The First Corpus-driven Lexical Database of Lithuanian as L2

Jolanta KOVALEVSKAITĖ, Loïc BOIZOU, Agnė BIELINSKIENĖ,
Laima JANCAITĖ, Erika RIMKUTĖ
Vytautas Magnus University
jolanta.kovalevskaite@vdu.lt, lboizou@gmail.com,
agne.bielinskiene@vdu.lt, laima.jancaite@vdu.lt,
erika.rimkute@vdu.lt

The Lexical Database is a new resource to support A2-B2 learners of Lithuanian with the **encoding abilities**:

- ❑ pronounciation information
- ❑ morphological information (e.g., inflected forms for nouns)
- ❑ headword usage patterns and examples
- ❑ derivatives related to each meaning of the headword, if it has more than one meaning

To develop the headword list and to study word patterning the **written part (618,637 tokens)** of the automatically morphologically annotated **Pedagogic Corpus** (Boizou et al. 2020) was used:

1) Lithuanian language coursebooks (17.2%) and
2) a variety of authentic Lithuanian material (82.8%) selected using the criteria of learner-relevant communicative function and genres: news portals, popular science books, advertisements, public information (travelling, health care and other services), etc.

The size of the database is appr. **3.500 headwords**:

- ❑ single words:
  - nouns
  - verbs (except auxiliary and modal verbs)
  - adjectives
  - adverbs (except deictic adverbs)
  - some numerals (hundred, thousand, million)
- ❑ MWEs (idioms, two-word compounds, and proverbs),
- ❑ word formations (derivatives and compounds).

Types of records in the database:

- ❑ **full-record** (words (and derivatives) with frequency 100 and above): usage patterns, examples and derivatives related to particular word meanings;
- ❑ **short-record** (derivatives with frequency below 100): examples and derivatives.

Headword list development procedure

- ❑ is based on the word frequency distribution in the ***Lithuanian Pedagogic Corpus***.
- ❑ **a relative core vocabulary** were identified, i.e. words that appear in each level or at least in three levels (appr. 7700 items) but for *Corpus Pattern Analysis* (CPA) we took only those **with corpus frequency 100 and above (appr. 700 items).**

Core vocabulary+their extensions (MWEs, word formations) = headword list of 3.500 items.

The entry structure in the XML database MONGO

- ❑ organizational data (status, comment, and editor),
- ❑ frequency data from each sub-corpus A1-B2,
- ❑ a phonetic container (pronunciation and transcription, the accentuation type for nominal words),
- ❑ a usage container (word meanings with corpus patterns and examples; derivatives related to particular word meaning)
- ❑ a morphological container (the part of speech of a headword, inflected forms that appear in the corpus and the frequency for each form).

RECOGNIZING AND DESCRIBING CORPUS PATTERNS (partly automated)

For the description of word usage, we adopted the **inductive procedure of Corpus Pattern Analysis** (CPA; Hanks 2004; 2012).

Meaning 'indicates, signifies'
Pattern [Sub] [REIKŠTI] [Obj_acc]
Geltona spalva reiškia saulę (The yellow colour means the sun.)

Meaning 'has a value':
Pattern [Sub] [Obj_dat] [REIKŠTI] [Obj_acc]
Ką Tau reiškia pokalbis? (What does a conversation mean to you?

❑ STEP 1

- ❑ frequent syntagmatic pattern(s) for each word were identified by means of the specially designed **Lithuanian Sketch Grammar** (https://www.sketchengine.eu/):

14 dual syntactic relations were defined using such categories as the part of speech and case, with verb forms (infinitive, participle) and neutral gender for adjectives playing an auxiliary role.

The rules are based on expected typical dependents for given parts of speech:

- **for verbs**: nouns/pronouns in different cases (except vocative), adjective (for the verb būti 'to be'), preposition, infinitive, conjunctions;
- **for nouns**: preposed adjectives/participles with case agreement, preposed nouns in genitive, some left dependents in dative or genitive (e.g., *įtaka kam*, 'influence on sth/sb') or related through a conjunction (e.g., *klausimas, ar…* 'the question whether…') or a preposition (e.g., *priemonė nuo ko* 'measure against sth/sb');
- **for adjectives**: prepended adverbs, some left dependents in instrumental or genitive (e.g., *išdidus kuo* 'proud of sth/sb'), infinitive (e.g., *svarbu matyti* 'important to see'), preposition (e.g., *greitesnis už ką* 'faster than sth/sb') or related through a conjunction (e.g., *keista, kad…* 'it is strange that…', for neutral adjectives only);
- **for adverbs**: prepended adverbs (e.g., *labai akivaizdžiai* 'very obviously').

❑ STEP 2

- ❑ To examine **collocates** in each grammatical pattern, and to sort collocates into **lexical sets** – a group of words that share one or more semantic features, e.g., collocates *wedding, festival, concert* form a lexical set, which is then used to define a **semantic type** "event" of one of the arguments in a particular pattern.

Meaning 'to phone'
Pattern "gramForm": "[Sub] [SKAMBINTI_imp] [Obj_ins]",
"semForm": "[person] [Pred] [device]",
"collocates": "[ ] [SKAMBINTI_imp] [telephone]"

Meaning 'to play'
Pattern "gramForm": "[Sub] [SKAMBINTI]|[Pred+SKAMBINTI][Obj_ins]",
"semForm": "[person] [Pred]|[modal+V] [musical instrument]",
"collocates": "[ ] [SKAMBINTI]|[+SKAMBINTI] [piano]"

We use a predefined finite set of semantic types: 3 types for verbs (active, state, independent) and 3 types for adjectives (physical, classifying, evaluative). For nouns, following the bottom-up approach, the list of semantic types was non-finite.

Model for pattern description

- ❑ **Multilevel description of patterns:** "gramForm, "semForm", "collocates" and „examples";
- ❑ **morphological categories,** marked using ***Leipzig glossing rules***,
- ❑ **syntactic categories,** marked by international abbreviations (Sub, Obj, Pred, etc.), taken from the syntactically annotated **Lithuanian corpus ALKSNIS** (Rimkutė et al. 2017).
- ❑ The variability in the pattern is indicated with a vertical slash „|" ('either – or').

Linking patterns to examples

- ❑ The examples were sorted according to different corpus patterns.
- ❑ The approximate frequency of a pattern can be seen from the number of examples.

In order to manually select **good examples**, we:

- follow the grammatical, lexical and semantic components of a pattern (an example for each collocate);
- avoid rare words, figurative usage, field-related terms;
- slightly edit some examples (shorten or clarify anaphora).

Concluding remarks

For a broader application of CPA in (learner) lexicography, more tools could be used in the process of both pattern recognition and description (e.g., Baisa et al. 2015).

It would be important to do more research in the future to evaluate the extent to which this headword list represents the basic vocabulary of Lithuanian. One of promising approaches could be the one demonstrated by Brezina et al. 2015.

Word patterns may provide valuable data for language learning and teaching, but application possibilities depend on the functionalities of the user interface which is now under development. The lexical database will be freely available for users on kalbu.vdu.lt in 2021.

Acknowledgements

This on-going lexicographic work is a part of the project "Lithuanian Academic Scheme for International Cooperation in Baltic Studies" (http://baltnexus.lt/en/baltic-studies-project).

References

1. Baisa V., El Maarouf I., Rychlý P., Rambousek A. Software and data for corpus pattern analysis. In: Horáck A., Rychlý P., Rambouseket A., (eds.) Proceedings of the 9th Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2015; 2015 Dec 4-6; Brno, Czech Republic: Tribun EU; 2015. p. 75-86.

2. Boizou L., Kovalevskaitė J., Rimkutė E. Lithuanian Pedagogic Corpus: correlations between linguistic features and text complexity. In: Proceedings of the 9th International Conference Human Language Technologies – the Baltic Perspective, Baltic HLT; 2020 Sep 22-23; Kaunas, Lithuania. p. 233-240.

3. Brezina V., Gablasova D. Is there a core vocabulary? Introducing the New General Service List. Applied Linguistics. 2015 Feb;36(1):1-22.

4. Hanks P. Corpus pattern analysis. In: Williams G., Vessier S., editors. Proceedings of the 11th EURALEX International Congress. Vol. 1; 2004 Jul 6-10; Lorient, France: Université de Bretagne-Sud; 2004. p. 87-97.

5. Hanks P. How people use words to make meanings: semantic types meet valencies. In: Boulton A., Thomas J. (eds.) Input, process and product: developments in teaching and language corpora. Brno, CZ: Masaryk University Press; 2012.

6. Rimkutė E., Bielinskienė A., Kovalevskaitė J., Boizou L., Aleksandravičiūtė G. Lithuanian Treebank ALKSNIS, CLARIN-LT digital library in the Republic of Lithuania [Data file]. Kaunas, Lithuania; 2017. [cited 9 Jul 2020]. Available from: http://hdl.handle.net/20.500.11821/10