Quantile-based approach to estimating cognitive text complexity

Maksim Eremeev, Vorontsov Konstantin

(Moscow State University • Moscow Institute of Physics and Technology)

Moscow • June 2020

Readability indices

Readability indices measure the perception complexity of a text

- Most common:
  — *Flesch-Kincaid Index*,
  — *Automated Readability index (ARI)*,
  — *SMOG-index*
- Combine simple parameters:
  — length of words and syllables,
  — mean number of letters in a syllable
  — mean umber of words in a sentence
- Comprise of large number of language-dependent constants

Flesch, R. How To Test Readability. 1951.
Senter, R.J. and Smith, E.A. Automated Readability Index. 1967.

Readability indices

Automated Readability index

$$ARI(d) = 4.71 \times \frac{c}{w} + 0.5 \times \frac{w}{s} - 21.43,$$

c — total number of letters in the document d,
w — total number of words,
s — total number of sentences in d.

Flesch-Kincaid readability test

$$Flesh(d) = 206.835 - 1.015 \times \frac{w}{s} - 84.6 \times \frac{syl}{w},$$

syl — total number of syllables.

Reading order to rank search results

Reading Order — order of documents from most general and simple to most specific and complex.

Applications:

- Ranking the exploratory information search results
- Personalizing learning pathways

Topic Modeling-based approach:

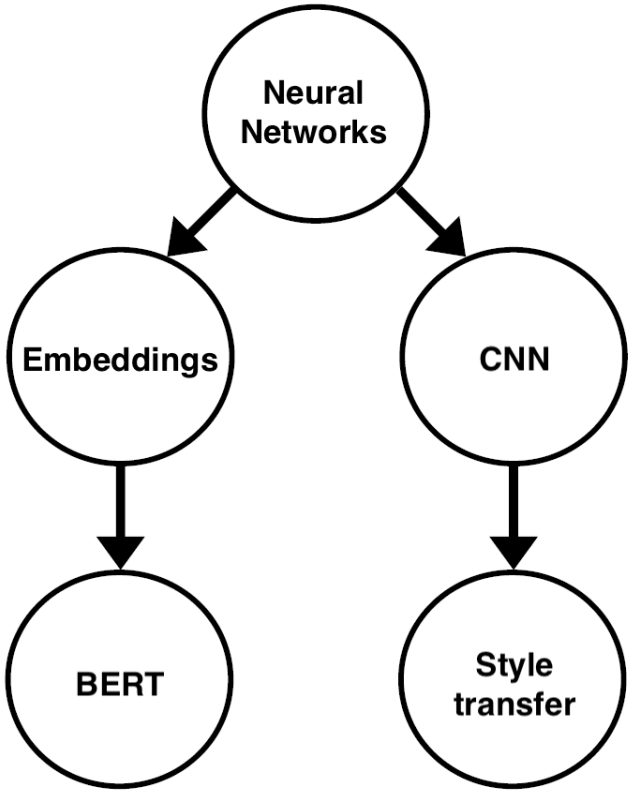
- Generality score of a document — enthropy of the topic vector
- Overlap score of two documents — cosine distance between their topic embeddings
- Reading sequence builds from highly-overlapped documents in order of decreasing generality

---

Georgia Koutrika, Lei Liu, Steven Simske. Generating Reading Orders over Document Collections. 2015.

Reading order to rank search results

Reading sequence can not only be a flat list but even a tree.



Calculating generality and overlap scores is not enough to build complete reading sequences. There is a need for more advanced parameters.

Psychophysiology pressure decoding a speech

What is a «hard text» and how is it different from an easy one?

Example. Frequency of the letter р in Russian is 0.04, but in this case it is increased to 0.17:

Conjectures from neuropsychology of speech:

- Decoding of each language element (token) puts pressure on a specific zone of the cortex
- The zone needs *refractoriness time* to recover
- The established frequency distributions for each token are strongly uneven (Zipf's law)
- For frequent language elements refractoriness time is less

A.Birkin. Code of speech. Saint-Perterburg.: Hyppocrat, 2007.

Ideas of multi-level complexity estimation

Foundations of the suggested approach:

- *language levels*: morphological, lexical, syntactic, discursive
- on level h text can be considered as a sequence of *tokens* from the alphabet A_h
- *text complexity* on level h is a percentage of abnormally complex tokens
- token complexity is *abnormally high*, if it exceeds the 95%-quantile of its complexity in the reference collection
- *a reference collection* comprises of moderately complex documents for the selected class of readers

---

M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

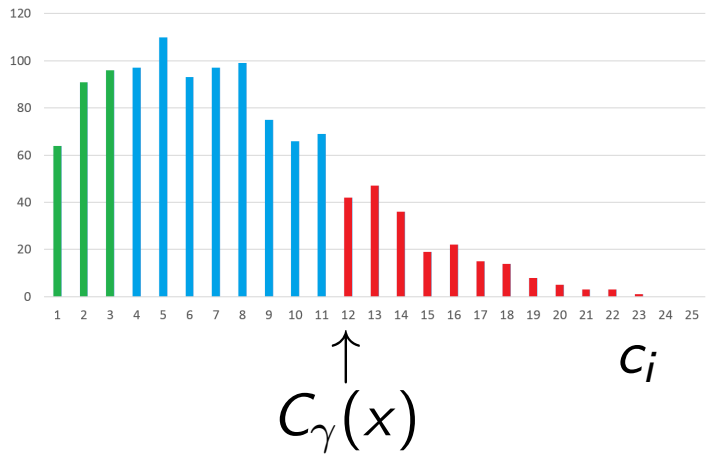
Quantile-based approach to estimating text complexity

$x_1, \ldots, x_n$ — sequence of tokens from A_h in the document d;
c_i — *complexity* of the token x_i
w_i — *complexity weight* of the token x_i

Text complexity score — total weight of complex tokens:

$$W(d) = \sum_{i=1}^{n} w_i \left[c_i > C_\gamma(x_i)\right]$$

$C_\gamma(x)$ denotes the γ-quantile of the complexities distribution of the token x in the *reference collection* of simple texts



M.Eremeev, K.Vorontsov. Lexical quantile-based text complexity measure. RANLP-2019.

Distance-based and counter-based models

Distance-based complexity function

- for each token $a \in A_h$ we construct an empirical distribution of its complexity scores $\{c_i : x_i = a\}$ over the reference collection
- $c_i = f(r_i)$ — decreasing function of the distance r_i of the token x_i to its previous occurrence
- examples: distances for letters, n-grams, words

Counter-based token complexity function

- considering single-element alphabet $A_h = \{a\}$, we construct an empirical distributions of all complexities $\{c_i\}$ over the reference collection
- c_i — numerical feature of the token x_i
- example: lengths of words, sentences, syntactic dependency

Phonetic level, letters as tokens

r_i — distance from token x_i to its previous occurrence:

$$\ldots \quad \boxed{x_{i-r_i} = a} \quad \underbrace{x_{i-r_i+1} \quad x_{i-r_i+2} \quad \ldots \quad x_{i-2} \quad x_{i-1} \quad \boxed{x_i = a}}_{r_i} \quad \ldots$$

If there is no previous occurrence, we redefine r_i, so that the sum of r_i over all tokens $x_i = a$ equals the document length.

token	t	h	e	g	r	e	a	t	g	a	t	s	b	y
r_i	–	–	–	–	–	3	–	7	5	3	3	–	–	–
redefined r_i	4	15	11	9	14	3	11	7	5	3	3	14	14	14

Examples of c_i:

$$c_i = C - r_i, \qquad c_i = 1/r_i$$

where C is an arbitrary constant.

A. Birkin. Nature of speech. Moscow.: Likbez, 2009.

Morphological level, *n*-grams as tokens

Tokens can be defined as:

- syllables,
- letter *n*-grams,
- letter-wise sorted letter *n*-grams,
- morphemes (prefix, root, suffix)

Example.

1-grams	t	h	e	g	r	e	a	t	g	a	t	s	b	y

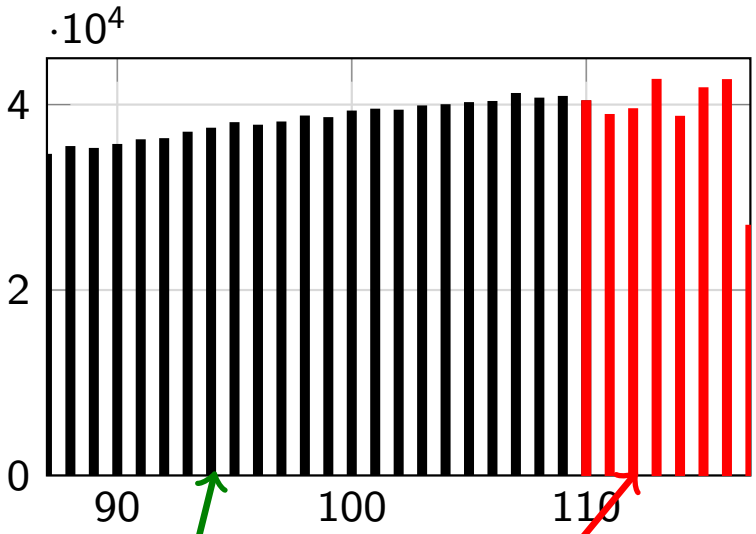
unsorted *n*-grams

2-grams	th	he		gr	re	ea	at		ga	at	ts	sb	by	
3-grams	the			gre	rea	eat			gat	tsb	sby			

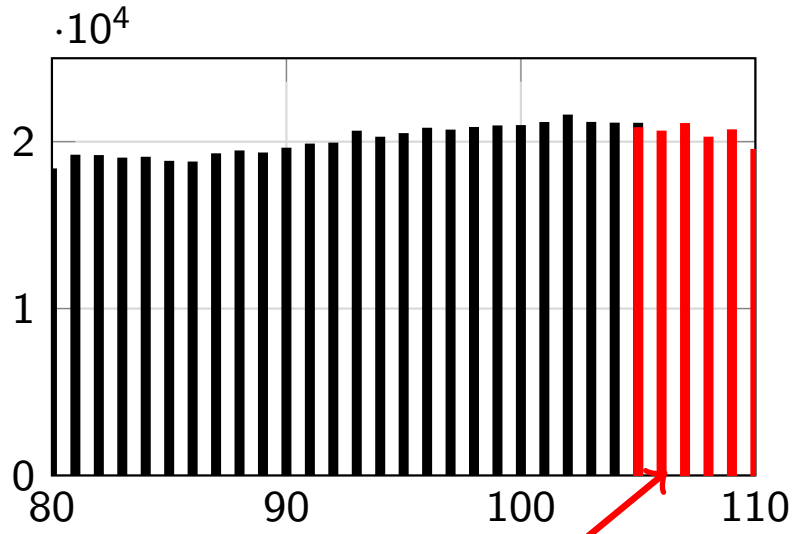
sorted *n*-grams

2-grams	ht	eh		gr	er	ae	at		ag	at	st	bt	by	
3-grams	eht			egr	aer	aet			agt	bst	bsy			

Example of the lexical distance-based model trained on Russian Wikipedia



Complexity score c_i for syllable «Да»



Complexity score c_i for syllable «Ши»

Да же ше ю, да же у ши ты испачкал в черной ту ши . Становись скорей под душ. Смой с ушей под ду шем тушь.

Lexical level, words as tokens

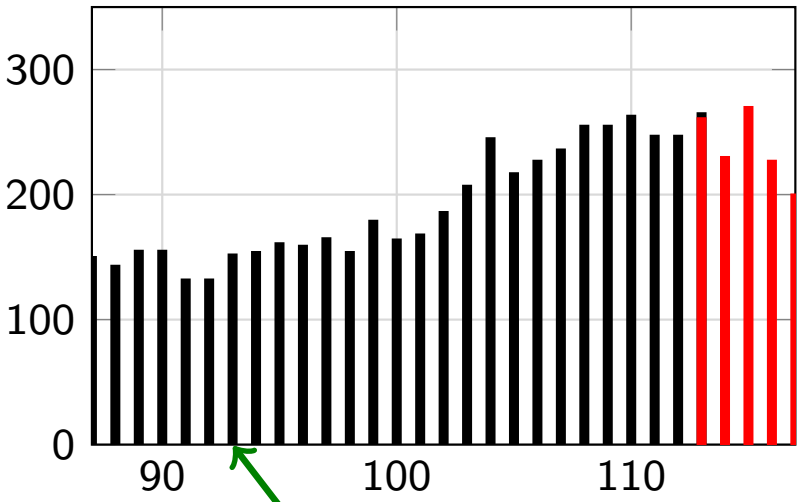
Distance-based models:

- words,
- lemmatized words,
- collocations,
- specific terms

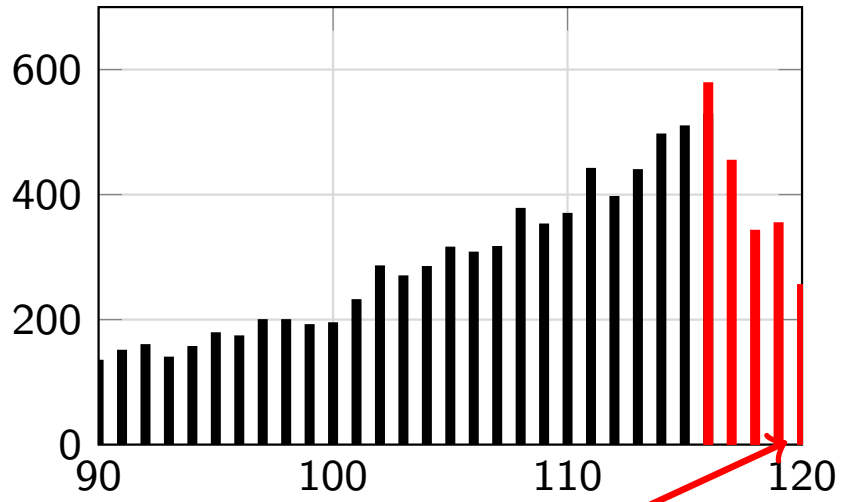
Counter-based models:

- c_i — word length
- $c_i = \frac{1}{\text{count}(x_i)}$ — rarity of the word in the reference collection, $\text{count}(x_i)$ denotes the words frequency

Example of the lexical distance-based model trained on Russian Wikipedia



Complexity score c_i for word «Физика»



Complexity score c_i for word «Уравнение»

Параболические уравнения, уравнения теплопроводности и эллиптические уравнения составляют основу исследований в математической физике ...

Syntactic level

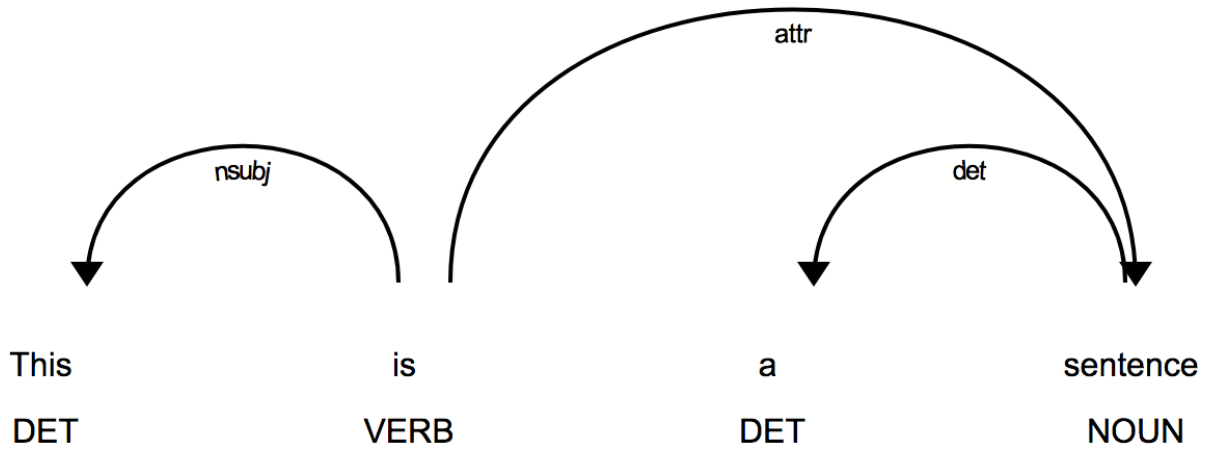
We acquire dependencies from sentences using UDPipe parser

Distance-based models:

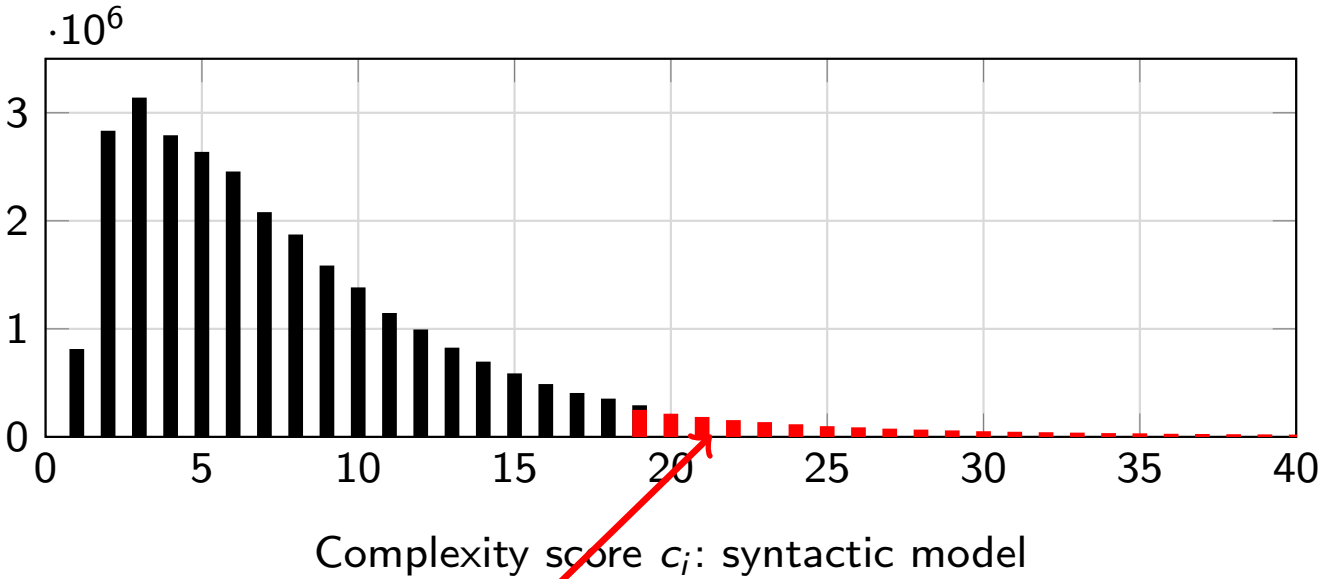
- grammatical structures — *syntgams*, i.e. pairs of part of speech and type of sentence part

Counter-based models: Tokens — sentences

- c_i — max length of the syntactic dependency in a sentence
- c_i — mean length of the syntactic dependency in a sentence



Example of the syntactic counter-based model trained on Russian Wikipedia

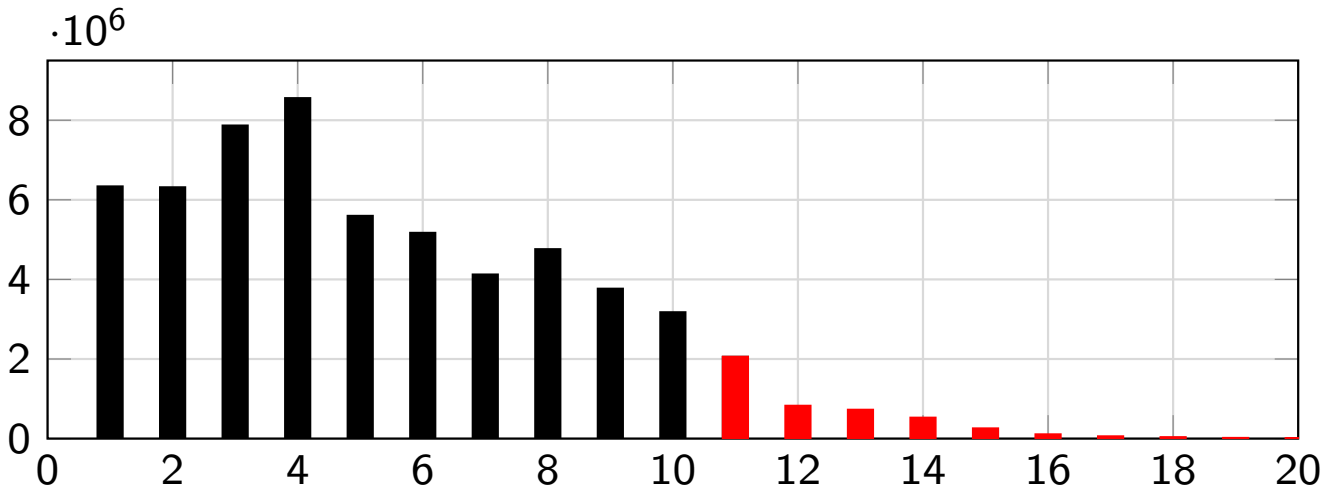


Потом долгие годы он любил думать, что был в очень большой опасности во время этого ужасного потопа, но единственная опасность угрожала ему только в последние полчаса его заключения, когда Сова уселась на ветку и, чтобы его морально поддержать, стала рассказывать ему длиннейшую историю про свою Тетку, которая однажды по ошибке снесла гусиное яйцо, и история эта тянулась и тянулась (совсем как эта фраза)...

Discursive level

Counter-based models: Tokens — sentences

- c_i — number of words in a sentence
- c_i — number of *connector words* in a sentence (и, или, значит, который, чтобы, . . . for Russian)



Complexity score c_i: discursive model

Supervised cognitive complexity model

Let $W_k(d), k = 1, \ldots, K$ be different complexity models.

Linear aggregated complexity score with params α_k:

$$W(d, \alpha) = \sum_{k=1}^{K} \alpha_k W_k(d), \quad \alpha_k \geqslant 0.$$

Dataset of labeled document pairs:
$d \prec d'$ — document d' is more complex than document d.

Loss function for the aggregated complexity model:

$$\sum_{d \prec d'} \mathscr{L}\Big(\underbrace{W(d', \alpha) - W(d, \alpha)}_{\text{pair-wise margin}}\Big) \to \min_{\alpha},$$

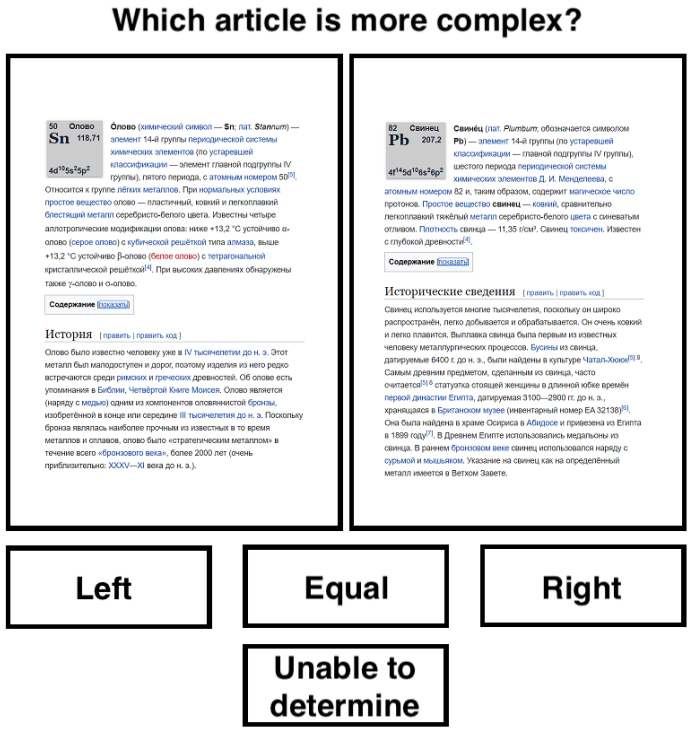
where $\mathscr{L}(M)$ denotes a smooth non-increasing margin function M.

Gathering assessors' labels

To evaluate the quality we generated a set of document pairs from Russian Wikipedia from math, physics, chemistry and computer science domains.

Assessors were asked to choose which document is more challenging to comprehend, or indicate that given documents lie in different areas.

Accuracy — percentage of pairs, where the model's result aligns with the label.



Experiment 1. Wikipedia as a reference collection

Model Class	Model	Accuracy
Readability Indexes	Automated Readability Index	50.5%
	Flesch-Kincaid Grade	44.7%
	Gunning FOG	44.4%
	Flesch Reading Ease	50.7%
	Linsear Write	45.2%
	Coleman-Liau	**52.1%**
Morphological	Letter	63.7%
	Syllables	70.9%
	Sorted Syllables	73.1%
Lexical	Length	42.4%
	Distance	**75.0%**
	Counter	71.2%
Syntactic	Length	62.0%
	Syntgam	64.2%
Discursive	Connectors	62.5%

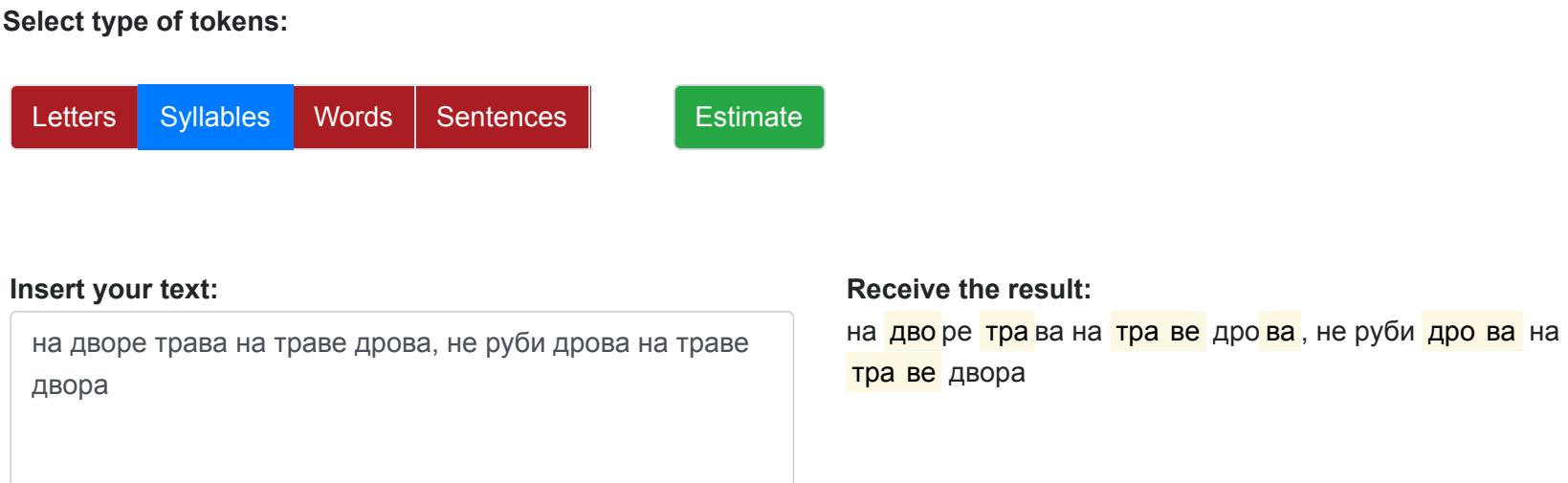
Experiment 2. Aggregated models

Model	Margin Function	**Accuracy**
Coleman-Liau	-	52.1%
Morphological Sorted Syllables	-	73.1%
Lexical Distance	-	75.0%
Syntactic Syntgams	-	64.2%
Connectors	-	62.5%
Ensemble	Negative SE	**88.1%**
Ensemble	Negative sigmoid	84.6%
Ensemble	Negative AE	85.1%

- Ensembling models improves the overall quality
- All quantile-based models outperform the readability indices
- Lexical distance-based model performs best among the proposed models

Web-resource TextComplexity.Net

Interactive web-system highlights abnormally complex tokens on every language level for a given text.
TextComplexity.net was implemented using the original open-source library *Cognitive-Complexity*.



https://github.com/maks5507/cognitive-complexity
http://textcomplexity.net

Conclusions

- We *proposed* a quantile-based approach to estimating cognitive text complexity. It is uniformly applicable for all language levels .
- We *suggested* the way to evaluate the complexity models.
- We *introduced* a supervised linear aggregation method, combining models from all levels.
- We *implemented* a demo web-resource TextComplexity.net, allowing to test the models on real texts.
- We *developed* an open-source library Cognitive-Complexity, featuring all code used in this research work.
- Experiments *show* high target metric value and competitiveness of our approach.