A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity (2016)

by **Ildiko Pilan, Sowmya Vajjala, Elena Volodina**

Подготовила Наталья Исупова

Цель: предсказать сложность текста на шведском для изучения его иностранцами

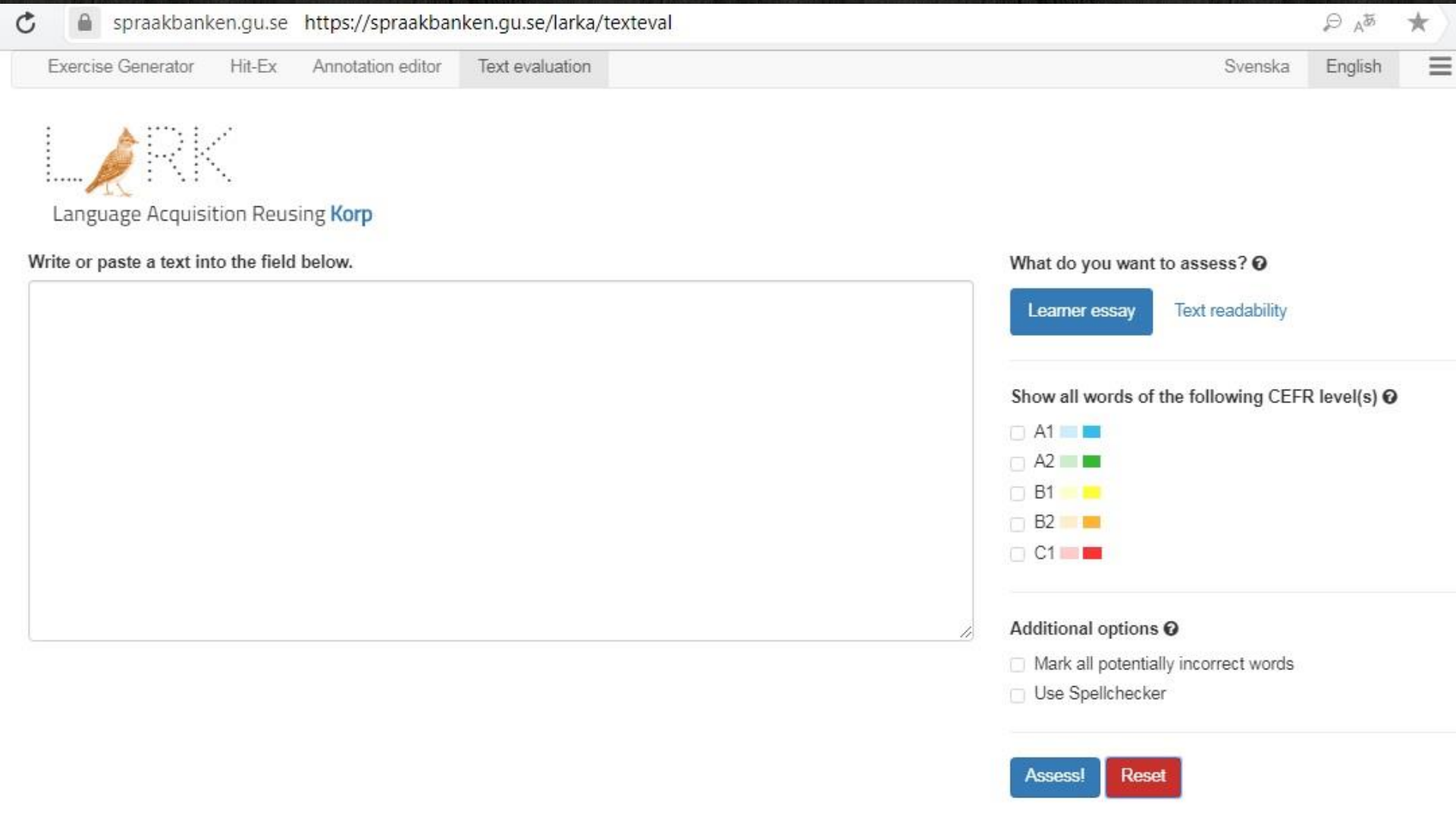
Результат: модель с точностью 81.3% и F-мерой 0.8

Läsbarthetsindex, 'Readability index'

$$LIX = \frac{A}{B} + \frac{C \times 100}{A}$$

A – количество слов в тексте,
B – количество предложений в тексте,
C – количество слов длиннее 6 букв

< 30 – детские тексты, 30-40 – газетные статьи, 40-50 – журнальные статьи, 50-60 – официальные тесты, > 60 - законы


spraakbanken.gu.se https://spraakbanken.gu.se/larka/texteval
Exercise Generator
Hit-Ex
Annotation editor
Text evaluation
Svenska
English
LARK
Language Acquisition Reusing Korp
Write or paste a text into the field below.
What do you want to assess?
Learner essay
Text readability
Show all words of the following CEFR level(s)
A1
A2
B1
B2
C1
Additional options
Mark all potentially incorrect words
Use Spellchecker
Assess!
Reset

Данные

- COCTAILL corpus
- Уровни A1-C1
- 12 книг (4 издательства)
- Тексты аннотированы: POS и dependency tags
- 867 текстов
- Основаны на диалогах

Данные

CEFR	Books	Publ.	Texts	Mean num. sentences
A1	4	3	49	14.0
A2	4	3	157	13.8
B1	5	3	258	17.9
B2	4	3	288	26.6
C1	2	2	115	42.1
Total	12	4	867	

Features

Nr.	Feature Name	Nr.	Feature Name
	Length-based		*Morphological*
1	Sentence length	30	Modal verbs to verbs
2	Average token length	31	Particle IncSc
3	Extra-long words	32	3SG pronoun IncSc
4	Number of characters	33	Punctuation IncSc
5	LIX	34	Subjunction IncSc
	Lexical	35	S-verb IncSc
6	A1 lemma IncSc	36	S-verbs to verbs
7	A2 lemma IncSc	37	Adjective IncSc
8	B1 lemma IncSc	38	Adjective variation
9	B2 lemma IncSc	39	Adverb IncSc
10	C1 lemma IncSc	40	Adverb variation
11	C2 lemma IncSc	41	Noun IncSc
12	Difficult word IncSc	42	Noun variation
13	Difficult noun and verb IncSc	43	Verb IncSc
14	Out-of-Kelly IncSc	44	Verb variation
15	Missing lemma form IncSc	45	Nominal ratio
16	Avg. Kelly log frequency	46	Nouns to verbs
	Syntactic	47	Function word IncSc
17	Average dependency length	48	Lexical words to non-lexical words
18	Dependency arcs longer than 5	49	Lexical words to all tokens
19	Longest dependency from root node	50	Neuter gender noun IncSc
20	Ratio of right dependency arcs	51	Con- and subjunction IncSc
21	Ratio of left dependency arcs	52	Past participles to verbs
22	Modifier variation	53	Present participles to verbs
23	Pre-modifier IncSc	54	Past verbs to verbs
24	Post-modifier IncSc	55	Present verbs to verbs
25	Subordinate IncSc	56	Supine verbs to verbs
26	Relative clause IncSc	57	Relative structure IncSc
27	Prepositional complement IncSc	58	Bilog type-token ratio
	Semantic	59	Square root type-token ratio
28	Avg. nr. of senses per token	60	Pronouns to nouns
29	Noun senses per noun	61	Pronouns to prepositions

Length-based features

1. Длина предложений
2. Средняя длина токена
3. Длинные слова (длиннее 13 символов)
4. Количество символов
5. LIX

Lexical features

KELLY Project (KEywords for Language Learning for Young and adults alike)

Вместо процентов – IncSc (incidence score) на 1000 слов: “The IncSc of a category was computed as 1000 divided by the number of tokens in the text or sentence multiplied by the count of the category in the sentence”.

Lexical features

1. A1 lemma IncSc
2. A2 lemma IncSc
3. B1 lemma IncSc
4. B2 lemma IncSc
5. C1 lemma IncSc
6. C2 lemma IncSc
7. Difficult word IncSc (выше уровня текста)
8. Difficult noun and verb IncSc (выше уровня текста)
9. Out-of-Kelly IncSc (нет в списке)
10. Missing lemma form IncSc (лемматизатор не распознал)
11. Avg. Kelly log frequency

Morphological features

1. Modal verbs to verbs
2. Particle IncSc
3. 3SG pronoun IncSc
4. Punctuation IncSc
5. Subjunction IncSc
6. S-verb IncSc
7. S-verbs to verbs
8. Adjective IncSc
9. Adjective variation (отношение категории к остальным категориям (сущ., прил., глаг., нареч.)
10. Adverb IncSc
11. Adverb variation
12. Noun IncSc
13. Noun variation
14. Verb IncSc
15. Verb variation
16. Nominal ratio
17. Nouns to verbs
18. Function word IncSc

Morphological features

19. Lexical words to non-lexical words (отношение к другим категориям)
20. Lexical words to all tokens
21. Neuter gender noun IncSc
22. Con- and subjunction IncSc
23. Past participles to verbs
24. Present participles to verbs
25. Past verbs to verbs
26. Present verbs to verbs
27. Supine verbs to verbs
28. Relative structure IncSc (consisted of relative adverbs, determiners, pronouns and possessives)
29. Bilog type-token ratio (Log all types / Log all token)
30. Square root type-token ratio (all types / root of all token),
31. Pronouns to nouns
32. Pronouns to prepositions

Syntactic features

1. Average dependency length (MaltParser)
2. Dependency arcs longer than 5
3. Longest dependency from root node
4. Ratio of right dependency arcs
5. Ratio of left dependency arcs
6. Modifier variation
7. Pre-modifier IncSc (e.g. adjectives and prepositional phrases)
8. Post-modifier IncSc (see 7)
9. Subordinate IncSc
10. Relative clause IncSc (E.g.: It is John (whom) Jack is waiting for.)
11. Prepositional complement IncSc

Semantic features

1. Avg. nr. of senses per token (использовали SALDO — шведская альтернатива Princeton WordNet)
2. Noun senses per noun

Эксперименты и результаты

Использовали WEKA (Waikato Environment for Knowledge Analysis), параметры по умолчанию (на Java)

1. a multinomial logistic regression model with ridge estimator (polytomous LR, multiclass LR, softmax regression, multinomial logit, the maximum entropy (MaxEnt) classifier, conditional maximum entropy model)
2. a multilayer perceptron,
3. support vector machine learner, Sequential Minimal Optimization (SMO),
4. decision tree (J48).

Эксперименты и результаты

Type	Number	Accuracy %	F	RMSE
Majority	-	33.2	0.17	0.52
LIX	1	34.9	0.22	0.38
Lex	**11**	**80.3**	**0.80**	**0.24**
All	**61**	**81.3**	**0.81**	**0.27**

multinomial logistic regression model with ridge estimator

Експерименты и резултаты

Accuracy %

Type	Number	Perceptron	SMO	J48
Lex	11	**77.4**	42.1	55
All	61	62.2	52.7	50.5

Эксперименты и результаты

Confusion matrix

Predictions						
A1	A2	B1	B2	C1		Label
37	12	0	0	0	A1	
12	121	18	5	1	A2	
4	11	206	24	13	B1	
0	5	21	238	24	B2	
0	0	0	12	103	C1	