

LEXICOGRAMMATICAL PATTERNS OF LITHUANIAN PHRASES

Rūta Marcinkevičienė, Gintarė Grigonytė
Vytautas Magnus University, Kaunas, Lithuania

Abstract

The paper overviews the process of compilation of the first corpus-based Dictionary of Lithuanian Phrases. Phrases are transformed from collocational strings which were extracted from the corpus of contemporary Lithuanian language of 100 million running words applying a new statistical method called Gravity counts. The paper presents theoretical approach towards the most relevant notions of collocation, collocational string, phrase, part of speech, grammatical and lexicogrammatical pattern. Statistical method of extraction of collocational strings is shortly presented together with the initial output of raw collocational strings. Types of transformations of collocational strings into phrases and other manual procedures are described in a nutshell while primary results of patterning of the Lithuanian phrases as well as future steps are presented in greater detail.

Keywords: collocation, collocational string, Gravity counts, fragment of text, POS pattern, grammatical pattern

1. Introduction

The compilation of the Dictionary of Lithuanian Phrases includes three main phases: extraction of collocational strings from the corpus of present day Lithuanian language, transformation of collocational strings into phrases, and patterning of all the phrases. Each phase is based on certain theoretical approaches as well as notions of collocation, collocational string, phrase, and pattern, presented here.

Collocation is a fuzzy term embracing a great variety of notions. The definition of a collocation differs according to researcher's standpoint and the method of extraction. There are two different perspectives on the notion of collocation from the point of view of its form and structure. One group of authors (J.Firth, J.Sinclair, M.Stubbs, among others) prefers contextual or statistical definition of collocation. It could be generalised as follows: one item collocates with another that appears somewhere near it in a given text. The assumption underlying collocation is based on its structure: collocation consists of a node word and its collocates, so the search of a collocation starts with the node word. Thus statistical definition highlights lexical relationship between two or more items that tend to co-occur. However, it does not allow one to detect multi-word collocations as they appear in the texts and to define their boundaries. Statistical collocations are usually presented as lemmas for node words and their collocates.

Another group of authors (G.Kjellmer, G. Williams, etc.) pursue a lexicographic approach and include grammatical well-formedness and grammatical acceptability in the list of criteria for collocations. They see collocation as a fragment of text, not as a list of collocates for the previously selected node words.

A compiler of a dictionary of collocations has to choose between the two approaches, since the attitude towards collocation predetermines the method of extraction and the method of presentation. From the perspective of the Lithuanian language the lexicographic approach is more acceptable. It provides a lexicographer with authentic strings of words or fragments of texts obtained by applying statistical tools. These strings contain collocating grammatical forms presented in their natural word order (and not isolated lemmas) which are of paramount importance for the highly inflected Lithuanian language. Such *collocational strings* can be sorted out with their grammatical autonomy in mind but they do not have to be reconstructed from a mere list of nodes and their collocates.

Finally, this approach allows us to avoid making a pre-selected list of node words and to process the entire corpus from the first to the final word. It presents, therefore, a full-text approach to language and utilises the entire corpus, i.e. every sentence it contains, not merely concordances derived from the corpus on the basis of a previously compiled list of node words. Consequently this approach allows us to determine the amount of text that is formed on the idiom principle (Sinclair 1991: 109-121). The choice of the lexicographic approach as opposed to the statistical one determines the choice of a particular method for the extraction of collocations.

Collocational strings after they are extracted from raw texts do not always coincide with grammatically well formed and semantically sufficient word combinations, therefore they have to be transformed into such autonomous phrases either by cutting off irrelevant or adding their missing parts. By *phrase* we understand as a frequently used autonomous fragment of text. Our theoretical standpoint does not allow us to interfere with the inner structure of a phrase and to change its word order or morphological form. The last phase in the compilation of the dictionary is patterning of phrases. The concept of a pattern is borrowed from corpus linguistics where it is conceived as a juncture of most prominent lexical and grammatical features of a phrase (Hunston et al. 2000). Traditionally patterns are centered either around a lexical item (in lexicography) so that they reveal its usage and meaning or they are centered around a part of speech (e.g. verbal, nominal, adjectival patterns in grammar). In the first case patterns are too concrete and specific, in the second case they are too abstract and general.

We apply a holistic approach and aim at combining of a) lexical, b) semantic and c) grammatical features into one pattern, e.g.:

verb of motion (b) + preposition "link" (a) + concrete noun in Genitive (b-c).

Our basis of patterning is different from the existing corpus or traditional approaches due to the source material. It is based on the list of phrases of various lengths instead of invented examples or node word concordances. Since phrases and collocations represent the most frequent and significant fragments of the Lithuanian language, patterns derived from these phrases can be regarded as basic. As such they can be of paramount importance for the probabilistic parser and other related tools.

2. Method of extraction of collocational strings

Collocational strings were extracted from the corpus of Lithuanian language with the help of a statistical method called Gravity counts. It adopts a linear approach of

consecutive counts of words in a text, and of all the texts in a corpus, based as it is on the combinability counts of each pair of words in the corpus irrespective of their hierarchical status, i.e. there is no *a priori* list of node words for which collocates are obtained from the corpus. Each word in the corpus is processed as the node word; its gravity by reference to the pairing word and the next two words in the span of three words is calculated using the formula below (for more about the method see Daudaravičius et al. 2004):

$$G(x, y) = \log\left(\frac{f(x, y) \cdot n(x)}{f(x)}\right) + \log\left(\frac{f(x, y) \cdot n'(y)}{f(y)}\right)$$

Gravity Counts are based on an evaluation of the combinability of two words in a text that takes into account a variety of frequency features, such as individual frequencies of words, the frequency of a pair of words and the number of different words in the selected span. Gravity Counts highlight habitual co-occurrence of two words in a text within the chosen span, in our case the span of three words. If the first word *x* is used more habitually than expected in front of the second word *y*, and the second word *y* is used more habitually than expected after the first word *x*, then *x* and *y* form a minimal collocational string.

Gravity Counts are also based on word order, so that for each first word *x* in a pair the frequency of the following three words is taken into consideration, while for each second word *y* of a pair the frequency of the three preceding words is computed. Therefore *n(x)* is the number of different words to the right of *x* and *n'(x)* words to the left of *y*; *f(x)* and *f(y)* is the frequency of *x* and *y* in the corpus.

The method of Gravity Counts and the detection of collocation boundaries helps to identify segments of texts as statistically significant collocational strings. These strings can be said to be always natural since they present authentic fragments of a text. Nevertheless, statistical collocational strings differ from the point of view of their grammatical and lexical autonomy, which is the most relevant feature in our analysis. Certain collocational strings are self-sufficient and can be regarded as autonomous and grammatically well-formed phrases. Other kinds of collocational strings are somewhat deficient and have to be transformed into a phrase. In order to differentiate between autonomous and deficient collocational strings obtained from the corpus using the method of Gravity counts, as well as to define their ratio, a manual analysis is performed. A fairly high percentage, i.e. 82 % of collocational strings are found to be autonomous and clear-cut phrases.

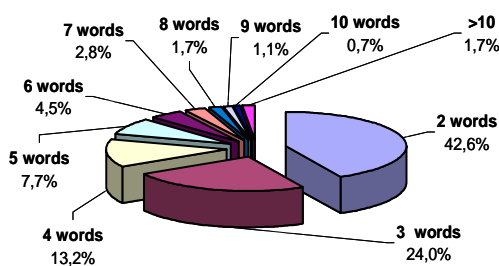


Figure 1. Distribution of initial collocational strings by their length

3. The output and its transformations

Application of Gravity Counts for the corpus of Lithuanian language resulted in processing of 110 935 000 pairs of words in the corpus of 100 million running words (1,7 million different word forms). Some pairs of words were joined into multi-word collocational strings, thus the initial list of collocational strings consists of 19,878,281 items. The list of different collocational

strings is of 10,147,250 items. All the collocations cover 68.1 % of the corpus.

The output of the calculations is the list of collocational strings of varying length. The general tendency for the length of collocational strings is the same as for the frequency of words, i.e. the longer strings are less frequent than the shorter ones. The majority of strings (8,462,626 items which form 42 % of all the list) are made up of two words. The list of three-word strings is twice as short (4,760,991 items, 24 % of the list), the same can be said of the four word strings (2,629,953 items, 13 % of the list) and the five word strings (1,532,370 items, 8 % of the list). The decrease in number for the longer strings is somewhat less (see Figure 1). A typical long collocational string is taken from governmental decrees and consists of 34 words (for a more detailed description see Marcinkevičienė 2004).

The manual processing of the raw output, i.e. transformation of statistical collocational strings into well-formed phrases, consists of several steps and procedures. The first step is to delete all rare strings, irrespective of their length (1 to 3 occurrences) and some more frequent strings depending on their length: two-word strings up to 19 occurrences, three word strings up to 9 occurrences, four word strings up to 8 occurrences, five word strings up to 4 occurrences. This arbitrary decision was based on the considerable amount of noise in these particular items. Besides, only these collocational strings were left that contained at least one noun. This was an arbitrary decision cause by manually unmanagable length of the initial output, i.e. ca 20 million items.

The remaining or intermediate list of 88 562 collocational strings of different length was processed applying three different procedures: lexically well-formed and grammatically autonomous collocational strings were included without changes. Some anomalous, structurally and semantically insufficient strings were deleted (e.g. parts of the string belonging to a different clause, or strings containing proper names, numbers, misprints, consisting exclusively of a noun plus conjunction, a pronoun or one of the forms of the verb “to be”) while some of them were changed. The changes include: a) shortening of grammatically irrelevant parts of long collocations, b) addition of missing words from concordances to deficient strings, mostly two or three word combinations consisting of nouns and prepositions.

The first stage of transformation, i.e. the deletion of deficient strings, left the compilers of the dictionary with the final list of 68,600. i.e. ca 74 % of the intermediate list.

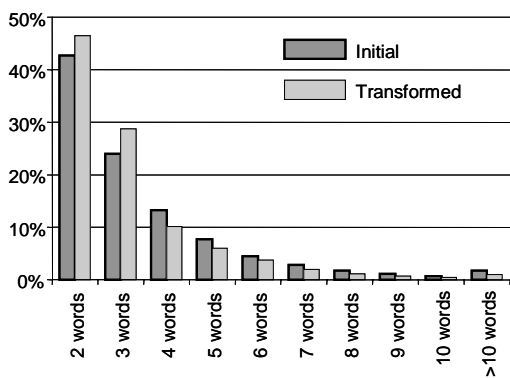


Figure 2. Distribution of initial and transformed collocational strings by their length (in number of words)

Addition of missing words to the deficient collocational strings (mostly consisting of prepositional phrases) did not affect the length of the list, only the length of the specific strings, e.g. some two-word strings were transformed into three or four-word strings. Circa 10% of collocations were lengthened by adding 2-4 words to both sides of a collocation. In some cases additions consisted not of words but of parenthesis which demonstrated the gap to be filled in by a specific lexical item, e.g. a numeral, proper noun, etc.

Figure 2 depicts how the length of

collocational strings was affected by the process of transformations. Two-word and three-word strings were lengthened, four and five-word strings were shortened, six-word and longer strings remained almost intact.

4. Patterns of phrases

Manually processed list of phrases revealed different nature depending on the length of a phrase. The whole list consists of three groups: phrases with more than 6 words could be called *fragments of text* due to their length, two-word phrases, on the contrary, are typical word combinations or traditional *collocations*, while the middle group, i.e. from six to three words, resembles traditional *phrases*. Four-word phrases were chosen to be patterned and described first of all here as the most intuitively recognisable and typical items, least affected by manual transformations, especially additions. The number of different items in this list is 3895, the overall frequency of all four-word phrases in the corpus is equal to 62854.

Four-word phrases were morphologically annotated using the system for morphological annotation of the Lithuanian language "Lemuoklis". The outcome is a list of 1511 POS patterns. Majority of these patterns contained several interpretations for the same word therefore a manual disambiguation was necessary. The first step of disambiguation was to classify all four-word phrases into four groups on the bases on the number of nouns in a phrase. The result is the number of ambiguous patterns for each structural group: 152 four-noun POS patterns, 391 three-noun POS patterns, 576 two-noun POS patterns and 393 one-noun POS patterns.

Disambiguation and clustering of patterns was carried out inside the groups. The outcome of this process revealed an obvious correlation between the number of nouns in the patterns and the number of POS patterns. Four-noun phrases were identical from the point of view of their morphological structure since all the phrases consisted of nouns exclusively. Three-noun phrases presented ca 40 POS patterns (a more exact number of patterns is still the topic of ongoing discussions), POS patterns for two-noun phrases were three times more numerous, i.e. 120 POS patterns, while one-noun phrases manifested the biggest variety of possible combinations of different parts of speech – ca 300 POS patterns. POS pattern of a one-noun phrase is exemplified below:

- (1) Smulkios ir vidutinės įmonės (small and medium enterprises)
(adj) (conj) (adj) (noun)

Further steps in the process of patterning of Lithuanian phrases include detailed analyses of morphological forms of different parts of speech as they are presented in the authentic non-lemmatised phrases. Additional morphological characteristics will give a more finely grained and therefore more numerous lists of grammatical patterns, e.g.:

- (2) Smulkios ir vidutinės įmonės
(adj pl nominative) (conj „ir“) (adj pl nominative) (noun pl nominative)

Last but not least, grammatical patterns will be enriched with specific lexical items for auxiliary parts of speech (e.g. prepositions, particles, conjunctions) and semantic features for the groups of notional parts of speech (e.g. nouns and verbs):

- (3) Smulkios ir vidutinės įmonės
(adj pl nominative) (conj „ir“) (adj pl nominative) (concrete noun pl nominative)

5. Conclusions

Extraction of collocational strings from 100 million word corpus of the Lithuanian language and their transformation into phrases revealed three main findings: a) almost 70 % of all the corpus consists of collocational strings and is based on idiom principle, b) the method of extraction is suitable since relatively small number of collocational strings were found deficient and had to be transformed into phrases manually, c) the number of POS patterns and their structural variety is big, esp. in those phrases that contain less nouns.

Acknowledgements

We are grateful to the State Commission of the Lithuanian Language for supporting the compilation of the Dictionary of Lithuanian Phrases in 2004-2005.

References

- Daudaravičius, Vidas; Marcinkevičienė, R. 2004. Gravity counts for the boundaries of collocations. In: In: Teubert, W.; Johansson, Stig (eds.) *International Journal of Corpus Linguistics* 9/2. 321-348.
- Hunston, Susan; Francis, G. 2000. Pattern grammar. A corpus-driven approach to the lexical grammar of English. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Marcinkevičienė, Rūta. 2004. Dictionary of Lithuanian phrases. In: Williams, G.; Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*, Lorient: UBS.741-751.
- Sinclair, John. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

RŪTA MARCINKEVIČIENĖ is head of the Department of Lithuanian language and the Centre of Computational Linguistics, Vytautas Magnus University, Kaunas. She received her doctor degree (Lithuanian language) at the University of Vilnius, dealing with comparative semantics of English and Lithuanian verbs and her degree of habilitated doctor at Vytautas Magnus University, dealing with Lithuanian corpus linguistics. Her research interests concern corpus linguistics and corpus-based lexicography, lexical semantics, pragmatics, and text linguistics (generic approach). As a visiting lecturer, she has taught Lithuanian language at the University of Stockholm, Lithuanian culture and the Theory of Genre at the University of Illinois, Chicago. She is the member of the board of the Nordic School of Language technologies. E-mail: ruta@hmf.vdu.lt.

GINTARĖ GRIGONYTĖ is an engineer-programmer of the Centre of Computational Linguistics at Vytautas Magnus University. She is doing her master studies at Kaunas Technology University, Software engineering programme. Her research interests concern computational linguistics, software engineering, automatic syntactical analysis. E-mail g.grigonyte@hmf.vdu.lt.