LITHUANIAN PEDAGOGIC CORPUS: CORRELATIONS BETWEEN LINGUISTIC FEATURES AND TEXT COMPLEXITY

Loïc BOIZOU, Jolanta KOVALEVSKAITĖ and Erika RIMKUTĖ

Vytautas Magnus University

lboizou@gmail.com, jolanta.kovalevskaite@vdu.lt, erika.rimkute@vdu.lt

Introduction

This research discusses the problem of automatic CEFR level assignment to texts. Specifically, we will address the linguistic features (morphological, syntactic and lexical) of the Lithuanian Pedagogic Corpus (sub-corpus of written texts from coursebooks) and their correlation with text complexity.

The study of the correlation between grammatical forms and language levels was based on the CEFR materials designed for levels A1, A2, B1, and B2 (see Stumbrienė 2016; Ramonienė et al. 2006; Ramonienė et al. 2016a ; Ramonienė et al. 2016b).

Lithuanian Pedagogic Corpus

The Lithuanian Pedagogic Corpus is a monolingual specialized corpus (669,000 tokens) which provides material relevant to learning and teaching Lithuanian as a foreign language.

The Lithuanian Pedagogic Corpus

Text level	Tokens/Language type		Total: 669,000
	Written	Spoken	
A1-A2	96,000	15,000	111,000
B1-B2	523,000	35,000	558,000

Sub-corpus of written texts

Text level	Total: 618,637
A1	6.93%
A2	8.52%
B1	10.99%
B2	73.56%

The sub-corpus of written texts includes two types of texts: 1) coursebooks of the Lithuanian language (17.2%); 2) authentic Lithuanian material: news portals, popular science books, advertisements, public information (travelling, health care, and other services), etc. (82.8%).

In total, the corpus includes 29 genres (texts from news portals, stories, fairy tales, advertisements, letters, songs, and others).

Correlations between lexical, morphological and syntactic features and the CEFR levels

Lexical surface features

Text level/lexical features	Average word length (in characters)	Coverage of the 3075 most frequent word forms	Type/token ratio
A1	5.39	**69.13%**	0.26
A2	5.59	61.05%	0.39
B1	5.95	55.86%	**0.43**
B2	**6.16**	54.23%	0.37

Morphological features

Finite and non-finite verb forms

Text level/ verb forms	Finite forms	Infinitives	Participles	Adverbial participles	Half participles
A1	**80.94**	15.98	2.83	0.09	0.16
A2	79.45	14.75	5.42	0.13	0.25
B1	68.34	**16.91**	12.86	0.97	0.90
B2	64.02	16.18	**16.51**	**1.63**	**1.63**

Verb mood

Text level/mood	Indicative	Imperative	Subjunctive
A1	90.34	6.16	3.50
A2	85.16	**9.63**	**5.21**
B1	91.28	4.08	4.64
B2	**93.08**	2.82	4.10

Tense of finite forms

Text level/tense	Present	Simple past	Past frequentative	Future
A1	**78.34**	13.91	0.16	7.60
A2	54.22	26.61	5.51	**13.66**
B1	58.15	25.66	**5.72**	10.46
B2	43.31	**47.10**	4.44	5.14

Voice and tense of participles

Text level/ voice and tense	Active present	Active simple past	Active past frequ-entative	Active future	Passive present	Passive past	Passive future	Nece-ssity
A1	1.61	33.06	0.00	0.00	**49.19**	15.32	0.00	0.81
A2	3.53	17.06	0.00	**0.59**	47.06	**30.59**	0.00	**1.18**
B1	**15.99**	21.88	**0.55**	0.18	33.09	27.57	0.18	0.55
B2	10.81	**29.58**	0.00	0.00	29.77	28.99	**0.39**	0.46

Noun case

Text level/ case	Nom.	Gen.	Dat.	Acc.	Ins.	Loc.	Voc.	Ill.
A1	**38.75**	27.59	1.46	**19.02**	3.53	**8.35**	1.24	0.04
A2	31.56	33.21	2.79	18.65	5.98	6.47	**1.35**	0.00
B1	30.05	36.03	2.77	16.88	**6.43**	7.38	0.40	**0.06**
B2	27.24	**39.32**	**3.29**	17.08	6.00	6.81	0.22	0.03

Type of numerals

Text level/type of numerals	Cardinal	Multiple	Collective	Ordinal
A1	26.75	0.56	0.00	**72.69**
A2	91.05	**1.43**	0.00	7.52
B1	**92.97**	0.67	0.00	6.35
B2	85.73	1.72	**0.08**	12.48

Degree of adjectives and adverbs

Text level/ degree	Adj. positive	Adj. comparative	Adj. superlative	Adv. positive	Adv. comparative	Adv. superlative
A1	**96.55**	1.44	2.00	**94.73**	4.02	1.25
A2	89.71	2.90	7.39	92.58	4.85	2.56
B1	87.95	3.35	**8.70**	87.56	**7.87**	**4.57**
B2	89.52	**4.21**	6.27	90.04	6.89	3.06

Syntactic features

Text level/syntactic features	Number of sentences	Average sentence length (in words)
A1	**5591**	8.08
A2	2864	10.12
B1	2575	14.44
B2	4599	**15.94**

Conclusions

The linguistic features described in the article have revealed that the automatic text classification applied earlier in Grigonytė et al. 2018 was not sufficiently precise; therefore, non-coursebook texts in the corpus should be reclassified.

In order to determine the text level automatically, it is worth considering the link between the language level and these properties:

- indicating more complex forms (participles, adverbial participles and half participles) in comparison with all verb forms;
- the usage of finite forms of past frequentative tense;
- the usage of present and past simple tense participles of the active voice;
- the usage of multiple and collective numerals;
- the usage of dative and instrumental for nouns in comparison with other cases;
- the usage of comparative and superlative degree;
- length of a sentence,
- word length,
- type/token ratio;
- the distribution of the most frequent words of the analyzed corpus.

As Grigonytė et al. 2018 suggested and as Pilán et al. 2016 demonstrated, a wider set of lexical information could strongly improve the quality of a renewed prediction on non-didactic materials.

References

Grigonytė G., Kovalevskaitė J., Rimkutė E. *Linguistically-Motivated Automatic Classification of Lithuanian Texts for Didactic Purposes*. In: Muischnek K, Müürisep K, editors. Proceedings of the Eighth International Conference Baltic HLT 2018; 2018 Sep 27-29; Tartu (Estonia). Frontiers in Artificial Intelligence and Applications, vol. 307. Amsterdam, Berlin, Tokyo, Washington, DC: IOS Press. P. 38-46.

Pilán I., Vajjala S., Volodina E. A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. International Journal of Computational Linguistics and Applications. 2016 7(1):143–59.

Ramonienė M., Pribušauskaitė J., Vilkienė L. *Pusiaukelė* (Waystage – A2). Europos Taryba; 2006.

Ramonienė M., Pribušauskaitė .J, Vilkienė L. *Slenkstis* (Threshold – B1). Vilnius: Vilniaus universiteto leidykla; 2016a.

Ramonienė M., Pribušauskaitė J., Vilkienė L. *Aukštuma* (Vantage – B2). Vilnius: Vilniaus universiteto leidykla; 2016b.

Stumbrienė V. *Lūžis* (Breakthrough – A1). Vilnius: Vilniaus universiteto leidykla; 2016.

Acknowledgements

The Lithuanian Pedagogic Corpus was collected in the project *Lithuanian Academic Scheme for International Cooperation in Baltic Studies*: http://baltnexus.lt/en/baltic-studies-project.

The corpus will be freely available on https://kalbu.vdu.lt/ in 2021.