Evaluation of Statistical Readability Measures on Slovene Texts

Tadej Škvorc,[*†] Simon Krek,[†◦] Senja Pollak,[†] Špela Arhar Holdt,[*◦]
Marko Robnik-Šikonja[*]

[*] University of Ljubljana, Faculty of Computer and Information Science
[◦] University of Ljubljana, Faculty of Arts
[†] Jožef Stefan Institute

20. 9. 2018

Readability in theory

“Today is Thursday.”

“Whenever the President transmits to the President pro tempore of the Senate and the Speaker of the House of Representatives his written declaration that he is unable to discharge the powers and duties of his office, and until he transmits to them a written declaration to the contrary, such powers and duties shall be discharged by the Vice President as Acting President.”

Readability in practice

- “How easy a text is to read”
- Simple way: Readability formulas
- Harder way: more complicated methods

Does this work in other languages?

Reasons why it might not:

- Formulas calibrated to the American school system.
- No list of easy words for Slovene.
- They do not take into account morphology.

Our approach

Evaluate simple readability measures on a corpus of Slovene texts from different categories and see if the measures are able to differentiate between the categories.

Dataset

Subcorpus	#docs	Avg. #words / doc
Children's magazines	125	5,488
Pop magazines	247	33,967
Newspapers	14,011	12,881
Computer magazines	163	110,875
National Assembly	35	58,841

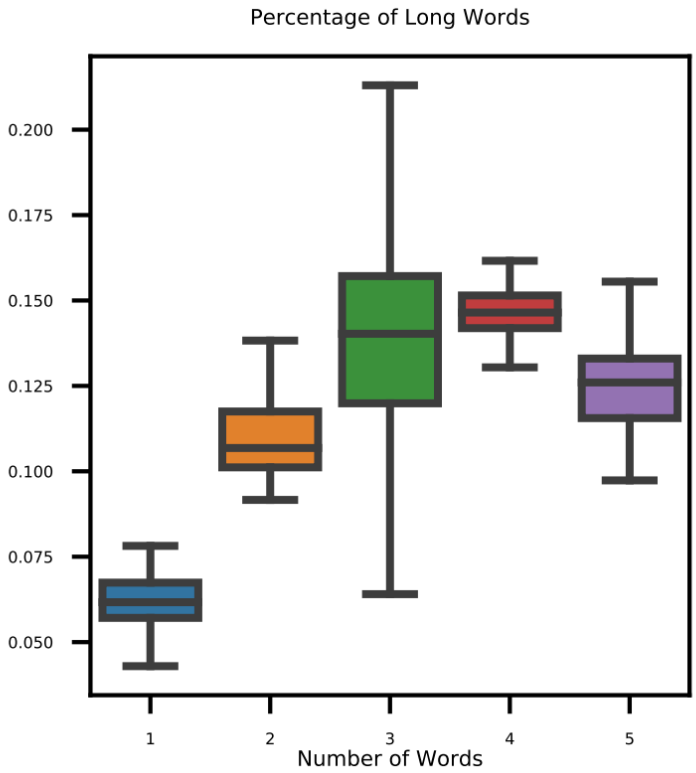
Readability measures

- RIX (long words per sentence)
- Gunning fog index (sentence length, long words)
- Flesch reading ease (sentence length, syllables per word)
- Dale-Chall readability formula (sentence length, complex words)
- Spache readability formula (sentence length, unique complex words)

Other readability criteria

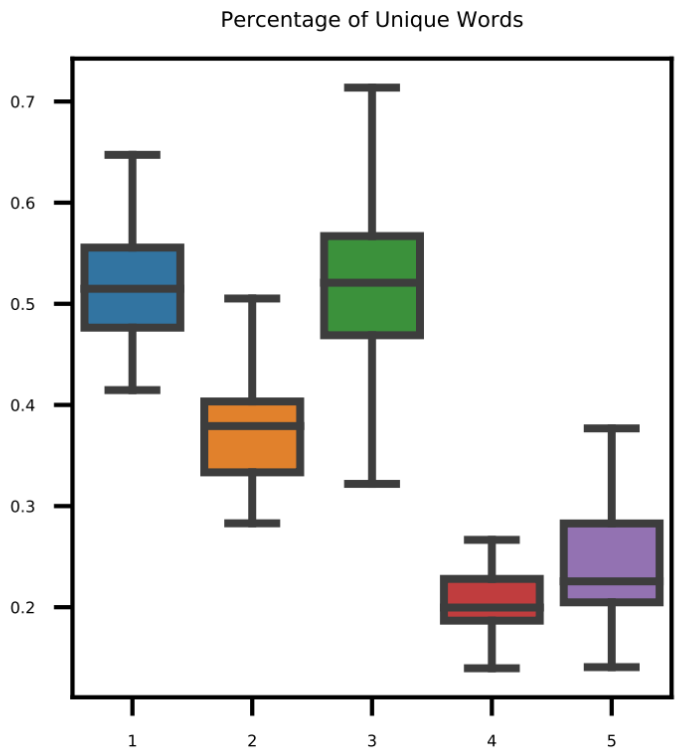
- Simple criteria (sentence length, long words, verbs, adjectives, vocabulary size)
- Context of difficult words
- “Morphological difficulty”

What we want



- Most readability formulas, to various extents
- Percentage of long words
- Average morphological difficulty
- Average sentence length
- Percentage of long words
- Context of difficult words

What we do not want



- Dale-Chall Readability Formula (based on word list)
- Spache (based on word list)
- Percentage or number of unique words

Results

- Most formulas are good at differentiating different groups.
- Word list has problems.
- Some simple criteria work well.
- National assembly texts are an exception.

Future work

- Evaluate more advanced methods.
- Evaluate on a corpus of textbooks.
- Evaluate coherence and cohesion methods.