Methods for Text Complexity Assessment in Russian

RuSTEP
September 2022

Vladimir Ivanov, Innopolis University

Text complexity

Definition / Multilevel Phenomenon

- TC has been studied at various levels of linguistic units
- whole texts
  (Crossley et al. (2008); Collins-Thompson and Callan (2005); Heilman et al. (2008))
- sentence level
  (Schumacher et al. (2016); Iavarone et al. (2021)
- individual words
  (Shardlow et al. (2021, 2020))

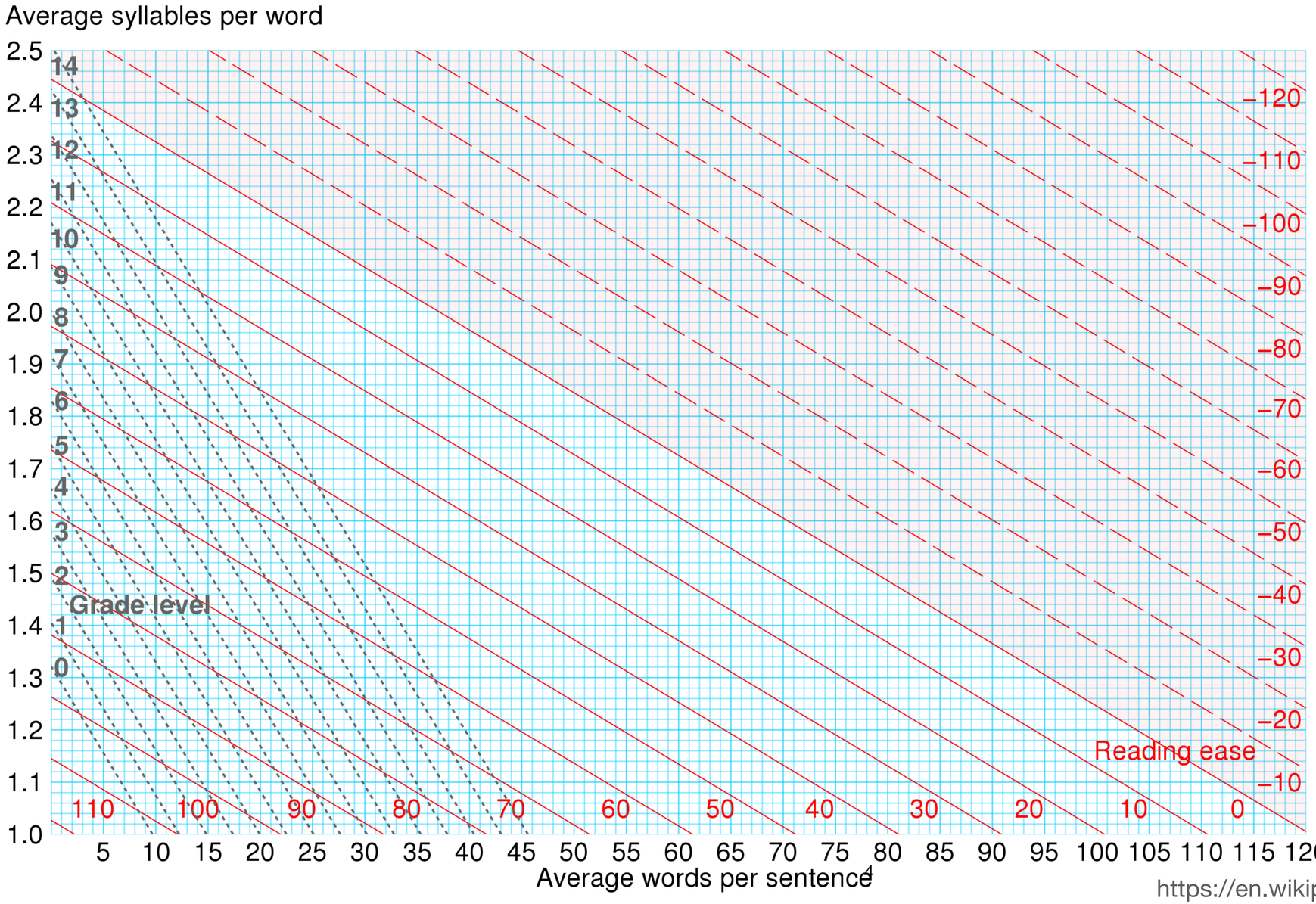
Text Complexity as Readability Ease

Readability indices / formulas

- readability formulas tools to match texts and readers

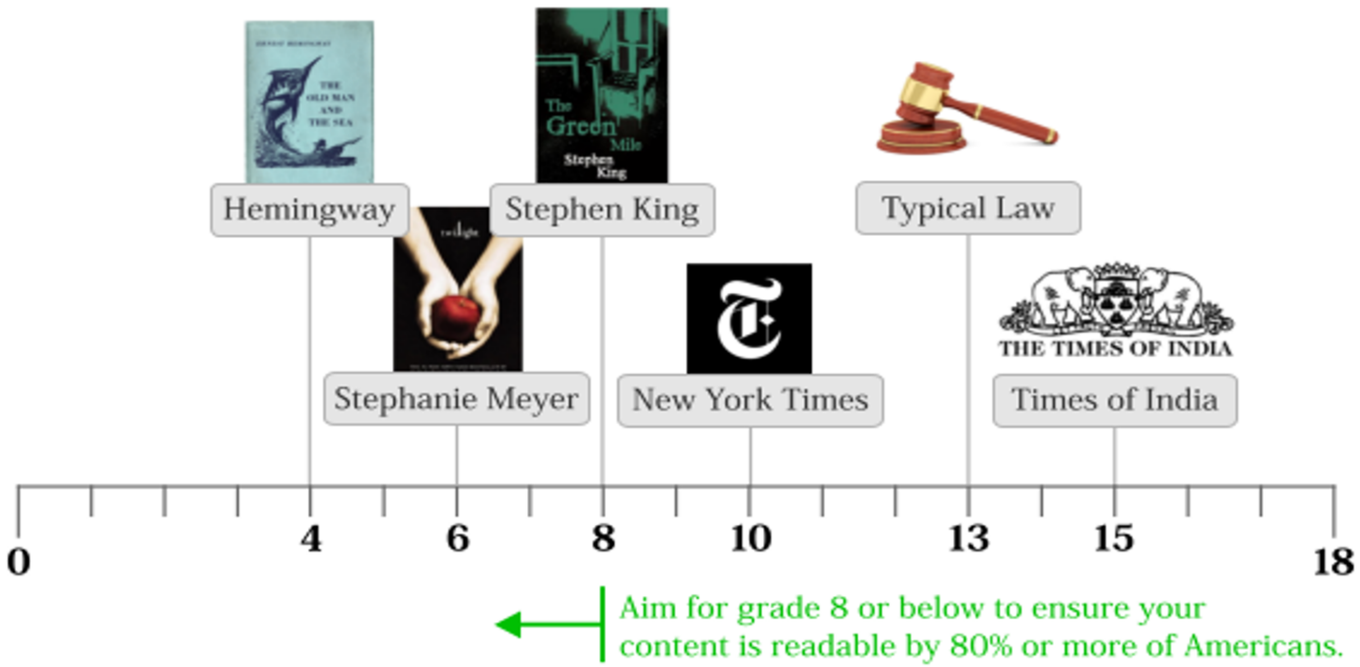
Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

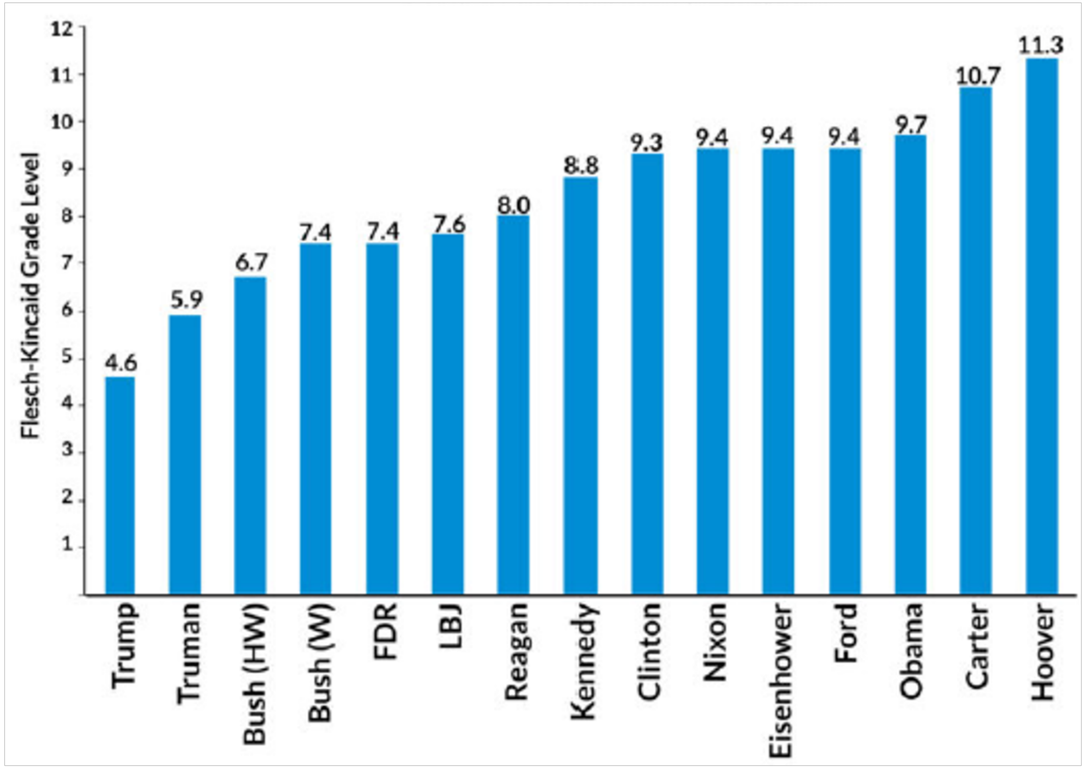
FK reading ease



https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests

Applications





Flesch-Kincaid

1975

- Flesch-Kincaid (English)
  The Flesch Reading Ease (FRE), and the Flesch–Kincaid Grade Level (FKG)

$$FRE = 206.835 - 1.015 \times ASL - 84.6 \times AWS$$

$$FKG = 0.39 \times ASL + 11.8 \times AWS - 15.59$$

- Adopted by I.Oboroneva (2006), for Russian

$$FRE = 206.835 - 1.52 \times ASL - 65.14 \times AWS$$

Other indices / readability formulae

- Gunning Fog Index
- Coleman Liau Index
- SMOG Index
- Dale-Chall Readability Formula
- …

https://readable.com/features/readability-formulas/

Modified version of FKG for Russian

Solovyev, Solnyshkina, Ivanov, Batyrshin (2018)

- Linear regession fitted on School Textbooks

$$FKG = 0.36 \times ASL + 5.76 \times AWS - 11.97$$

Studying of Groups of Features

An extended feature set for the text explored:

- Features based on length and frequency of words and sentences
- Features based on Part-of-Speech tags
- Features based of syntactic dependencies

Features based on length and frequency

- **ASL** is an average number of words per sentence
- **ASW** is an average number of syllables per word
- **FREQ** is a cumulative frequency of content words

Features based on POS tags

- **NOUNS** is a number of nouns per sentence
- **VERBS** is a number of verbs per sentence
- **ADJ** is a number of adjectives per sentence
- **PRONOUNS** is a number of pronouns per sentence
- **PERONAL PRONOUNS** is a number of personal pronouns per sentence
- **NEG** is a number of negations per sentence

Features based of syntactic dependencies

1.AVERAGE_PATH is the quotient of the number of nodes and the number of leaves in a sentence

2.**AVERAGE_SOCHIN_LENGTH** is the average length of coordinating constructions

3.DEEPRICH_RATE is the average number of verbal participles

4.DEEPRICH_V is the average span of a verbal adverb phrase

5.LEAVES_NUMBER is the average number of 'leaves‘ in a sentence

6.LONGEST_PATH is the average length of the longest branch

7.NOUNS_DEP is the average number of modifiers in a nominal group; coordinating and explanatory links are ignored

8.PODCHIN_NUMBER is the ratio of sentences in which there is at least one subordinate conjunctions or relational links

9.PODCHIN_RATE is the average number of subordinate links

10.PRICH_RATE is the average number of participial construction; participial constructions are defined as a participle that has at least one dependent

11.PRICH_V is the average span of a participial construction is the quotient of the number of nodes that depend on the participle

12.SENTSOCH_NUMBER is the average number of compound sentences

13.SOCHIN_NUMBER is defined as the average number of coordinating chains

14.PATH_NUMBER is defined as the average number of sub-trees (in a sentence)

15.VERBS_DEP is defined as the average number of finite dependent verbs and is calculated as the sum of nodes directly dependent on the finite verb divided by the number of finite verbs; coordinating and explanatory links were ignored.

Correlation between features and grade level

	Feature name	Correlation			Feature name	Correlation
1	ASL	0.94		13	NOUNS	0.82
2	ASW	0.94		14	VERBS	0.74
3	SOCHIN_NUMBER	0.93		15	NEGATIONS	0.7
4	PRICH_RATE	0.91		16	PRONOUNS	0.7
5	NOUNS_DEP	0.88		17	PODCHIN_RATE	0.64
6	AVERAGE_SOCHIN_LEN	0.87		18	PODCHIN_NUMBER	0.62
7	PATH_NUMBER	0.87		19	DEEPRICH_V	0.52
8	LONGEST_PATH	0.84		20	PERS_PRONOUNS	0.47
9	FREQ	0.84		21	DEEPRICH_RATE	0.44
10	LEAVES_NUMBER	0.84		22	VERBS_DEP	0.43
11	AVERAGE_PATH	0.84		23	PRICH_V	0.33
12	ADJ	0.82		24	SENTSOCH_NUMBER	0.03

Significance of features

	Feature	Absolute value of Coefficient in Ridge Regression
1	ASL	0.506
2	ASW	0.125
3	SOCHIN_NUMBER	0.119
4	PRICH_RATE	0.106
5	LONGEST_PATH	0.089
6	PATH_NUMBER	0.079
7	LEAVES_NUMBER	0.075
8	AVERAGE_SOCHIN_LE	0.071
9	NOUNS_DEP	0.071
10	FREQ	0.034
11	NEGATIONS	0.01
12	AVERAGE_PATH	0.007
13	PERS_PRONOUNS	0.003
14	VERBS	0.001
15	ADJ	0.001
16	NOUNS	0.0

Study of fragment size

V. Solovyev, V. Ivanov, M. Solnyshkina (2017)

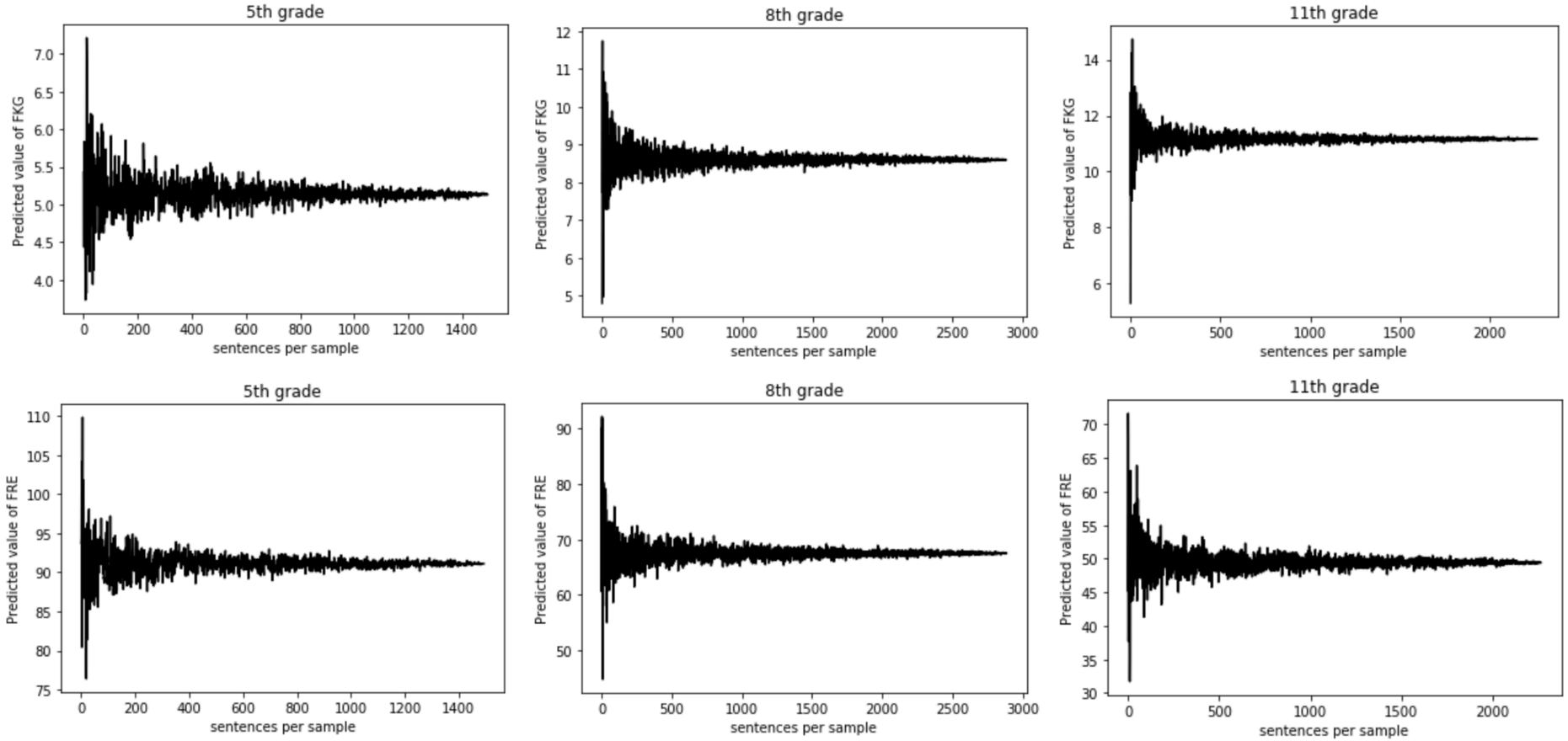
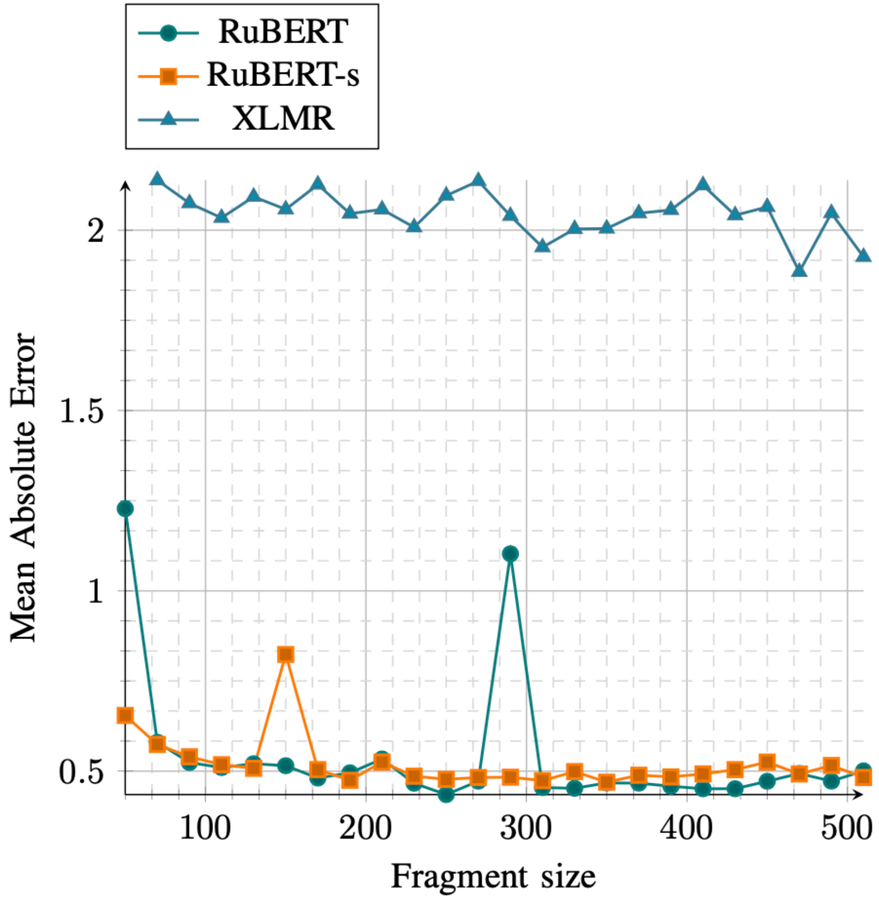
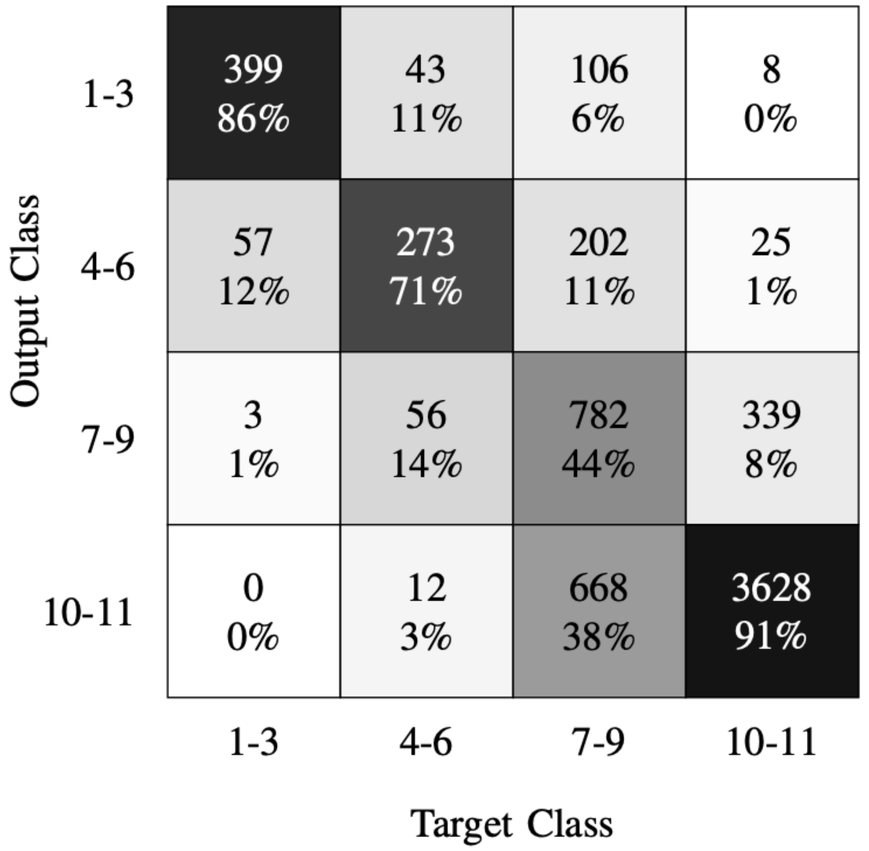


Figure 3: FKG_{mod} and FRE_{mod} values for different number of sentences sampled from a document.

Study of fragment size





Beyond Readability formulae

Analysis of Academic text Complexity

in collaboration with V. D. Solovyev and M.I. Solnyshkina (KFU)

- Classic ML (linear regression, classifiers)
  - 2018: Analysis of sets of features and fragment size
  - 2019: Analysis of semantic-level features
- Deep learning
  - 2020-21: Appication of neural networks to texts
  - 2021-22: Complexity of sentences and words

Corpora of Academic texts

- NIK + BOG Corpus (14 books)
- + Textbooks in History (approx. 5 books)
- + Textbooks in Biology (approx. 5 books)
- + Elementary school (>100 books)

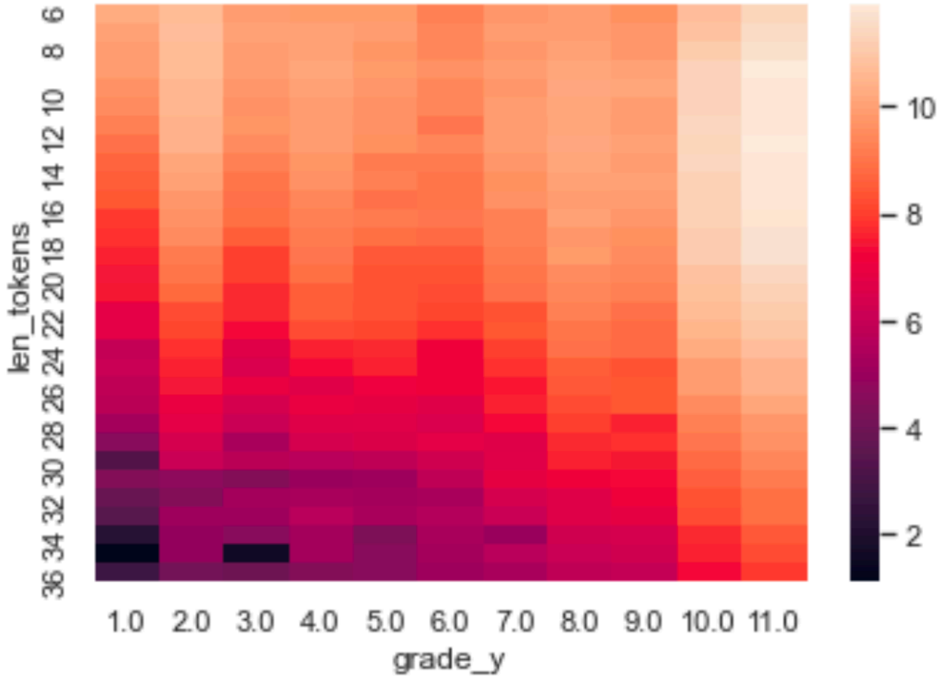
tokenization, splitting text into sentences, excluded all extremely long sentences and short sentences

Grade level	Tokens		Sentences		ASL		ASW	
	BOG	NIK	BOG	NIK	BOG	NIK	BOG	NIK
5-th	-	17 221	-	1 499	-	11.49	-	2.35
6-th	16 467	16 475	1 273	1 197	12.94	13.76	2.56	2.71
7-th	23 069	22 924	1 671	1 675	13.81	13.69	2.84	2.70
8-th	49 796	40 053	3 181	2 889	15.65	13.86	2.96	2.88
9-th	42 305	43 404	2 584	2 792	16.37	15.55	3.04	3.00
10-th	75 182	39 183	4 468	2 468	16.83	15.88	3.07	3.12
10-th*	98 034	-	5 798	-	16.91	-	3.05	-
11-th	-	38 869	-	2 270	-	17.12	-	3.11
11-th*	100 800	-	6 004	-	16.79	-	3.19	-

Corpora of Academic texts

Sentence level

- Dataset of sentence pairs
- Dataset of sentences



File Name (book)	Grade Level	Label	Category	ASL	AWL	Total Sentences
year_bog_11p.txt	11.50	11	high	18.63	7.13	5,648
year_petrov_11.txt	11.00	11	high	15.65	6.71	5,029
year_guryan_11.txt	11.00	11	high	16.01	6.72	5,848
year_nik_11.txt	11.00	11	high	17.99	6.95	2,078
year_ponom_11.txt	11.00	11	high	15.71	6.65	2,190
year_plenko_11.txt	11.00	11	high	16.66	6.63	3,470
year_bog_10p.txt	10.50	11	high	19.07	6.92	5,579
year_sobol_10.txt	10.00	10	high	15.93	6.75	5,236
year_unk_10.txt	10.00	10	high	16.19	6.68	2,000
year_nik_10.txt	10.00	10	high	17.81	6.91	2,271
year_bog_10.txt	10.00	10	high	18.27	6.78	3,145
year_klimov_10.txt	10.00	10	high	17.09	6.76	3,967
year_bog_09.txt	9.00	9	high	17.88	6.68	1,710
year_nik_09.txt	9.00	9	high	16.90	6.79	2,480
year_bog_08.txt	8.00	8	medium	17.49	6.72	2,999
year_nik_08.txt	8.00	8	medium	15.74	6.41	1,821
year_nik_07.txt	7.00	7	medium	15.41	6.14	1,509
year_bog_07.txt	7.00	7	medium	15.00	6.46	1,632
year_nik_06.txt	6.00	6	medium	15.94	6.18	1,029
year_bog_06.txt	6.00	6	medium	15.13	5.86	985
year_nik_05.txt	5.00	5	medium	13.11	5.57	1,566
year_vah_4pu.txt	4.50	5	medium	15.78	5.86	1,174
year_vah_4u.txt	4.00	4	low	13.78	6.17	1,423
year_ben_4u.txt	4.00	4	low	12.72	6.38	604
year_gor_4u.txt	4.00	4	low	14.75	6.56	833
year_rud_3u.txt	3.00	3	low	14.15	5.67	1,319
year_vah_2pu.txt	2.50	3	low	11.13	5.75	1,005
year_uch_2pu.txt	2.00	2	low	14.14	5.73	1,559
year_uch_2u.txt	2.00	2	low	12.00	6.05	1,621
year_vah_2u.txt	2.00	2	low	11.10	5.73	1,100
yead_rud_2u.txt	2.00	2	low	13.40	5.44	619
year_vah_1pu.txt	1.50	2	low	11.22	5.69	292
year_rag_1u.txt	1.00	1	low	8.76	5.95	74
year_rog_1u.txt	1.00	1	low	10.33	5.95	468
year_rud_1u.txt	1.00	1	low	12.20	5.17	495
year_lut_1u.txt	1.00	1	low	9.74	6.36	390
year_kur_1u.txt	1.00	1	low	9.86	6.15	200
year_vah_1u.txt	1.00	1	low	11.17	5.38	139

Tasks and model types

depending on the dataset we have different task setups

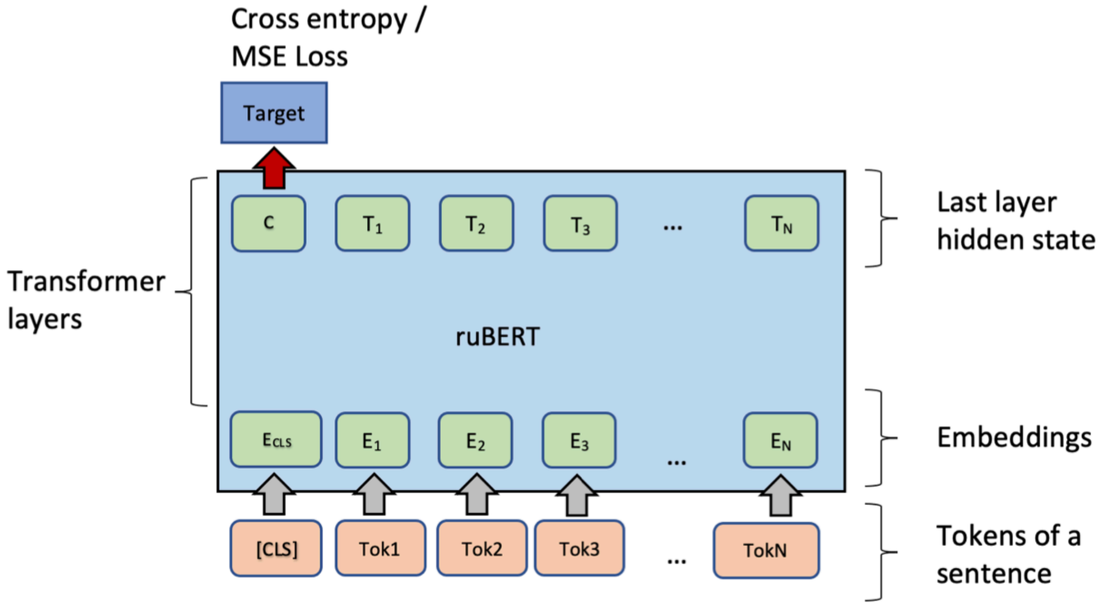
Dataset	Regression	Binary Classification	Multiclass Classification
1-s dataset	grade level (value)	-	11 or 3 categories
2-s dataset	difference between grade levels	complex / simple sentence	-

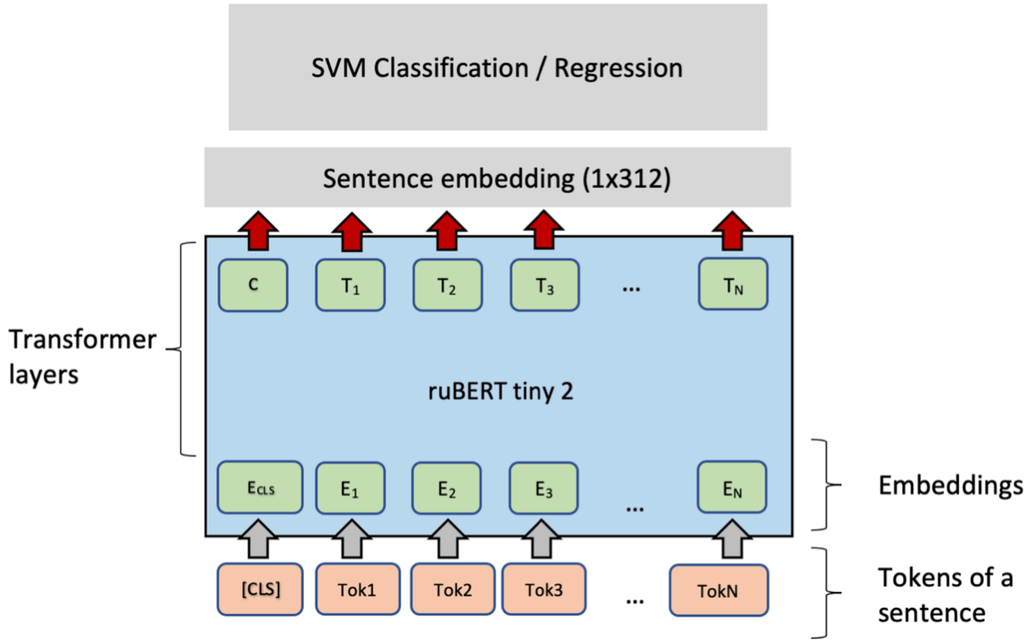
Types of Models:

- Linear regression
- Transformer-based
- GNN-based

Model architectures

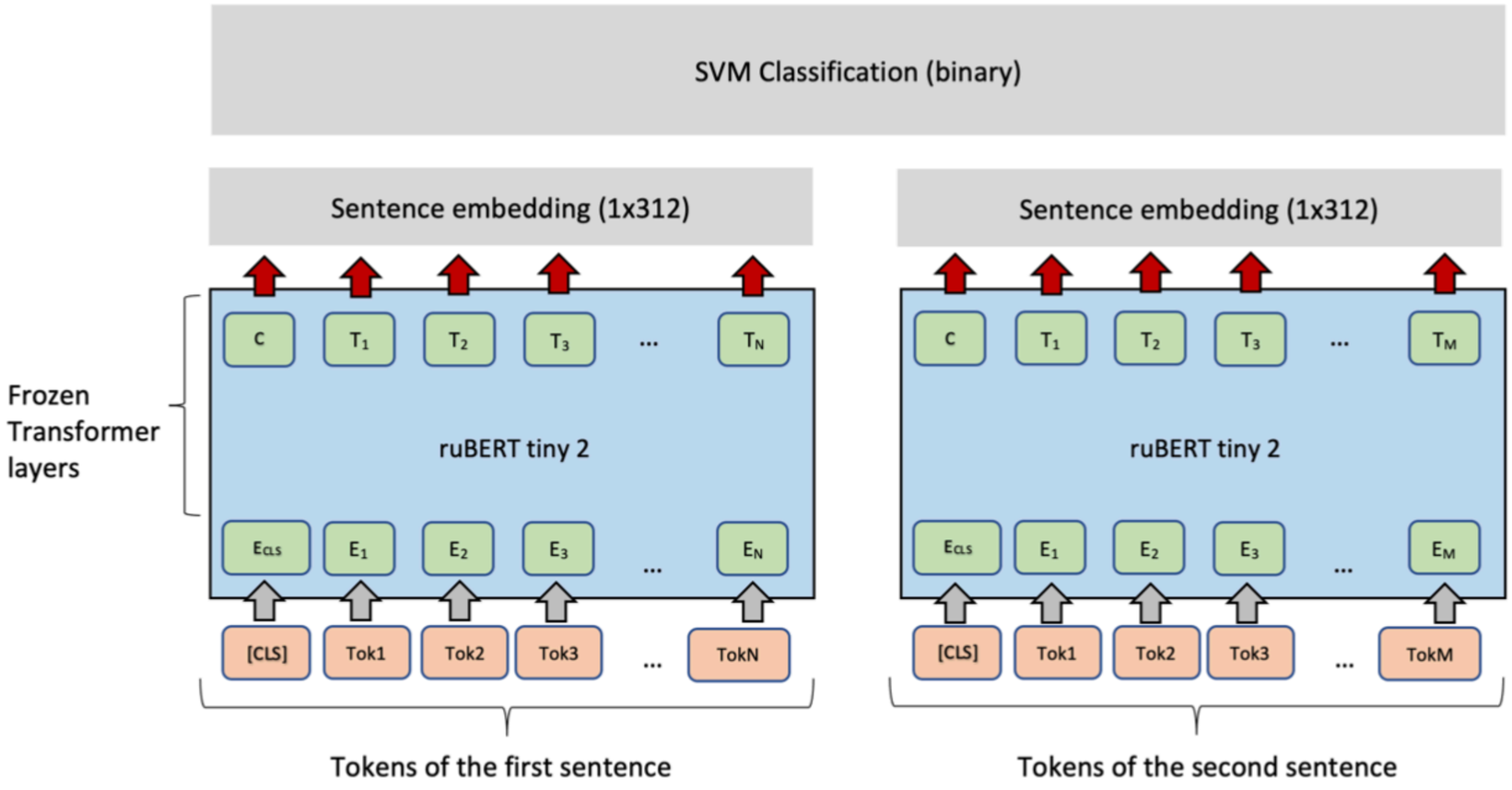
for single sentence complexity prediction





Model architectures

for classification of pairs of sentences

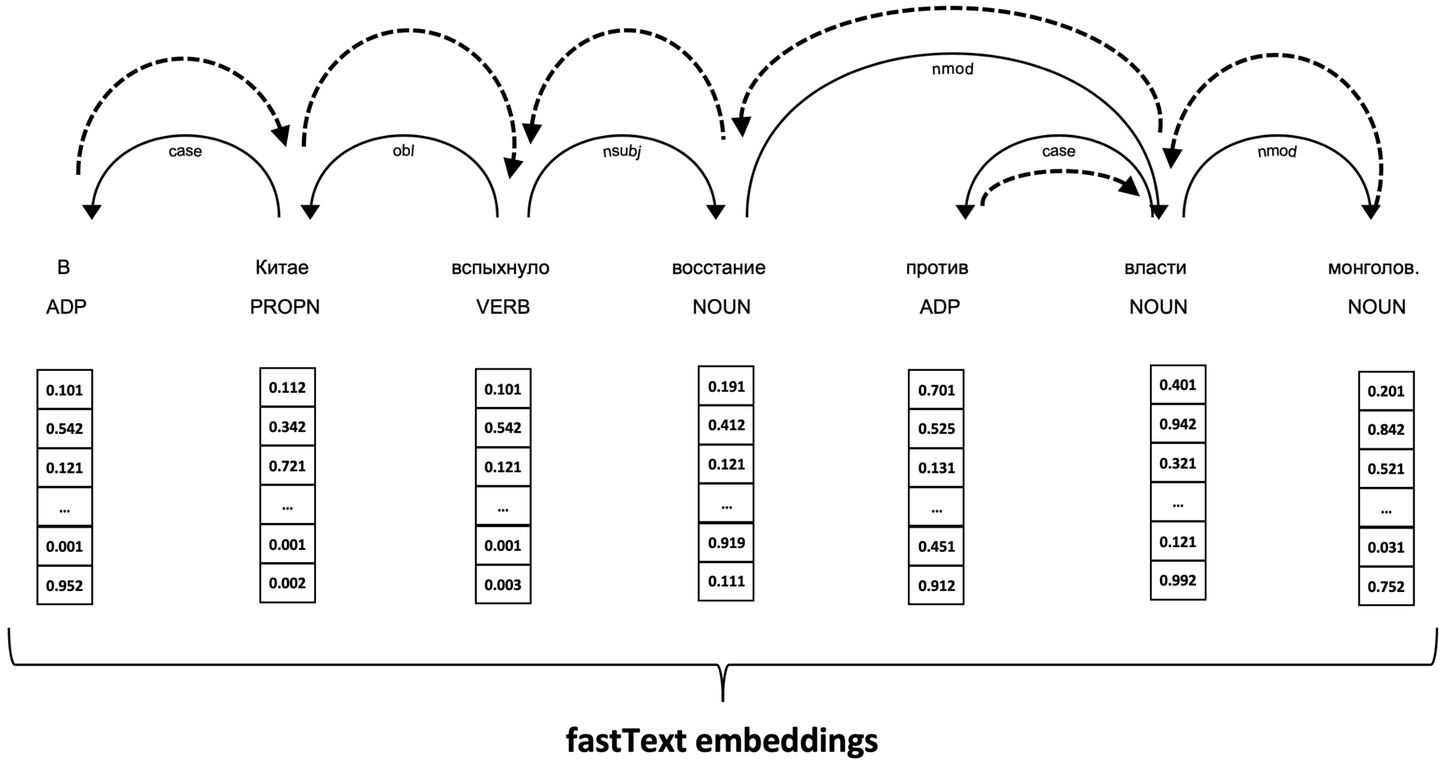


Model architectures

Graph Neural Networks

- Multi-layer Graph Convolutional Network (GCN) applied for Dependency Tree

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} h_j^{(l)} W^{(l)}\right)$$



Results

Regression

	1-s data			2-s data		
Model	R^2	MSE	MAE	R^2	MSE	MAE
Linear Reg. on 2 parameters (SL+AWL)	0.13	8.34	2.32	-	-	-
Linear Reg. on Sentence Embeddings	0.65	3.39	1.37	0.73	10.55	2.56
SVR on Sentence Embeddings	0.71	2.79	1.11	0.77	8.96	2.28
Fine-tuned RuBERT	0.80	1.96	0.80	0.98	0.79	0.56
GNN	0.73	2.58	1.10	0.97	1.15	0.75

Results

Classification

2-s dataset (with margin=3)				
Model	Accuracy	F1	P	R
SVM on Sentence Embeddings	0.93	93.18	93.18	93.18
Fine-tuned RuBERT	0.98	98.47	98.44	98.50
GNN	0.97	96.60	96.60	96.60

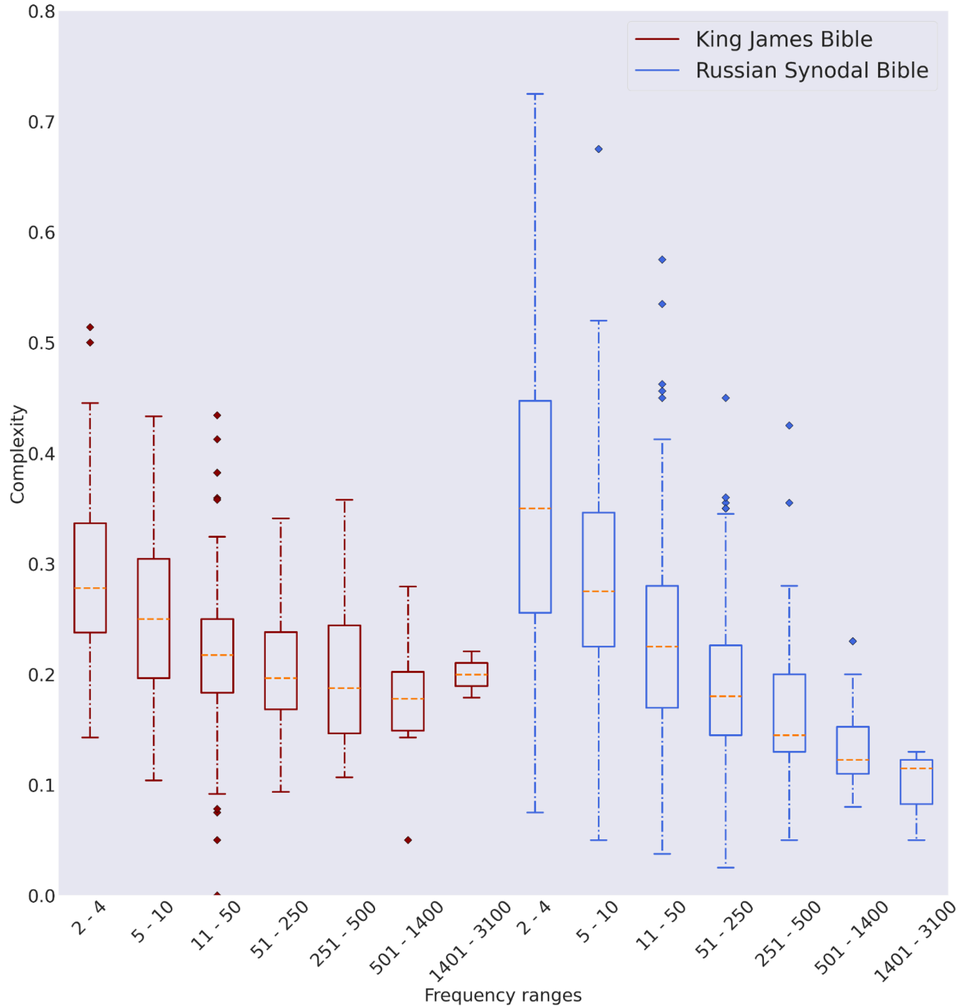
Complexity of words

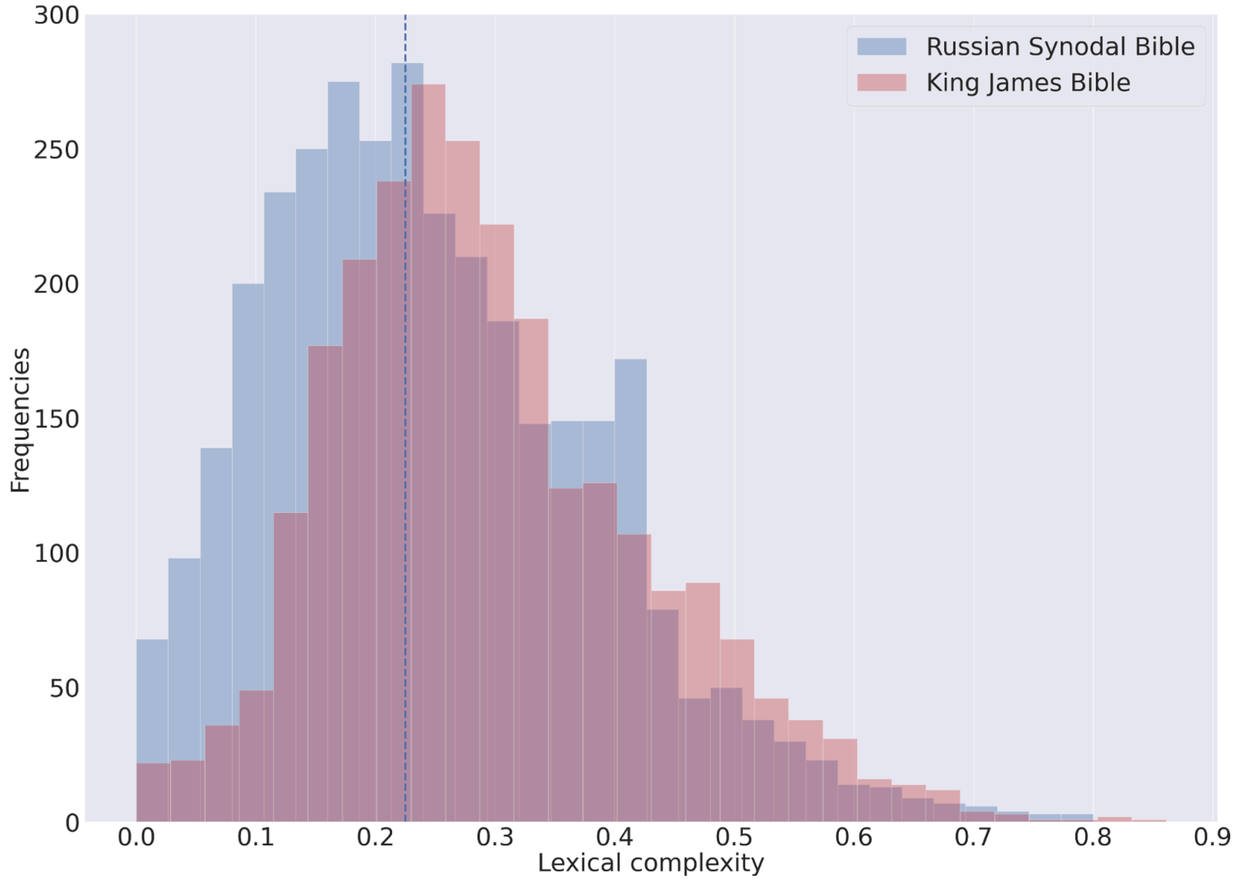
Lexical complexity prediction

- Commonly this task is referred to as Complex Word Identification (CWI) or Lexical Complexity Prediction (LCP).
- Following (Paetzold, 2016; Zampieri 2017; Yimam 2018; Shardlow, 2021)
- Results for **Russian** (Abramov, Ivanov, 2022)
  - a corpus consisting of 931 distinct words that occurred within 3,364 different contexts
  - We evaluate a linear regression model as a baseline
  - handcrafted features, fastText and ELMo embeddings of target words.

Annotation Results

from Holy Bible





Modeling Results

Table 2: The result of linear regression on handcrafted features (HC), Fasttext and ELMo embeddings and concatenated features

	Handcrafted	**Fasttext**	**ELMo**	**Fasttext+HC**	**ELMo+HC**
MAE	0.102	0.084	0.099	0.084	0.099
Pearson correlation	0.342	0.614	0.498	0.619	0.501

Table 3: Pearson correlation between handcrafted features

	Frequency	**Word length**	**Number of syllables**
Frequency	1	-0.206	-0.172
Word length	-0.206	1	0.819
Number of syllables	-0.172	0.819	1

Summary

- Text complexity is a multilayer phenomenon
  - whole text, passage, sentence, word
  - frequency / statistical features, syntactical features, semantic, contextual features
- Domain / genre dependent
- Multilingual text complexity analysis

References

- Abramov, Aleksei V., and Vladimir V. Ivanov. "Collection and evaluation of lexical complexity data for Russian language using crowdsourcing." Russian Journal of Linguistics 26.2 (2022): 409-425.
- Solovyev, Valery, et al. "Prediction of reading difficulty in Russian academic texts." Journal of intelligent & fuzzy systems 36.5 (2019): 4553-4563.
- Solnyshkina, Marina, Vladimir Ivanov, and Valery Solovyev. "Readability formula for Russian texts: a modified version." Mexican International Conference on Artificial Intelligence. Springer, Cham, 2018.
- Shardlow, Matthew, Michael Cooper & Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI). 57-62.
- Shardlow, Matthew, Richard Evans, Gustavo H. Paetzold & Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). 1-16. DOI: 10.18653/v1/2021.semeval-1.1
- Iavarone, B., Brunato, D., and Dell'Orletta, F. (2021). Sentence complexity in context. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Online: Association for Computational Linguistics), 186–199. doi:10.18653/v1/2021.cmcl-1.23
- Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. Journal of the american society for information science and technology 56, 1448–1462
- Crossley, S. A., Greenfield, J., and McNamara, D. S. (2008). Assessing text readability using cognitively based indices. Tesol Quarterly 42, 475–493
- Schumacher, E., Eskenazi, M., Frishkoff, G., and Collins-Thompson, K. (2016). Predicting the relative difficulty of single sentences with and without surrounding context. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (Austin, Texas: Association for Computational Linguistics), 1871–1881. doi:10.18653/v1/D16-1192
- Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In Proceedings of the third workshop on innovative use of NLP for building educational applications. 71–79
- Paetzold, Gustavo & Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 560-569. DOI: 10.18653/v1/S16-1085
- Zampieri, Marcos, Liling Tan & Josef van Genabith. 2016. Macsaar at semeval-2016 task 11: Zipfian and character features for complex word identification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 1001-1005. DOI: 10.18653/v1/S16-1155
- Zampieri, Marcos, Shervin Malmasi, Gustavo Paetzold & Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017). 59-63.