Quantitative features in science-popular texts: text difficulty

Kseniia Samoylenko

Hypothesis

- texts about physics, math and space (‘exact’ science) are more difficult, than about humanities
- difficulty of media sources is differents

Data

Sources

N+1 π GT

Индикатор

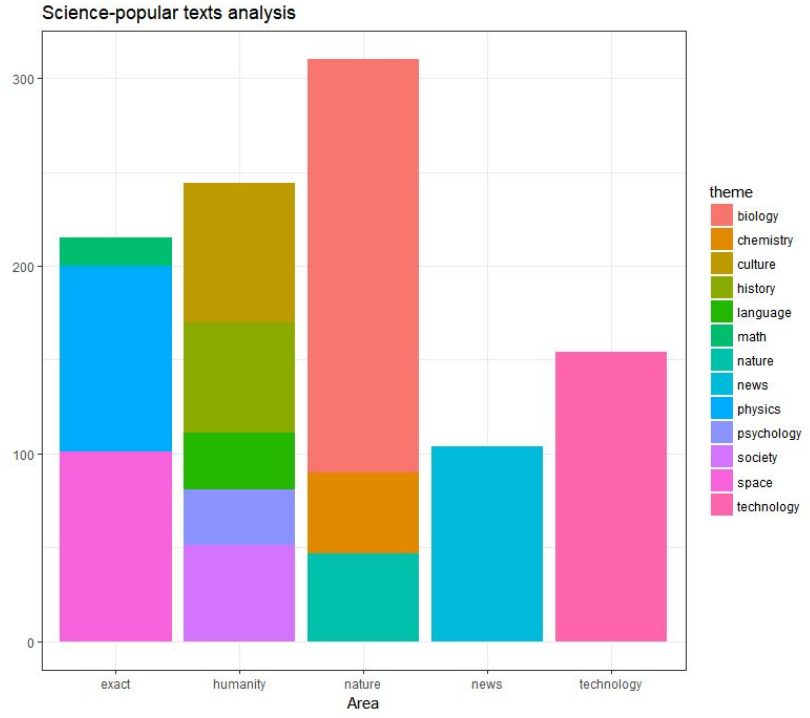
ПОЛИТ.РУ

научно-популярный портал Чердак

Variables

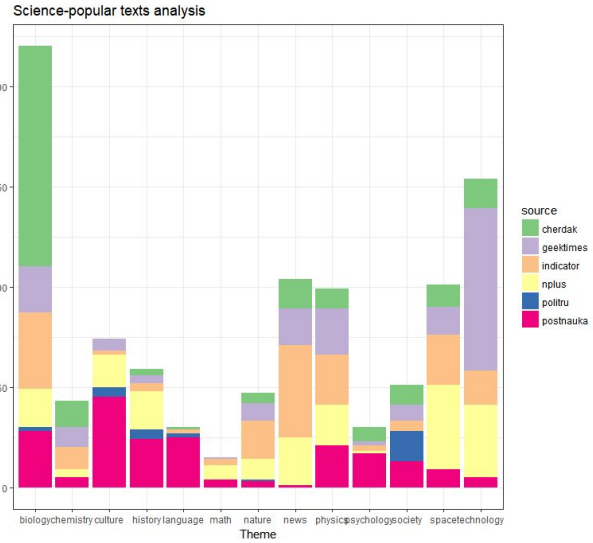
- quantitative features:
- readability indexes
- dictionaries
- POS-tags

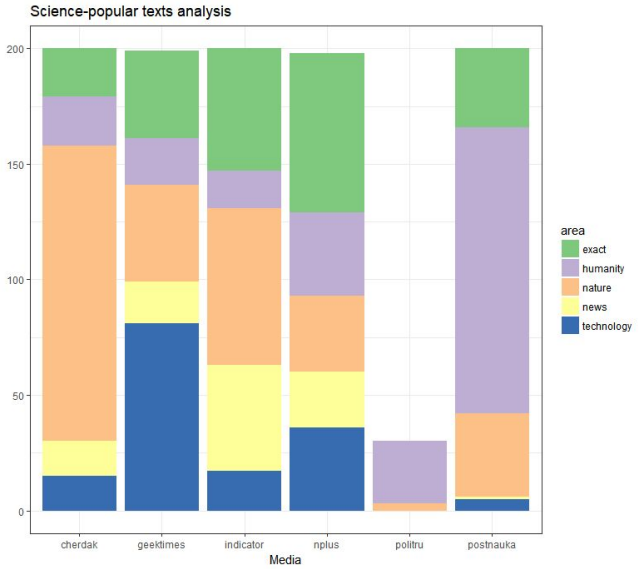
Areas and thems

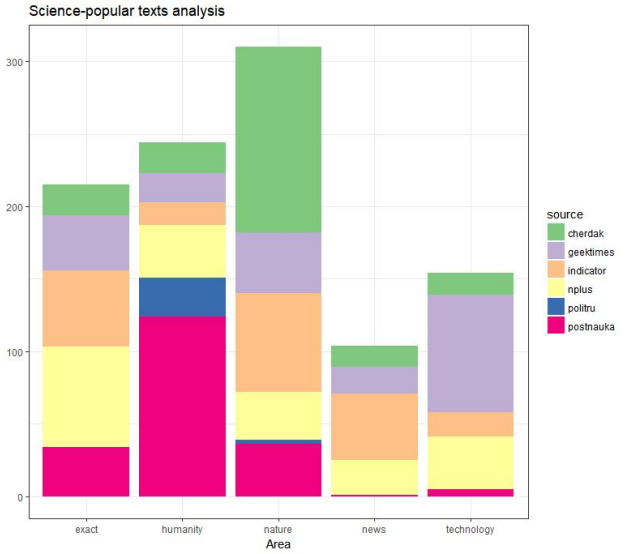


Descriptive analysis

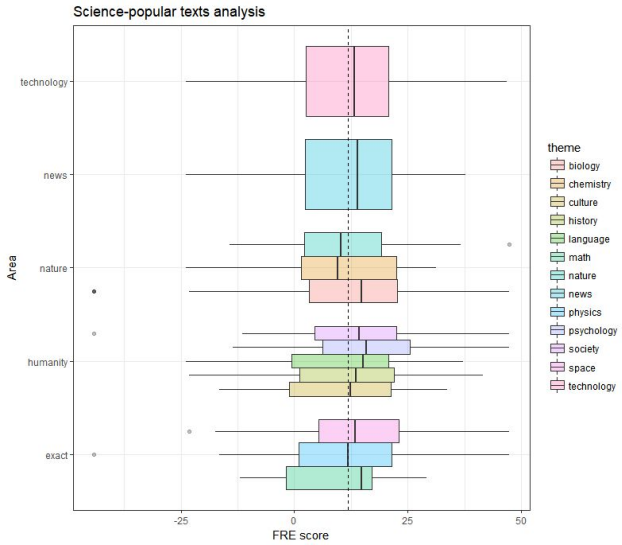
Thems and sources distribution

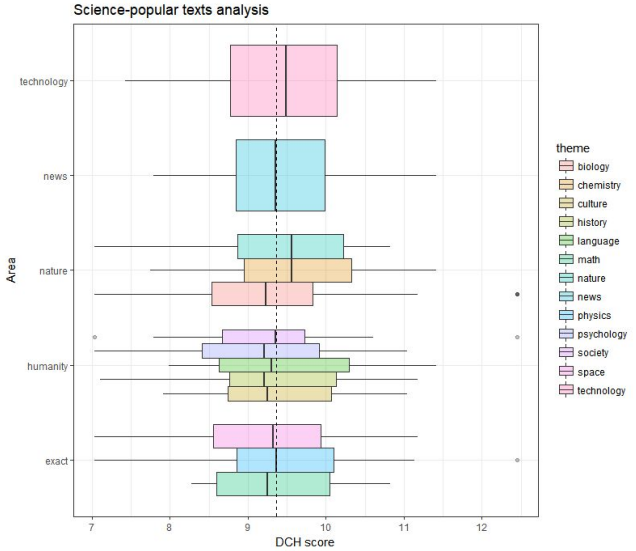


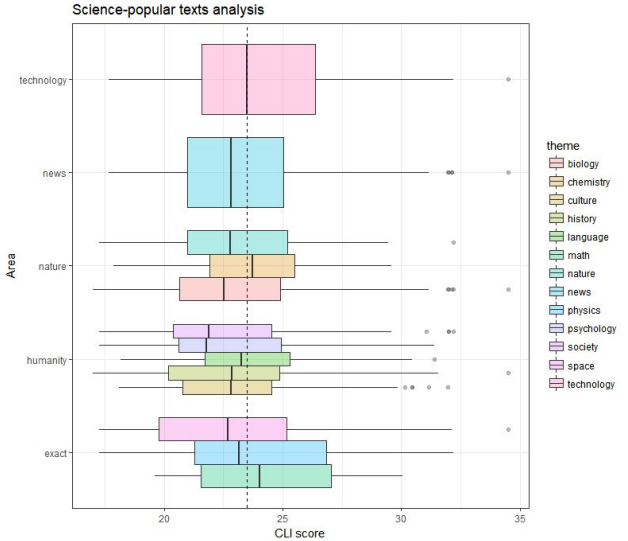




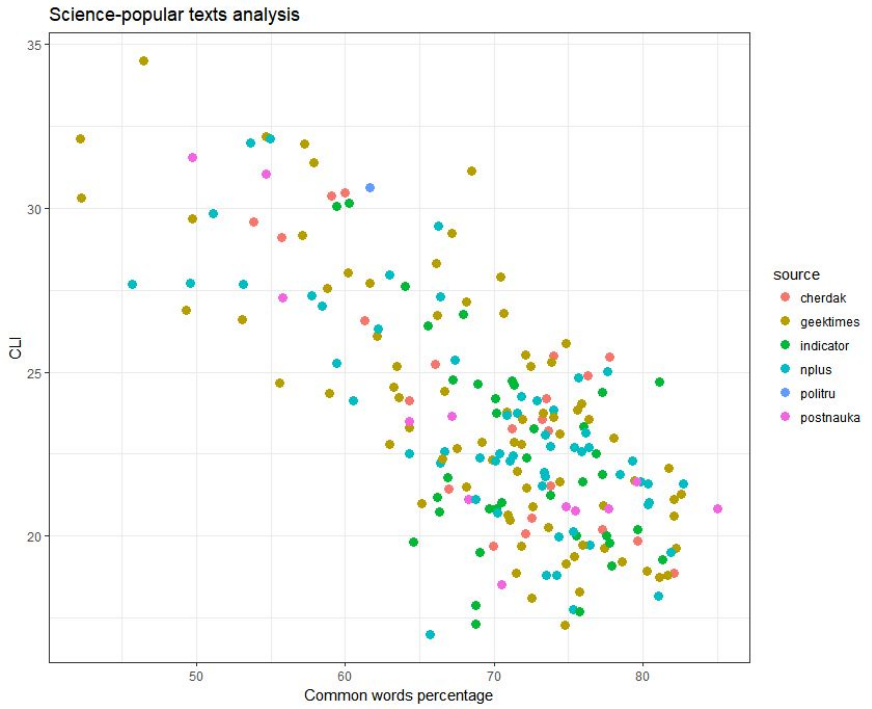
Difficulty distribution

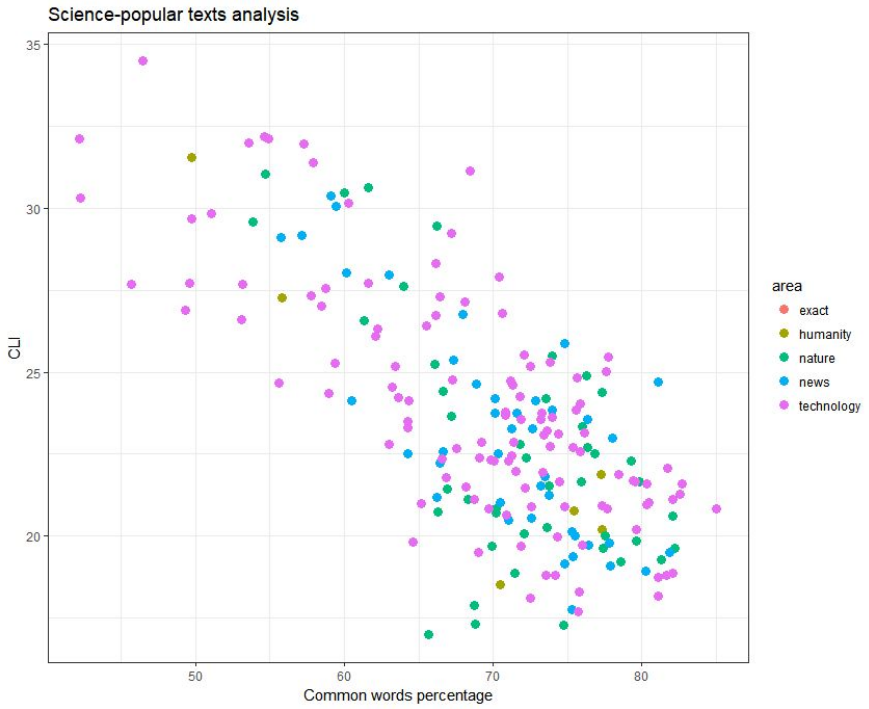




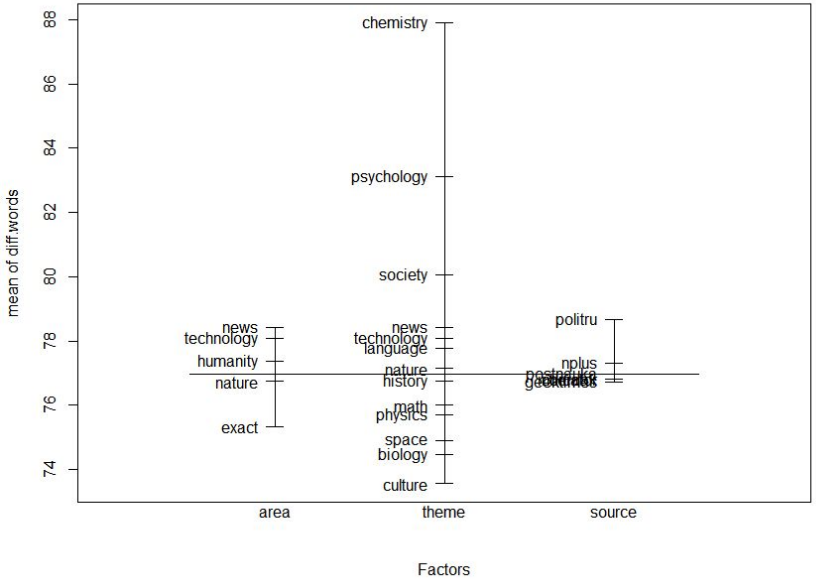


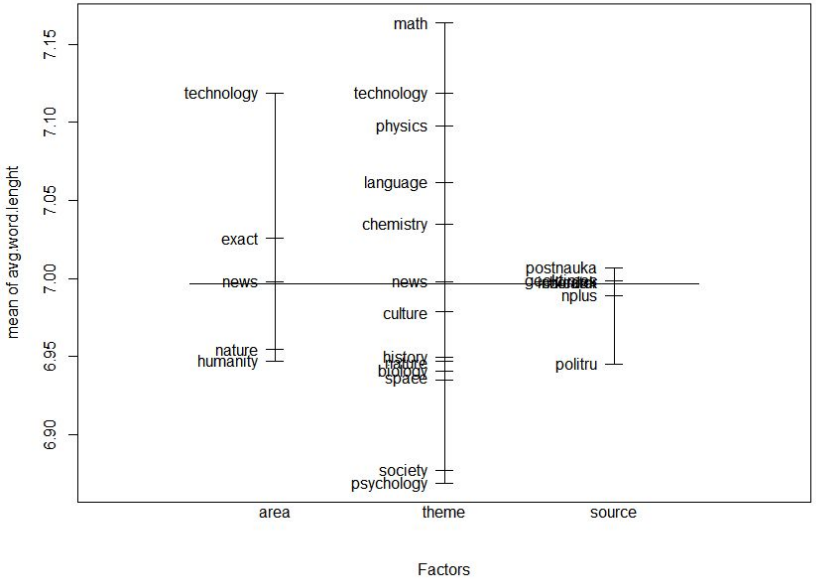
Common words in science and media



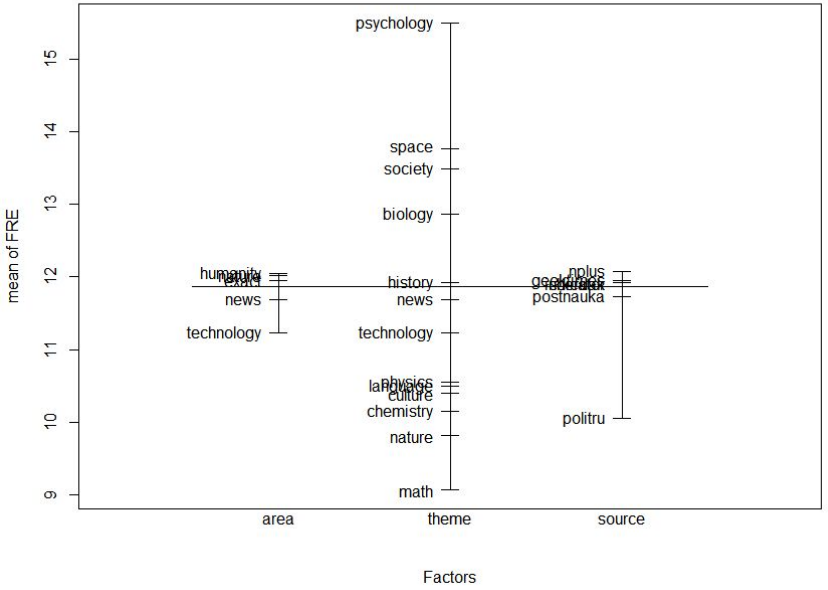


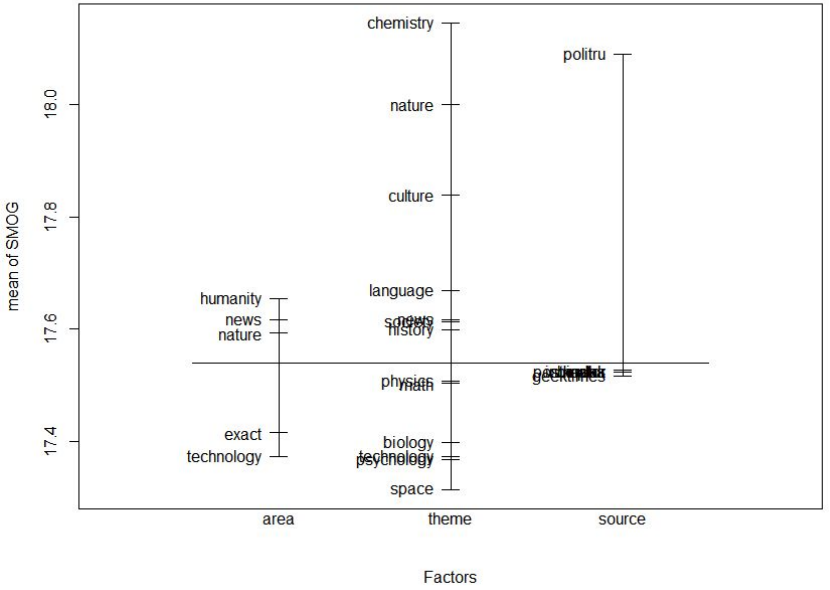
Data summary



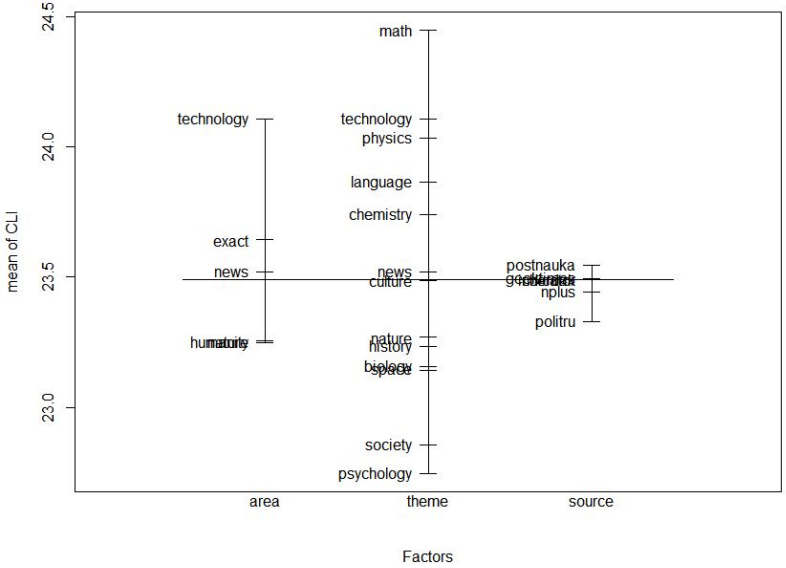


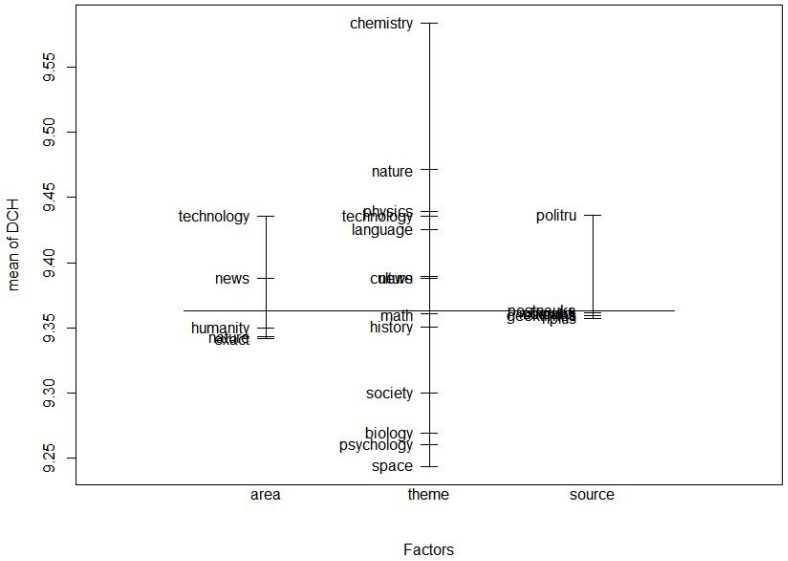
Data summary



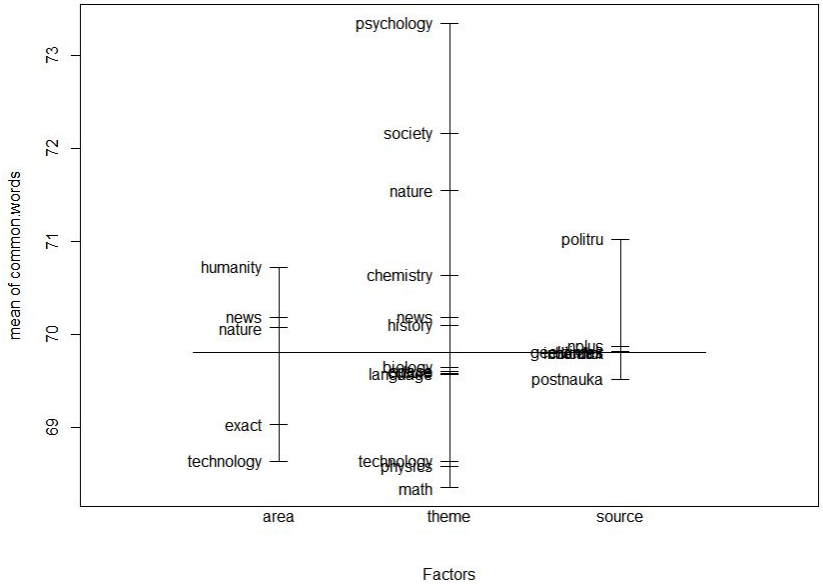


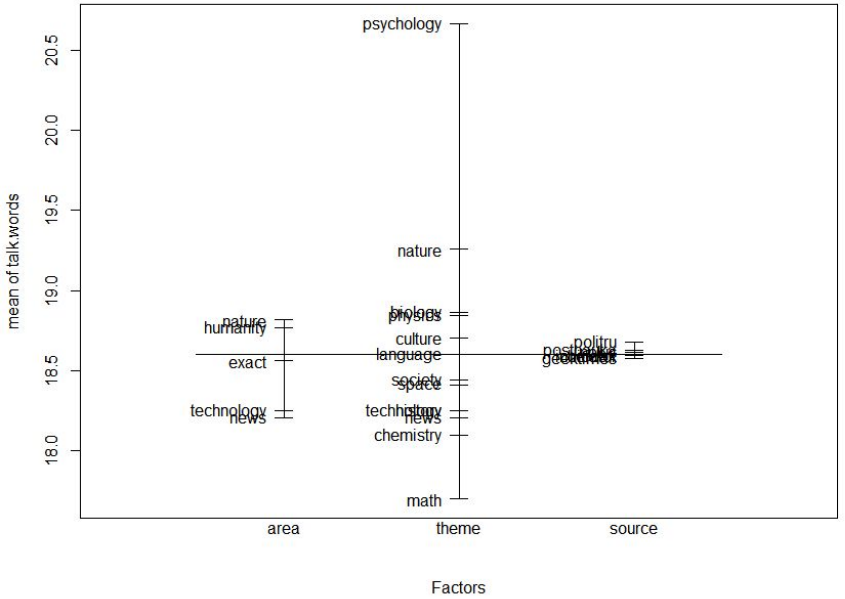
Data summary



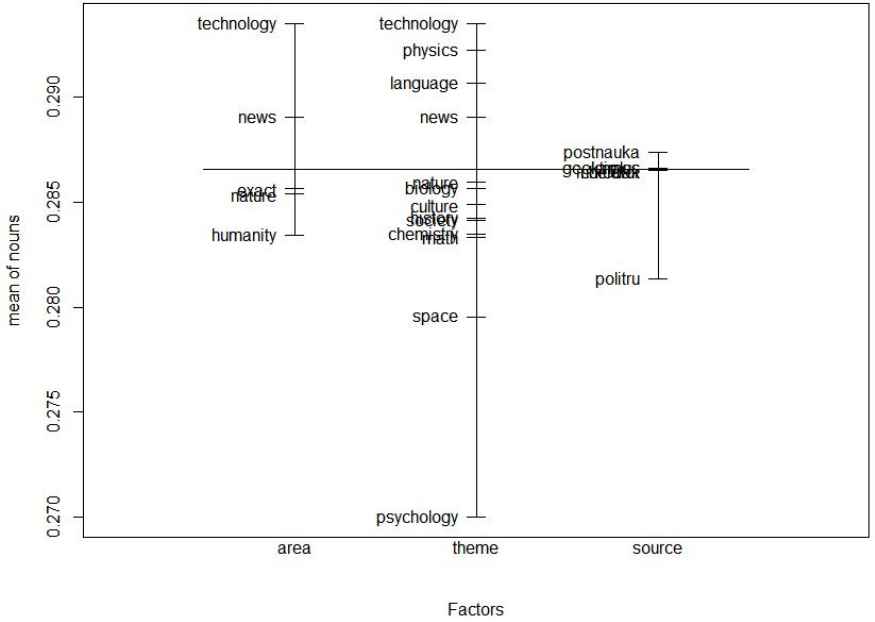


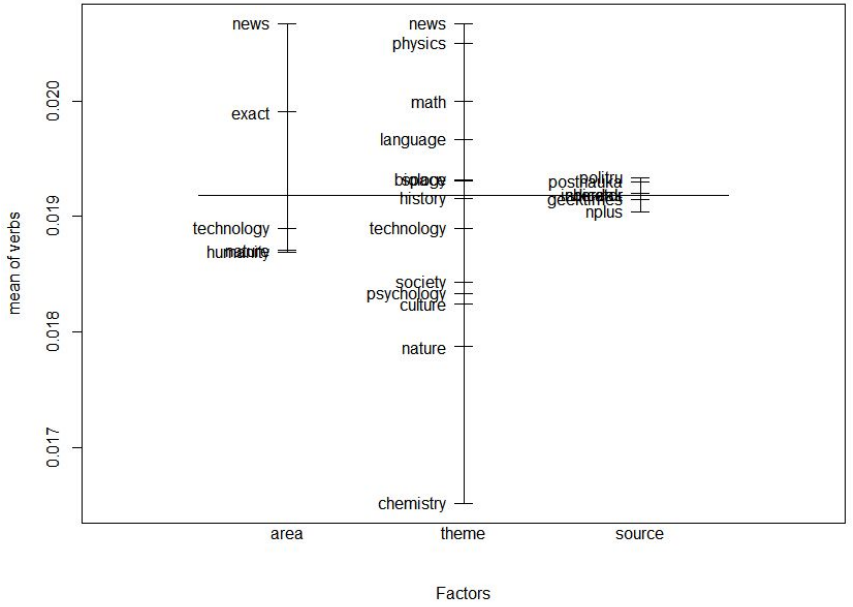
Data summary



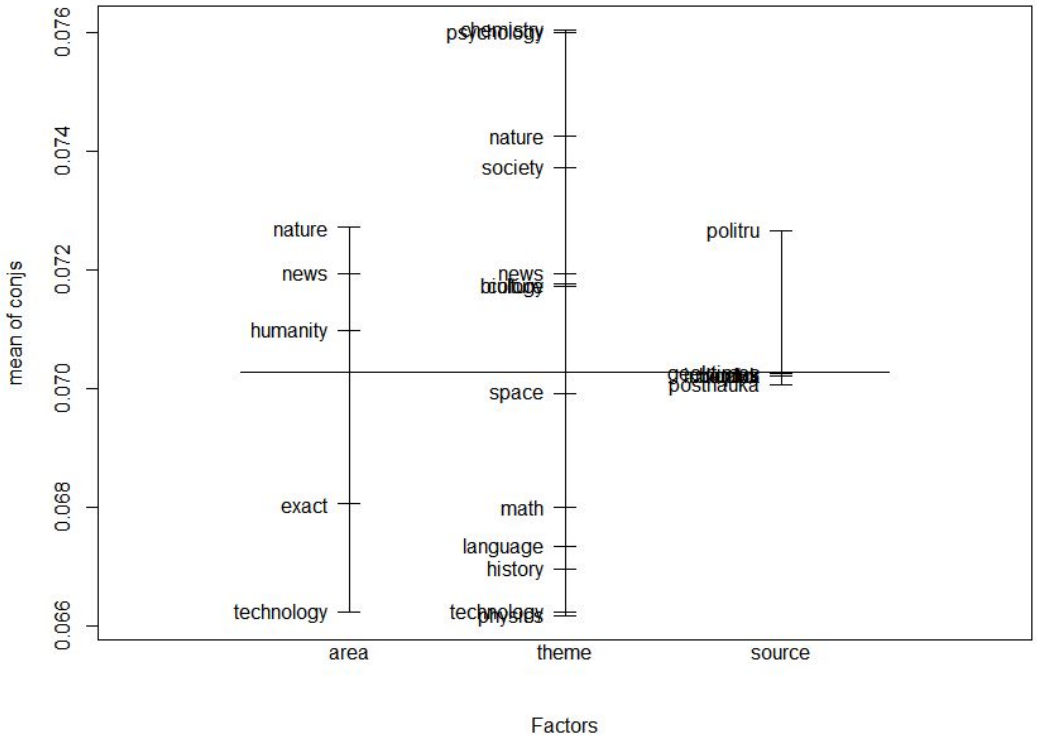


Data Summary

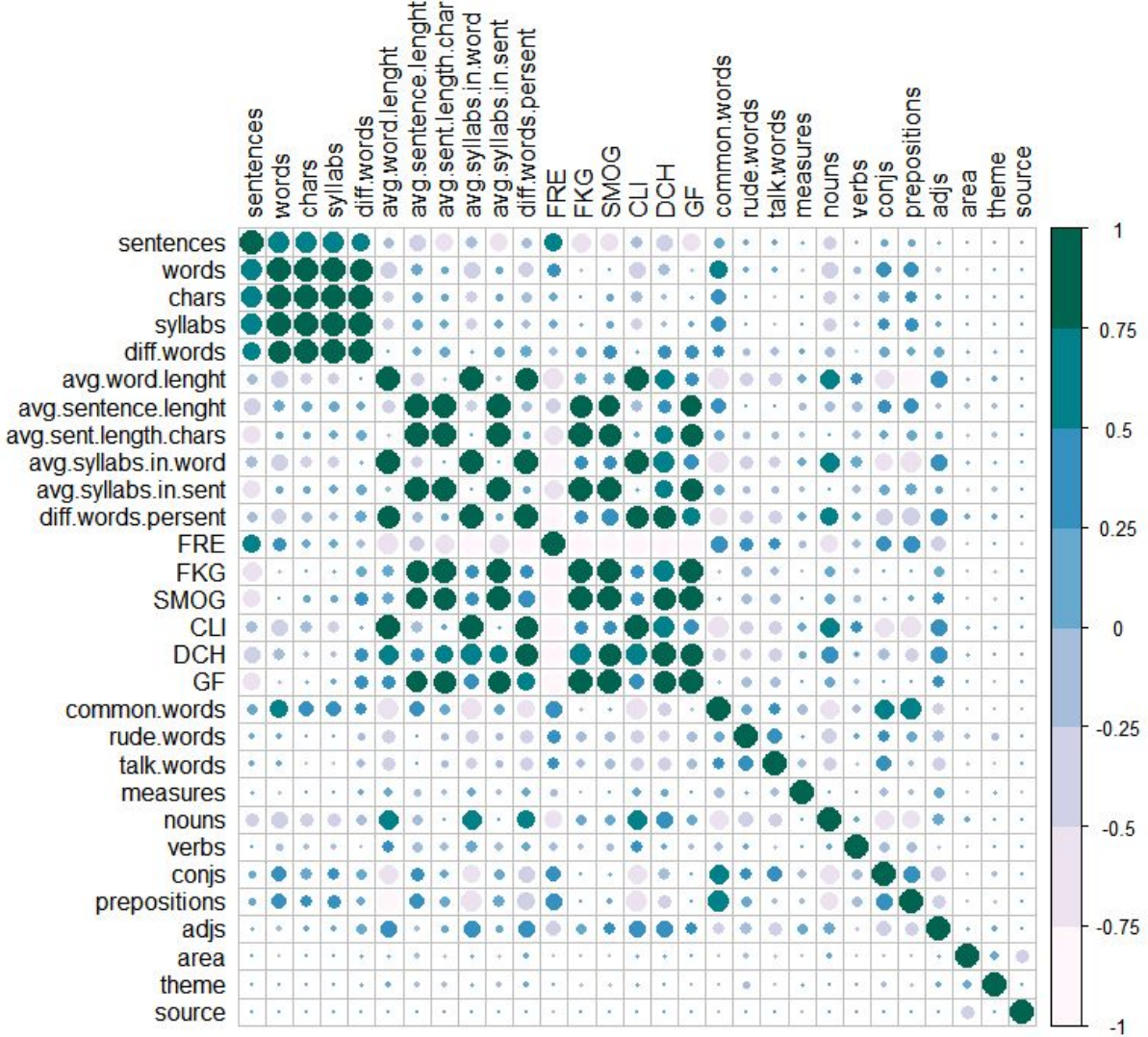




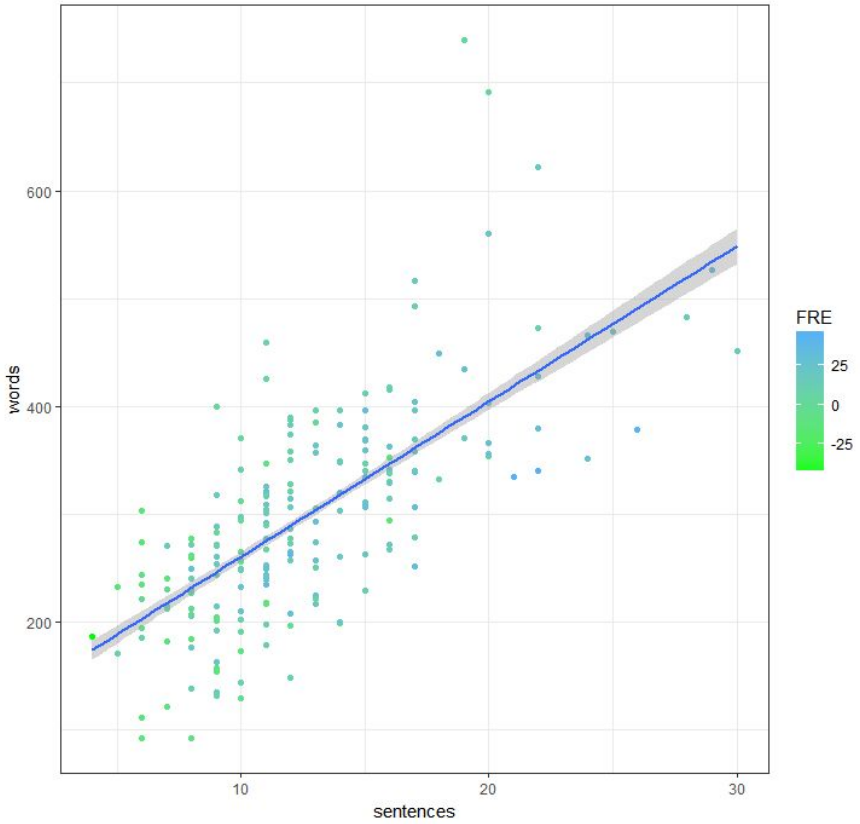
Data summary

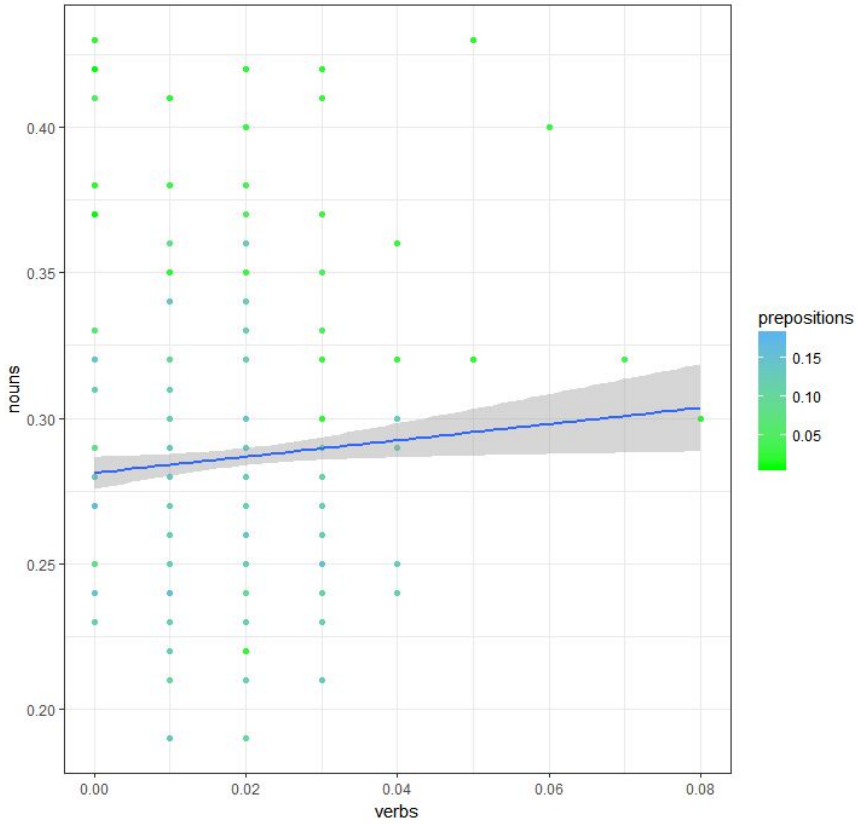


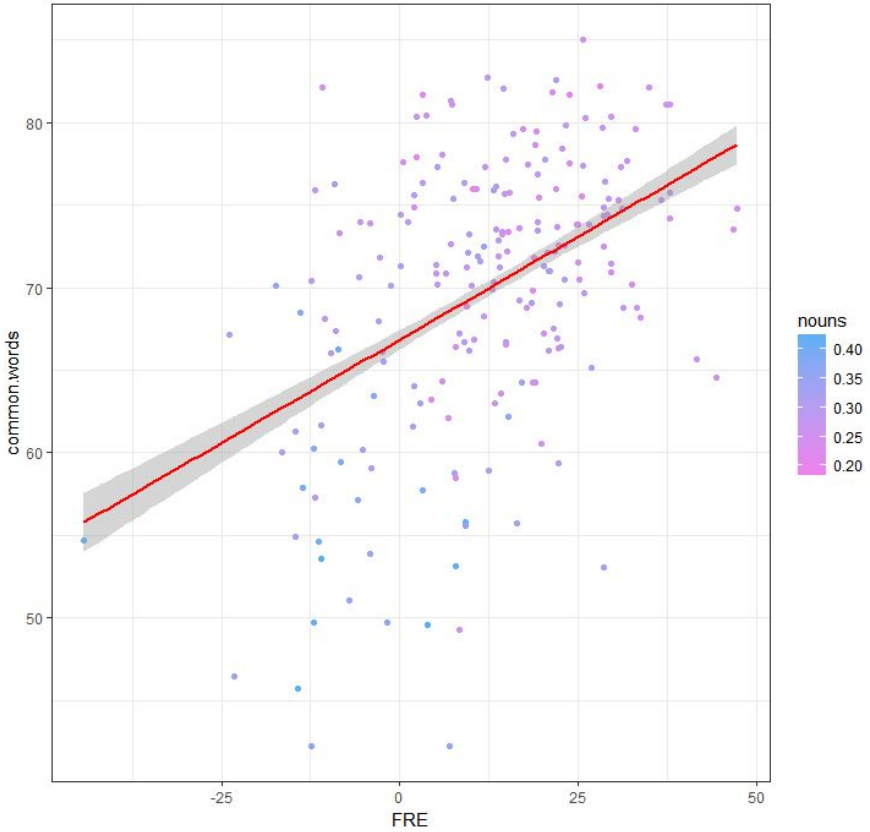
Correlations

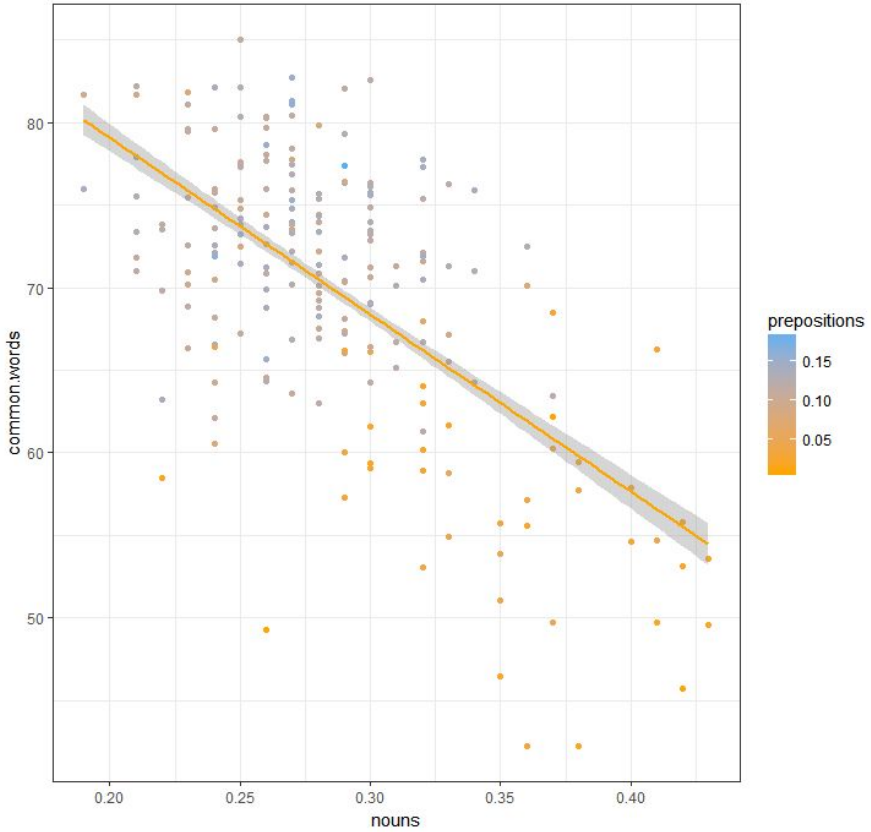

sentences
words
chars
syllabs
diff.words
avg.word.lenght
avg.sentence.lenght
avg.sent.length.chars
avg.syllabs.in.word
avg.syllabs.in.sent
diff.words.persent
FRE
FKG
SMOG
CLI
DCH
GF
common.words
rude.words
talk.words
measures
nouns
verbs
conjs
prepositions
adjs
area
theme
source
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.75
-1

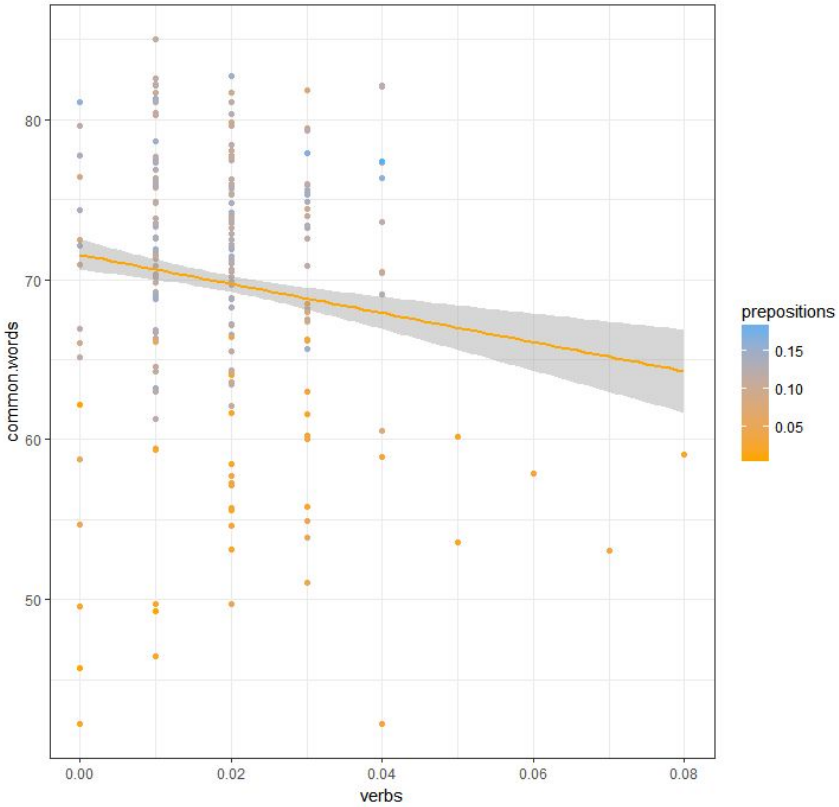
Regressions

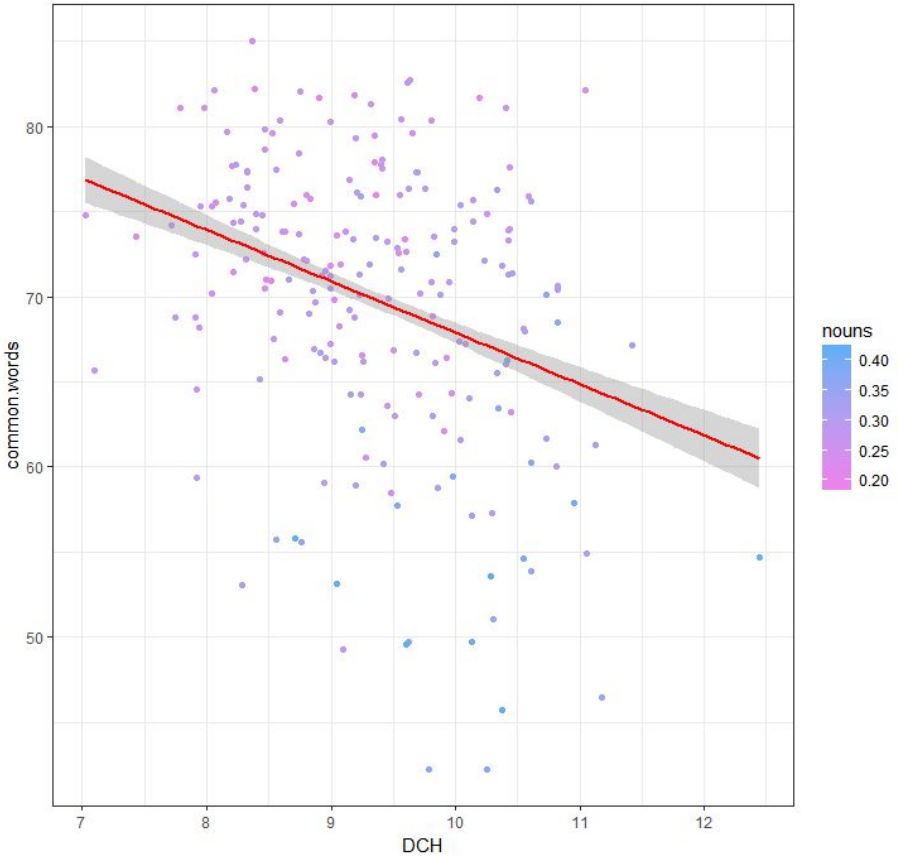

words
600
400
200
10
20
30
sentences
FRE
25
0
-25


nouns
0.40
0.35
0.30
0.25
0.20
0.00
0.02
0.04
0.06
0.08
verbs
prepositions
0.15
0.10
0.05

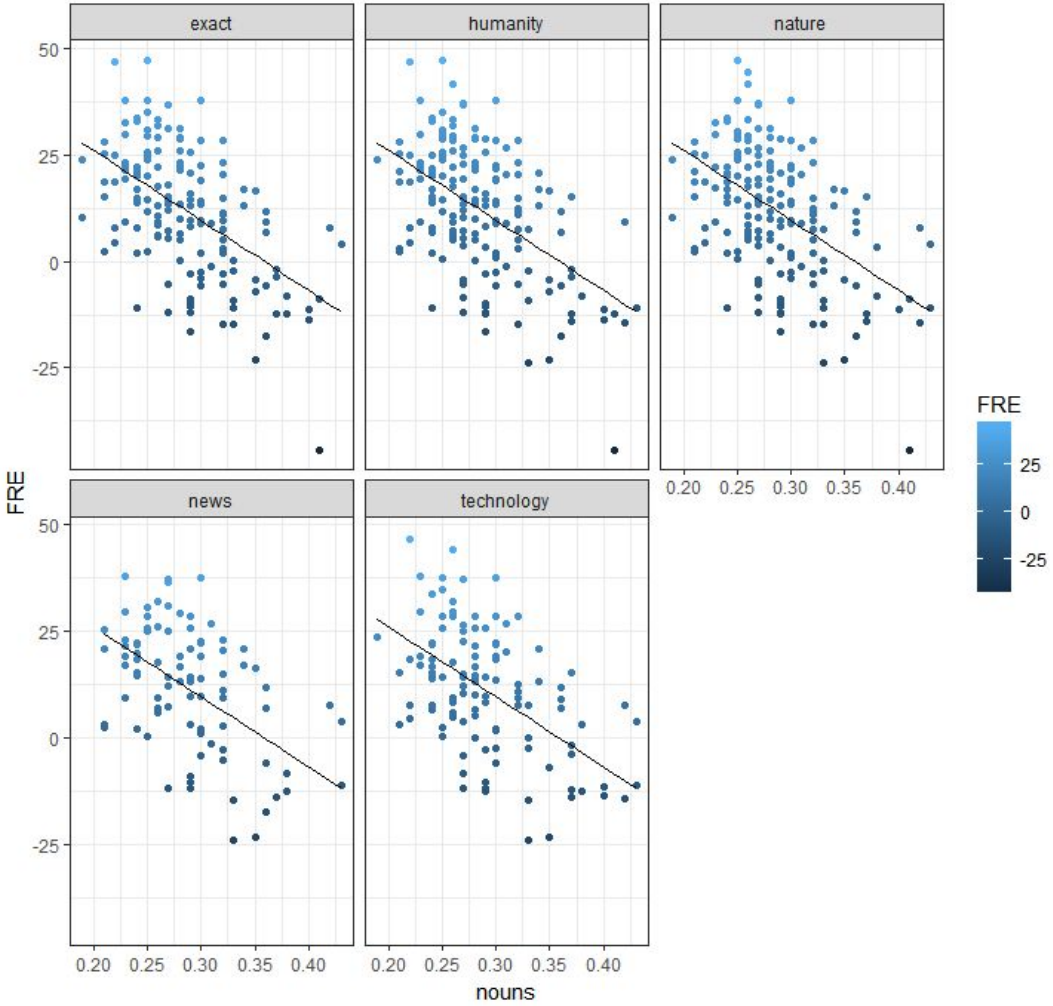

common.words
80
70
60
50
-25
0
25
50
FRE
nouns
0.40
0.35
0.30
0.25
0.20

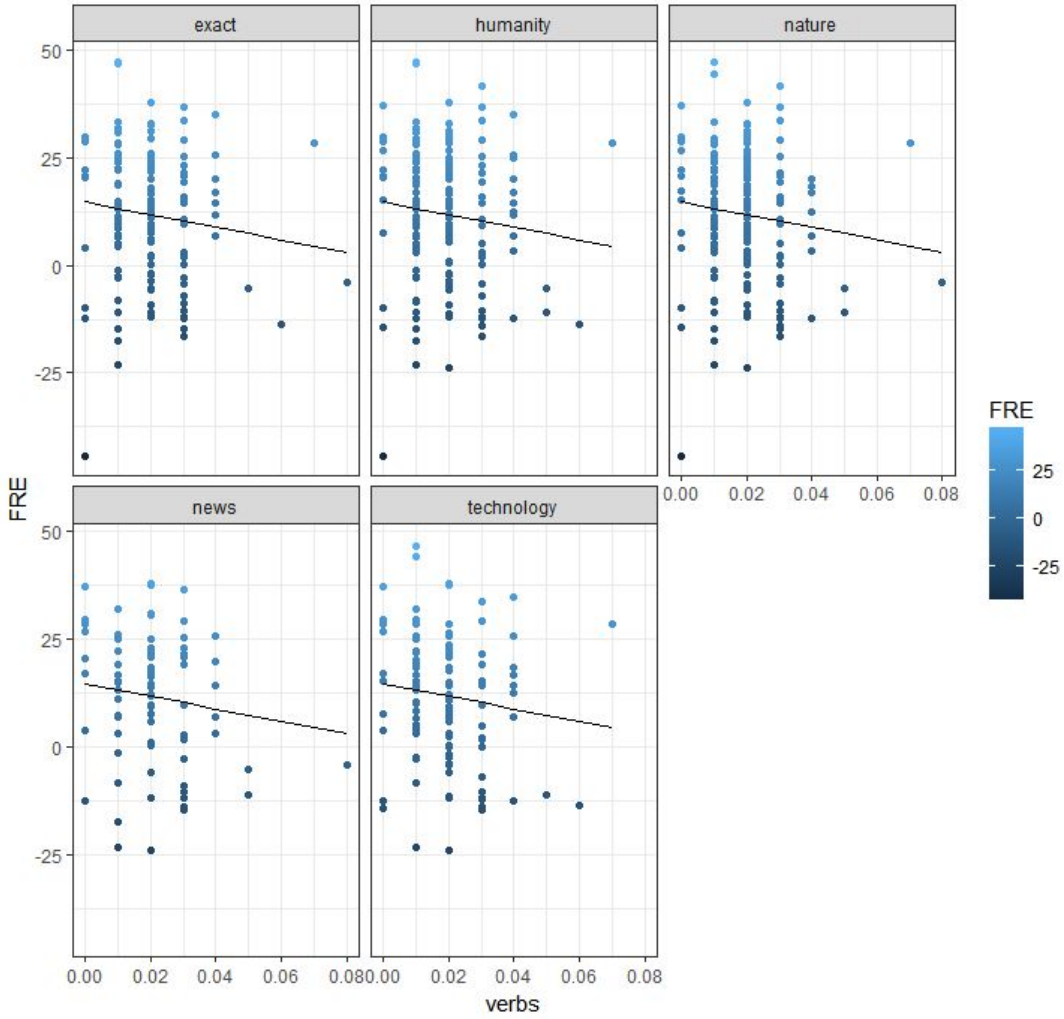

common.words
80
70
60
50
0.20
0.25
0.30
0.35
0.40
nouns
prepositions
0.15
0.10
0.05

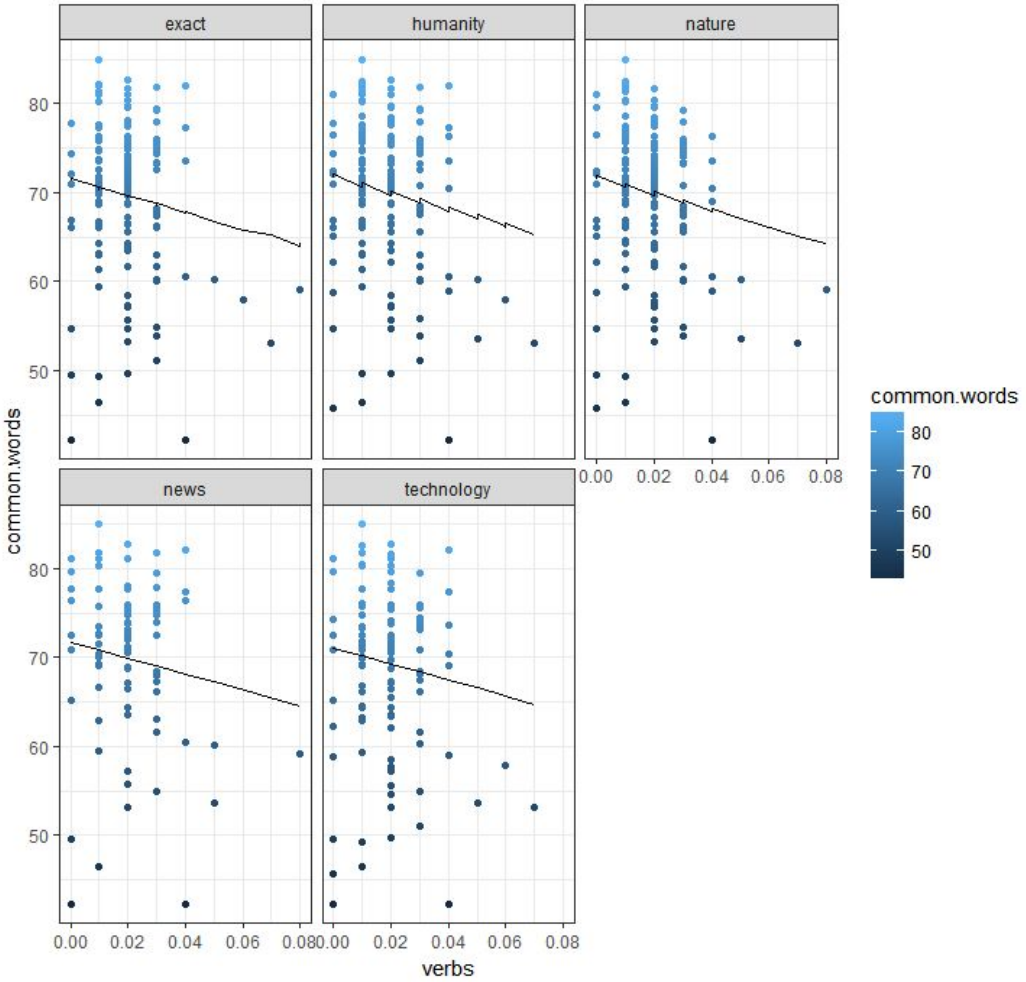

common.words
80
70
60
50
0.00
0.02
0.04
0.06
0.08
verbs
prepositions
0.15
0.10
0.05


common.words
80
70
60
50
7
8
9
10
11
12
DCH
nouns
0.40
0.35
0.30
0.25
0.20

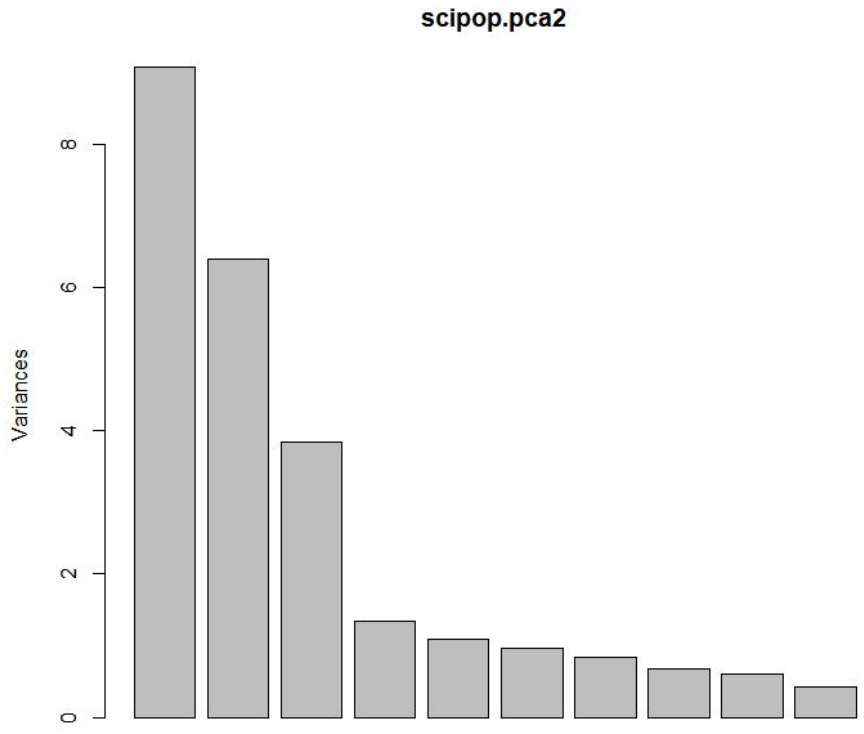
Mixed-effect model

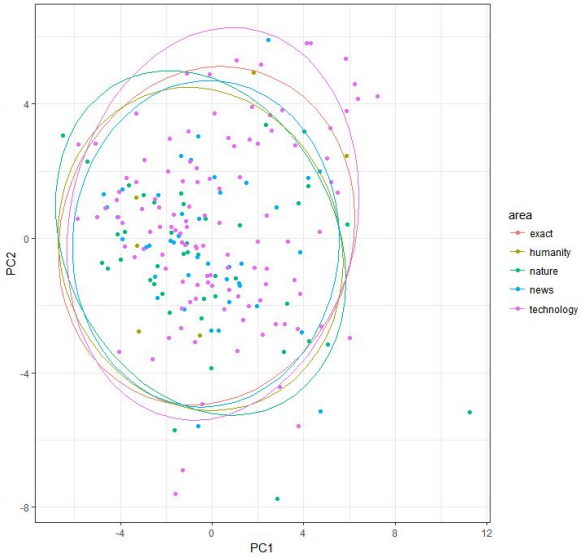

exact
humanity
nature
news
technology
FRE
nouns
50
25
0
-25
0.20
0.25
0.30
0.35
0.40
FRE
25
0
-25

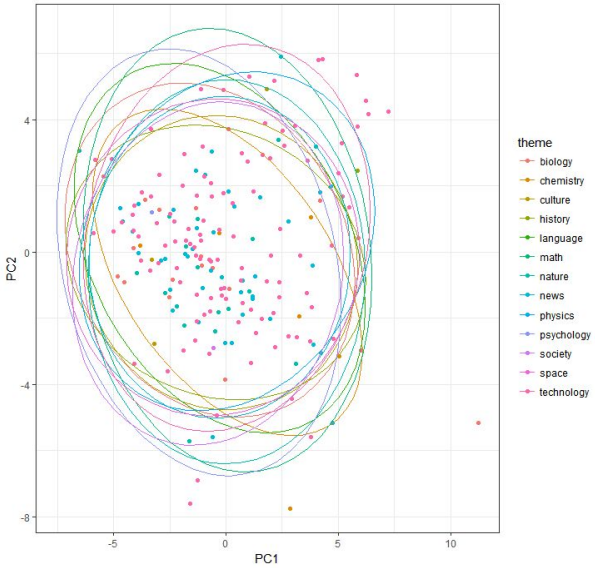

exact
humanity
nature
news
technology
FRE
verbs
50
25
0
-25
0.00
0.02
0.04
0.06
0.08
FRE
25
0
-25

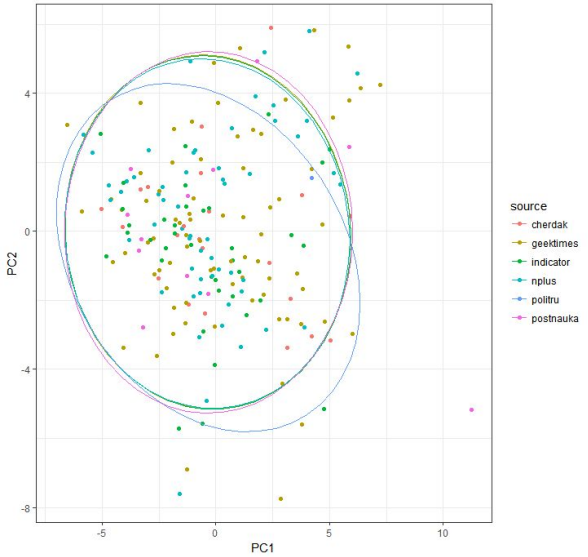

exact
humanity
nature
news
technology
common.words
verbs
80
70
60
50
0.00
0.02
0.04
0.06
0.08
common.words
80
70
60
50

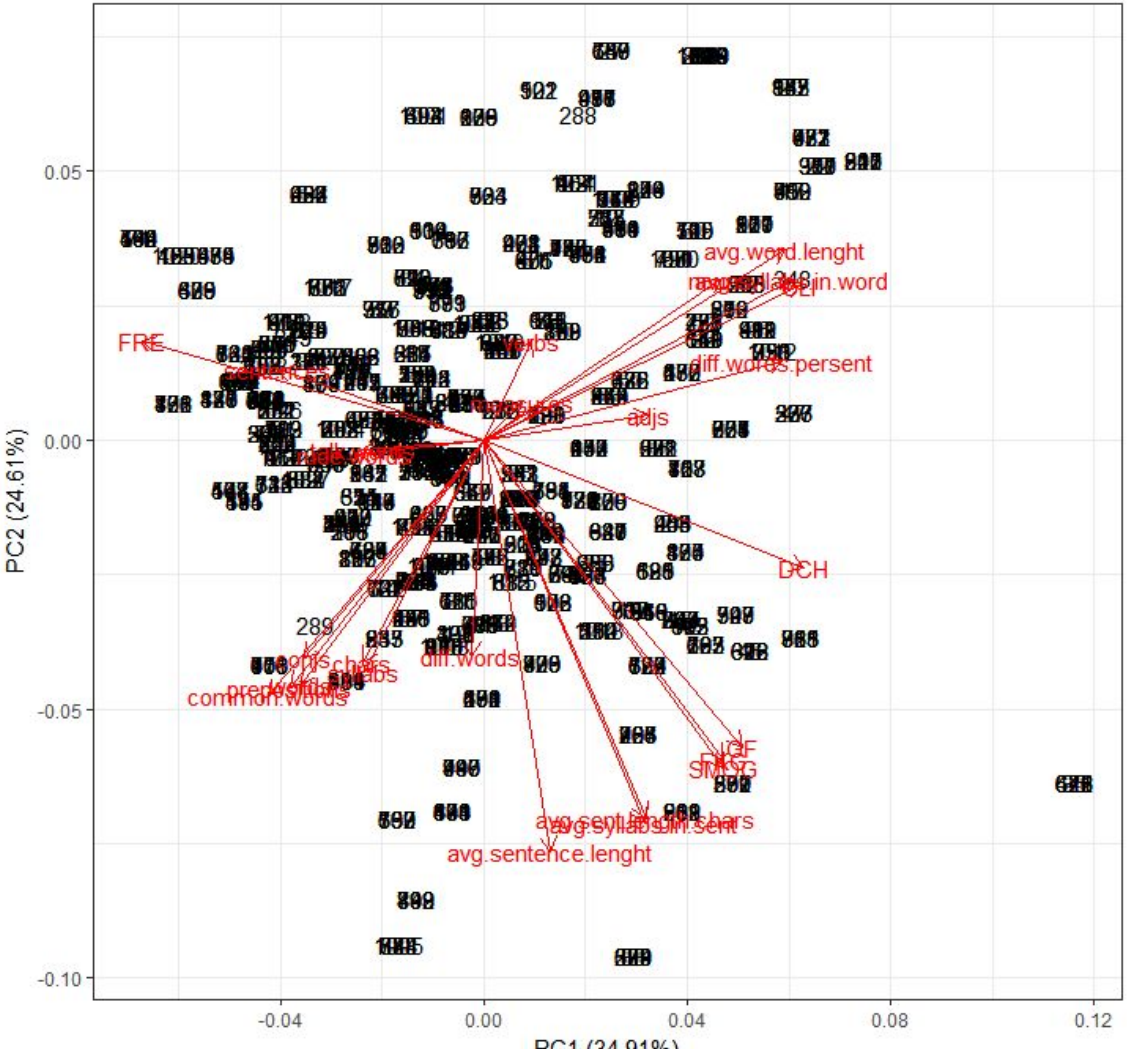
PCA


scipop.pca2
Variances
0
2
4
6
8


PC2
PC1
4
0
-4
-8
-4
0
4
8
12
area
exact
humanity
nature
news
technology


PC2
PC1
4
0
-4
-8
-5
0
5
10
theme
biology
chemistry
culture
history
language
math
nature
news
physics
psychology
society
space
technology


PC2
PC1
4
0
-4
-8
-5
0
5
10
source
cherdak
geektimes
indicator
nplus
politru
postnauka


288
289
avg.word.lenght
diff.words.persent
adjs
DCH
diff.words
common.words
avg.sentence.lenght
PC2 (24.61%)
PC1 (24.91%)
0.05
0.00
-0.05
-0.10
-0.04
0.00
0.04
0.08
0.12