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

doi:10.3233/FAIA200613

Corpus-Based Methods for Assessment of Traditional Dictionaries

Virginijus DADURKEVIČIUS[a,1] and Rūta PETRAUSKAITĖ[b]
[a] *Vilnius University*
[b] *Vytautas Magnus University*

Abstract. The paper presents the investigation of *The Dictionary of Modern Lithuanian* (6th edition) from the point of view of its coverage in comparison with a Joint Corpus of Lithuanian. Resources, methods and procedures are described together with the results revealing that only 81 % of the dictionary lemmas have their counterparts in the corpus.

Keywords. Corpus-based lexicography, updates of traditional dictionaries, Hunspell platform, comparison of dictionaries and corpora

1. Introduction

From its very start, corpus linguistics was used for different purposes of lexicography. At first, raw corpora served as sources of authentic data, then annotated corpora provided different patterns of usage, and finally, lists of entry headwords for newly compiled dictionaries were derived from corpus-based frequency lists. There were other numerous applications of corpora and corpus-based methods of language description, however, they were applied for the compilation of new dictionaries and not for updating the old ones. Nevertheless, traditional dictionaries can be updated and made more efficient with the help of corpora and computational linguistics. This paper presents methods and procedures exploiting corpora for the update of traditional dictionaries, specifically, the list of their entry words. A case study of *The Dictionary of Modern Lithuanian* (6th edition, hence, the DML6) [1] and the Joint Corpus of Lithuanian (hence, JCL) serve as an example.

2. Resources and Procedures

JCL is a merge of three corpora (see Table 1 below): Vilnius university corpus (VU) representing the Lithuanian internet content from 2014 and primarily used for machine translation, a legal document corpus in a form of wordlist (courtesy of the Office of the Seimas of the Republic of Lithuania, 2011, hence, LRSK) and a balanced corpus of present-day Lithuanian of Vytautas Magnus University (VMU). The terms "*tokens*" (all

[1] Corresponding Author: Virginijus Dadurkevičius; Vilnius University Faculty of Physics Institute of Photonics and Nanotechnology, M. K. Čiurlionio str. 29, 03100 Vilnius, Lithuania; E-mails: virginijus.dadurkevicius@tmi.vu.lt, dadurka@gmail.com.

words, including repeated), "*types*" (only distinct words) and "*Type to Token Ratio*" (TTR) are used while discussing corpora, comparing them, assessing their size, scope and representativeness [2]. Usually, TTR is expressed as percentage and tends to decrease as the corpus gets larger. With reference to these terms, the overall size of JCL is 1,334,845,080 tokens, 4,968,125 types, and 0.37 % TTR. The size of JCL is approximately equal to 10,000 books, i.e. the number of books published in Lithuanian in three years.

Table 1. Composition of JCL

Specific corpus	Tokens	Types	TTR	Contribution to JCL
VU	779,154,268	3,958,963	0.51 %	58.4 %
LRSK	443,114,936	1,092,473	0.23 %	33.2 %
VMU	112,575,876	1,778,259	1.58 %	8.4 %

DML6 [1] contains ~600,000 entries with ~86,000 lemmas. The difference in numbers can be explained by the fact that only part of naturally existing lemmas is presented as entry headwords, others are explicitly mentioned in the entries while some of them are not mentioned at all. The latter are called implicit lemmas based on regular word formation patterns. In the Introduction to the dictionary, they are described as belonging to the regular derivational patterns therefore assumed "by default". Thus, the entry with the headword "gailėti" contains 13 lemmas:

a) explicit lemmas of

1. the verb "gailėti" from the derivational paradigm "gaili, gailėjo";
2. the verb "gailėti" from the derivational paradigm "gailėja, gailėjo";
3. the noun "gailėjimas";
4. the noun "gailėjimasis";
5. the verb "gailėtis" derivational paradigm "gailisi, gailėjosi" (a hint of the existence of such a reflexive form is given in the entry "|| sngr.");

b) implicit lemmas of

1. the prefixed derivative verbs "negailėti", "tegailėti", "nebegailėti", "tebegailėti" (regular derivational pattern of the above form is discussed in the Introduction of the dictionary, hence, these forms are not presented in the respective entries);
2. reflexive forms of the above prefixed verbs "nesigailėti", "tesigailėti", "nebesigailėti", "tebesigailėti".

Lithuanian is a synthetic language rich of flexions. First, for the comparison of the dictionary with the corpus, all inflected forms which could be theoretically derived from dictionary lemmas and morphological information provided there had to be generated. As a tool for this task, the Hunspell platform [3] has been chosen. The primary goal of this platform is spelling, but after substantial modification [4], it can also be successfully applied to morphological analysis and synthesis. Successful application of Hunspell platform for Lithuanian was described by Dadurkevičius [5]. Using Hunspell formalism, the scope of a particular language is represented in two files: affixes (morphological rules) and dictionary (words with references to its rules). In our case, the Hunspell dictionary was built by obtaining all the possible lemmas from DML6 entries (both explicitly stated and implied). That made about 200,000

entries in total. The file of morphology rules is used to generate all the theoretically possible word forms. In our case, these rules (about 5,000 items) were based on the *Grammar of Modern Lithuanian* [6]; they are described in detail by Dadurkevičius [5]. References from the Hunspell dictionary to the rules were derived on the basis of information provided in DML6 entries. More than 50 million word forms of DML6 can be generated combining a Hunspell dictionary and its rules. This is how the tool is made suitable for both spelling and morphological analysis based on DML6.

Assessing the coverage of DML6 of the contemporary Lithuanian language represented by JCL, two research questions were asked:

1. What part of JCL is covered by DML6? As a measure for such assessment, the percentage of JCL tokens overlapping with grammatical word forms of DML6 (50+ millions of possible word forms) was calculated. Looking at this facet of the assessment, 100 % overlap would mean a perfect dictionary, able to identify every single word of a corpus. To simplify and speed up the calculation processes, we used the spelling feature of the Hunspell platform to find out if the token in JCL has the matching word form in DML6. A correctly spelled token means that it can be derived from DML6 content. An incorrectly spelled token means a failure to find the match in DML6 and would mark a possible lexical gap in the dictionary. The list of possible gaps [7] could be a valuable resource for updating DML6.

2. How up to date the full list of headwords and other explicit entry lemmas of DML6 really is? A measure for such assessment is the percentage of DML6 explicit lemmas having counterparts (any form, at least one occurrence) in JCL. 100 % would mean a perfect dictionary, with every single headword being used in the corpus that covers a major part of the present-day Lithuanian language. To make this estimation, the list of JCL types has been lemmatized using the functionality of Hunspell platform; implicit lemmas have been ignored. The number of DML6 lemmas having counterparts in the corpus has been compared to the total number of lemmas in DML6. Failure to find DML6 lemma in JCL would mark presently unused words. The fact of such a failure cannot be sufficient to state that headwords, absent in JCL, are out of use nowadays. Nevertheless, the list of unused headwords [7] should be tested applying other methods, e.g. linguistic experiment or introspection.

3. Results

In reply to the first research question concerning lexical gaps and the coverage of DML6, the results, provided below, were obtained. DML6 based Hunspell spell-checker accepted 1,191,815,754 tokens (89.3 %) and 1,252,370 (25.2 %) types of JCL. See Tables 2 and 3 for the distribution of the results in the constituent parts of JCL.

Table 2. Corpora tokens covered by DML6

Corpora	Number of tokens covered by DML6	Total number of tokens in the corpora	%
VU	694,405,495	779,154,268	89.1
LRSK	393,344,588	443,114,936	88.8
VMU	104,065,671	112,575,876	92.4
JCL	1,191,815,754	1,334,845,080	**89.3**

Table 3. Corpora types covered by DML6

Corpora	Number of types covered by DML6	Total number of types in the corpora	%
VU	1,081,818	3,958,963	27.3
LRSK	426,958	1,092,473	39.1
VMU	789,982	1,778,259	44.4
JCL	1,252,370	4,968,125	**25.2**

The reply to the second research question concerning unused lemmas in DML6 provides information about the lemmatization of the corpus that allows to identify 81.1 % of DML6 lemmas. Thus, about one fifth of DML6 lemmas can be regarded as presently unused lexis. See Table 4 for a detailed part of speech analyses of the overlapping lemmas in the compared resources.

Table 4. Number of overlapping lemmas and their POS features in the compared resources

Part of speech	Number of explicit lemmas in DML6	Number of explicit lemmas present in JCL	Number of explicit lemmas absent in JCL	% of the DML6 lemmas having their counterparts in JCL
Adjective	7,398	6,885	513	93.1
Adverb	3,063	2,591	472	84.6
Noun	49,801	37,503	12,298	75.3
Numeral	85	82	3	96.5
Proper noun	2,717	2,706	11	99.6
Pronoun	59	59	0	100.0
Verb	22,020	19,161	2,859	87.0
Other	927	826	101	89.1
TOTAL	86,070	69,813	16,257	**81.1**

A detailed qualitative analysis of the lexical gaps of DML6 as well as its unused dictionary lemmas is planned as the next stage of this research hoping that it should help lexicographers to update the dictionary.

References

[1] The Dictionary of Modern Lithuanian. Edited by Keinys S. 6th (3 electronic) edition of the Dabartinės lietuvių kalbos žodynas. 2006.
[2] Scott, M. WordSmith Tools version 8, Stroud: Lexical Analysis Software, 2020.
[3] Hunspell platform https://hunspell.github.io
[4] Németh L, Trón V, Halácsy P, Kornai A, Rung A, Szakadát I. Leveraging the Open Source Ispell Codebase for Minority Language Analysis. SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages. Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation. Edited by Julie Carson-Berndsen, 2004:56-59.
[5] Dadurkevičius V. Lietuvių kalbos morfologija atvirojo kodo "Hunspell" platformoje [Lithuanian Morphology in the "Hunspell" Framework]. Bendrinė kalba. 2017:1-15.
[6] Lithuanian Grammar. Edited by Ambrasas V. (in English). 1997.
[7] Dadurkevičius V. Assessment Data of the Dictionary of Modern Lithuanian versus Joint Corpora, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/36. 2020.