Corpus based methods for assessment of the traditional dictionaries

Virginijus DADURKEVIČIUS and Rūta PETRAUSKAITĖ

Vilnius University, Vytautas Magnus University

Corpora and lexicography

Corpora for the compilation of the **new** dictionaries:

- The source of authentic language data
- The source of patterns of usage (if annotated)
- The source of entry headwords derived from frequency lists

Corpora for the assessment / recompilation / updating of the **old** dictionaries – the case of Lithuanian language:

- The Dictionary of Modern Lithuanian, 6th ed. (DML6)
- The Joint Corpus of Lithuanian (JCL)

Resources

JCL is a merge of three corpora (see the table below):

- Vilnius university corpus (VU) compiled out of the Lithuanian internet content from 2014 and primarily used for machine translation
- Legal document corpus in a form of wordlist (courtesy of the Office of the Seimas of the Republic of Lithuania, 2011, hence LRSK)
- Balanced corpus of present day Lithuanian of Vytautas Magnus University (VMU).

Overall size of JCL is 1 334 845 080 tokens, 4 968 125 types and 0,37 % Type to Token Ratio (TTR) [2]

DML6 [1] contains ~60 0000 entries with ~86 000 lemmas.

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 %

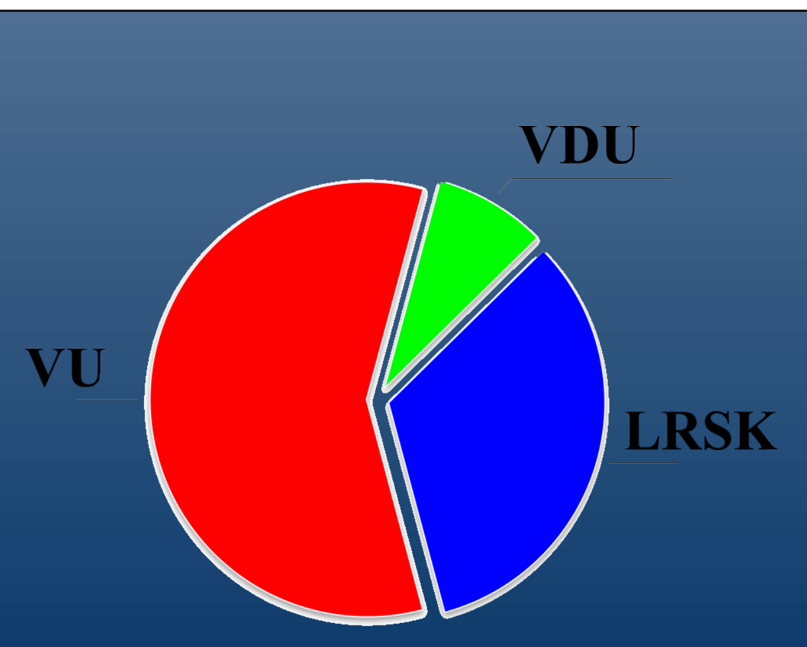
Research Questions

No 1.
What part of JCL is covered by DML6?

No 2.
How up to date the full list of headwords and other explicit entry lemmas of DML6 really is?

Procedures

- Main tool for DML6 digital representation – Hunspell platform. Using Hunspell formalism more than 50 million possible word forms of DML6 can be generated. [3], [4], [5], [6]
- We used the spelling feature of the Hunspell platform to find out if the token in JCL has the matching word form in DML6.
- JCL has been lemmatized using functionality of Hunspell platform. 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.



Results

- In reply to the first research question, concerning lexical gaps and 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 table 2 and 3 for the distribution of the results in the constituent parts of JCL.
- The reply to the second research question concerning unused lemmas in DML6 provides information about the lemmatization of the corpus that allow to identify 81 % of DML6 lemmas. So, 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.
- 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.

Table 2. Corpora tokens covered by DML6

Corpora	Number of tokens covered by DML6	Total number of tokens in the corpora	%
VU	694405495	779154268	89,1 %
LRSK	393344588	443114936	88,8 %
VMU	104065671	112575876	92,4 %
JCL	1191815754	1334845080	**89,3 %**

Table 3. Corpora types covered by DML6

Corpora	Number of types covered by DML6	Total number of types in the corpora	%
VU	1081818	3958963	27,3 %
LRSK	426958	1092473	39,1 %
VMU	789982	1778259	44,4 %
JCL	1252370	4968125	**25,2 %**

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	7398	6885	513	93,1 %
Adverb	3063	2591	472	84,6 %
Noun	49801	37503	12298	75,3 %
Numeral	85	82	3	96,5 %
Proper noun	2717	2706	11	99,6 %
Pronoun	59	59	0	100,0 %
Verb	22020	19161	2859	87,0 %
Other	927	826	101	89,1 %
TOTAL	86070	69813	16257	**81,1 %**

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 repository, 2020, https://clarin.vdu.lt/xmlui/handle/20.500.11821/36.