Statistical Analysis of Sentence Structure in Lithuanian Texts

Neringa Bružaitė

Academic Supervisor: Dr. Tomas Rekašius

Vilnius Gediminas Technical University

2017 04 24

Content

Object of research

Examined 92 text files from morphologically annotated corpus MATAS.
Corpus contains 1641263 words and 138123 sentences.

Example of morphologically annotated corpus:

```
<word="Pasaulio" lemma="pasaulis" type="dktv vyr.gim vnsk K">
<space>
<word="pabaiga" lemma="pabaiga" type="dktv mot.gim vnsk V">
<sep=":">
<space>
<word="apsakymai" lemma="apsakymas" type="dktv vyr.gim dgsk V">
```

Coding Sentences

Texts from corpus are coded keeping order of sentences:

- I – in the sentences remain only nouns (“D”) and verbs (“V”), and any other part of speech are replaced by the symbol “-”. Several consecutive “-” symbols are combined;
- II – obtained from the code of type I, by joining several consecutive nouns (“D”) or verbs (“V”).

Structure of sentence	DDDDBD
I encoding	DDDD-D
II encoding	D-D

Code of sentence „Lietuvių kalbos sakinių struktūros statistinė analizė“ for different encodings. Symbol “B” stands for adjective.

Repetition frequencies of codes

In the table below are showed counts of structure codes with various frequencies.

Frequency	1	2	3	4	5	6	7	...	149	161	196
I encoding	954	77	34	17	13	2	11	...	0	0	1
II encoding	632	86	26	18	14	8	7	...	1	1	1

For codes following rule applies: there are few very high frequency codes, and many low frequency codes. This distribution approximately follows mathematical form known as Zipf's law.

History of Zipf's law

- George Zipf was not the first one who noticed such phenomena as the unequal distribution of words in the text.
- Jean-Baptiste Estoup was the first person who noticed and mathematically formulated this law in his book *Gammes sténographiques* (3d ed. 1912).
- Edward Condon was the first person, who graphed the Zipf's law (which then was not called Zipf's law) in 1928.

Zipf's law

Zipf's law (1949 m.)

Empirical law, which describes relation between rank of frequency and frequency of a word f_z, which has rank z

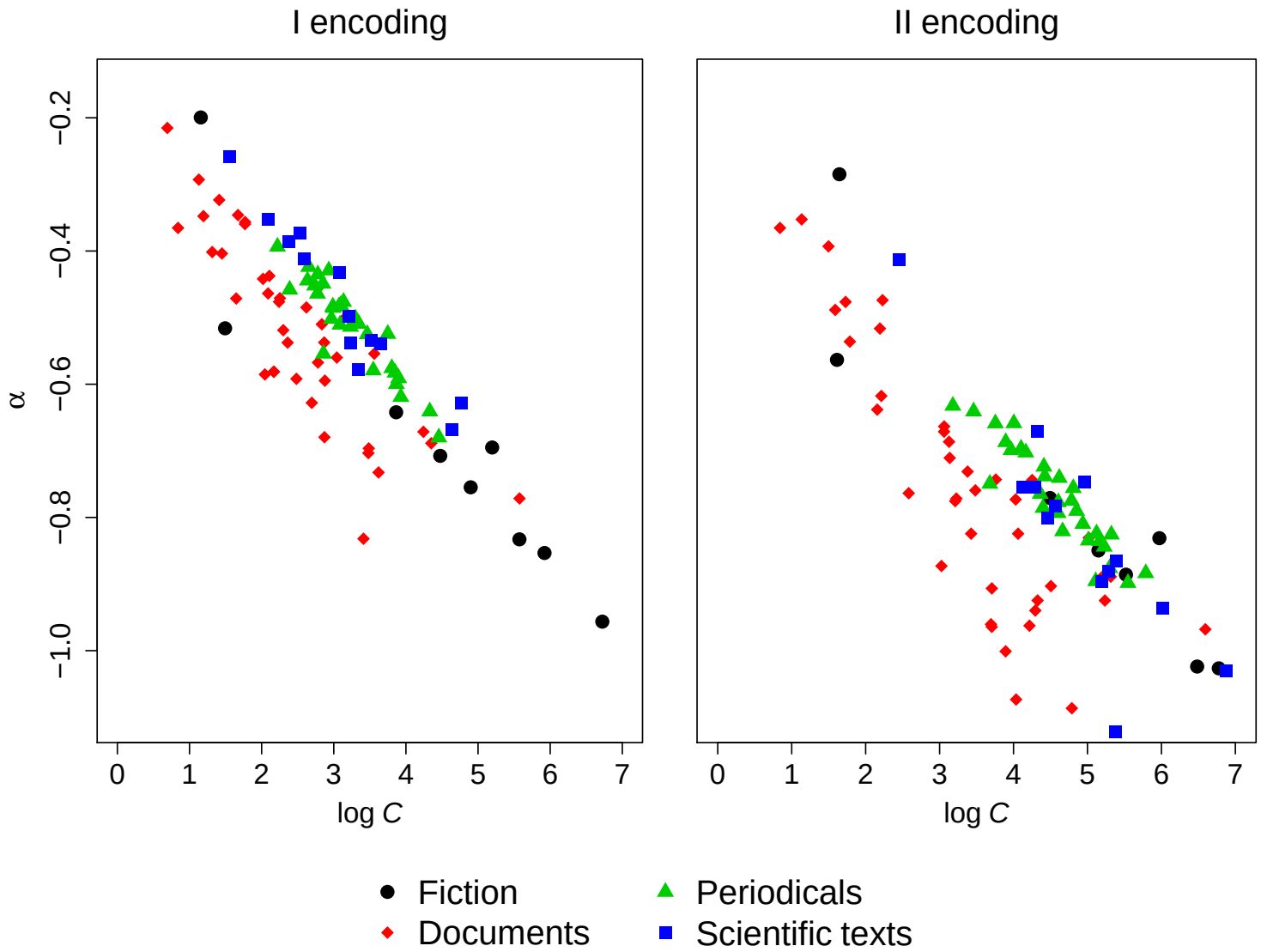
$$f_z = \frac{C}{z^\alpha}, \tag{1}$$

here z – rank, $\alpha > 0$, C – const.

Taking logarithms at both sides, the linear relation between $\log f_z$ and $\log z$ follows immediately:

$$\log f_z = \log C - \alpha \log z. \tag{2}$$

Parameter estimates of Zipf's law



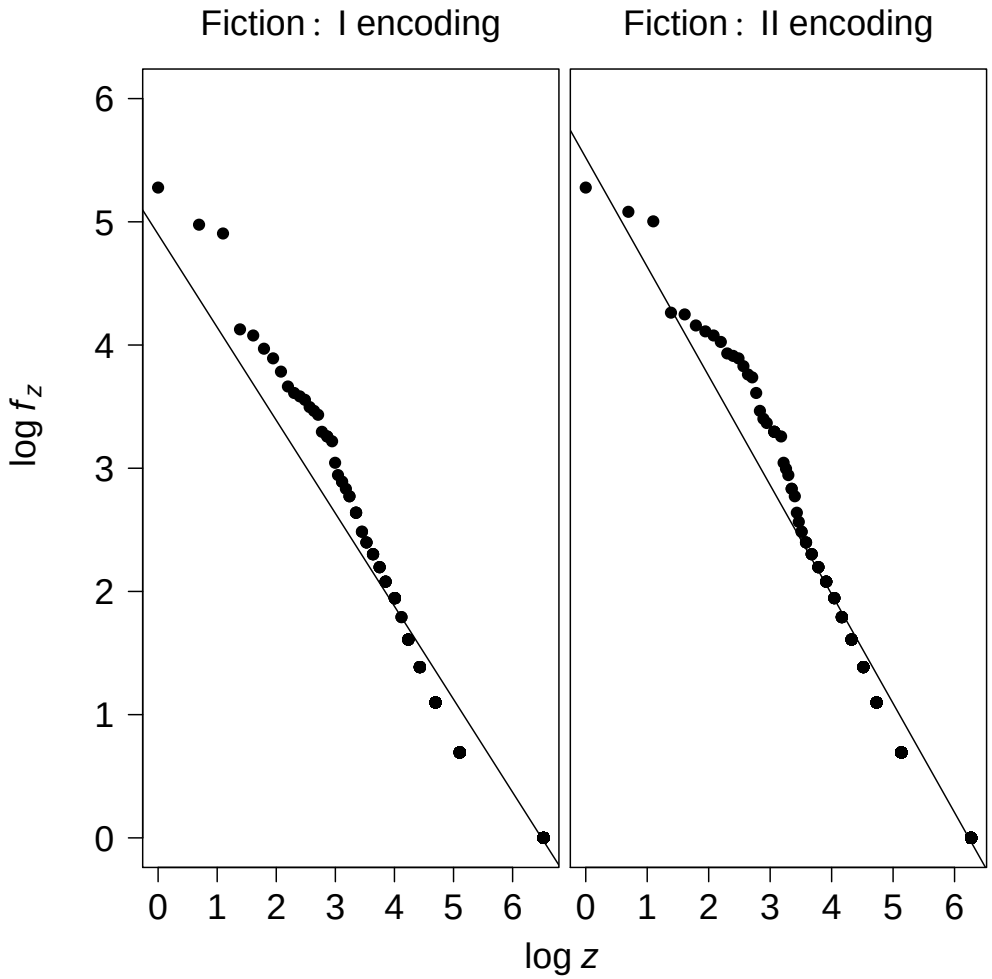
Parameter estimates of Zipf's law

	TRUE	FALSE
$\alpha_I > \alpha_{II}$	91	1
$\log C_I > \log C_{II}$	1	91

In the table above α_I and $\log C_I$ are Zipf's law parameters for I encoding sentences and α_{II} and $\log C_{II}$ – for II encoding sentences.

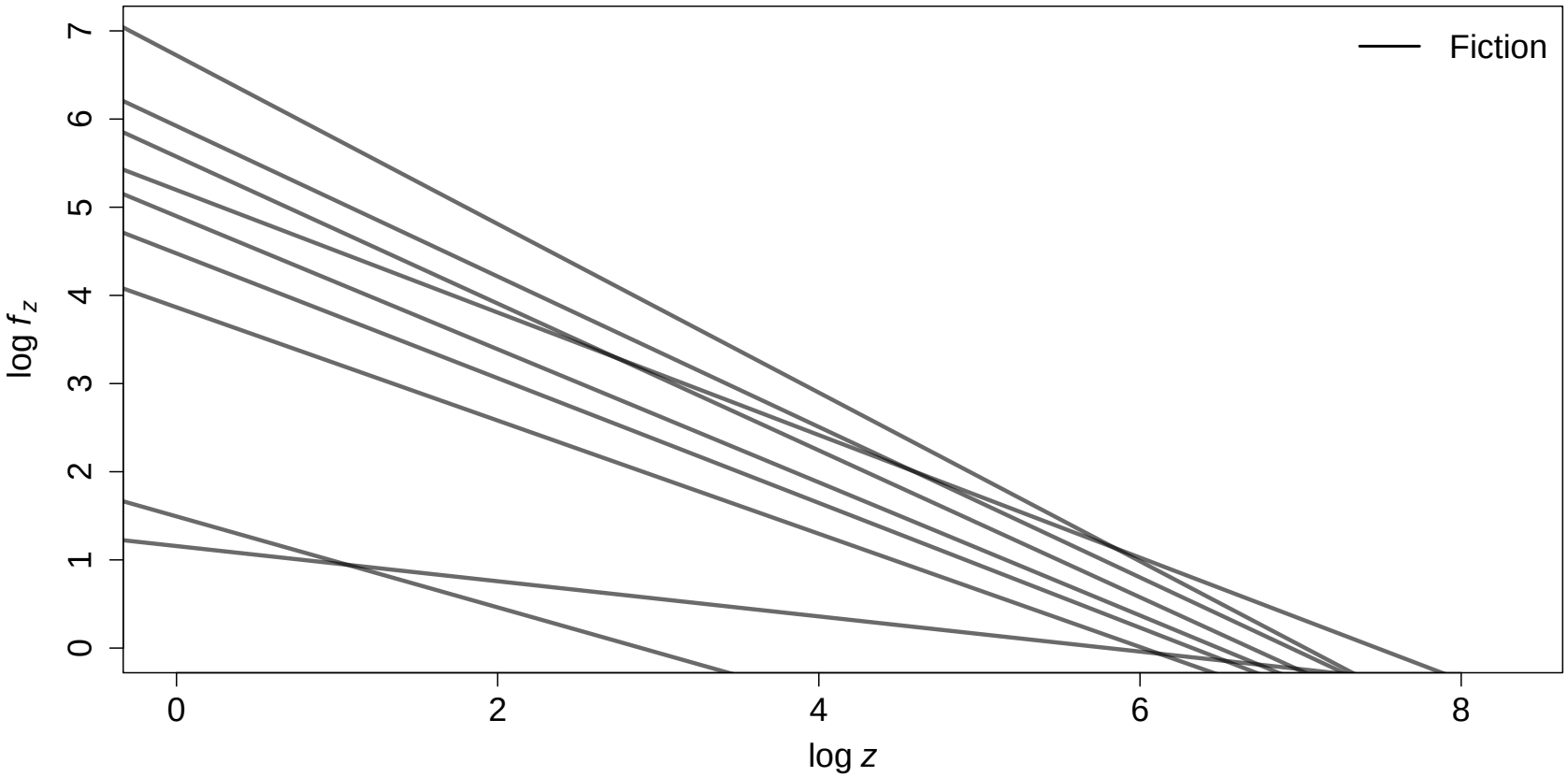
There is one text, for which Zipf's parameter estimates are the same for both encodings.

Log-log graphs of sentence structure code frequency

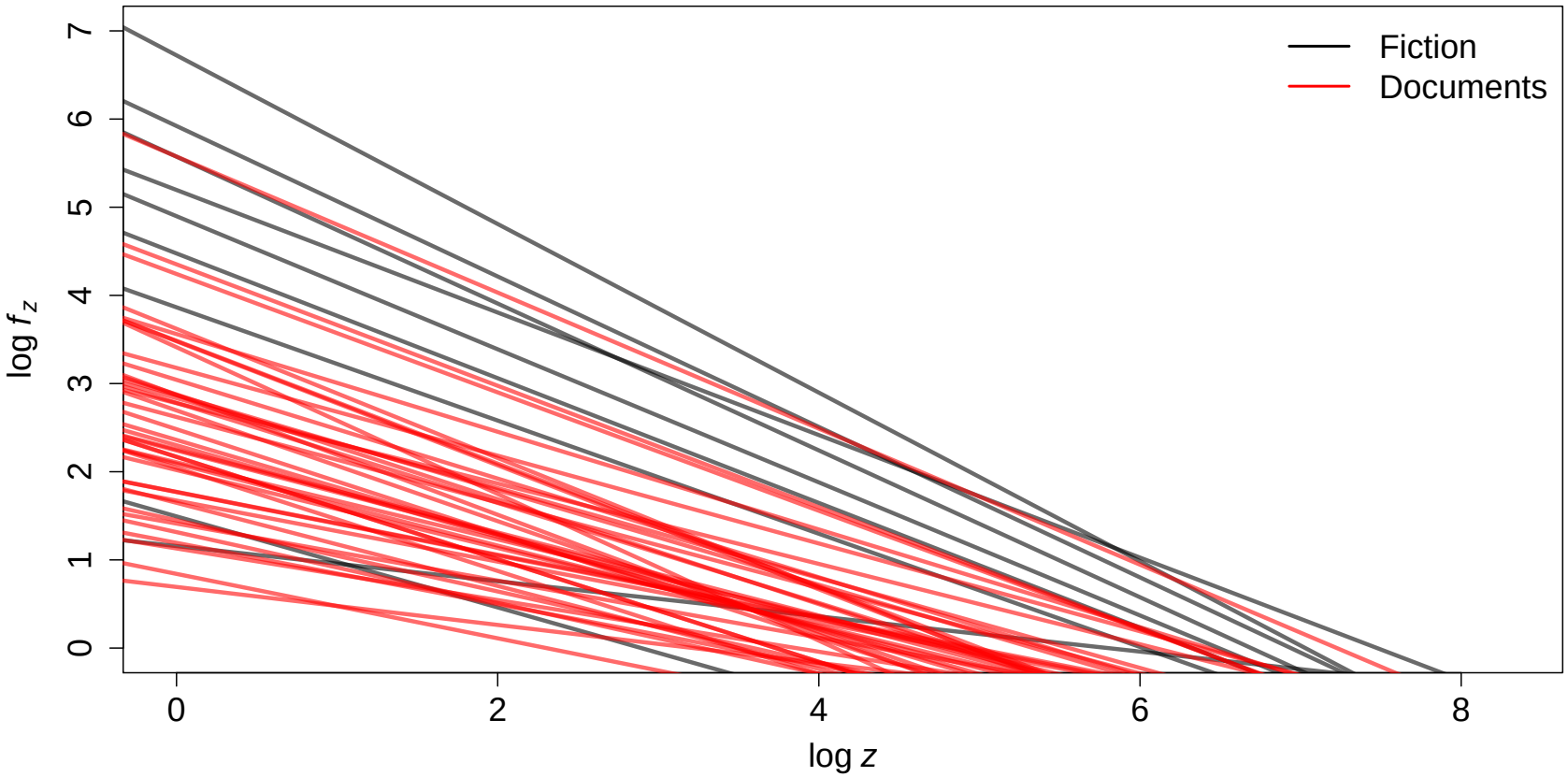


Code frequency f_z as a function of Zipf rank z in the log-log plane for one fiction text, which has $N = 2886$ sentences.

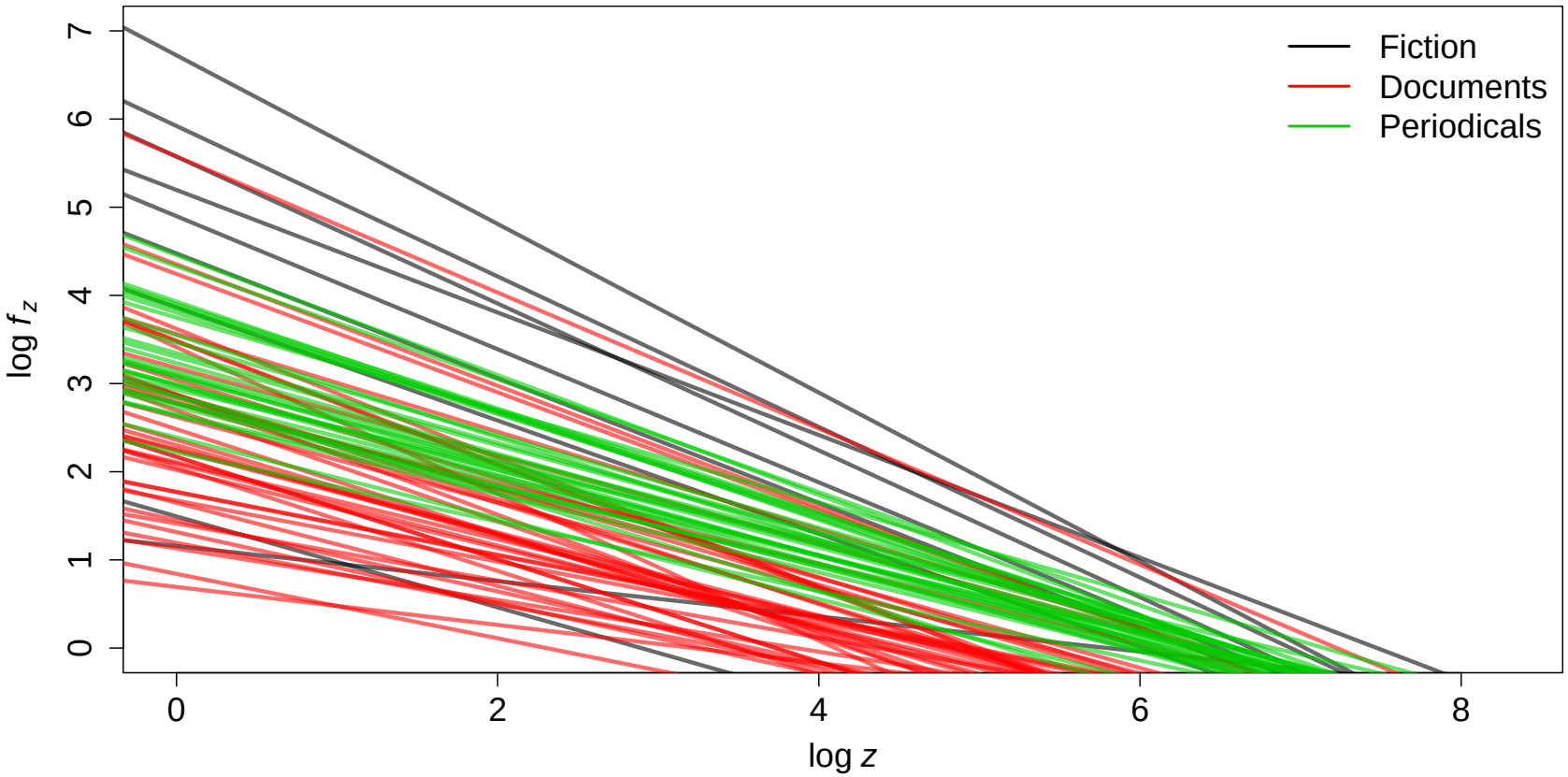
Zipf's law: predicted frequencies



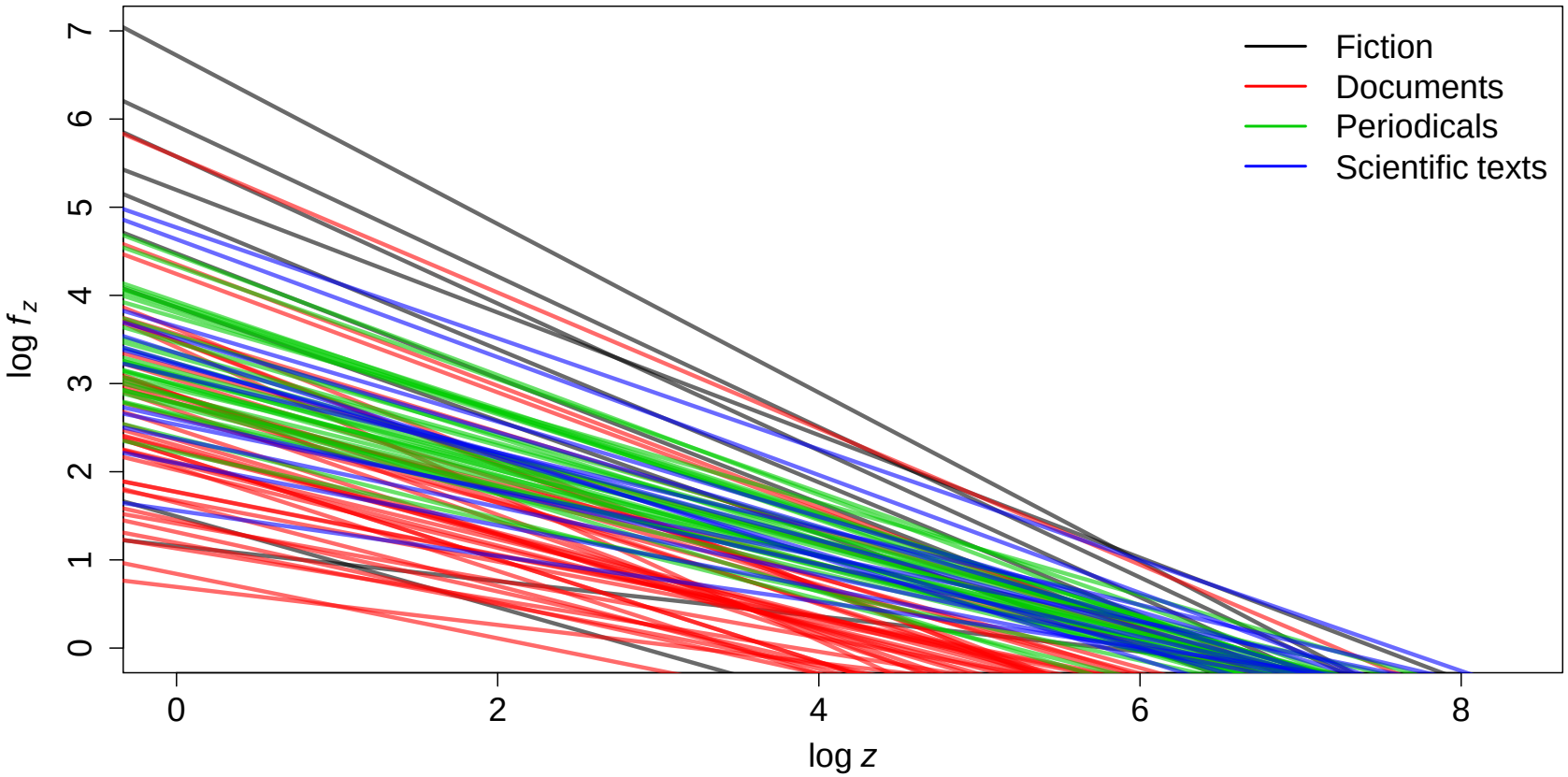
Zipf's law: predicted frequencies



Zipf's law: predicted frequencies



Zipf's law: predicted frequencies



Conclusions and results

- A large part of encoded sentences has a standard structure. On the other hand, there are a lot of encoded sentences which are unique.
- The encoded sentences are described by Zipf's law quite well.

Conference materials (slides, computing code, corpus, literature) can be found using the link below:
https://github.com/neringabr/conference-ktu