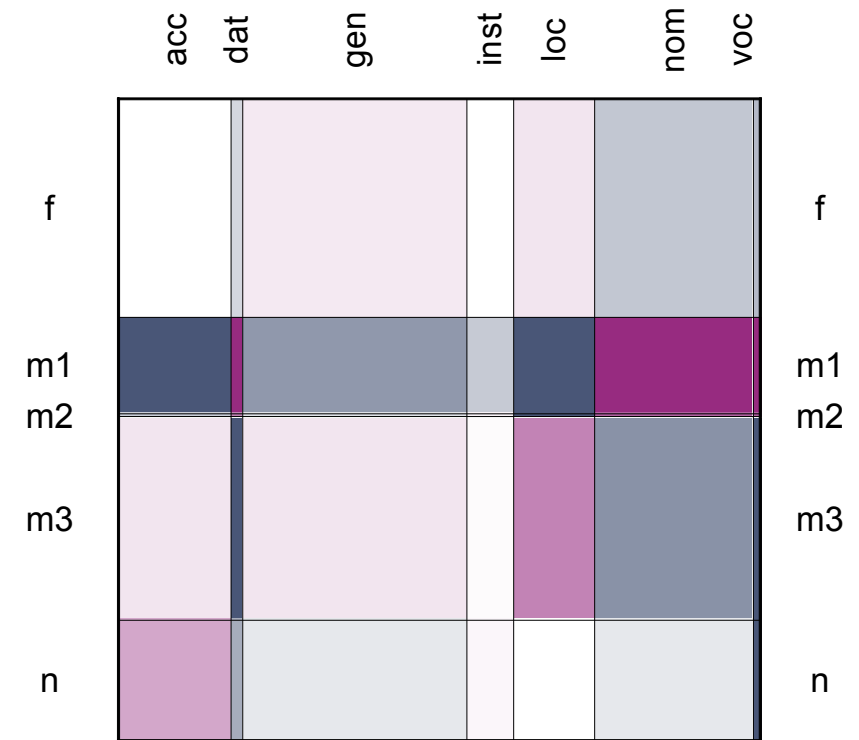


acc dat gen inst loc nom



253
2083
3456
4828
6201
7573
8946
10318
11691
13063
14436
15808
17181
18553
19926
21298

mapa nadreprezentacji



acc dat gen inst loc nom voc



0.6
0.65
0.69
0.74
0.79
0.84
0.89
0.95
1.01
1.07
1.14
1.21
1.29
1.37
1.45
1.55

Ścisła definicja nadreprezentacji

- częstość dla komórki (i,j): p_{ij}
(częstości sumują się do 1)
- suma częstości komórek w i-tym wierszu: p_{i+}
- suma częstości komórek w j-tej kolumnie: p_{+j}
- nadreprezentacja komórki (i,j): $p_{ij}/p_{i+}p_{+j}$



Ścisła definicja GCA

Przestawmy wiersze i kolumny macierzy częstości tak, aby zmaksymalizować ρ Spearmana, czyli

$$\rho = 3 \sum_{j=1}^k \sum_{i=1}^m (S_{i-1} + S_i - 1)(T_{j-1} + T_j - 1) p_{ij}$$

gdzie dystrybuanty dla wierszy i kolumn

$$S_i = p_{1+} + p_{2+} + \dots + p_{i+},$$

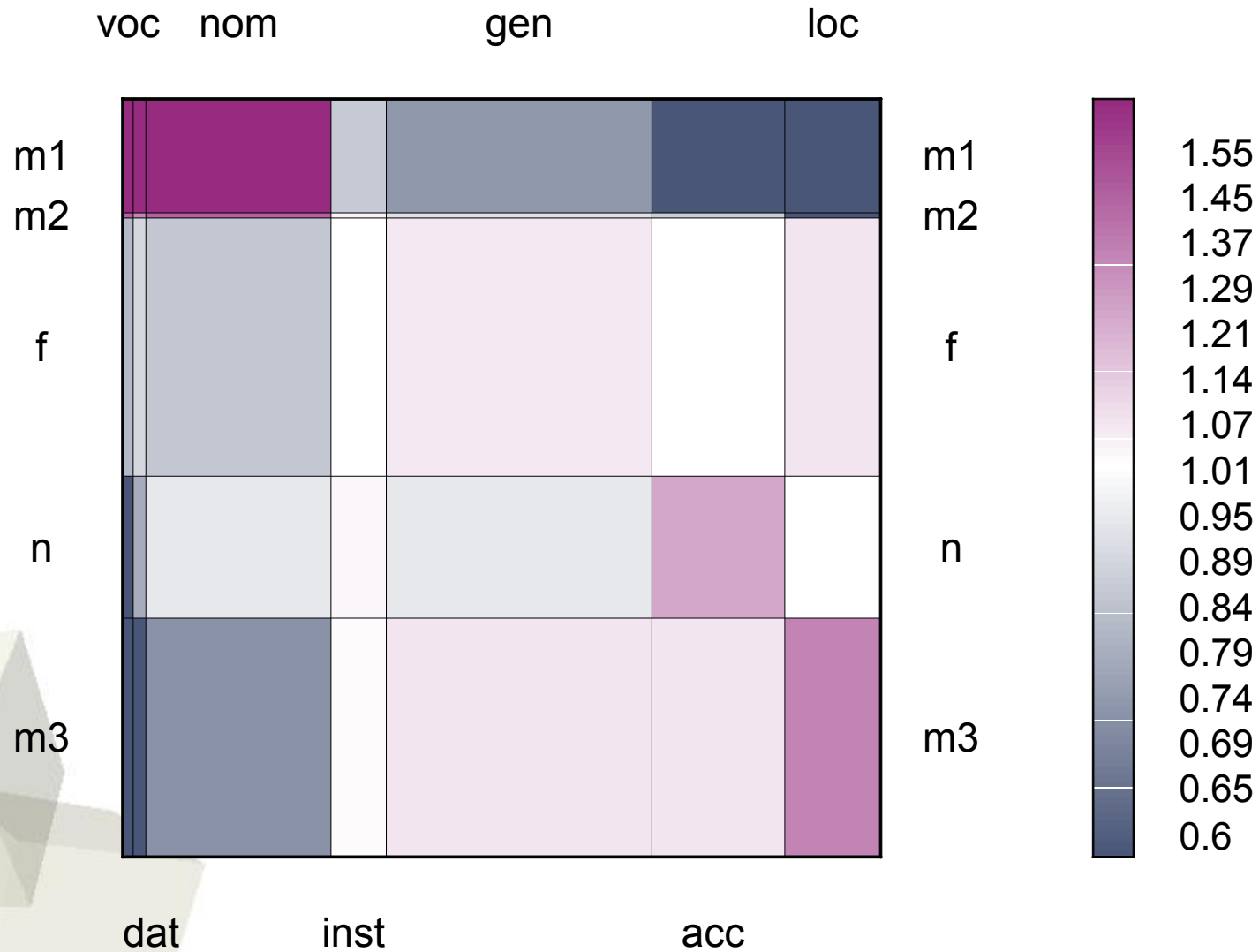
$$T_j = p_{+1} + p_{+2} + \dots + p_{+j}.$$

zależą od ich kolejności.



Gradacyjna analiza odpowiedniości

mapa nadreprezentacji po GCA
(grade correspondence analysis)





- Narzędzie i dane:
GradeStat i Korpus SFPW
- Mapy nadreprezentacji
- Analiza odpowiedniości (GCA)
- Trzy przykłady
 - klasyfikacja słów nieodmiennych
 - współwystępowanie rzeczowników i liczebników
 - deklinacja rzeczowników





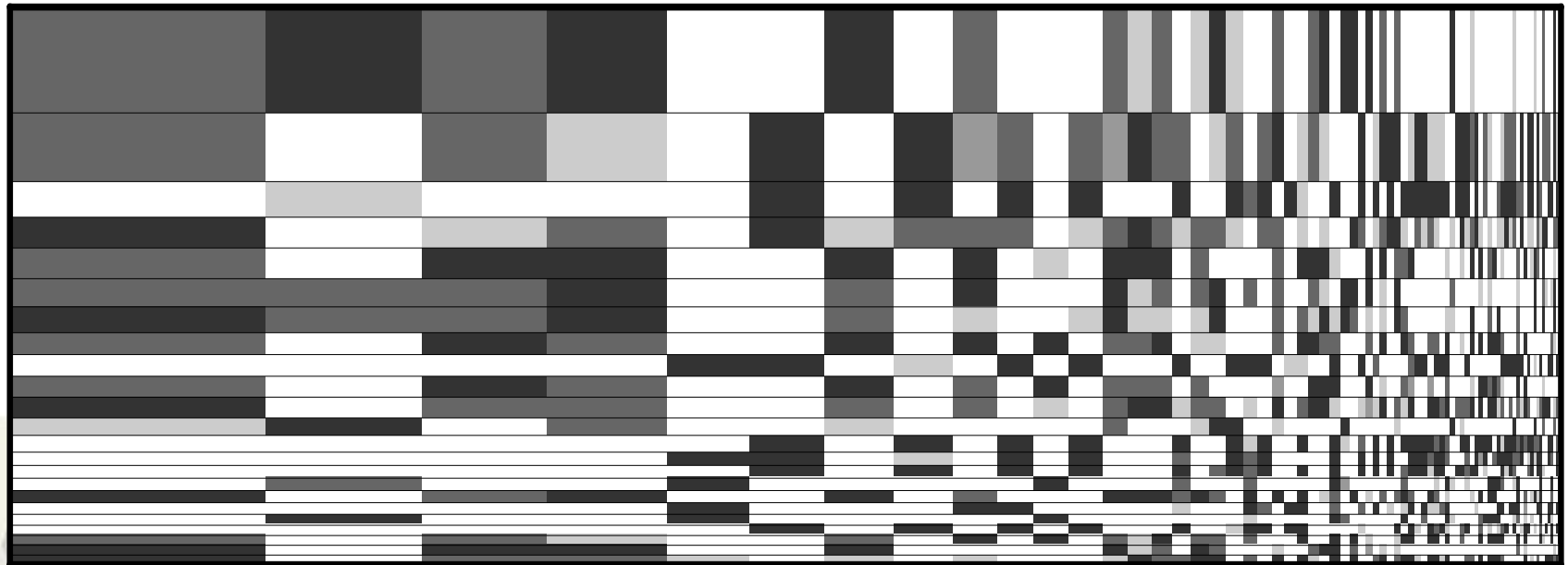
Klasyfikacja słów nieodmiennych

Przykład 1

73 najczęstsze słowa a otaczające części mowy

w i na z nie ze do a o ale się jak od po za to przez dla czy tak bo przy tuż ob pod przyś
mieszki je nie bez pamiłki rólki z dła to n

subst:subst
interp:subst
interp:qub
interp:adj
qub:subst
subst:adj
adj:subst
fin:subst
interp:fin
praet:subst
conj:subst
adj:adj
interp:prep
interp:praet
interp:ppron
subst:in
qub:adj
interp:inter
subst:praet
inter:conj
fin:adj
ppas:subst
conj:adj

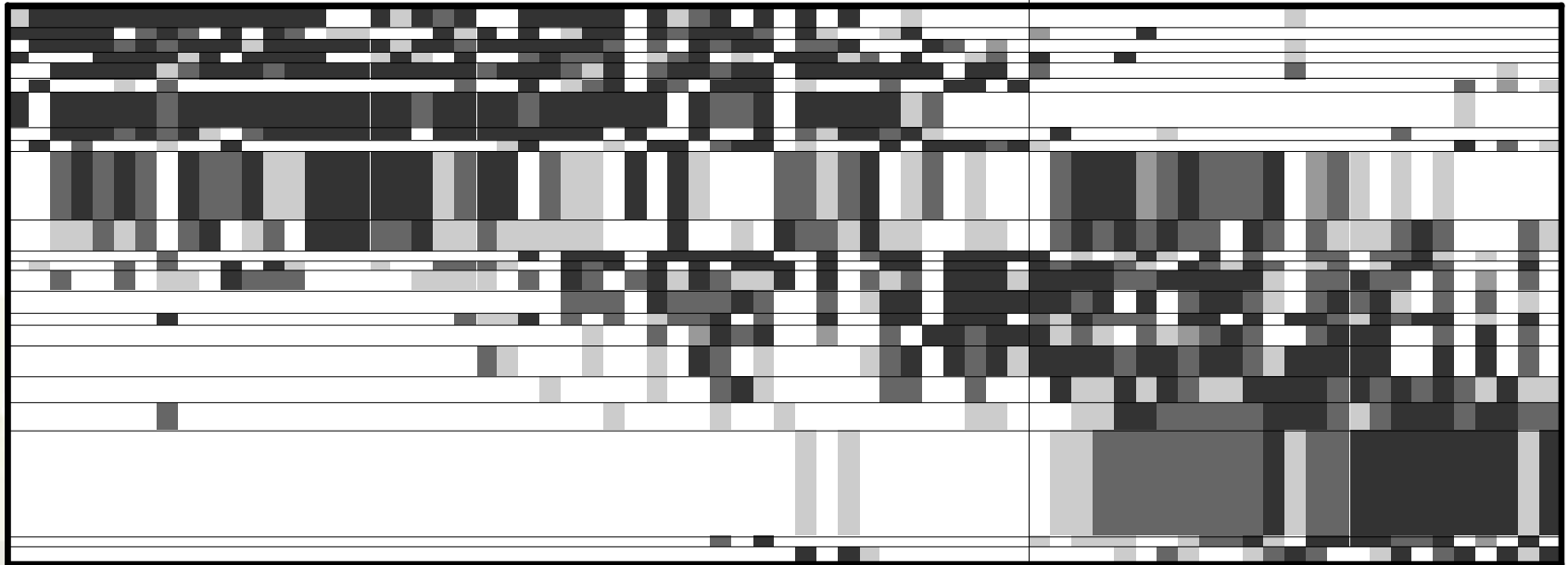




Dwa skupienia (po GCA)

no
nie
dlaczego
gdzie
ale
jeśli
bo
tak
kiedy
lecz
potem
jeżeli
dlatego
może
że
żeby
aby
gdy
a
iż
co
to
gdyby
jakby
zawsze
jak
przecież
teraz
chyba
niech
tu
natomiast
wiec
tam
własnie
już
mimo
albo
nawet
czy
ani
jeszcze
tylko
by
też
jednak
również
się
także
u
po
bez
jako
o
wśród
przy
za
przed
w
wobec
niż
od
na
pod
do
dla
między
z
i
nad
oraz
przez
lub

interp:fin
interp:inter
interp:praet
interp:conj
interp:prep
subst:fin
interp:qub
interp:pron
subst:praet
interp:subst
interp:adj
fin:adj
conj:adj
conj:subst
fin:subst
qub:adj
praet:subst
qub:subst
adj:subst
subst:adj
subst:subst
ppas:subst
adj:adj

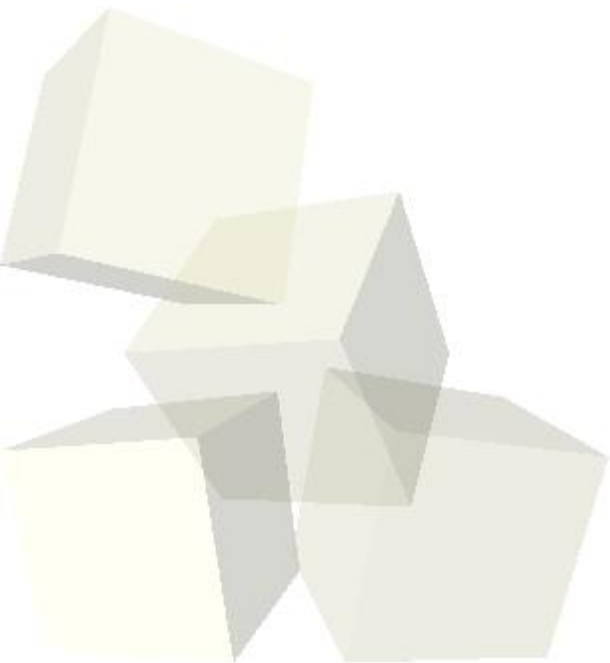




Kolumny najbardziej odstające

- Posortowane według AvgDistA:

właśnie, tam, się, mimo, by, czy, albo, zawsze,
też, tu, już, chyba, niech, natomiast, niż, iż, ...





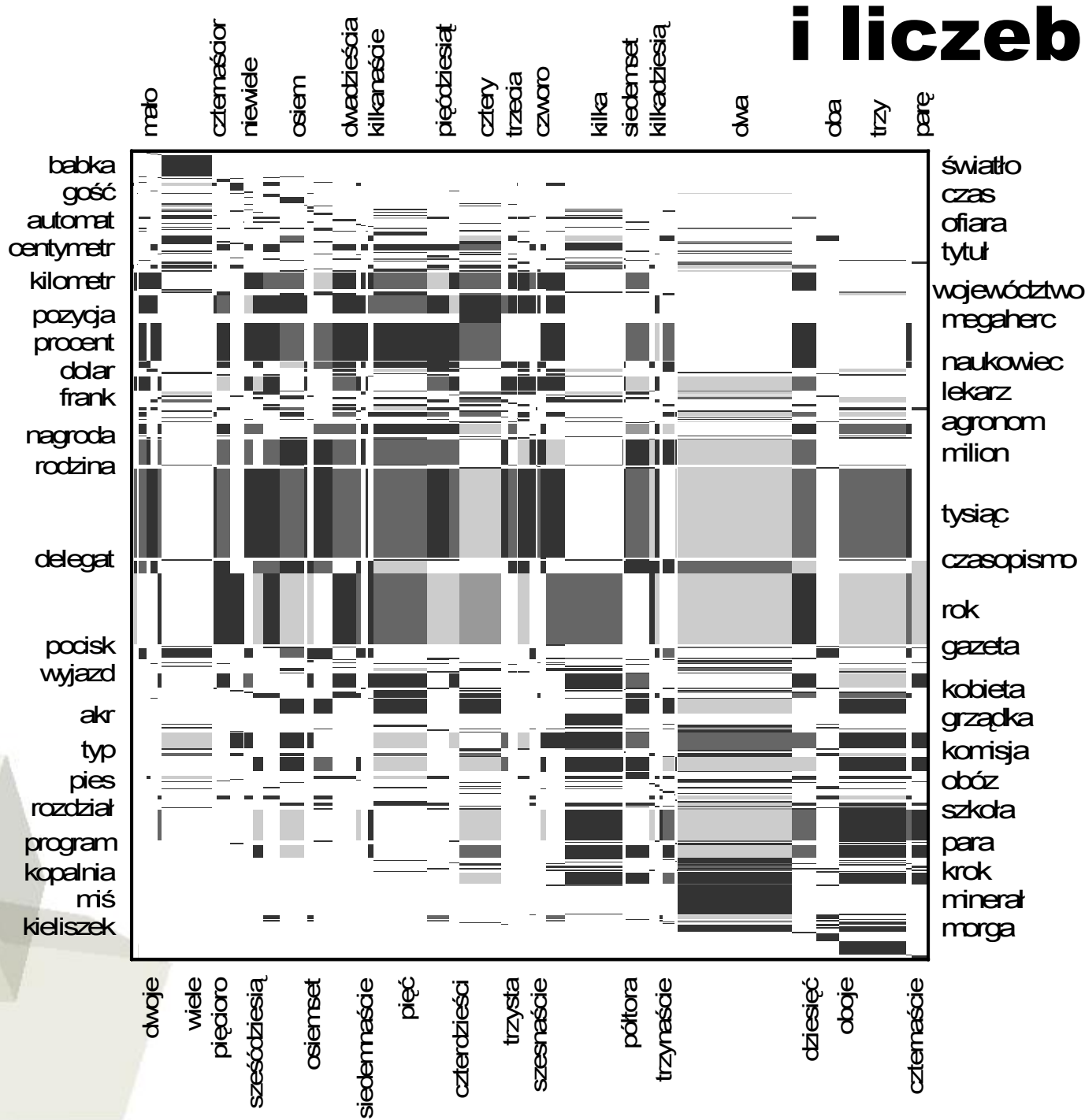
- Narzędzie i dane:
GradeStat i Korpus SFPW
- Mapy nadreprezentacji
- Analiza odpowiedniości (GCA)
- Trzy przykłady
 - klasyfikacja słów nieodmiennych
 - **współwystępowanie rzeczowników i liczebników**
 - deklinacja rzeczowników





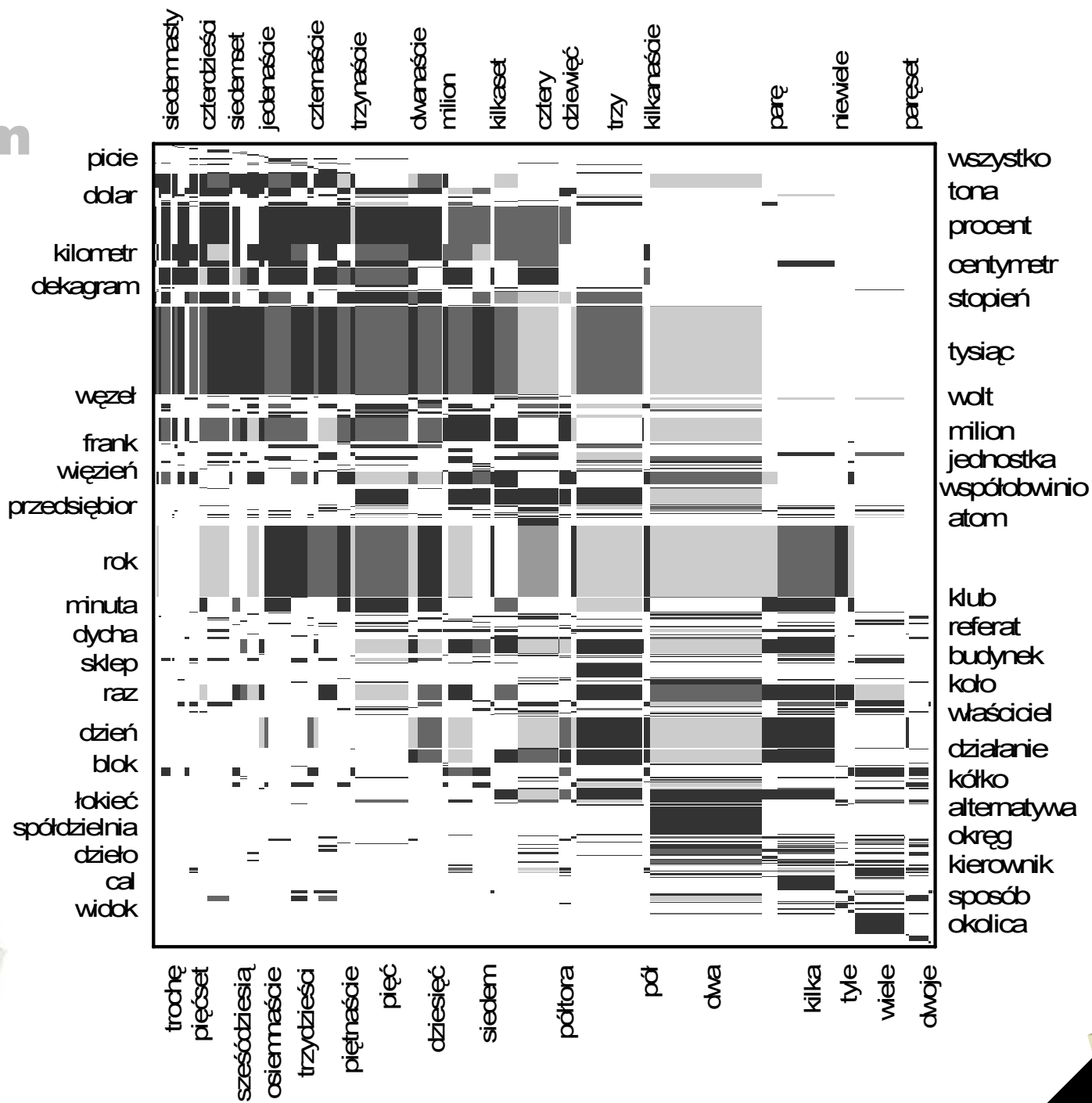
Współwystępowanie rzeczowników i liczebników

Przykład 2



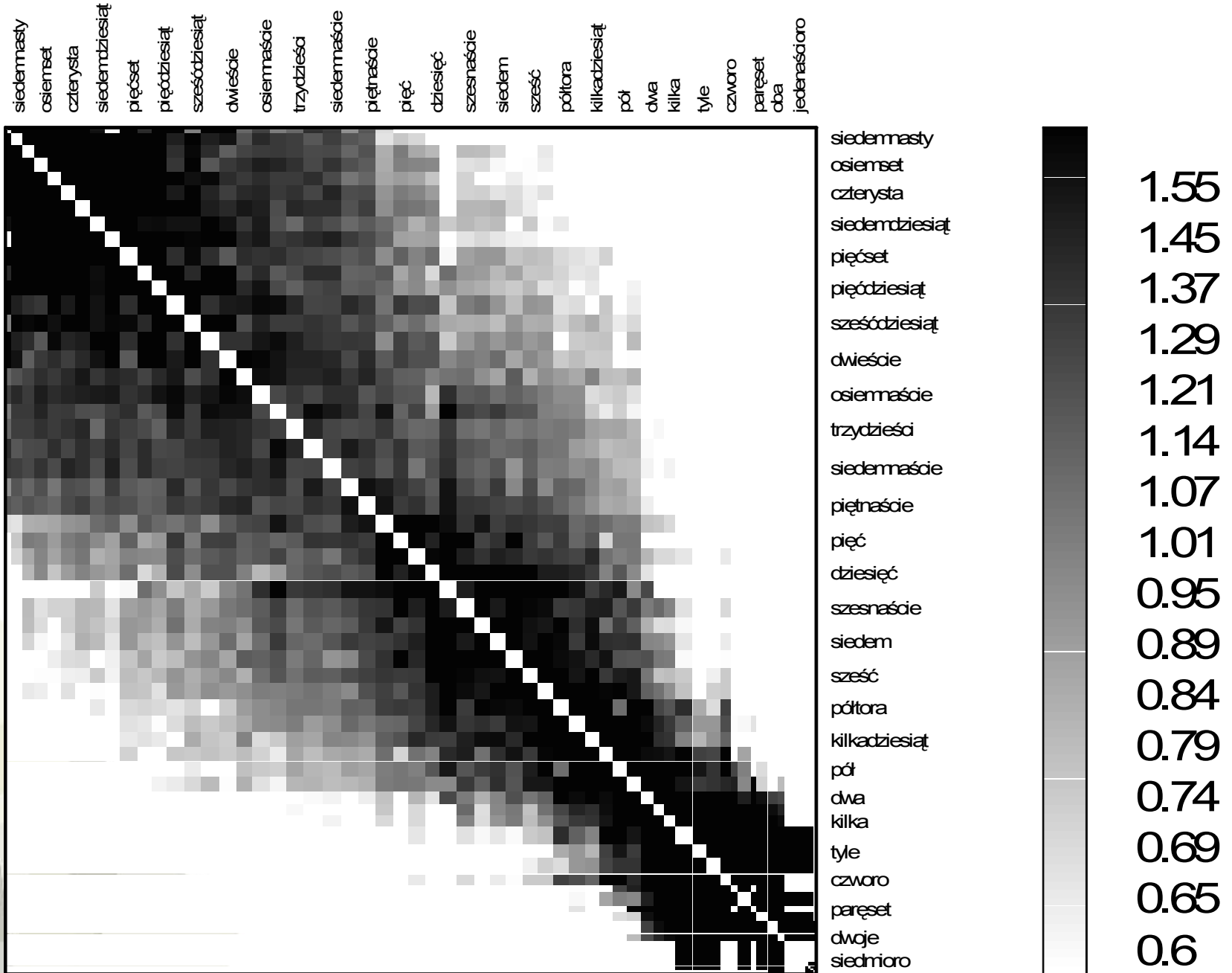


Czy »trochę« jest elementem odstającym?





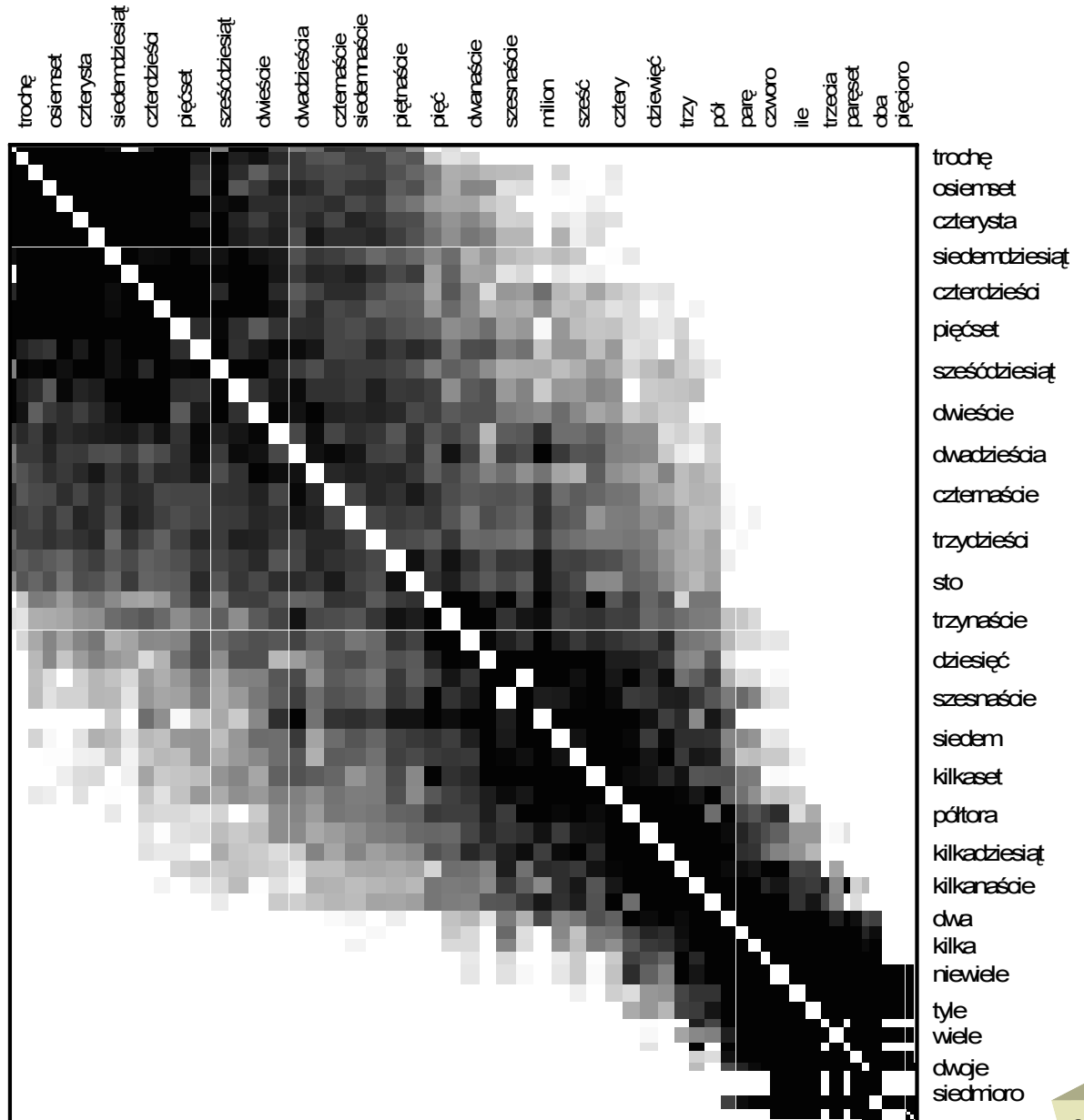
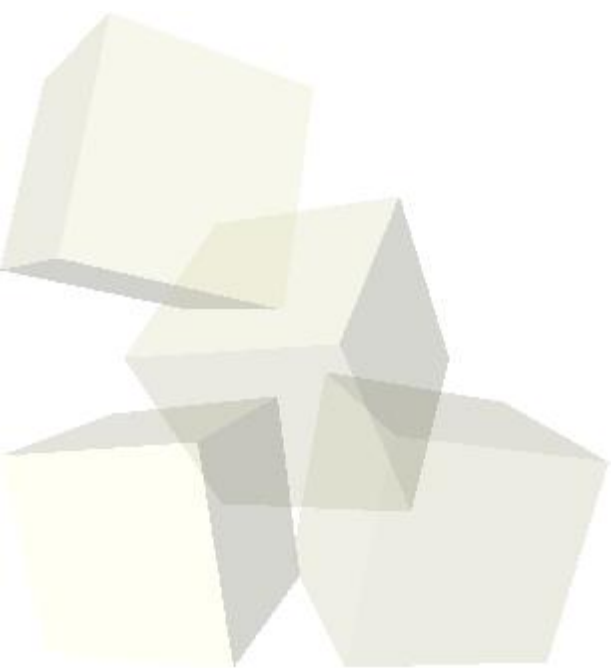
Odstępstwa od regularności dla kolumn





GCA na odstępstwach od regularności

Odstępstwa od regularności
znalezionej przez GCA
również są regularne.

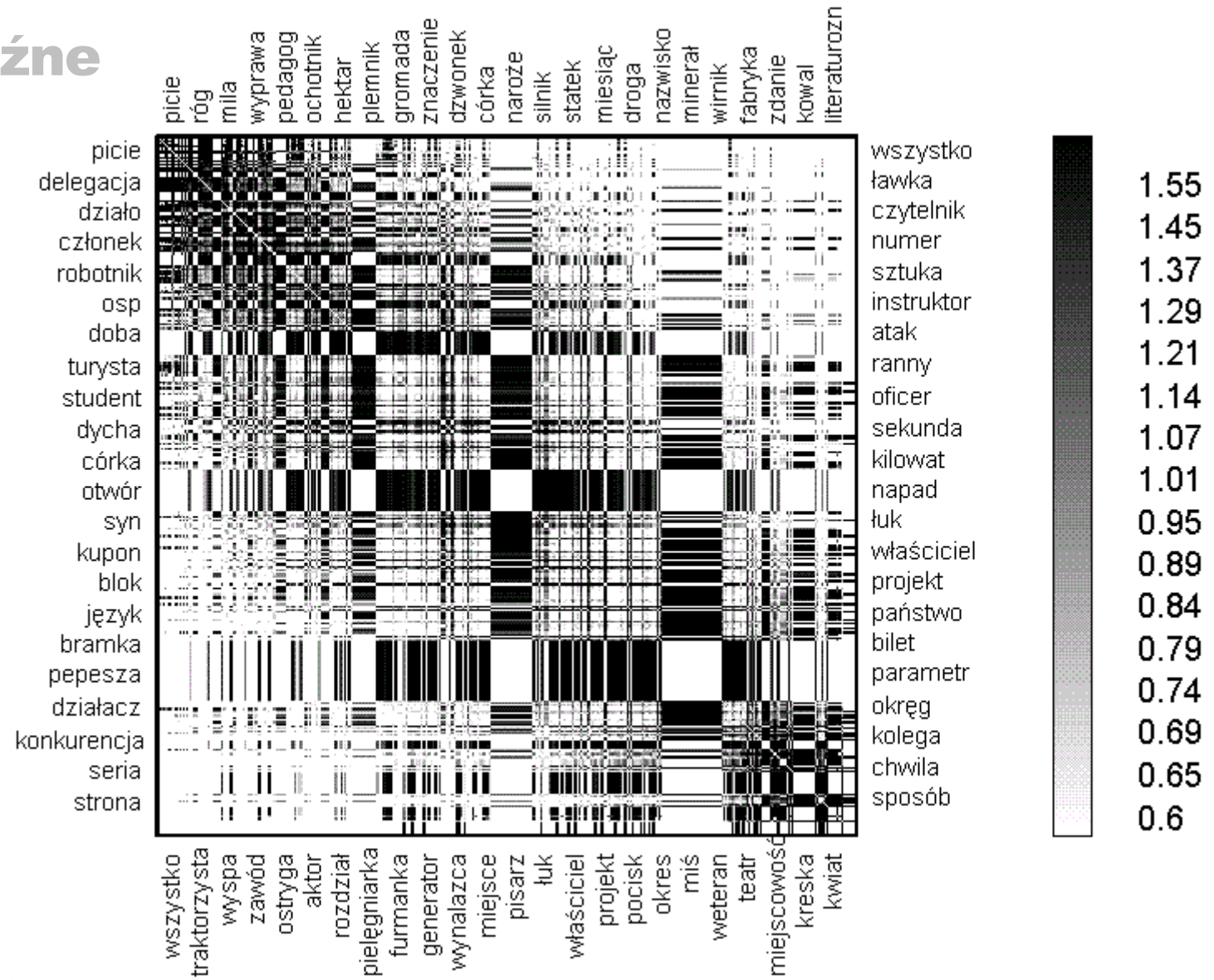


Co widać na mapie odstępstw?

- Dwa skupienia liczebników:
 - ♦ precyzyjne określenia: *pięćset, trzydzieści,*
 - ♦ nieprecyzyjne określenia: *wiele, kilkanaście.*
- *Trochę* jest elementem odstającym: występuje jako skrajny przykład określenia precyzyjnego.
- Słowa o szerokim zastosowaniu (np. *tyle, ile, wiele*) występują najczęściej i w podobnym kontekście, co liczebniki określające małe wielkości.

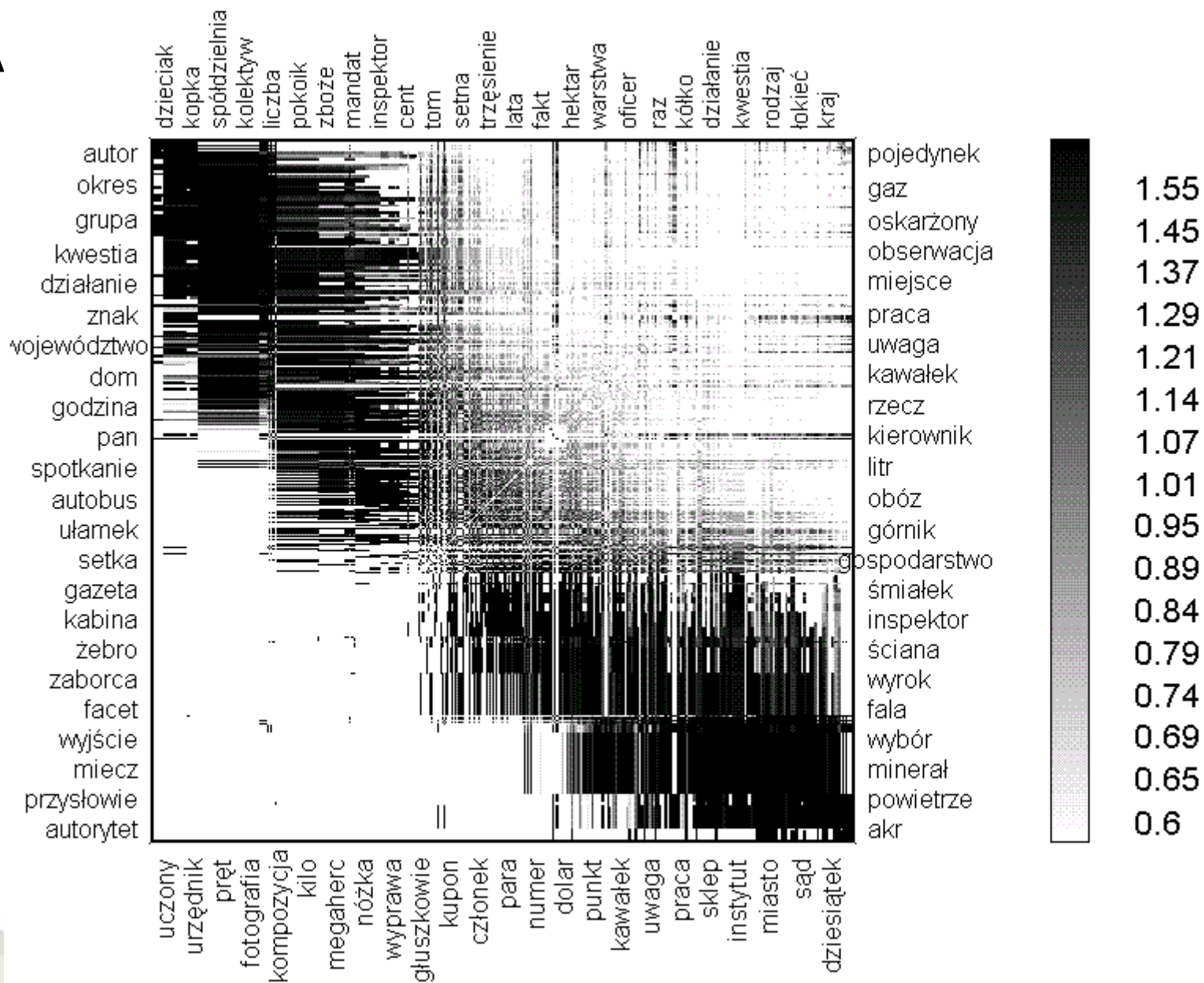
Odstępstwa od regularności dla wierszy

Czy widać wyraźne skupienia?



Odstępstwa wierszy od regularności

po GCA





Co widać na mapie odstępstw?

- Skupienia rzeczowników o podobnym rozkładzie współwystępowania z liczebnikami – można uporządkować liniowo.
- Odstawanie od liniowego porządku (łatwa zmiana miejsca w kolejnych iteracjach) – słowa o wielu znaczeniach.
- Czy współwystępowanie z liczebnikami jest dobrym kryterium klasyfikacji rzeczowników?



- Narzędzie i dane:
GradeStat i Korpus SFPW
- Mapy nadreprezentacji
- Analiza odpowiedniości (GCA)
- Trzy przykłady
 - klasyfikacja słów nieodmiennych
 - współwystępowanie rzeczowników i liczebników
 - **deklinacja rzeczowników**





Deklinacja rzeczowników

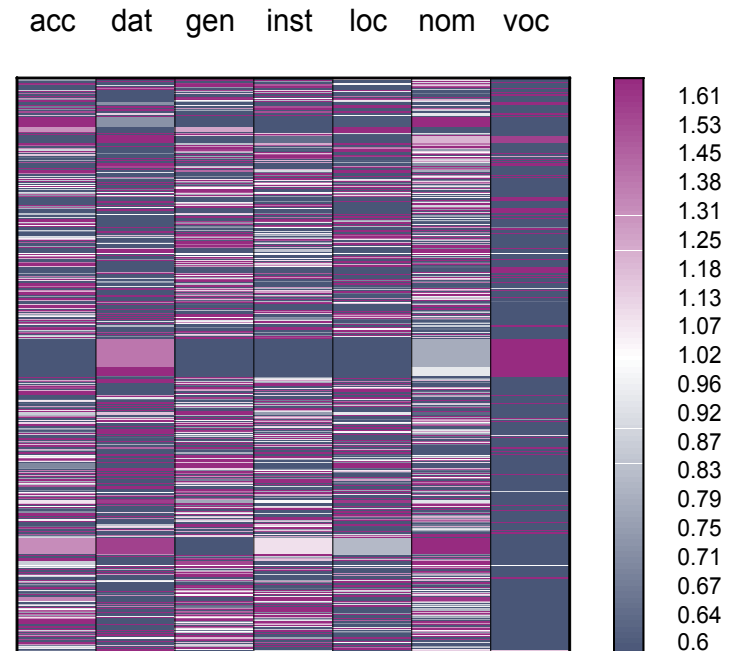
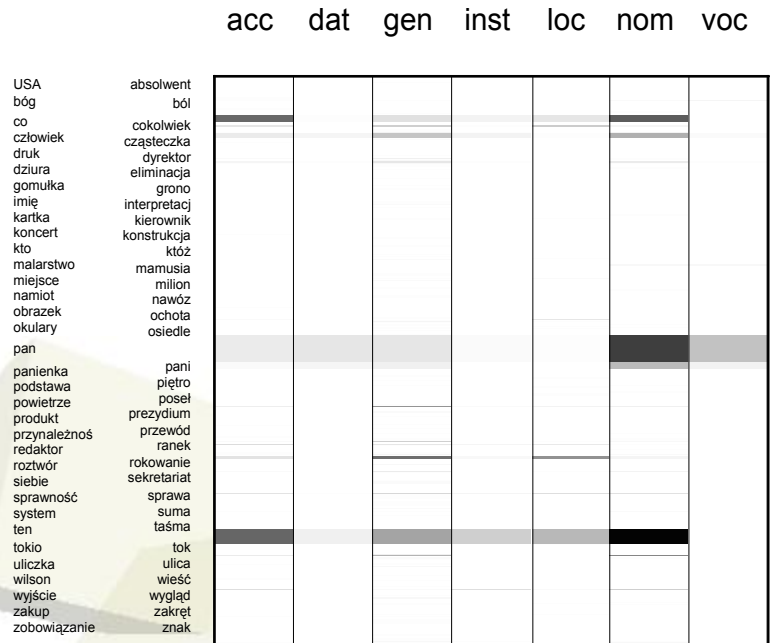
Przykład 3

- z Korpusu zostały wyciągnięte informacje o częstościach występowania rzeczowników we wszystkich przypadkach
- przypadki zostały rozdzielone do oddzielnych grup, co w uproszczeniu oznacza, że brzeg ignorowana jest informacja o tym, jak często wystąpił rzeczownik w danym przypadku w całym korpusie; przypadki są więc potraktowane jako równie ważne przy uporządkowaniu tabeli
- czy w latach 60 wszystkie przypadki były rzeczywiście równoważne?



Deklinacja rzeczowników

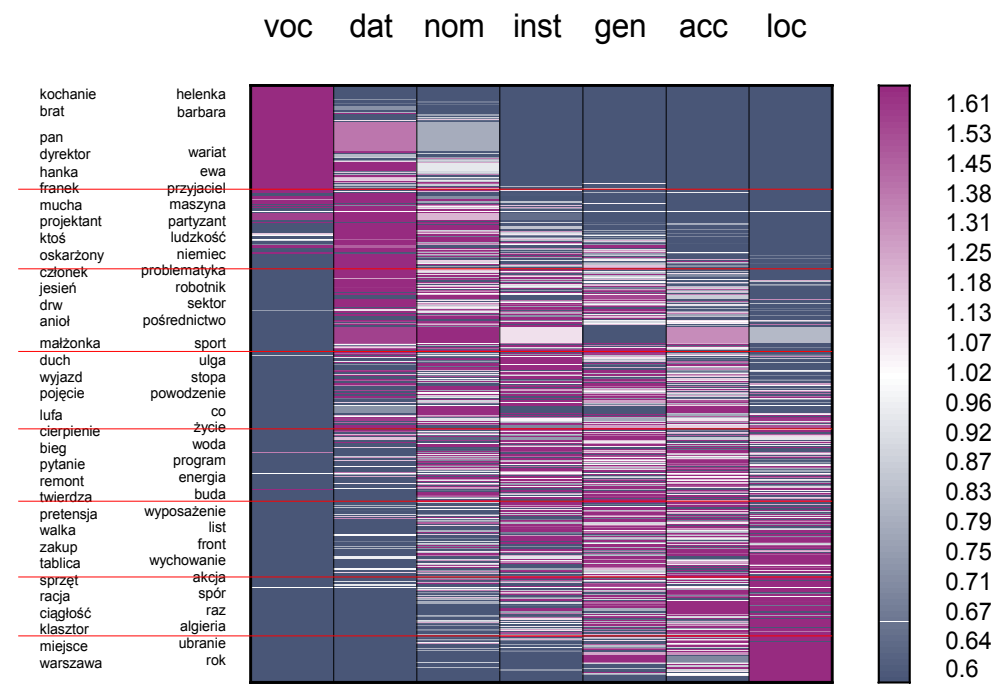
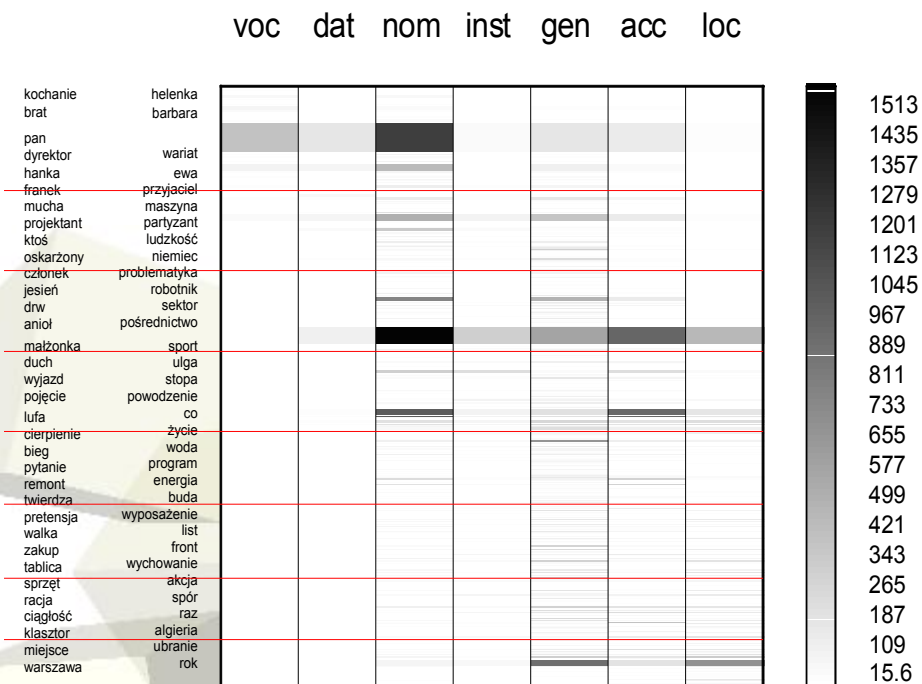
- mapy danych surowych i nadreprezentacji przed posortowaniem zgodnie z GCA





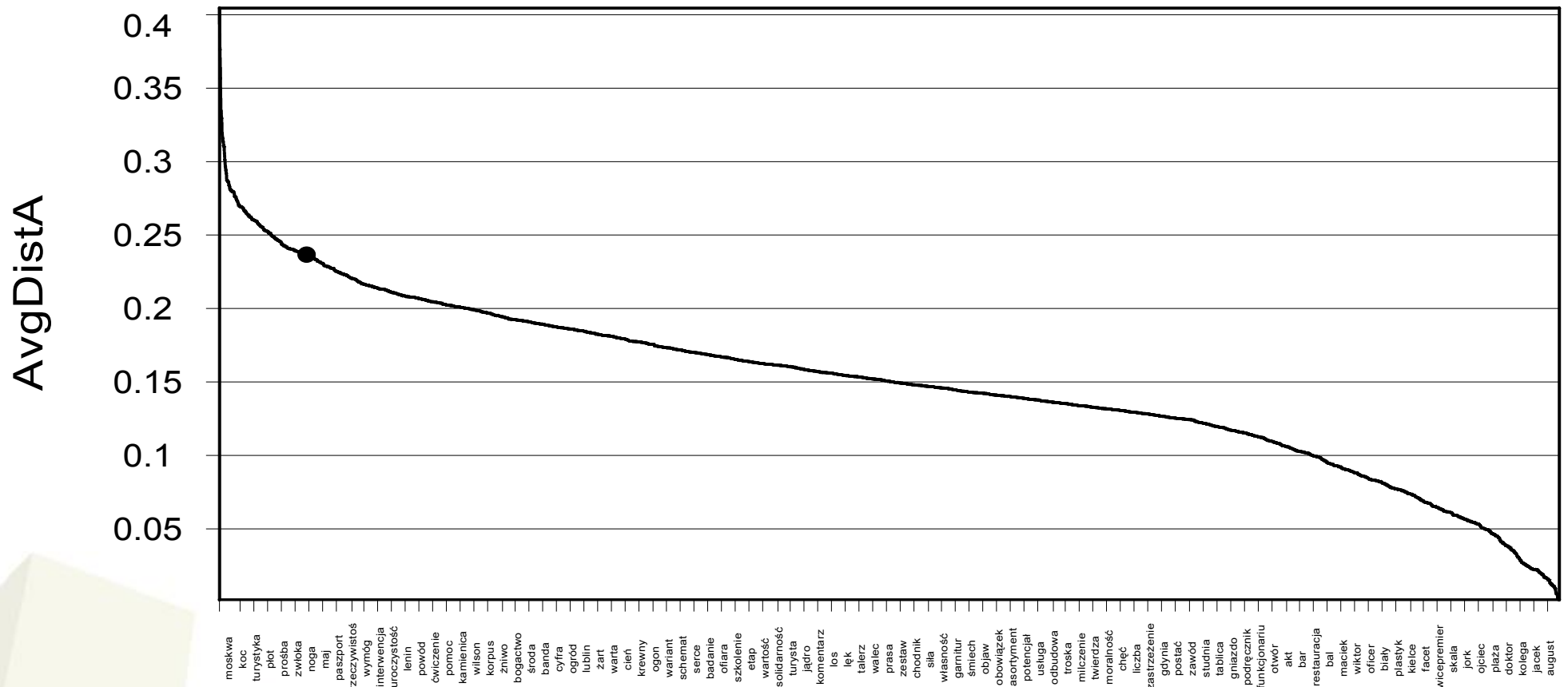
Deklinacja rzeczowników

- mapy danych surowych i nadreprezentacji posortowane zgodnie z GCA, wysokie zróżnicowanie ($\rho^* = 0.69$)
- wciąż są elementy odstające





Deklinacja rzeczowników

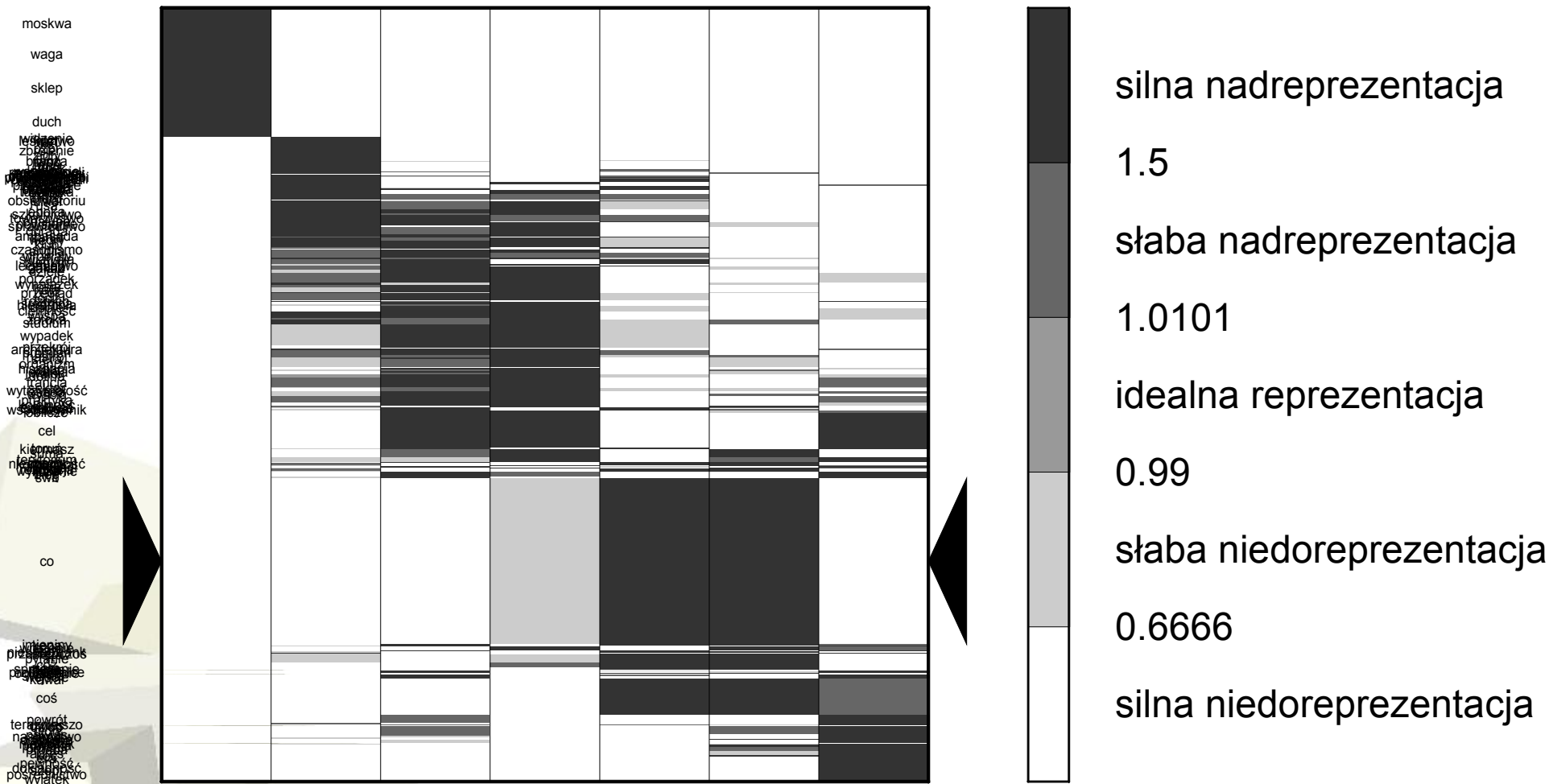


- rzeczowniki na lewo od czarnej kropki zostały przeniesione do grupy elementów odstających (czarna kropka to również element odstający, wyraz co)



Elementy odstające

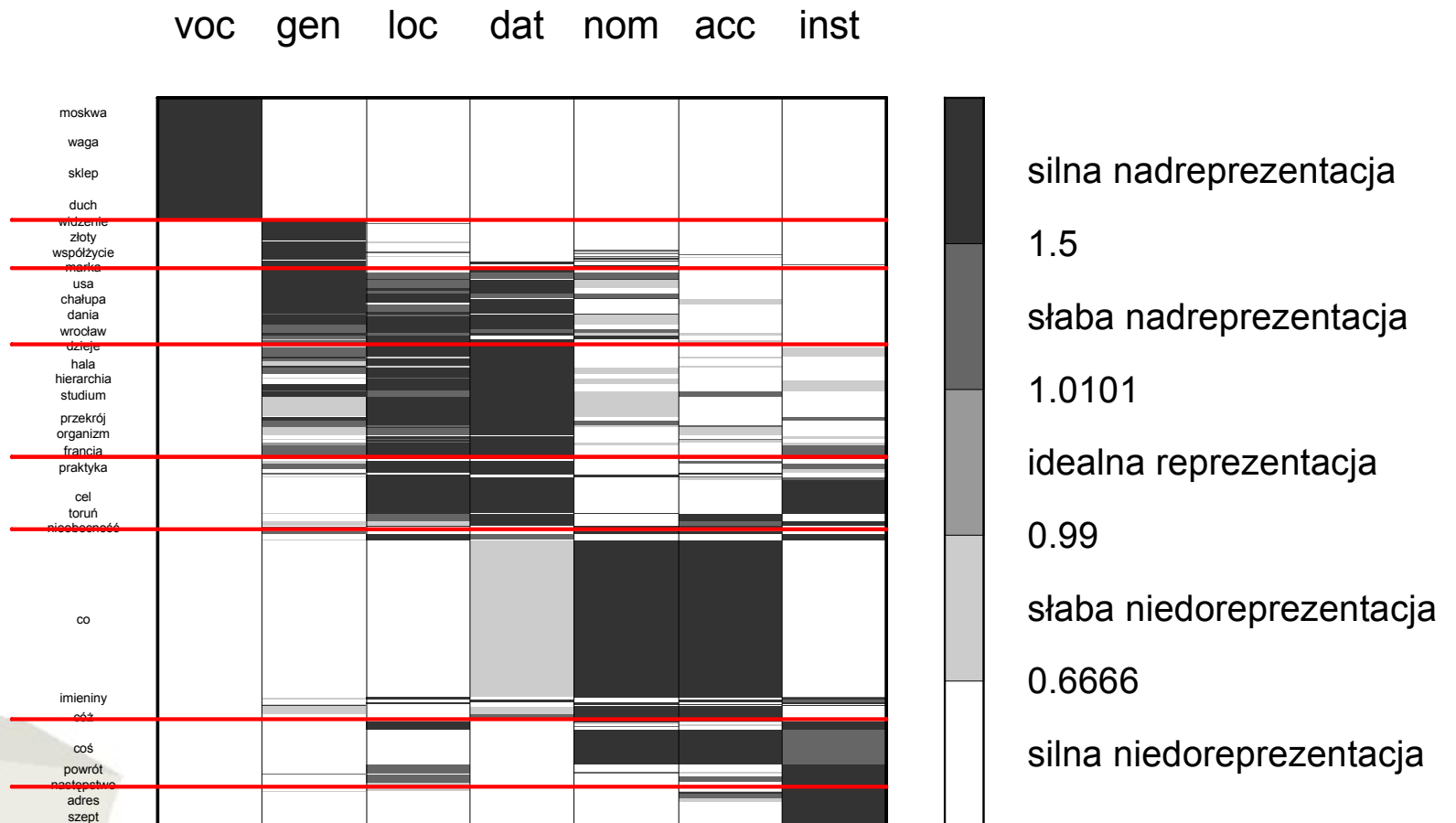
voc gen loc dat nom acc inst





Podział na 8 skupień

- skupienie 1 – po prostu rzeczowniki, które wystąpiły chociaż raz w wołaczu: *moskwa*, *waga*, *sklep*, *duch*





Ustawienie po GCA – bez wołacza

inst acc nom loc dat gen

	inst	acc	nom	loc	dat	gen
wyjątek			szept			
dokładność			era			
koc			prośba			
powieka			gest			
niewolnik			żądanie			
nawias			dziś			
dziób			terazniejszo			
owies			powrót			
wejście			plot			
przekształce			coś			
kawał			sprzężenie			
staw			kepka			
świt			imieniny			
msza			sam			
pytanie			niespodziank			
bieg			pech			
jury			cóż			
rezolucja			linka			
kiermasz			toruń			
zeszyt			nieobecność			
częstość			współczynnik			
zwłoka			waga			
suma			cel			
praktyka			sklep			
wyścig			francja			
dolina			ghana			
dania			plenum			
rozruch			osada			
dzieje			średnica			
łódź			porządek			
wyspa			anglia			
hiszpania			moskwa			
opole			nastrój			
sektor			studium			
willa			jezdnia			
zakaz			walor			
klub			obserwatori			
hierarchia			towarzystwo			
świątynia			przeгляд			
wynalazek			hala			
czaszka			junior			
ambasada			powstanie			
biuro			usa			
szkolnictwo			przewóz			
marka			planowanie			
ekonomia			departament			
mo			turytyka			
nato			wynagrodzeni			
prawdopodobni			reuter			
oświata			oddziaływani			
egzekutywa			tysiąclecie			
placa			użytek			
rwpg			tona			
złoty			funt			
zbrojenie			kc			
widzenie			kpzr			
pośrednictwo						
pewność						
adres						
ołówek						
następstwo						
spód						
odrobina						
winda						
buda						
ogłoszenie						
pięta						
minus						
przestępczoś						
duch						
co						
wyliczenie						
wrażenie						
komunikat						
tokio						
oblicze						
terytorium						
kolejność						
wypadek						
ciemność						
promień						
czasopismo						
sfera						
chałupa						
wytrzymałość						
wrocław						
architektura						
przekrój						
organizm						
rosja						
ewolucja						
kongo						
zatoka						
epoka						
węgry						
połów						
lecznictwo						
obrada						
sprawiedliwo						
brygada						
csrs						
politechnika						
rzeczpospoli						
współzycie						
kp						
mrn						
branża						
leśnictwo						
pzpr						
wrn						



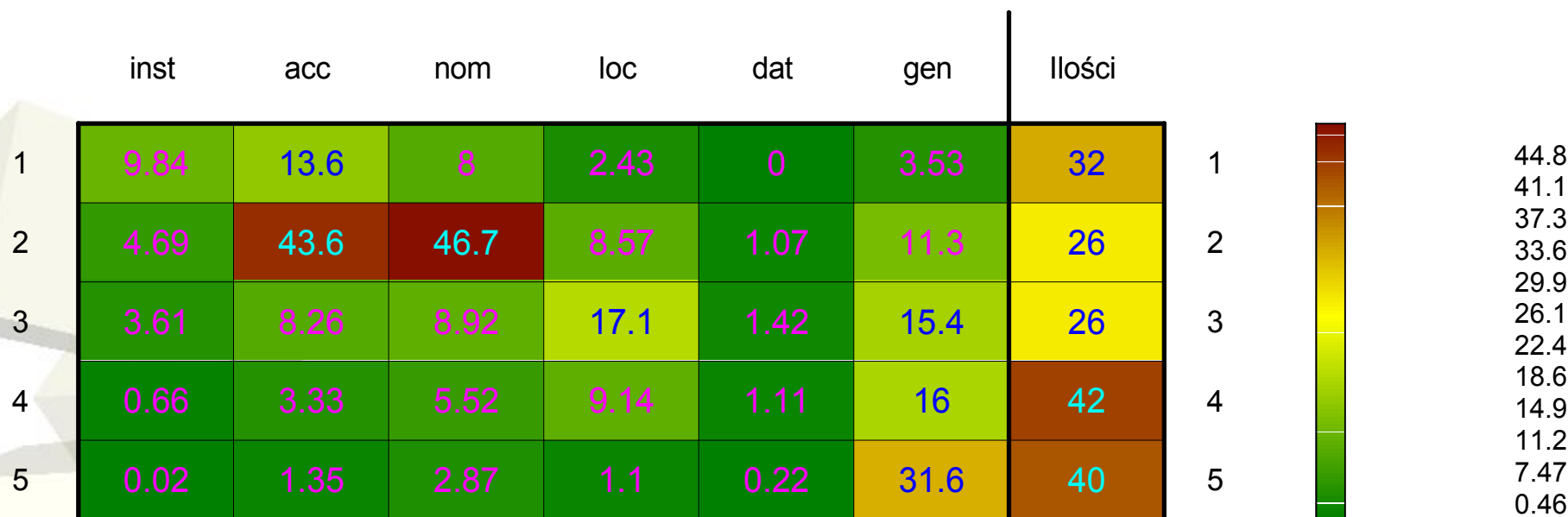
Czyżby za ułożeniem stała opozycja
rzeczywista informacja - nowomowa?

- ♦ skupienie 1 – *wyjątek, pośrednictwo, dokładność, adres, prośba, żądanie, wejście, teraźniejszość, ogłoszenie, niewolnik, coś*
- ♦ skupienie 5 – *biuro, obrada, sprawiedliwość, planowanie, departament, mo, csrs, turystyka, nato, rzepospolita, reuter, oświata, kp, współżycie, mrn, rwpg, pzpr, złoty, zbrojenie, kc, kpzr...*



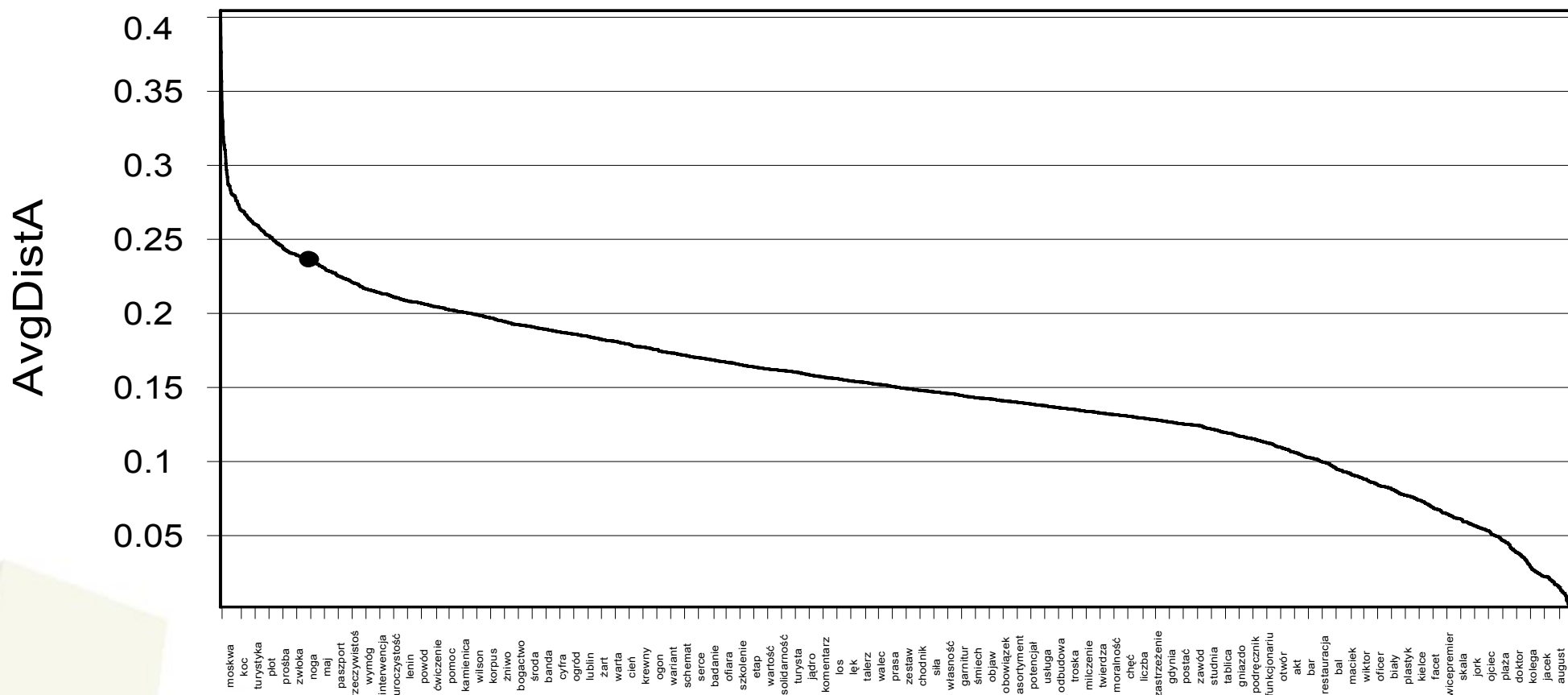
Informacja a nowomowa

- rzeczowniki ze skupienia 1 (*rzeczywiście informujące?*) częściej występowały w narzędniku
- rzeczowniki ze skupienia 5 (*nowomowa?*) znacznie częściej występowały w dopełniaczu (*planowania, departamentu, kc...*)





Deklinacja rzeczowników



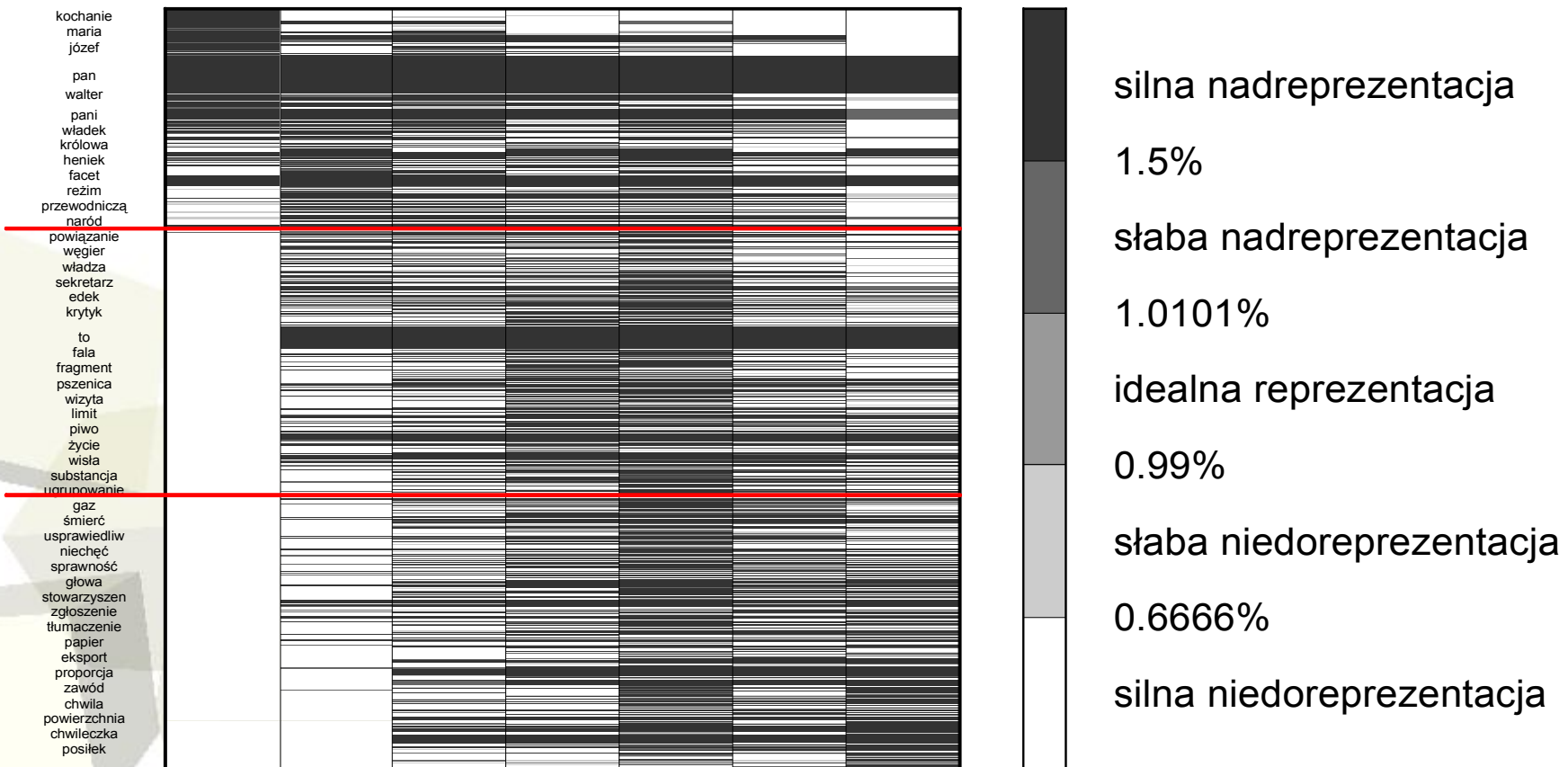
- druga grupa to rzeczowniki na prawo od czarnej kropki (zgodne z wyłonioną po GCA regularnością w macierzy)



Rzeczowniki z grupy „regularnej”

- także tym razem wołacz najsilniej wpłynął na kolejność wierszy i kolumn (skupienie 2 i 3 to rzeczowniki, które nigdy nie wystąpiły w wołaczu)

voc dat nom inst gen acc loc

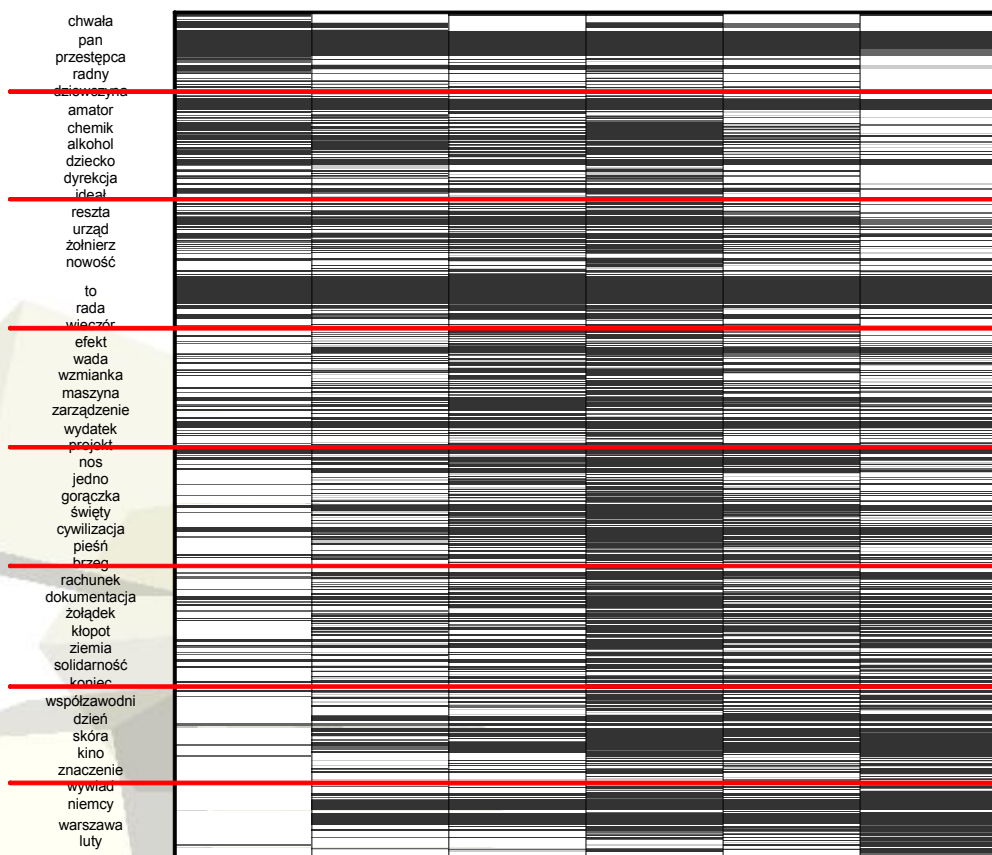




Wołacz wyłączony

- po wyłączeniu wołacza i przeprowadzeniu GCA ustaliła się identyczna kolejność kolumn, $\rho^* = 0.557$

dat nom inst gen acc loc



silna nadreprezentacja

1.5%

słaba nadreprezentacja

1.0101%

idealna reprezentacja

0.99%

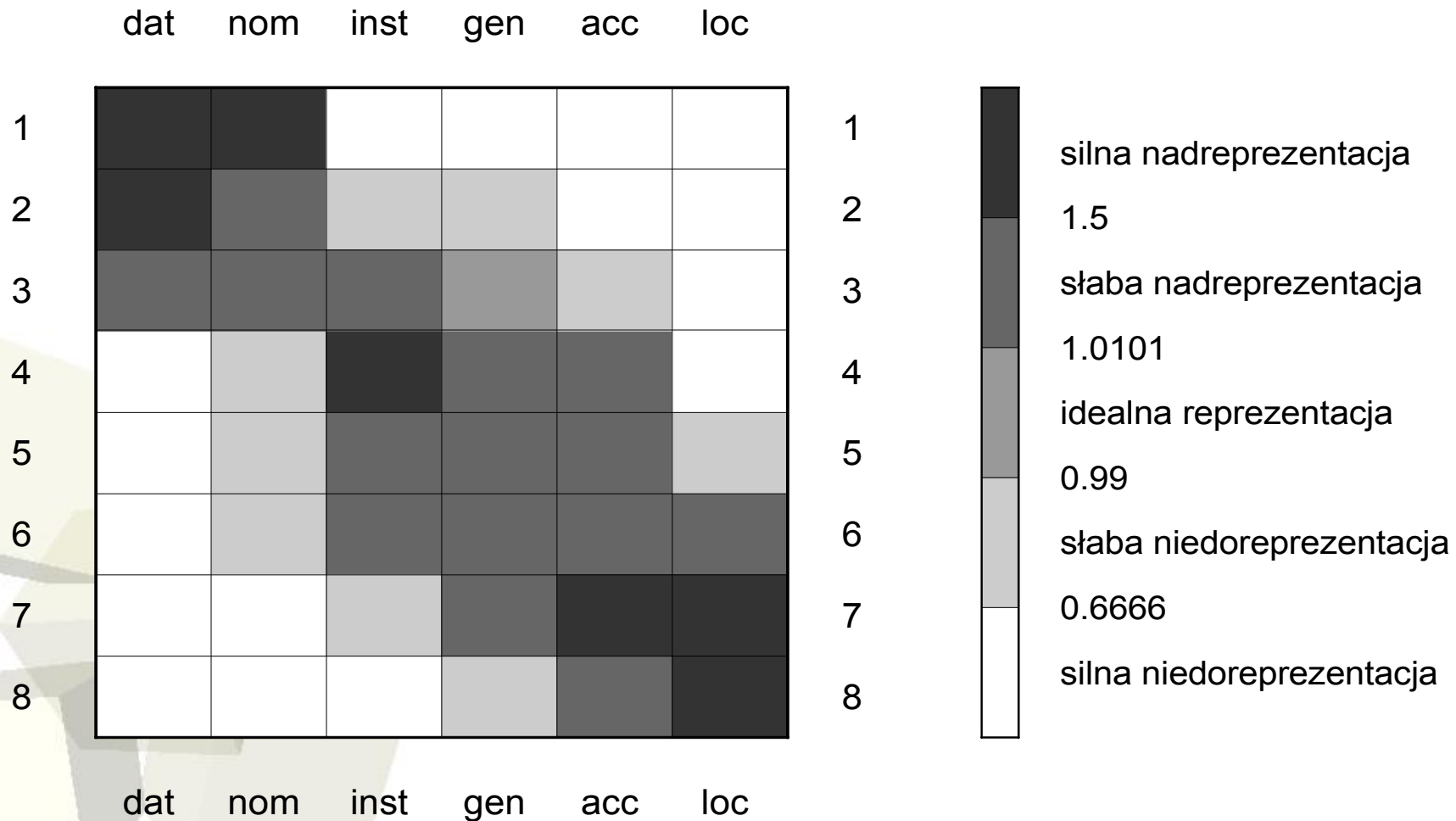
słaba niedoreprezentacja

0.6666%

silna niedoreprezentacja

Nadrepzentacje dla agregacji

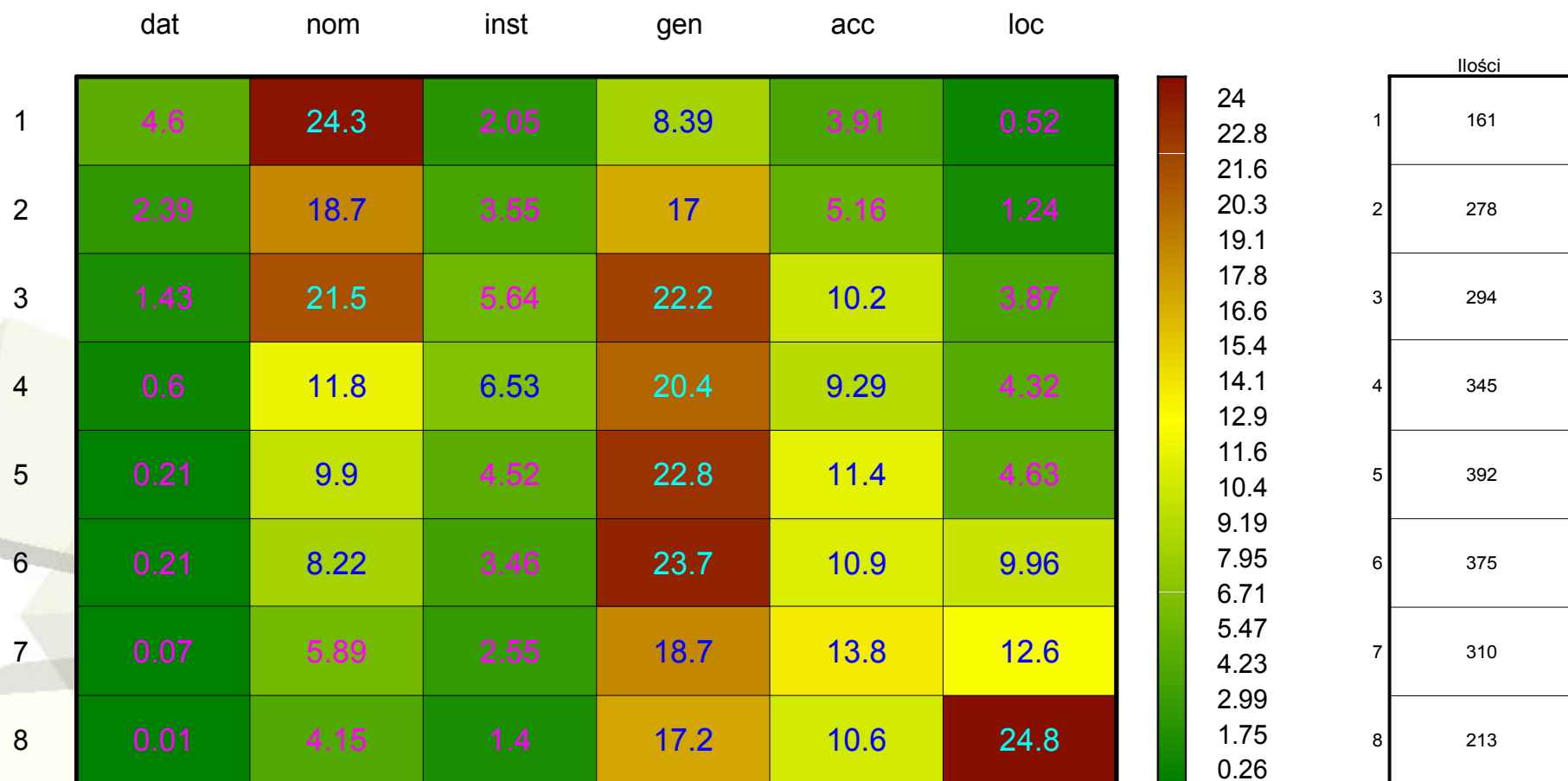
- czym różni się 8 wydzielonych skupień





Mapa danych surowych

- średnie częstotliwości wystąpienia danego przypadku dla rzeczowników w każdym ze skupień





Podsumowanie analizy przykładu 3

- skupienie 1 – pan, pani, siebie, nikt, ojciec, minister, naród, członek
- skupienie 2 – człowiek, dziecko, państwo, kobieta, zmiana
- skupienie 3 – to, tysiąc, rada, problem, rząd, organizacja, liczba
- skupienie 4 – sprawa, wszystko, życie, pomoc, siła
- skupienie 5 – praca, nic, oko, woda, szkoła, rzecz
- skupienie 6 – kraj, związek, świat, miasto, ręka, warunek, głowa, ziemia
- skupienie 7 – raz, dzień, chwila, przykład, droga, dom, polska, sposób, strona
- skupienie 8 – rok, czas, miejsce, godzina, okres



Dziękujemy!

Zapraszamy na nasze strony:

<http://korpus.pl>

<http://gradeostat.ipipan.waw.pl>