Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200628

Lithuanian Pedagogic Corpus: Correlations Between Linguistic Features and Text Complexity

Loïc BOIZOU, Jolanta KOVALEVSKAITĖ and Erika RIMKUTĖ[1]
Vytautas Magnus University, Lithuania

Abstract. This paper discusses the problem of automatic CEFR[2] level assignment to texts. We address the correlations between the lexical, morphological and syntactic features and the different CEFR levels of the texts in the Lithuanian Pedagogic Corpus. Only the texts from coursebooks showed the correlation of investigated linguistic features with text complexity. In the coursebook sub-part of the corpus, we observed that higher language proficiency levels are associated with more complex linguistic features: their number increases in texts of higher CEFR levels from A1 to B2 (e.g., non-finite verb forms, participles, adverbial participles and half participles, dative and instrumental noun cases or longer sentences).

Keywords. Lithuanian language, Lithuanian Pedagogic Corpus, automatic text classification, text complexity, linguistic features, Common European Framework of Reference for Languages (CEFR)

1. Introduction

This paper discusses the problem of automatic CEFR level assignment to texts. Specifically, we address the linguistic features of the Lithuanian Pedagogic Corpus[3] and their correlation with text complexity. The Lithuanian Pedagogic Corpus is a small monolingual specialized corpus which provides material relevant to learning and teaching Lithuanian as a foreign language. The corpus consists of 669,000 tokens and includes 111,000 tokens of A1-A2 level texts (96,000 tokens of written and 15,000 tokens of spoken samples) and 558,000 tokens of B1-B2 level texts (523,000 tokens of written and 35,000 tokens of spoken samples) [1]. For this study, only the sub-corpus of written texts (618,637 tokens) has been used (in the corpus, A1 level texts make up 6.93 %, A2 – 8.52 %, B1 – 10.99 %, and B2 – 73.56 %).

The data for this sub-corpus was collected from 1) coursebooks of the Lithuanian language (17.2 %) and 2) a variety of authentic Lithuanian material (82.8 %): news portals, popular science books, advertisements, stories, fairy tales, letters, songs, public information (travelling, health care, and other), etc. In total, the corpus includes 29 genres.

[1] Corresponding Author: Erika Rimkutė; Vytautas Magnus University, V. Putvinskio st. 23-216, Kaunas LT-44243, Lithuania; E-mail: erika.rimkute@vdu.lt.
[2] CEFR – Common European Framework of Reference for Languages: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.
[3] The project *Lithuanian Academic Scheme for International Cooperation in Baltic Studies*: http://baltnexus.lt/en/baltic-studies-project.

In previous research [2], texts taken from coursebooks were used to automatically predict the CEFR level of texts from other sources. The procedure involved the classification of texts into two (A1-A2 and B1-B2) or four (A1, A2, B1, B1) CEFR levels using machine learning (ML) methods (as described by [2]). Different experiments with various ML methods were carried out with a combination of surface quantitative features (number of sentences, average sentence and word length, ratio between longer and shorter items, etc.) and linguistic features (word length and type/token ratio for selected parts of speech; proportions of selected morphological features). The best results were obtained using the logistic regression. According to [2], the difference between the most beneficial and the least beneficial features was very small. In addition, the training allowed to reclassify texts that were previously defined in broader categories (A1-B1, B1-B2) into four categories. The relatively low efficiency of automatic classification (cross validation score of about 0.6 with four CEFR levels) did not allow to be confident about the results. Additional examination of the data reinforced the reservations about the validity of classification for non-didactic material.

In this study, we aim to reassess the results of the automatic text classification experiment reported by [2] and to get a better understanding of the representativeness of the Lithuanian Pedagogic Corpus and its sub-parts. Furthermore, we want to re-evaluate the previously analysed linguistic features important for text complexity assessment and to identify additional useful features.

After the first step of the research, the distribution of morphological, syntactic and lexical features has shown considerable discrepancies between the coursebooks and the texts from other sources that confirmed the weakness of the classification for non-didactic material (non-coursebook texts did not show the correlation between moving from lower to higher levels and the growing complexity of lexis and grammar). For this reason, the scope of the analysis was restricted to texts from coursebooks. In this paper, we first provide information about the distribution of linguistic features presented in the materials for the teachers of Lithuanian as L2 (i.e., which grammatical features and how they are described in the materials for relevant language levels) (Section 2); and compare the number and manner of the distribution of linguistic features in the Lithuanian Pedagogic Corpus (Section 3).

2. Grammatical Features in the Teaching Material of Lithuanian as L2 Prepared According to CEFR: Learning, Teaching, and Assessment

The discussion of the connections between grammatical forms and language levels is based on the CEFR materials designed for levels A1, A2, B1, and B2, see respectively [3], [4], [5], [6].

Noun gender is introduced in level A1, while more complex cases of expressing notions are discussed in level A2. Number[4] is also explained in level A1; cases are introduced in level A1 (only the most frequent meanings are taught, e.g., locative to indicate location and time, nominative to indicate a thing, a person, a phenomenon or a state, vocative to address a person or an animal) as well as in level A2, while shortened forms of cases are introduced in level B2.

[4] The number of other words that agree with nouns is not specified, because CEFR emphasizes agreement; other agreement categories – gender and case – are discussed separately, as their acquisition differs and is more difficult.

Adjective gender, cases, semantic classification into qualitative and relational (this distinction determines some grammatical features), and degree are discussed in level A1; pronominal forms are introduced in level A2.

Pronoun cases are introduced in level A1 (only forms important for learner communication at level A1 are taught, e.g., *manęs* ('me' Gen.), *tavęs* ('you' sg. Gen.), *jus* ('you' pl. Acc.), *mane* ('me' sg. Acc.), only in nominative, genitive, and accusative); the topic is continued in level A2; pronominal forms are instructed in level B2.

Numeral cases are introduced in level A1 (only forms relevant to learner communication are taught, e.g., *Reikia dviejų stiklinių miltų* ('I need two glasses of flour'); *Turiu penkis eurus* ('I have five euros'); *Yra keturios kėdės* ('There are four chairs'): only nominative, genitive, and accusative) as well as in level A2. Cardinal numerals are introduced in level A1; multiple and ordinal numerals are taught in A2 level, while in level B2, learners get acquainted with collective cardinal numerals and fractions. The structure of numerals (simple, combination, and compound) is described in level B2. The numeral governance over other words (e.g., in *dešimt vyrų* ('ten men'), with the noun in genitive because of the numeral) is explained in level A1.

Adverb degree is introduced in the material for level A1.

Verb tenses are explained in different levels: present, past simple and future tense of finite forms in A1; past frequentative of finite forms and compound tense forms in A2. The notion of mood is introduced in level A1 (only the politeness aspect of the subjunctive is explained) and level A2. Reflexive forms are discussed in levels A1 (only the forms that are part of phrases to be learned by heart[5]), A2 and B2 (reflexive participles). Participles are introduced in levels A1 (as multi-word lexical units[6]) and A2; pronominal forms of participles are taught only in level B2; participle voices are described in levels A2 (with more emphasis only on passive forms due to easier declining; while future participles are not included), B1 (active participle forms of all tenses are also presented) and B2 (passive future participles are introduced). Syntactic features of participles (predicative, half predicative and attributive usage) are discussed in the material for level A2. Verb transitivity is introduced in levels A1 (only the fact that verbs usually govern genitive and accusative is mentioned) and A2; aspect is discussed in level A2. Non-finite forms are introduced from levels B1 (half participles, present and past adverbial participle) and B2 (adverbial participle of past frequentative tense, and necessity participles).

The formation of various parts of speech, i.e., derivatives is introduced in coursebooks for level A2. Thus, starting from level A2 learners are taught morphologically more complex and longer words. **Sentence types, word order and sentence parts** are also discussed in the material for level A2.

3. Correlations between Lexical, Morphological and Syntactic Features and the CEFR Levels

We start the analysis by showing the correlation between the morphological features and the different CEFR levels and continue with the discussion of the results of syntactic and

[5] E.g., *Kaip sekasi?* ('How are you?') *Mokausi Vilniaus universitete.* ('I am studying at Vilnius University.') *Man patinka maudytis.* ('I like swimming.').

[6] E.g., *rūkyta žuvis* ('smoked fish'), *rauginti agurkai* ('pickled cucumbers'), *pavargęs* ('tired' (masc.)), *ištekėjusi* ('married' (fem.)).

lexical features. The assessed syntactic features include sentence length and lexical features as type/token ratio, word length (this correlates with the fact that higher language proficiency levels contain more complex derivatives) and the lexical coverage of the most frequently used vocabulary.

3.1. Morphological Features

Morphological features are especially important as Lithuanian is a highly inflected language: learners are taught many grammatical categories and inflected forms of the noun, verb and other parts of speech. All texts in the corpus were morphologically annotated automatically. For this reason, some inaccuracies can be found; however, they are not numerous and their quantity does not invalidate the general tendencies.

In this sub-section we discuss the distribution of some forms of verbs and nominal words.

3.1.1. Verb Forms

Table 1. 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**

As we can see in Table 1, the distribution of finite and non-finite verb forms correlates with a language level: it is obvious that **finite forms** prevail in lower levels (in level A1 texts – 80.94 %), while the number of these forms decreases with the rising language level where they are replaced by grammatically more complex forms. We have also noticed a significant difference between the least used finite verb form cases (64.02 % in B2) and the mentioned largest number of usage instances (80.94 %).

Infinitives are used quite consistently: from 14.75 to 16.91 % (the highest frequency is for level B1 texts). However, the infinitive in the Lithuanian language is one of the fundamental forms acquired in order to be able to use a verb. Moreover, the infinitive is used in diverse areas (not only as a predicate, but also as an object, a subject, an attribute or an adverbial). For these reasons, the usage of infinitives cannot accurately reflect the complexity or simplicity of the language.

Participles, on the other hand, can be considered as an important indicator of the higher language proficiency; as we see in Table 1, their number varies considerably: from 2.83 % (A1) to 16.51 % (B2). Although **adverbial participles** and **half participles** are not abundant, they are mostly used in level B2 texts which shows the growing complexity of the grammar.

Table 2. Verb moods

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

The pedagogic corpus reflects a high usage of the feature typical to both spoken and written Lithuanian – **indicative** forms with variation ranging from 85.16 % to 93.08 %

across language levels. The frequencies of **imperative** reveal that it is the most frequent in level A2 texts (9.63 %). This can be explained by the fact that these forms are common to dialogues, which make up a major part of lower level texts. The distribution of **subjunctive** forms does not show a significant variation across language levels (from 3.5 to 5.21 %).

We can draw a conclusion that the data of verb moods confirms the general tendencies of the Lithuanian grammar features and does not provide reliable information about the correlation between complex forms and higher language levels.

Table 3. Tenses 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

The usage of present and simple past tenses cannot indicate the correlation between the complexity of grammatical forms and the language level because these forms are very common and their distribution in other corpora (e.g. MATAS[7]) is very diverse. On the other hand, the forms of past frequentative suggest the following correlation: because of their complexity, they are less frequent in lower level texts and more frequent in higher levels (with the highest usage in level B1). The fact that the most future forms (13.66%) appear in A2 texts is not enough to show the correlation of these forms with a language level – a larger corpus to highlight this correlation should be used.

Table 4. 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	Necessi-ty
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

Due to the low numbers of instances, we cannot draw conclusions about the usage of past frequentative and future tenses of active participles, passive future and necessity participles, as only several cases or none of these occurred in the texts of every level.

According to CEFR, the usage of participles should increase from level A2. In this level, the focus is put on passive participles, because they are simpler than active ones. This fact is supported by the data in Table 4: A1-A2 level texts contain more passive than active participles. Admittedly, high frequency of passive present participles in A1 texts is surprising – even 49.19 %. This could be explained by the necessity for learners, even at the beginning, to acquire certain multi-word lexical units containing passive present participles, e.g., *rašomasis stalas* (‘a desk’), *valgomasis šaukštas* (‘a tablespoon’).

The largest number of active present participles in level B1 (15.99 %) and active simple past participles in A2 texts (29.58 %) allows the presumption of a correlation between grammatically more complex forms and a higher language level.

[7] Lithuanian morphologically annotated corpus MATAS: https://clarin.vdu.lt/xmlui/handle/20.500.11821/33.

3.1.2. Nominal Forms

We paid a particular attention to such grammatical features of nominal words as noun cases, numeral types, pronominal forms, and adjective and adverb comparative forms, because CEFR quite clearly prescribes when and which numerals should be used or when pronominal forms are taught. Although according to CEFR, all noun cases are introduced in level A1, we can presume that texts of lower levels will contain fewer instances of rarer cases (especially dative or instrumental).

Table 5 provides only **noun cases**. The distribution of other parts of speech was not analysed as most adjectives, pronouns, numerals and participles agree with nouns, thus the choice of their cases (as well as gender and number) depends on the form of a noun.

Table 5. Noun cases

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

We assume that the three most frequent cases (nominative, genitive, and accusative) will not reveal the correlation between the grammatical complexity and language level. The correlation is not indicated by very rare cases – vocative and illative (a type of locative, not included into the grammar system of Modern Lithuanian). Even though locative is not a frequent case, it is inevitable even at the beginning of language acquisition, because one has to learn to indicate a place or time. For this reason, it is not surprising that most locative instances occur in texts for level A1 (8.35 %).

The link between the complexity of grammar and language level can be demonstrated by two rarely used cases: dative and instrumental. These cases are usually used to express the facultative valency; they can be often replaced by prepositional constructions with frequently used cases of genitive and accusative. Based on the data, we can state that our earlier hypothesis was confirmed and both cases show the relation between the growing grammatical complexity and language level: most dative forms occur in level B2 texts (3.29 %), while instrumental in level B1 texts (6.43 %).

Table 6. Types 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

As to the **numeral** usage, we can maintain that even though multiple and collective numerals are not common, their higher frequency in level B2 suggests the correlation between more difficult grammatical forms and a higher language level. It was surprising though to see that most ordinal numerals are used in level A1 texts (72.69 %). Especially common are the same first ordinal numerals: *pirmas* (‘the first’), *antras* (‘the second’), and *trečias* (‘the third’). Presumably, they were learned as individual lexical items. We cannot draw any conclusions about cardinal numerals, because we analyse only numerals written in words and exclude numerals written in a numerical form.

Table 7. Degrees of adjectives and adverbs

Text level/ degree	Adj. positive	Adj. comparat.	Adj. superlative	Adv. positive	Adv. comparat.	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

We can see that **degree** might be important in determining the relation between the grammatical complexity and the language level, because the positive degree is more frequent in lower level texts, while more complex forms – the comparative and superlative degrees – in higher level texts.

3.2. Syntactic and Lexical Surface Features

Table 8 shows the number of **sentences** and their length in the analysed part of the pedagogic corpus. The average sentence length is 12.15 words. The sentence length substantially correlates with the complexity of texts: A1 level texts contain the shortest sentences – 8.08 words, while the longest sentences are found in level B2 – 15.94 words.

Table 8. Length of sentences

Text level/syntactic features	Number of sentences	Average sentence length (in words)
A1	5,591	8.08
A2	2,864	10.12
B1	2,575	14.44
B2	4,599	**15.94**

Although the **word length** (in terms of the number of letters) is not very diverse, it is evident that words in higher levels are longer, thus, morphologically or derivationally more complex (see Table 9).

Table 9. Lexical surface features

Text level/lexical features	Average word length (in characters)	3,075 most frequent word forms (coverage)	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

In this study, the most **frequent vocabulary**, i.e., the 3,075 most frequent word forms of the corpus, was integrated into the assessment of text complexity. The results confirm that vocabulary is larger in higher level texts: 69.13 % of words are from the most frequent vocabulary list for A1 level texts; in B2 level texts, the most frequent vocabulary comprises 54.23 % of all words. For future work, it will be important to have a better-defined reference word list, since most experiences on automatic level assignment stress the primary importance of the lexicon for this task.

Type/token ratio also indicates the correlation between higher level texts and higher lexical diversity. However, as we can see in Table 9, the highest diversity, as one might expect, is not found in level B2 texts but, rather, in level B1 texts (0.43). This might be explained by the repetition of similar topics in level B2, e.g., Lithuania, customs, holidays, the same famous people; thus, the lexical diversity becomes lower. Furthermore, several coursebooks included in the corpus were of transitional level, e.g.,

B1-B2. Such texts were automatically classified into B1 or B2 according to the experiment described in [2]. This might have influenced the fact that the highest lexical diversity was not in level B2 texts, although these texts are characterized by the most complex grammar.

4. Conclusions

In this work, as well as in the experiments described by [2], we focused mostly on the morphological features. During the study [2], proportions of various morphological features were calculated, and the obtained data was used to assign the language level to text. Also, lexical features, specifically, which part of all vocabulary is covered by the most frequent words, were considered.

The linguistic features described revealed that the automatic text classification applied earlier by [2] was not sufficiently precise; therefore, non-coursebook texts in the corpus should be reclassified. As [2] suggested and as [7] demonstrated, a wider set of lexical information could strongly improve the quality of a renewed prediction on non-didactic materials.

We can state that in order to determine the text level automatically, it is worth considering the correlation described in this article – the link between the language level and 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. It is also important to consider the length of a sentence, word length, type/token ratio and the distribution of the most frequent words of the analysed corpus. Nevertheless, in order to determine clear values of each aforementioned linguistic properties in automatic text level assignment, more texts and additional experiments are needed.

References

[1] Kovalevskaitė J, Rimkutė E Mokomasis lietuvių kalbos tekstynas: naujas išteklius lietuvių kalbos besimokantiesiems (Pedagogic Corpus of Lithuanian: a New Resource for Learning and Teaching Lithuanian as a Foreign Language). Sustainable multilingualism (forthcoming). 2020.

[2] Grigonytė G, Kovalevskaitė J, Rimkutė E. Linguistically-Motivated Automatic Classification of Lithuanian Texts for Didactic Purposes. In: Muischnek K, Müürisep K, editors. Proc. of the 8th International Conference Baltic HLT 2018; 2018 Sep 27-29; Tartu (Estonia). Frontiers in Artificial Intelligence and Applications, vol. 307. Amsterdam: IOS Press. P. 38-46.

[3] Stumbrienė V. Lūžis (Breakthrough – A1). Vilnius: Vilniaus universiteto leidykla; 2016. 67 p.

[4] Ramonienė M, Pribušauskaitė J, Vilkienė L. Pusiaukelė (Waystage – A2). Europos Taryba; 2006. 148 p.

[5] Ramonienė M, Pribušauskaitė J, Vilkienė L. Slenkstis (Threshold – B1). Vilnius: Vilniaus universiteto leidykla; 2016. 244 p.

[6] Ramonienė M, Pribušauskaitė J, Vilkienė L. Aukštuma (Vantage – B2). Vilnius: Vilniaus universiteto leidykla; 2016. 247 p.

[7] 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.