An overview of Lithuanian Internet media n-gram corpus

Ieva Bumbulienė
Vytautas Magnus University, Lithuania
Baltic Institute of Advanced Technology, Lithuania
e-mail: ieva.bumbuliene@bpti.lt

Justina Mandravickaitė
Vilnius University, Lithuania
Baltic Institute of Advanced Technology, Lithuania
e-mail: justina@bpti.lt

Loïc Boizou
Vytautas Magnus University, Lithuania
e-mail: l.boizou@hmf.vdu.lt

Tomas Krilavičius
Vytautas Magnus University, Lithuania
Baltic Institute of Advanced Technology, Lithuania
e-mail: t.krilavicius@bpti.lt

***Abstract*—This paper describes construction and properties of the open 70 million words Lithuanian Internet media n-gram corpus. Due to copyright limitations often contemporary media based resources availability is restricted, while n-grams corpora (e.g., Google N-gram viewer/corpus) solve the problem. Lithuanian language is under-resourced, hence n-gram corpus of Lithuanian media is designed to contribute to publicly available ready-to-use lexical resources. In this paper we report corpus construction procedure, preprocessing, corpus statistics and possible areas of application.**

Keywords—corpus; Internet media; Lithuanian, n-grams

I. Introduction

This paper describes the construction, and properties of the Lithuanian Internet media n-gram corpus. N-grams have been used in the variety of Natural Language Processing (NLP) tasks. Besides, n-gram language models have become popular as a general resource for data-driven algorithms in many areas such as speech recognition, text tagging, spelling correction, named entity recognition, word sense disambiguation, lexical substitution [1] and especially statistical machine translation [2], [3]. A bit less common application of n-grams is sentiment analysis or polarity mining, i.e., identifying and extracting subjective information from text [4], developing a grammar checker without grammar rules [5]. Also, n-grams were reported to improve classification tasks, e.g., email-act classification, where n-grams provide contextual information for better characterization of distinct classes [6], legal text classification [7], [8]. This versatile applicability of n-grams is one of the main motivations for construction of Lithuanian Internet media n-gram corpus.

A number of papers (e.g., [9], [10]) report that effectiveness of NLP techniques and methods is sensitive to the data size. Therefore there have been significant efforts to create larger datasets and thus at least several large n-gram corpora became available. Among others, Google Books N-gram Corpus is worth mentioning in this context. Primarily designed for building better language models for machine translation and other NLP applications, Google Books N-gram Corpus is a database of n-grams up to 5 words with the frequency distribution of each sequence/unit in each year from 1500 [5]. In preparation of the corpus, n-grams were filtered by eliminating unigrams with less then 200 occurrences and 2-5-grams with less than 40 occurrences. If there were n-grams in the corpus that had a token with less than 200 unigram instances, they were represented with a special token [11].

Such filtering and the latter preprocessing, although performed for technical reasons were pointed-out as the drawbacks of the corpus ([2], [12]) as "pruning" makes the frequency counts unsuitable to estimate a language model, e.g., using popular and successful Kneser-Ney smoothing algorithm [2]. Moreover, after eliminating low-frequency n-grams, it is impossible to obtain reliable frequency estimates when pooling data [12].

However, Google Books N-gram Corpus was successfully used in many applications, e.g., for research of cultural or semantic change regarding meaning shifts of concepts [13], [14] as a resource for developing rule-less grammar checker for Spanish [5], measuring cultural complexity [15], temporal analysis of language change [16], etc.

Lithuanian resources for NLP tasks and applications are limited, therefore n-gram corpus of Lithuanian media is designed to contribute to publicly available ready-to-use lexical resources. The corpus is constructed from media, i.e., news portal texts[1] and has 72 million words. The texts are taken from 11 categories (see below). N-gram corpus consists of unigrams, bigrams, trigrams and tetragrams. Certain meta-information and statistical information, e.g., n-gram aggregated frequencies are included. This corpus will be made publicly available under CC (Creative Commons) license in the official PASTOVU project website http://mwe.lt.

II. Data source and initial corpus

Initial corpus is constructed of articles from a popular Lithuanian news portal delfi.lt and covers time period from 2014 March to 2016 November, inclusive. Articles and their meta-information were crawled and stored to a database. The database consists of the following fields:

1) category,
2) date,

[1] delfi.lt

3) title,
4) author,
5) link (url),
6) word count and
7) article itself.

Articles were crawled from the following 11 categories:

1) people,
2) projects,
3) science,
4) auto,
5) sport,
6) life,
7) news,
8) citizen,
9) business,
10) fit,
11) other category, which consists of articles that did not fit into any of the other mentioned categories or which category was not recognized.

Video category was ignored because the articles it contained were too short. Thus initial corpus is constructed of 190000 articles, it has 72 million words and requires for a database of 0.48 GB. Such initial corpus was used as a base to build an n-gram (where n ranges from 1 to 4) corpus consisting of 4 n-gram collections. Development and characteristics of the corpus are described in the following sections.

III. Preprocessing and n-gram generation

Preprocessing and n-gram generation were separated into several phases:

1) documents symbols analysis,
2) sentence splitting,
3) n-grams generation according to the rules,
4) n-grams frequency calculation.

During the first step of preparation for n-gram generation, analysis of symbols that form news articles was performed. It was found out that articles are constructed not only from Lithuanian letters, digits and main punctuation marks but also of other miscellaneous symbols, e.g., letters of different foreign languages (e.g., Cyrillic for Russian, some Hebrew, Greek letters, etc.), currency symbols, degree symbol, accented symbols, etc. Moreover, some of the symbols were similar and had similar purpose, but were encoded differently, e.g., there were 14 types of dashes and 8 types of double quotes. While the usage of acronyms in the articles was not analysed, it was decided to keep words of foreign languages during n-gram generation process as they are important for the meaning of a sentence. Discarding these words might be incorrect action because the sequence of words in the sentence cannot be changed while generating n-grams.

The second step was sentence splitting. Each file was split into sentences by a rule-based Haskell segmentation library developed at the Centre of Computational Linguistics (Vytautas Magnus University) since 2013 in order to process rough or morphologically annotated data. Older versions have been used to prepare data for Sketch Engine and for the training of the morphological analyser of semantika.lt web service. In late 2016, the current version of the library contributed to the processing of Lithuanian Parseme data.

The sentence limits in the files used to generate n-gram lists are marked by empty lines. Tokenisation rules for n-grams were created based on the analysis of symbols in the previous step. Thus n-gram splitting consisted of the following steps:

1) Every sentence was split by spaces.
2) All text was converted to lowercase characters.
3) Analysis of each created token was performed.
   a) If a token consisted only of letters or digits, it was declared a word.
   b) A token of one symbol which was not a letter and not a digit was ignored as it was considered to be a punctuation mark or some symbol that can be thrown out.
4) Predefined patterns where symbols should not be deleted were searched:
   a) URL addresses,
   b) email addresses,
   c) float numbers and a special word structure consisting of a number, symbol and a part of word, e. g. "*32-tasis (the 32nd)*", "*101-osiose (in the 101st)*", "*1965-ais (in the year 1965)*".
5) A dictionary of abbreviations [17] was used to find shortened forms of the words in order to keep such tokens as words without discarding full stops because in the later stages all full stops were removed.
6) When a token in format "*digit:digit*" occurred, it was split into two digits which were then saved as two separate tokens. This type of token, e.g., could refer to a result of a sport game match.
7) All the punctuation marks that occurred at the start or at the end of a token were removed while the remaining part was saved.
8) Finally, tokens that contained no letters and no digits were removed. The analysis of the removed content revealed it to be various meaningless sequences of symbols that mostly were punctuation notations, e.g., <...>.

After creating and preprocessing tokens, n-grams were generated for each sentence and stored in database. Additionally, some meta-information about n-grams was saved as well. It included:

1) separate words that formed n-gram,
2) information about the document n-gram was extracted from,
3) sentence number,
4) index of every word,
5) length of every word.

The result was 4 collections of n-grams (unigrams, bigrams, trigrams and tetragrams) in the database. These collections were aggregated to count frequencies of different n-grams and construct final corpora. Statistics of constructed n-gram corpus is presented in the next section.

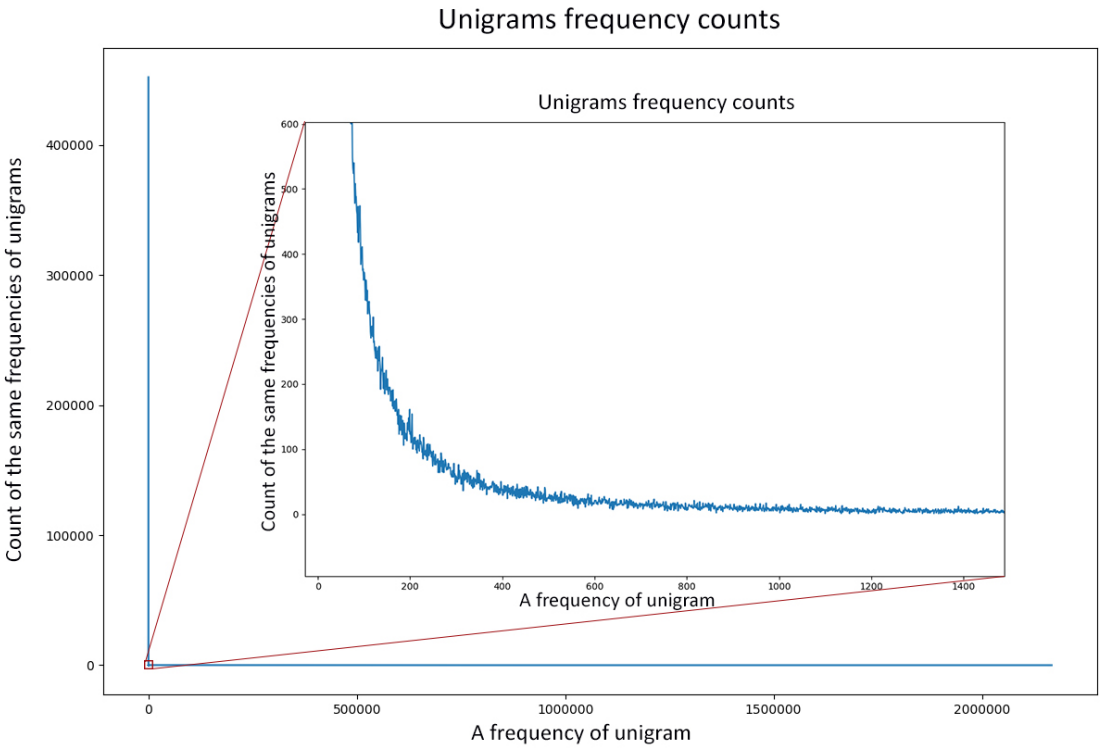

Figure 1. Unigram frequency counts

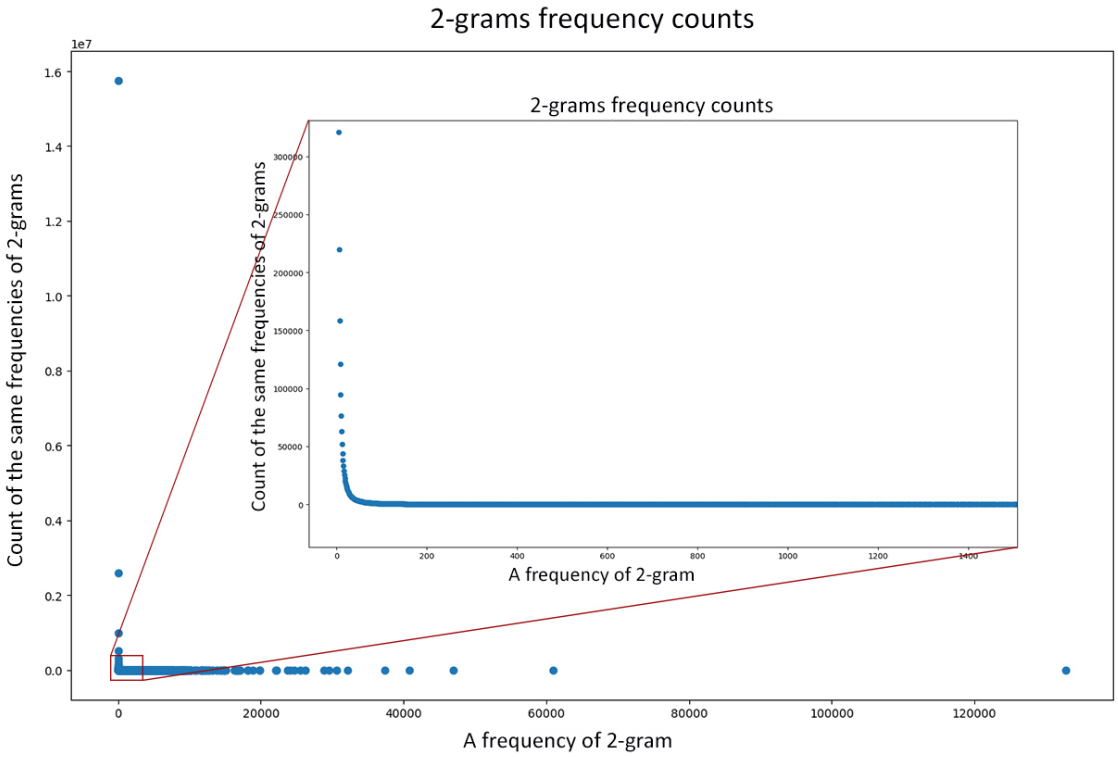

Figure 2. 2-gram frequency counts

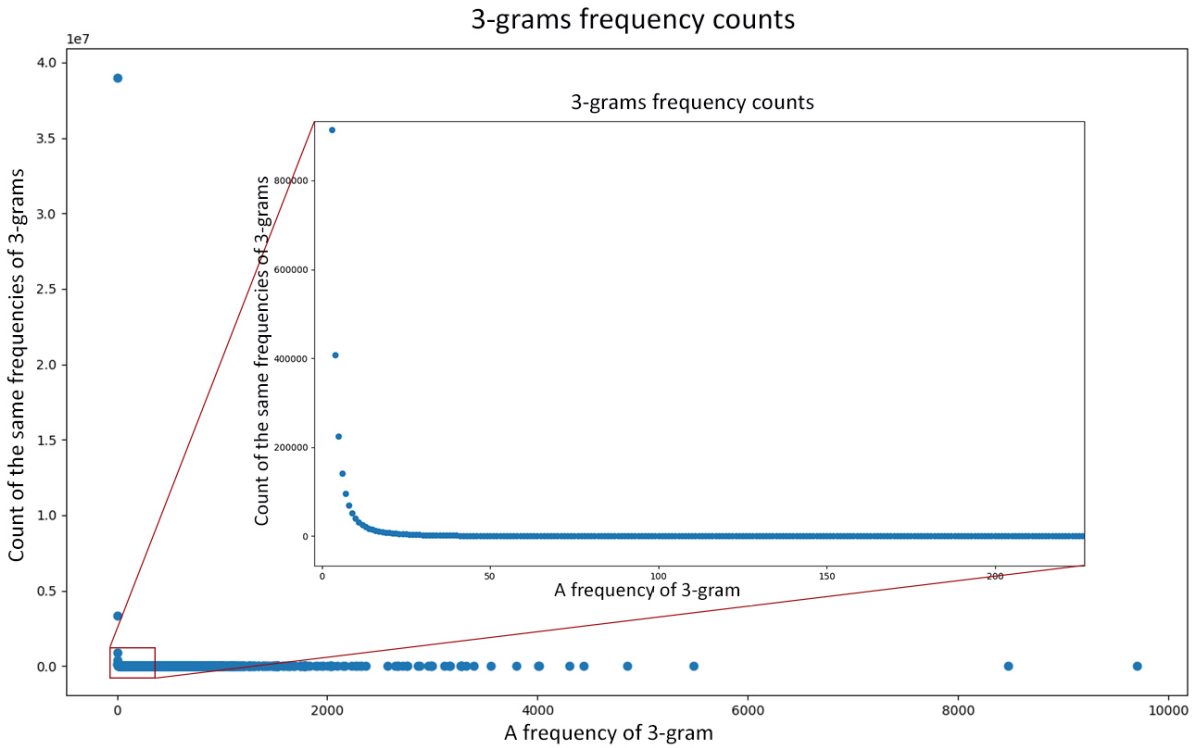

Figure 3. 3-gram frequency counts

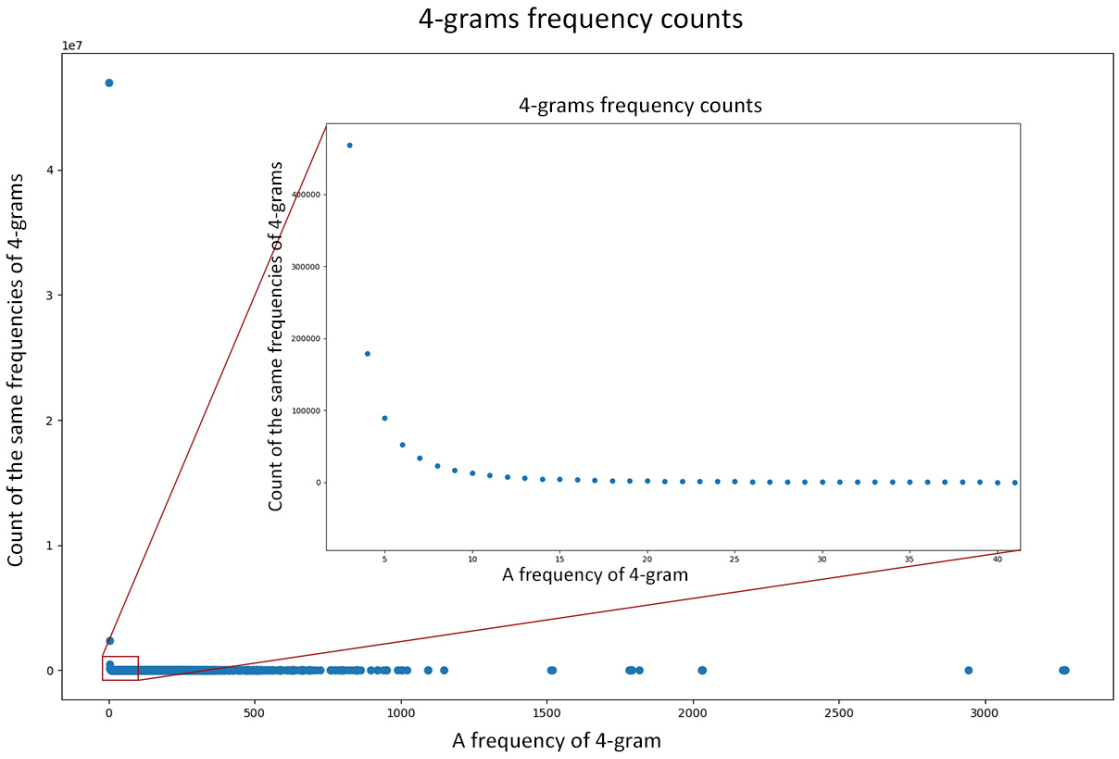

Figure 4. 4-gram frequency counts

Table I. N-GRAM CORPUS BASIC STATISTICS

	unigrams	2-grams	3-grams	4-grams
In total	72 918 468	67 588 522	62 371 336	57 285 915
Unique (units)	1 063 473	21 500 526	44 545 633	50 310 978
Unique (%)	1.46	31.81	71.42	87.82
N-grams with freq. of 1 (units)	452 062	15 746 450	39 011 992	47 005 072
N-grams with freq. of 1 (%)	0.62	23.30	62.55	82.05
Maximum frequency of 1 n-gram	2 167 384	132 780	9 698	3274

IV. CORPUS STATISTICS

Summary of the basic statistics of n-gram corpus is presented in the Table I. The percent of unique n-grams varies from 1.45 for unigrams to 87.82 for 4-grams. In addition, the percent of n-grams which occur in corpus once (*hapax legomena*) is 0.62 for unigrams and 82.05 for 4-grams whereas values for 2-grams and 3-grams are distributed in mentioned intervals. On average, unique n-grams consist of 45.13% and n-grams with frequency 1 make 39.29% of all the n-grams in the corpus.

Figures 1-4 present variability of n-gram frequencies in the corpus. X axis shows frequencies of every n-gram that occurs in the corpus whereas y axis represents a count of different n-grams that are in the corpus with a specified frequency. Thus Figure 1 shows that when a value of unigram frequency is higher than ≈ 200, the amount of unigrams with the same frequency in the n-gram corpora is less than 100.

Meanwhile Figure 2 shows that counts of different 2-grams with frequency higher than ≈ 50 are much smaller in comparison to other n-grams with lower frequency. For 3-grams and 4-grams frequencies of different n-grams continue to drop as well. These values are approximately 25 for 3-grams and 10 – for 4-grams (see 3 and 4).

V. CONCLUSION

This paper describes construction and properties of the 70 million Lithuanian Internet media n-gram corpus. All important steps of preprocessing, and statistical details of the corpus are discussed. Initial texts were collected from the Lithuanian news portal delfi.lt (190 thousands articles; 72 million words). Then analysis of symbols that comprise extracted texts was performed for generation of tokenization rules. Later sentence splitting and tokenization using additional resources and patterns were executed. After that, n-grams (uni, bi, tri and tetragrams) were generated for every sentence. The following statistics were calculated for the newly constructed Lithuanian Internet media n-gram corpus for each n-gram collection: total number, number of unique n-grams, percentage of unique n-grams, percentage of n-grams with frequency 1, and frequency counts. N-gram corpus of Lithuanian Internet media is designed to contribute to publicly available ready-to-use lexical resources for various NLP, linguistic, etc. tasks, and soon will be available for all users.

ACKNOWLEDGMENT

This research was funded by the Research Council of Lithuania (No. LIP-027/2016)), see www.mwe.lt for more details.

REFERENCES

[1] M. Flor, "A fast and flexible architecture for very large word n-gram datasets," *Natural Language Engineering*, vol. 19(1), pp. 61-93, 2013.

[2] C. Buck, K. Heafield, B. Van Ooyen, "N-gram counts and language models from the common crawl," in *LREC*, vol. 2. Citeseer, p. 4, 2014.

[3] A. Pauls, D. Klein, "Faster and smaller n-gram language models," in *Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies-Volume 1*. ACL, pp. 258-267, 2011.

[4] D. Bespalov, B. Bai, Y. Qi, A. Shokoufandeh, "Sentiment classification based on supervised latent n-gram analysis," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 375-382, 2011.

[5] R. Nazar, I. Renau, "Google books n-gram corpus used as a grammar checker," in *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering*. ACL, pp. 27-34, 2012.

[6] V. R. Carvalho, W. W. Cohen, "Improving email speech acts analysis via n-gram selection," in *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*. ACL, pp. 35-41, 2006.

[7] V. Mickevičius, T. Krilavičius, V. Morkevičius, "Classification of short legal lithuanian texts," *BSNLP 2015*, p. 106, 2015.

[8] V. Mickevičius, T. Krilavičius, V. Morkevičius, A. Mackutė-Varoneckienė, "Automatic thematic classification of the titles of the seimas votes," in *Proc. of the 20th Nordic Conf. of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, B. Megyesi, Ed. Linköping University Electronic Press / ACL, 2015, pp. 225–231. [Online]. Available: http://aclweb.org/anthology/W/W15/W15-1828.pdf

[9] P. Norvig, "Statistical learning as the ultimate agile development tool," in *ACM 17th Conf. on Information and Knowledge Management Industry Event (CIKM-2008)*, 2008.

[10] M. Banko, E. Brill, "Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing," in *Proceedings of the first international conference on Human language technology research*. ACL, pp. 1-5, 2001.

[11] T. Hawker, M. Gardiner, A. Bennetts, "Practical queries of a massive n-gram database," in *Proc. of the Australasian Language Technology Workshop*, pp. 40-48, 2007.

[12] S. Evert, "Google web 1t 5-grams made easy (but not for the computer)," in *Proc. of the NAACL HLT 2010 sixth web as corpus workshop*. ACL, pp. 32-40, 2010.

[13] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant *et al.*, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331(6014), pp. 176-182, 2011.

[14] K. Gulordava, M. Baroni, "A distributional similarity approach to the detection of semantic change in the google books ngram corpus," in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. ACL, pp. 67-71, 2011.

[15] P. Juola, "Using the google n-gram corpus to measure cultural complexity," *Literary and linguistic computing*, vol. 28(4), pp. 668-675, 2013.

[16] Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, S. Petrov, "Temporal analysis of language through neural language models," *ACL 2014*, p. 61, 2014.

[17] "Abbreviations dictionary of lithuanian language," https://github.com/tokenmill/ltlangpack/blob/master/tokenizer/abbr-dictionary.xml, accessed: 24 02 2017.