Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-38

Linguistically-Motivated Automatic Classification of Lithuanian Texts for Didactic Purposes

Gintarė GRIGONYTĖ[a], Jolanta KOVALEVSKAITĖ[b] and Erika RIMKUTĖ[b1]
[a] *Stockholm University, Sweden*
[b] *Vytautas Magnus University, Lithuania*

Abstract. This paper presents an effort to provide a level-appropriate study corpus for Lithuanian language learners. The collected corpus includes levelled texts from study books and unlevelled texts from other sources. The main goal is to assign the level-appropriate labels (A1, A2, B1, B2) to texts from other sources. For automatic classification we use preselected surface features, based on text readability research, and shallow linguistic features. First, we train the model with levelled texts from study books; second, we apply the learned model to classifying other texts. The best classification results are achieved with Logistic Regression method.

Keywords. L2, Lithuanian language, linguistically-motivated automatic classification, levelled study corpus for learners

1. Introduction

This paper presents early results of the project "Lithuanian Academic Scheme for International Cooperation in Baltic Studies"[2]. The project aims at providing a level-appropriate study material for Lithuanian language learners. One outcome of the project is a levelled study corpus for learners of Lithuanian similar to existing corpora for German [1], Swedish [2] and Russian [3]. The need of a levelled study corpus is motivated by the fact that the general Corpus of the Contemporary Lithuanian[3] is not appropriate for learners of Lithuanian as a foreign language.

Creating corpora for this kind of levelled study corpus typically assumes collecting texts from various study books. Even though Lithuanian language resources have increased in the past years, there are not enough texts to create a representative levelled study corpus. Therefore, we tried to address this issue by including several resources: a) written text materials from study books, levelled by the book authors, who are working practitioners; b) texts collected from other resources. This corpus will provide teachers and learners with a level-based spoken and written language learning material (the corpus size is appr. 600.000 tokens). The corpus aims at representing 4 levels: A1,

[1] Corresponding author: Erika Rimkutė, Vytautas Magnus University, V. Putvinskio g. 23-216, Kaunas LT-44243, Lithuania; E-mail: erika.rimkute@vdu.lt.

[2] http://baltnexus.lt/en/baltic-studies-project.

[3] 208M tokens, http://corpus.vdu.lt/lt.

A2, B1 and B2: it will contain 100.000 tokens of language data appropriate for A1-A2 scale and 500.000 tokens of language data appropriate for B1-B2 scale according to CEFR (Common European Framework of Reference for Languages)[4].

The main practical problem to be solved in building the corpus is that of automatically assigning level-appropriate labels to texts which are collected from other resources. Having two sources of texts – books with level-appropriate labels and unlabeled texts from other resources – we need to assign labels to the latter. In this study, we tackle the problem step-by-step: 1) use "clear" labels and texts from books for training, 2) apply the learned model to re-classify the "in between" texts from books to "clear" classes, 3) finally apply the learned model to texts collected from other resources and automatically assign them into levels A1, A2, B1 and B2. The latter step has a potential to enable faster construction of corpora of Lithuanian learner material and creation of resources and services for Lithuanian language learners and researchers.

We present findings in automatically classifying the study data according to 4 CERF levels: A1, A2, B1 and B2. We approach the classification task, first, by the popular notion of text readability [4], [5] and second, by following some of the pedagogical recommendations for grammar topics to be covered in each of the 4 levels.

The paper is organized as follows: in Section 2, we describe the corpus we used for our study. Section 3 explains the selected features for training the ML models which are presented in Section 4. In Section 5, we present the results: in 5.1. – the performance of the classifier, in 5.2 and 5.3 – the practical application of the learned model for reclassification of "in between" texts and for classification of other texts.

2. Corpus

The written language corpus has been collected from several resources: a) texts from a series of Lithuanian language learning books; b) texts collected from other resources (texts from news portals, stories, fairy tales, advertising, letters, song texts etc.). Texts from the study books are levelled by the book authors, who are working practitioners. These texts are distributed over 6 levels: A1, A2, A1-A2, B1, B2, B1-B2[5].

Table 1. The corpus of the Lithuanian language study material

Corresponding CERF level	A1	A2	A1-A2	A2-B1	B1	B2	B1-B2
Number of words (of study books)	14126	13119	15147	-	11280	24234	35311
Number of words (of other texts)	4700	2618	17076	43442	-	29700	386001
Number of texts (of study books)	278	129	115	-	61	94	310
Number of texts (of other texts)	23	46	274	217	-	90	1210

[4] https://www.coe.int/en/web/common-european-framework-reference-languages.

[5] Note: there are "clear" and "in between" levels.

Number of genres (of study books)	10	8	7	-	7	4	7
Number of genres (of other texts)	2	3	5	4	-	1	8

The labels were given also to the texts collected from other resources, but these labels are not of the same reliability as labelling was performed by the linguists who collected the texts but are not working practitioners. All texts were processed with Lithuanian POS tagger [6][6]. The corpus details are presented in Table 1 which shows the composition of the corpus in terms of text number and type of genres (letters and SMS, dialogues, information texts, stories, recipes and menus, advertisements, greetings, timetables, questionnaires, prescriptions for medicine) we used for training the classifier.

3. Features

Experiments presented in this paper make use of several linear classifiers. We have employed two types of features: surface and shallow linguistic features. Surface features are inspired by research in readability (e.g. [9]) and focus on parameters like the number of word types or sentence length. Surface features are selected for text, sentence and word levels. The motivation for text length is directly connected to the observation that texts in lower level CERF material are short. Sentence-length is yet another strong indicator of the complexity of a text [9]. Finally, the research on word level in readability shows that long words as well as multi-syllable words indicate complex morphology and text specificity ([4], [10]). Thus, the **surface features** we have selected to work with are:

- sentence_number – the number of sentences in a text,
- max_sentence_len – the maximum number of words per sentence in a text,
- avg_sentence_len – the average sentence length in a text,
- avg_word_len – the average word length in a text,
- difficult_words_s – the proportion of words that are longer than 8 characters per sentence, calculated as average value in a text,
- long_short – the long and short word ratio in a text,
- long_perc – the percentage of long words in a text,
- short_perc – the percentage of short words in a text, our intuition for these three features is that texts for less advanced language learners will contain more short words,
- log_words – log10 normalized number of words per text, with this feature long and short texts are differentiated on a log-scale.

Shallow **linguistic features**. First, we selected several features from studies on text difficulty characteristics:

[6] Available through http://semantika.lt/; this tagger is based on the Hunspell open source platform supplemented with the statistical HMM method for the disambiguation task. Semantika.lt tagger achieved ~98.0%, ~95.3%, ~86.8% of accuracy on the lemmatization, part-of-speech tagging, and annotation of the morphological categories, respectively [7].

- ttr – type token ratio, this feature indicates lexical richness of a text. Higher CERF level texts are expected to have richer vocabulary.
- nq – nominal quotient indicates the ratio of nouns, prepositions and participles divided by pronouns, adverbs and verbs per document. [9] use this ratio to measure the information quantity in a text. The assumption is that higher CERF level texts contain more information.
- noun_pron_q – noun pronoun quotient indicates the ratio of nouns divided by pronouns per text. This feature is inspired by [9] who argue that nouns are a part of speech bearing high information whereas pronouns have a function of repeating previous information.
- adj_perc – the percentage of adjectives in a text.
- noun_perc – the percentage of nouns in a text.
- punct_perc – the percentage of punctuation in a text.
- pron_perc – the percentage of pronouns in a text.
- avg_len_noun – the average length of nouns in a text.
- avg_len_verb – the average length of verbs in a text.
- avg_len_adj – the average length of adjectives in a text.

Additionally, after examining available didactic material for teaching Lithuanian as a foreign language (recommendations for grammar syllabus for each particular level), and after interviewing working practitioners who teach Lithuanian as foreign language, we added several **linguistic features** as possible good indicators for a particular level:

- noun_TT – the diversity ratio of nouns in a text, calculated as the type/token ratio. The assumption is that higher CERF level texts have a higher variety of nouns.
- verb_TT – the diversity ratio of verbs in a text.
- adj_TT – the diversity ratio of adjectives in a text.
- dal_perc – the percentage of participles in a text.
- psd_perc – the percentage of half participles in a text.
- pad_perc – the percentage of adverbial participles in a text.
- advanced_cases_perc – the instrumental and dative case percentage in a text.
- cmp_perc – the percentage of comparative cases in a text.
- mood_perc – the percentage of verb forms in imperative and subjunctive mood in a text.
- neut_perc – the percentage of words with a neutral gender in a text.

[3] results show that classification may be done in a relatively accurate way by using simple features such as the proportion of familiar words (i.e. the top word frequency list of 5,000 lemmas as a threshold between simple and difficult vocabulary) and extending it with readability scores and other linguistically motivated features. Although we consider the core vocabulary as a promising descriptor for classification, we have not applied it in the study: according to recent recommendations for identifying the core vocabulary by taking into consideration both frequency and dispersion of lexical items [11], the available resource for Lithuanian core vocabulary was not representative enough. In further work with the levelled study corpus of Lithuanian, we plan to work with the core vocabulary for all levels separately.

4. Methods

We have used 5 ML models from the Scikit-learn library:

- k-Nearest Neighbor classifier[7] as described in Scikit-learn. kNN uses similarity in order to find data point located to the given instances.
- Support Vector Machines[8] as described in Scikit-learn. SVM makes use of separating points of a transformed problem space into two groups, i.e. one vs. remaining.
- Naive Bayes[9] as described in Scikit-learn. GNB makes prediction based on a conditional relationship between a label and each attribute.
- Decision tree classifier[10] as described in Scikit-learn. DTC mirrors human decision making and aims to divide data in the best way leading to leaves (classes).
- Logistic Regression[11] as described in Scikit-learn. LR makes prediction of the probability of event.

Test and training data have been divided into proportion 1:9. We applied standard 10-fold cross validation in our experiments.

5. Experimental Results

In this part we report the results of our experiments. Subsection 5.1 describes training and selecting the most suitable classifier by using study books texts with assigned clear labels: here we experimented with the classification into 4 and 2 classes. We also analyze the impact of each feature. In subsections 5.2 and 5.3, we report the application of the learned model in order to re-classify the "in between" texts from study books and to classify the other texts automatically assigning them into levels A1, A2, B1 and B2. During this stage of the experiment, we also applied manual evaluation.

5.1. Training of a Classification Model

Table 2 presents the experimental results with the study books texts from the corpus, the features presented in section 3 and 5 ML methods presented in section 4.

We had the following experimental settings: 5 methods with 4 labels (A1, A2, B1 and B2); 5 methods with 2 labels (A1 and A2 merged into A, B1 and B2 merged into B).

As seen in Table 2, the 4 class classification yields 0.607 as the best score by LR method, and 0.869 score by LR method if the classification is between 2 simplified levels, A and B. Initial experiments indicated that the corpus data is rather uneven and most of the texts end up in A1 or B2 classes (see Table 1 for text proportion at levels A1, A2, B1 and B2).

[7] http://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification.

[8] http://scikit-learn.org/stable/modules/svm.html#svc.

[9] http://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes.

[10] http://scikit-learn.org/stable/modules/tree.html#classification.

[11] http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

The higher accuracy with 2 classes evidences the difficulty of the classification into 4 levels. The classification task is less complicated if the distance between the texts is larger: e.g. [3] classify between simple (A1, A2, B1 levels) and more difficult texts (B2, C1, C2) and achieved an average accuracy of 0.91 with surface-oriented features complemented by vocabulary-based features including POS information.

Table 2. Experimental results

Experimental setup	Cross_val_score mean	Recall score mean	Precision mean	Accuracy score
KNB_4	0.582	0.582	**0.562**	0.582
SVM_4	0.547	0.546	0.502	0.547
GNB_4	0.564	0.564	0.513	0.564
DTC_4	0.575	0.574	0.485	0.575
LR_4	**0.607**	**0.607**	0.521	**0.607**
KNB_2	0.852	0.921	0.880	0.852
SVM_2	0.822	0.938	0.836	0.822
GNB_2	0.831	0.845	**0.915**	0.831
DTC_2	0.862	0.936	0.880	0.861
LR_2	**0.869**	**0.953**	0.876	**0.868**

Figure 1 shows the impact on LR model by leaving one feature out. The dotted line indicates the classification into 4 levels and the dash line – into 2 levels.

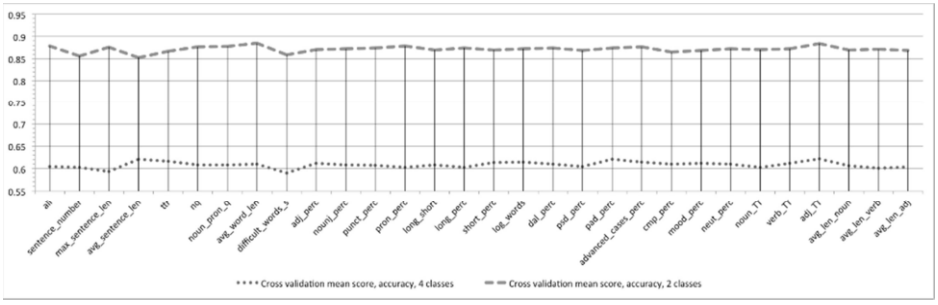

Figure 1. Feature ablation.

Feature ablation shows a rather consistent impact for both tasks: 4-level and 2-level classification. However, in some cases, for instance max_sentence_len and avg_sentence_len, it has the opposite effect: in the case of avg_sentence_len, the performace for the 4-level classification decreases and for 2-level classification increases.

The most beneficial features were: ttr, sentence_number, avg_sentence_len, difficult_words_s, punct_perc, noun_perc, long_perc, psd_perc, cmp_perc, mood_perc, neut_perc, and noun_TT. The least beneficial features for this experiment were: max_sentence_len, avg_word_len, adj_perc, pad_perc, and adj_TT, but as the score differences are rather slight, we did not exclude these features.

5.2. Application of the learned model for the reclassification of "in between" texts

In the 2nd stage of the experiment, the best performing model (LR_4) trained on 562 texts was used to re-classify the "in between" texts from the study books. For this part of the task, 425 "in between" level texts from the books were used. After the re-classification, manual evaluation was performed. During the manual evaluation, only those texts were considered whose labels did not match any of the "in between" label: for example, an "A2-B1" text needed a manual inspection if it was labelled A1 or B2. The manual evaluation of the classification in terms of accuracy is 0.812. Overall 425 "in between" texts were level labelled, 80 labelling errors were detected upon inspection.

The classification errors occurred mainly in the following genres: information texts, letters and SMS and advertisements. A closer inspection revealed that what was common for these texts is the abundance of 1 or 2-word slogans as well as densely used numerical data such as price, time, volume and quantity. This means that the surface form of these texts appears to be simple in terms of high percentage of short words and short sentences. At the same time, it also has a high variation of cases and verb forms and thus indicating the advanced level of text. These factors seem to appear in all 4 levels.

We have trained the classifier on the initial training data set and have used it to classify both: "in between" texts and unlevelled texts from other sources. One modification of the experiment was classifying the "in between" text first and then using them as an additional training material, however, it did not change the performance of the classifier significantly, and thus was not applied in the project.

5.3. Application of the learned model for the classification of other texts

During the classification of texts from other sources, they were automatically assigned into levels A1, A2, B1 and B2 using the best performing model (LR_4). During the manual evaluation, only those texts were considered whose labels did not match either of the "in between" labels. The manual evaluation of the classification in terms of accuracy is 0.926. Overall 1860 texts were level labelled, 137 labelling errors were detected upon inspection.

We considered lexical and grammatical criteria during the evaluation of text material of beginner or more advanced levels: lexical (difficult words, archaic, slang or dialectal words) and grammatical (archaic categories, e.g. illative, infrequent categories, e.g. participle, half-participle).

There were some observations worth mentioning: 1) different texts originating from the same book might be assigned to different levels, if they differ in word or sentence length statistics, the variety of grammatical forms, e.g. the book on a good behavior, where each chapter starts with an aphorism, which is assigned to level A1 or A2, and the rest of the chapter – to level B1 or B2. 2) In some cases, we have observed a correlation between genre and level: e.g. jokes were manually classified as B1-B2

level texts during the manual compilation the corpus. The automatic classification labelled the text of this genre as A1 or A2 level texts because they are short, include direct speech often containing simple verb forms like inflective forms, imperative mood, usually learned by beginners.

6. Conclusions

The experimental results indicate that using preselected surface features, shallow linguistic features and the available training corpus, the best classification results with 4 labels are achieved with LR (Logistic Regression) method – the mean accuracy score is 0.607. If compared to the baseline of assigning all texts to the largest class, the accuracy would be 0.494. It would be beneficial to collect more text samples for small classes, as imbalanced size classes in training data affects classification quality. The evaluation of the impact of each feature on LR model shows that the indicators for syntactic and lexical complexity seem to be useful for the evaluation of text difficulty: surface features such as the number of sentences per text, the average sentence number and the proportion of difficult words (longer than 8 characters), as well as some shallow linguistic features as the type-token ratio, the percentage of punctuation in a text, the diversity ratio of nouns.

The learned model was applied for reclassification of "in between" texts (study book texts) and for the classification of texts, collected from other sources. The manual evaluation of the classification in terms of accuracy is 0.812 (for "in between" texts) and 0.926 (for other texts). The major contributing factors to classification errors are related to uneven complexity of the same text. The challenge here is that text complexity is inconsistent and changes within a book, thus some texts in higher level are of the same complexity as texts in the lower level. Additionally, we observed a correlation between a particular genre and a level (e.g. information texts, letters and jokes). Therefore, as possible directions of improvement we suggest considering new features and adaptation to genre variation. As for the first one, token or type n-grams might account for a topic-oriented style of the learning material, i.e. each level books aim at introducing some topic vocabulary (job, travel, health etc.). For the second, we propose looking at genre first and then applying the suggested features and classification methods.

Corpora of CEFR related texts are used for formalizing CEFR descriptors, i.e. to study grammar and vocabulary features characteristic of different CEFR levels with the help of machine learning algorithms [2]. In this study, we employed the approach using grammatically and pedagogically motivated shallow text features to train the model which then was used for automatic text classification according to their difficulty [3]. Our approach is rather prescriptive; a descriptive approach might give additional insights concerning the features of the pedagogical texts and this experience in return might be used in automatic classification.

References

[1] P. Hoffstaedter, K. Kohn, Ein didaktisches Interview-Korpus als Ressource für explorative und kollaborative Deutschlernaktivitäten. *German as a Foreign Language* **2-3** (2012), 3-31.

[2] E. Volodina, J.S. Kokkinakis, Compiling a corpus of CEFR-related texts. *Proceedings of the Language Testing and CEFR conference*, 2013.
[3] D. Batinić, S. Birzer, H. Zinsmeister, Creating an extensible, levelled study corpus for learners of Russian. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)* (2016), 38-43.
[4] G. R. Klare, *The Measurement of Readability*. Ames, The Iowa State University Press, 1963.
[5] F. Edward, Readability versus leveling. *The reading teacher* **56.3** (2002), 286-291.
[6] V. Dadurkevičius, Lietuvių kalbos morfologija atvirojo kodo Hunspell platformoje. *Bendrinė kalba* **90** (2017), 1–15. http://www.bendrinekalba.lt/?90.
[7] J. Kapočiūtė-Dzikienė, E. Rimkutė, L. Boizou, A Comparison of Lithuanian Morphological Analyzers. *20th International Conference Text, Speech, and Dialogue (TSD 2017)* (2017), 47-56.
[8] E. Dale, J.S. Chall, A Formula for Predicting Readability. *Educational Research Bulletin* **27** (1948), 37-53.
[9] L. Melin, S. Lange, *Att analysera text*, Studentlitteratur, Lund, 1995.
[10] D. Biber, *Variation across speech and writing*, Cambridge University Press, New York, 1988.
[11] V. Brezina, D. Gablasova, Is There a Core General Vocabulary? Introducing the New General Service List, *Applied Linguistics* **36/1** (2015), 1-22.