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

doi:10.3233/FAIA200630

The First Corpus-Driven Lexical Database of Lithuanian as L2

Jolanta KOVALEVSKAITĖ [1], Loïc BOIZOU, Agnė BIELINSKIENĖ,
Laima JANCAITĖ, Erika RIMKUTĖ
Vytautas Magnus University, Lithuania

Abstract. The article presents a new resource for A2-B2 learners of Lithuanian as L2 to improve their lexical competence and language production skills. The lexical database is a lexicographic application of the Lithuanian Pedagogic Corpus which was used both to develop headword lists and to collect word usage information. For this study, we adopt the inductive procedure of Corpus Pattern Analysis which was partly automated using the Lithuanian Sketch Grammar in Sketch Engine. We explain the model for pattern recognition and description, sense division, the selection of examples and give some details concerning the user interface.

Keywords. Lithuanian, Lithuanian Pedagogic Corpus, learner lexicography, foreign language learning resources, corpus pattern, Lithuanian Sketch Grammar

1. Introduction

The paper aims to present a new lexical resource for learning and teaching Lithuanian – the first Corpus-driven Lexical Database (henceforth, the database). The target group of the database is A2-B2 learners of Lithuanian (according to *the Common European Framework of Reference for Languages* (henceforth, CEFR)). This on-going lexicographic work is a part of the project "Lithuanian Academic Scheme for International Cooperation in Baltic Studies"[2].

In the first part, we briefly describe the Lithuanian Pedagogic Corpus (henceforth, the corpus), which was used to study word patterning, and explain our strategy for headword list development. As the integration of available Lithuanian corpora in L2 classroom is rather limited, and explanatory Lithuanian dictionaries often lack data that reflects the modern language usage, the main purpose of the database is to provide learners and teachers with relevant material on language use. For the description of word usage, we adopted the inductive procedure of Corpus Pattern Analysis (CPA; Hanks [1, 2]), which consists of several steps: 1) preparation of the Lithuanian Sketch Grammar; 2) adoption of a model for pattern recognition, sense division and pattern description; 3) guideline setting for example selection (see Section 3).

All the information related to word usage was recorded in the XML database MONGO: the entry structure contains organizational/technical data (status, comment, and editor), frequency data from each A1-B2 sub-corpus, a phonetic container (pronunciation, transcription, and the accentuation type for nominal words), a

[1] Corresponding Author: Jolanta Kovalevskaitė; Vytautas Magnus University, V. Putvinskio st. 23-216, Kaunas LT-44243, Lithuania; E-mail: jolanta.kovalevskaite@vdu.lt.
[2] http://baltnexus.lt/en/baltic-studies-project.

morphological container (the part of speech of a headword, inflected forms that appear in the corpus, and the frequency for each form). In the usage container, the record depends on the headword frequency in the corpus: a) for words (and derivatives) with frequency of 100 and above, a full-record is prepared: with patterns, examples and derivatives related to particular word senses; b) for derivatives with frequency below 100, a short-record is prepared: with examples and derivatives related to particular word senses.

Due to financial project limitations, we chose to use open source and (at least, for the project period 2018-2020) freely available tools: the text analysis software Sketch Engine and the Dictionary Writing System Lexonomy. However, after the testing stage, we had to reject Lexonomy because of data buffering problems. Besides, the experience with some experimental Sketch Engine functionalities (e.g., OneClick Dictionary) showed that in a real lexicographic project, their use is (still) rather limited.

2. The Structure of the Database

In this part of the paper, we explain our procedure for the headword list development that is based on the word frequency distribution in the Lithuanian Pedagogic Corpus.

2.1. The Lithuanian Pedagogic Corpus

The size of this monolingual specialized corpus is 669,000 tokens. The corpus consists of two sub-parts: written language (A1-A2: 96,000 tokens; B1-B2: 523,000 tokens) and spoken language (A1-A2: 15,000 tokens; B1-B2: 35,000 tokens) (see [3]). The material for the database was collected from the written part of the corpus (618,637 tokens), thus, the database represents the usage of the written Lithuanian language. The data for the written part of sub-corpus was collected from 1) the 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.

Although the corpus we used is rather limited, the probable usage and meanings [2] of frequent words can be seen in small corpora as well. As pattern recognition for the low frequency words might be more problematic, we analyzed the concordances of at least 100 occurrences of the node word. Accordingly, word patterns were provided only for words with corpus frequency of 100 and above (see 2.2). The other shortcoming is related to the fact that although the corpus includes 29 genres (texts from news portals, stories, fairy tales, advertisements, letters, songs, and others), their distribution is unbalanced.

Some genres make up a bigger part of the corpus than others, and this can influence word patterning: e.g., advertisements amount to 6.4 % of the data, whereas public information texts (e.g., notices in public transport, public catering and healthcare institutions, etc.) are short, and their percentage in the corpus is low (0.3 %). As a result, patterns in advertisements are detected automatically because of frequent usage, and patterning in public texts may not be automatically recognized because of infrequent usage. Thus, frequency is important, but as less frequent usage patterns could also be worth lexicographers' attention, they should be verified by consulting a bigger corpus.

2.2. Core Vocabulary and the Headword List

Since we were aware that corpus division according to CEFR levels is problematic (see [4]), we decided to dismiss the approach of level-linking (e.g., CEFR-graded lexical resources [5]). Instead, we attempted to define a relative core vocabulary, i.e., words that appear in each level or at least in three levels (appr. 7,700 items). However, for the pattern analysis, only words of particular word classes with frequency of 100 and above were selected from this core vocabulary (see motivation in 2.1). Thus, our headword list for CPA analysis includes appr. 700 items that appeared in all CEFR levels from A1 to B2 or from A2 to B2. The headword list consists of nouns, lexical verbs (except auxiliary and modal verbs), adjectives, adverbs (except such deictic adverbs as *čia* 'here', *ten* 'there', *dabar* 'now', *kodėl* 'why') and some numerals (*šimtas* 'hundred', *tūkstantis* 'thousand', *milijonas* 'million'). In the case of primary verbs (e.g., *siekti* 'to seek', *imti* 'to take', *likti* 'to stay'), which can function both as auxiliary and lexical verbs, full record was prepared to show usage differences.

The headword list is being extended with two kinds of headwords: a) some multiword expressions (henceforth, MWEs) including an item from the headword list, b) some word formations (derivatives and compounds) from the core vocabulary related to the items from the headword list. With all MWEs and related word formations added to the headword list, we expect the final headword list of appr. 3,000 lexical items.

In the database, 3 types of MWEs are included: idioms, two-word compounds and proverbs (sayings and greetings). MWEs are identified manually and represented with a short record. The derivatives and compounds are selected from the core vocabulary, but as their frequency is below 100, their usage is demonstrated only with a short record. Word formations are selected manually from the headword list. Words that are not in the headword list are also sometimes indicated as base words and then they are labelled with a special symbol. Occasionally, the relation between a base word and its derivatives is just formal, but not semantic, e.g., *sakyti* 'to express thoughts in words' - *užsakyti* 'to commission, to arrange'; *eiti* 'to go from one location to the other on foot' - *apsieiti* 'to get by'. In such cases, both a base word and its derivative are assigned a special symbol.

3. Recognizing and Defining Corpus Patterns

For the description of word usage, one of the corpus-driven methods, Corpus Pattern Analysis (CPA) [1, 2], was adopted. CPA describes a pattern as a syntagmatic structure with semantic values for arguments, i.e., semantic types populated by lexical sets. CPA draws on the insights of the corpus-driven language analysis and the contextual and functional theory of meaning. Maintaining Sinclair's position that "not single words, but rather words in their contextual patterns are the true bearers of meaning" [6], the meaning of a word is associated with a specific lexical and grammatical environment.

Relying on a slightly modified CPA approach, in this project, the observation and definition of meaning-related patterning were performed using the Sketch Engine[3] and the specially designed Sketch Grammar for Lithuanian. During the analysis of the patterning of a word, we applied both automatic (grammatical patterns) and manual procedures (semantic type identification, sense identification, and example selection).

[3] https://www.sketchengine.eu/.

3.1. Lithuanian Sketch Grammar

The corpus had been previously automatically morphologically annotated with *Semantika.lt* analyser; therefore, it was possible to prepare a morphology-based Sketch Grammar for Lithuanian. Given the intended purpose to extract lexico-grammatical patterns, the aim was to capture some syntactic relations 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').

For Lithuanian Sketch Grammar, the following 14 dual relations were defined:

has_adj_modifier/is_adj_modifier_of	has_gen_noun/is_gen_noun_of
has_right_modifier/is_right_modifier_of	has_dat_noun/is_dat_noun_of
has_adv_modifier/is_adv_modifier_of	has_loc_noun/is_loc_noun_of
has_noun_modifier/is_noun_modifier_of	has_inf_compl/is_inf_compl_of
has_acc_noun/is_acc_noun_of	has_pred_adj/is_pred_adj_of
has_nom_noun/is_nom_noun_of	has_adp/is_adp_of
has_ins_noun/is_ins_noun_of	has_conj/is_conj_of

To reduce the number of false relations, the syntactic relations were defined inside the sentences and with a limited number of tokens inserted between the considered words. In general, verb-centered relations allow for a maximum of 3 token gap, whereas other structures – for a one-token gap or no gap at all. Another difference of verb-centered relations is that they are bilateral (the related word can be before or after a verb), while for other relations, the related word is expected either to the right or to the left (depending on the relation).

Such an approach has well-known shortcomings. Given the lack of syntactic annotation, some relations can be postulated on the basis of the Sketch Grammar between words that are not directly syntactically related. Furthermore, some cases are quite ambiguous, especially the genitive case, which is mostly used for attributes, but may also express a complement for negative verbs, as well as for several positive verbs. WordSketch selects only those combinations that are described by the rules in the Sketch Grammar. As a consequence, occasionally such automatic WordSketch analysis might not reveal some portion of typical usage (e.g., the parenthesis usage function of the adjective *aiškus* 'clear', combinations with numerals, e.g., *dveji metai* 'two years'). Nevertheless, the Sketch Grammar acts as a filter which prepares a WordSketch for a lexicographer that is then used for the manual analysis of headword patterning.

3.2. Corpus Pattern Analysis Applied

In the database, we provide a systematic description of word usage patterns formed by grammar and lexis while analyzing words with the frequency of 100 and above. A corpus pattern includes grammatical (syntactic functions and morphological categories of case and verb forms), lexical (words and collocations) and semantic (semantic types) components. A language feature (collocate, grammatical form, and syntactic function) has to occur at least 3 times in the corpus to be analyzed as a pattern element.

Pattern recognition. First, frequent syntagmatic pattern(s) for each word were identified by the Sketch Grammar. Word usage overview provided in the WordSketch helps lexicographers to make initial hypotheses about meaning-related patterning. Sometimes valency alone is sufficient to make the semantic distinction [2]: in the case of the verb *reikšti* ('to mean'), it is the object in dative which differentiates two verb meanings – *nurodo, žymi* ('indicates, signifies') and *turi vertę* ('has a value'):
[Sub_nom] [REIKŠTI] [Obj_acc]: Geltona spalva reiškia saulę (The yellow colour means the sun.)
[Sub_nom] [Obj_dat] [REIKŠTI] [Obj_acc]: Ką Tau reiškia pokalbis? (What does a conversation mean to you?)
Generally, collocations and semantic types are also needed for sense distinction, thus, the final sense binding to patterns is performed only at the second stage of the analysis.

As the WordSketch provides two-element grammatical patterns, a lexicographer decides where the boundaries of the pattern are. Usually, due to their government structure, verb patterns have more elements than noun, adjective and adverb patterns. On the other hand, it is known that some nouns are used as adverbs (e.g., *daugybė* 'multitude', *daugelis* 'most, many'), predicative adjectives are used as verbs (e.g., *vertas* 'worthy', *pilnas* 'full'), thus, the analysis and description of their patterns are different in comparison with those of nouns or adjectives.

Sometimes several elements function as one pattern component (adverbial, subject, object, attribute). In such cases, manual work is needed to identify them properly and integrate them into the pattern, e.g., *gero būdo žmogus* 'an easygoing person' was described by an attributive pattern which consists of two components [AtrA BŪDAS_sg.gen] [Mod]: attribute [AtrA BŪDAS_sg.gen] and modifier [Mod], which is realized semantically by a semantic type [human] and lexically – by a collocate *žmogus* 'a person'.

The information of morphological forms is very important in the identification of patterns, because sometimes the forms that realize a word in the corpus show that the word is only used as a parenthesis, e.g., only the form *vadinasi* ('it turns out') of *vadintis* ('refl. to call') is used.

After the grammatical patterns are sketched, the second part of the procedure begins: a lexicographer examines collocates provided by WordSketch in each grammatical pattern (unlike in CPA, we do not evaluate collocates by statistical significance), and sorts collocates into lexical sets – a group of words that share one or more semantic feature, 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.

Semantic types are often the main separators of meanings, especially when two verb senses are associated with the same grammatical pattern, e.g., when using the verb *skambinti*, both meanings 'to phone' and 'to play' are realized by the grammatical pattern [Sub] [SKAMBINTI] [Obj_ins], but the [Obj_ins] is a semantic separator, because for the first meaning it is realized by a semantic type [device] (to call by telephone) and for the second meaning – by a semantic type [musical instrument] (to play the piano).

Besides, the meaning 'to phone' is realized by 5 patterns, while the meaning 'to play' has one pattern.

The analysis of semantic types according to the CPA procedure has to be performed with a preliminary ontology. We did not use any ontology, but for collocates that are verbs and adjectives, we used 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: more types can be added depending on the context of a word.

Sometimes, as mentioned by the CPA practitioners [2], it is problematic to decide on the appropriate level of semantic generalization for a semantic type, e.g., too broad semantic types like [abstract] could be not sufficient to make important semantic distinctions. In our case, this problem is sometimes related with the size of the corpus: e.g., for adjective *naujas* ('new') we have to semantically categorize the collocates *kontaktai* 'contacts'*, giminės '*relatives*', augintinis '*a pet*', galerija* 'a gallery*', skonis '*a taste*'*, and because each of them goes to a different semantic type, it is difficult to generalize them semantically.

When lexical and semantic elements are integrated into the grammatical patterns, every pattern (or patterns) is linked to a specific sense. Sense division is based only on patterns, while the existing explanatory dictionaries of Lithuanian were used only in problematic cases.

Pattern description. We used the model for pattern description which consists of grammatical categories and some rules how to show separate elements and their variability.

While learning such a morphologically rich language as Lithuanian, it is important to master the cases and grammatical forms. In pattern description, it is necessary to indicate a case and, quite often, verb forms. For this reason, we provide a multilevel description of a pattern, i.e., grammatical (gramForm), semantic (semForm) and lexical (collocates) realizations are given separately, e.g.,

"gramForm": [ARBATA] [su AtrN_ins] / [TEA] [with AtrN_ins]
"semForm": [Mod] [maistas] / [Mod] [food]
"collocates": [ARBATA] [su citrina] / [TEA] [with lemon]

In the collocate line, collocates are indicated as a lemma; however, fixed forms, and multi-word constructions (cf. *su citrina*) can also be given.

We can see in the pattern above that the word under analysis is capitalized, and separate components in fields "gramForm, "semForm" and "collocates" are surrounded by square brackets. The variability in the pattern is indicated with a vertical slash '|' ('either – or'). For example, in [Pred] [CENTRAS_acc]|[CENTRAS_ins]: the object is expressed in accusative or instrumental. For the grammatical description of the pattern, morphological categories are marked using Leipzig glossing rules and syntactic categories are marked by international abbreviations (Sub, Obj, Pred, etc.), taken from the syntactically annotated Lithuanian corpus ALKSNIS [7].

Linking patterns to senses and examples. As already mentioned, each sense of a headword is represented with one (or more) pattern(s). The examples were sorted according to different corpus patterns. Lexicographers can provide a sense description for one or more patterns, but the database users will be provided only with patterns and examples.

Explanations of the meaning are not included for several reasons. First, this resource is meant to train learner's encoding skills. Findings from several studies presented by [8] support the idea that examples indeed seem to help language production. Second, there

seems to be no consistent correlation between learners' preferences of dictionary explanations and their success in encoding (see the discussion by [9]). Added equivalents could be a good option for resource development, as they can be used for the same purpose as explanations.

Our approach to example selection was not automated by GDEX facility of Sketch Engine, because our example selection principle was based on described corpus patterns. The grammatical, lexical and semantic components of a pattern help to collect corpus examples that are typical. To ensure that examples are informative and clear, we avoided rare words, figurative usage, and field-related terms. Some examples were slightly edited (shortened or with inserted explanations to clarify anaphora). Usually, the example is one sentence, but in some cases more, than one sentence is given – this helps to illustrate some MWEs (sayings or idioms) in a broader context.

The number of examples depends on the number of collocates for each semantic type; therefore, the approximate frequency of a pattern can be seen from the number of examples. Encoding examples which illustrate patterns and collocations dominate. On the other hand, decoding examples which contain contextual clues about the meaning are also included. As our headword list is not CEFR-level graded, we do not aim to select examples which correspond to the level of an item.

4. User Interface

As the development of user interface is now in progress, we will only provide its brief description. We plan to offer two search options – the search in the headword list and the search in the collocates list. In the pattern description, we provide frequently used collocates from each semantic type (see Section 3), but not all of those collocates are headwords, thus, the collocate search gives the user a possibility to see more collocational networks, which is important for the development of lexical competence. By selecting a particular word, the users will be provided with patterns and examples associated with each word sense. The current form of grammatical realization description may be difficult to understand for learners, thus, we are searching for options to simplify the representation for the end-user.

In the development of some possible end-user scenarios, we address two user groups – language learners (who have reached intermediate level A2) and teachers. For A2 learners, examples, pronunciation, and inflection could be the first relevant option. More advanced learners could be interested in corpus patterning, examples or derivatives. Meanwhile, teachers might benefit from both examples and corpus patterns: they can be used to explain lexical and grammatical environment related to a word (word sense) or to prepare learning material (e.g., exercises with pattern analysis, comparison, pattern-sense relation detection, etc.).

5. Conclusions

In the paper, we explained the application of CPA for Corpus-driven Lexical Database, and mentioned some problematic issues concerning its application. While recognizing and defining corpus patterns, a real challenge for lexicographers is to remain flexible in their observation task (not to be limited only to the repertoire of preselected categories) and, at the same time, to follow the guidelines. To ensure the consistency in pattern

description, we apply the cross-validation approach commonly used in dealing with corpus annotation – every full-record is checked by two lexicographers. Working with the morphologically annotated corpus, we partly automated the grammatical pattern recognition at least at the beginning of using the Sketch Grammar, but for the broader application of CPA in (learner) lexicography, more tools could be used in the process of both pattern recognition and description (e.g., [10]).

Given the limited scope of the project and the mentioned limitations of the corpus, we consider our approach for headword list as reasonable. Nonetheless, 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 as L2. One of the promising approaches could be the one demonstrated by [11].

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.

References

[1] 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.

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

[3] Kovalevskaitė J, Rimkutė E. Mokomasis lietuvių kalbos tekstynas: naujas išteklius lietuvių kalbos besimokantiesiems. (Pedagogic Corpus of Lithuanian: a new resource for learning and teaching Lithuanian as a foreign language.) Sustainable Multilingualism. Forthcoming 2020.

[4] 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. Forthcoming 2020.

[5] François Th, Gala N, Watrin P, Fairon C. FLELex: a graded lexical resource for French foreign learners. In: Calzolari N, et al., editors. Proceedings of International Conference on Language Resources and Evaluation, LREC 2014; 2014 May 26-31; Reykjavik, Iceland: European Language Resources Association (ELRA); 2014. p. 91-102.

[6] Sinclair J. Corpus, concordance, collocation. Oxford, UK: Oxford University Press; 1991.

[7] 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

[8] Frankenberg-Garcia A. Dictionaries and encoding examples to support language production. International Journal of Lexicography. 2015; 24(4):490-512.

[9] Moon R. Explaining meaning in learners' dictionaries. In: Durkin Ph, editor. The Oxford Handbook of Lexicography. Oxford, UK: Oxford University Press; 2016. p. 123-143.

[10] Baisa, V, El Maarouf I, Rychlý P, Rambousek A. Software and data for corpus pattern analysis. In: Horácк A, Rychlý P, Rambouseket A, editors. 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.

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