Morphemic Structure of Lithuanian Words

DOI 10.1515/opli-2016-0008
Received October 17, 2014; accepted January 31, 2016

Abstract: The Lithuanian language is a typical flectional language that has a very sophisticated system of grammatical forms and many means of derivation; it is also characterized by uncertain boundaries between morphemes. All this makes the morphemic analysis of the Lithuanian language very complex. The aim of this research is to define and describe morphemic structural models of inflective parts of speech (i.e. nouns, adjectives, numerals, pronouns, and verbs) and regularities of their usage in contemporary Lithuanian.

Keywords: morphemic model, morphemes, Lithuanian language, flectional language

1 Introduction

There are 11 parts of speech in the Lithuanian language, namely nouns, adjectives, pronouns, numerals, verbs, adverbs, conjunctions, particles, pronouns, interjections, and onomatopoeia. Conjunctions, particles, prepositions, interjections, and onomatopoeia are fixed and uninflected. They cannot be morphologically split, as synchronically they are formed only by the root, while adverbs have degree (e.g., gerai – geriau, geriausiai ‘good, better, the best’). Almost all nominal words, i.e. nouns, adjectives, pronouns, numerals are declined by gender, number, and case. Due to semantic reasons some nominal words do not have all of these categories. The verb is by far the most difficult part of speech. Lithuanian verbs can be uninflected (infinitive, adverbial participle), conjugated, and even declined (participles, and partially half participles).

Grammatical categories are typically expressed by inflectional or derivational affixes: prefix (e.g. nu-nes-ti), suffix (e.g. nam-eli, neš-dav-o), flexion (e.g. vaik-o, neš-g), and reflexive affix (e.g. nu-si-neš-a, neš-a-si). The distribution of grammatical categories across different parts of speech is presented in Table 1. Syntactic and semantic categories (transitivity and aspect) are not included in the table.

As it is seen in Table 1, Lithuanian grammatical categories are mostly expressed by inflections. Suffixes for nominal words are only used for adjectives to express the degree. There is quite a different situation with verbs, as suffixes are used for degree, tense, voice, and mood.

Lithuanian is a flectional language, where grammatical relations between words are expressed by inflections that may have multiple meanings. For example, the inflection -uose in the word ger-uose ‘good’ indicates the masculine gender, the plural number, and the Locative case. Hence, the majority of words consists of both the root and the inflection. An inflection may also function as a derivational affix, e.g. stal-ius ‘carpenter’ (derived from stal-as ‘table’), vilk-ė ‘she-wolf’ (derived from vilk-as ‘wolf’). Due to this reason, we do not use the term primary word, as the morphemic structure does not always distinguish between deri-

1 Morphemes are separated by hashes in all examples of the paper, the roots are in bold, and the analysed affixes are underlined. Grammatical categories are only given, if they are not clear from the translation. Translation of each morpheme is not seen as a viable solution, as not all Lithuanian morphemes are translatable into English.

*Corresponding author: Andrius Utka, Department of the Lithuanian Language, Centre of Computational Linguistics, Vytautas Magnus University, Kaunas, LT-44244, Lithuania, E-mail: andrius.utka@vdu.lt
Erika Rimkutė, Asta Kazlauskienė, Department of the Lithuanian Language, Centre of Computational Linguistics, Vytautas Magnus University, Kaunas, LT-44244, Lithuania

© 2016 Erika Rimkutė et al., published by De Gruyter Open. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.
ative and primary words, e.g. *nam-as* ‘house’ and *stal-ius* ‘carpenter’ have the same structure (root + flexion); however the former is primary, while the latter is a derivative word. We are using a more neutral term of *simple word* for words that do not have derivational affixes.

### Table 1. Grammatical categories across different parts of speech

| Part of speech / Grammatical categories | Noun | Adjective | Pronoun | Numeral | Declined verb forms | Final verb forms |
|----------------------------------------|------|-----------|---------|---------|---------------------|-----------------|
| Gender                                 | F²   | F         | F       | F       | F/F                | -               |
| Number                                 | F    | F         | F       | F       | F/F                | F               |
| Case                                   | F    | F         | F       | F       | S/F                | S/F³            |
| Tense                                  | S    | S         | S       | S       | S/F                | -               |
| Voice                                  | S    | S         | S       | S       | S/F                | -               |
| Reflexivity                            | d⁴  | d         | d       | d       | d/F                | -               |
| Degree                                 | S²  | S         | S       | S       | S/F                | -               |
| Definiteness                           | S    | S         | S       | S       | S/F                | -               |
| Person                                 | S    | S         | S       | S       | S/F                | -               |
| Mood                                   | S    | S         | S       | S       | S/F                | -               |

Prefixes and suffixes typically play the role of derivational affixes in Lithuanian, and they are instrumental in creating new words. The Lithuanian language has many derivational affixes – almost 600 suffixes and around 60 prefixes⁵. They show the huge derivational potential of the language. In Lithuanian suffixes also occur in inflections. To this group belong forms of derivational verbs, for example, *skait-o* and *skait-ant-is* (a present tense participle form with the inflectional suffix -ant). Degrees of adjectives are also expressed by means of inflectional suffixes, e.g. *ger-as*, *ger-esn-is*, and *ger-iaus-ias* ‘good, better, the best’. It should be noted, that suffixes perform a single function: they are either derivational or inflectional, and inflectional suffixes always denote a single grammatical category. In this respect inflections differ from suffixes, as they can be both derivational and inflectional, and the latter may denote more than one grammatical category.

Prefixes are rarely used in derivation (most commonly they are used to form reflexive participles, e.g. *be-si-suk-ant-is* ‘turning’; the usage of prefix *be-* makes the reflexive marker -si- come into the front of the word, thus avoiding the formation of the wordform *suk-ant-is-is*, which would be hard to pronounce).

Expression of reflexivity by the means of affixes is not usual, as it is conveyed by a reflexive affix whose position depends on whether a word contains a prefix or not; thus it is at the end in a prefix-less word (e.g. *praus-ia-si* ‘washing himself’), while in a word with a prefix the position of the reflexive marker is between the prefix and the root (e.g. *ne-si-praus-ia* ‘not washing himself’).

Moreover, Lithuanian compounds can also be formed by combining two or more roots into a single word, e.g. *darb-dav-ys* ‘employer’ or joining them with a connective vowel, e.g. *darb-o-tvark-ė* ‘agenda’. New words may also be derived from derivatives, for example, *žem-uog-ėtiav-im-as* ‘strawberry picking’ is formed from the compound noun *žem-uog-ė* ‘strawberry’.

Thus summing up the functions of Lithuanian affixes, one should consider that inflections of nominal words and declined verbs denote gender, number and case categories, while inflections of verbs denote tense, number and person. As it has been mentioned previously, sometimes inflections are used as derivational affixes in forming new wordforms. Inflectional suffixes of nominal words denote degree for nominal words, and tense, mood and voice for verbs (these categories also remain in nouns that are derived from verbs). Due to the ambiguity of such affixes, only inflectional and derivative suffixes can easily be distingu-

---

2 Abbreviations of affixes are explained in the Abbreviations section below the paper.
3 Different root allomorphs may be used in some verb tenses, e.g. *ei-ti* ‘to go’ (infinitive form), *ein-a* ‘goes’ (present tense, 3rd person), and *ėj-o* ‘went’ (Past simple tense, 3rd person).
4 Reflexivity is characteristic only to the nouns that are derived from verbs.
5 Such a large number affixes is due to the fact that it includes international affixes as well as phonetically different affixes. Thus for instance suffixes -ėti and -ėtė have been considered as different suffixes.
ished. Thus, in section 4, which is about the morphemic structure of different parts of speech, only suffixes are classified as derivational and inflectional and no such classification is used for other affixes.

Having this in mind, one can assume that since the morphemic word structure\(^6\) in Lithuanian is so diverse it may not show any regularities or have characteristic models. This very assumption was, in fact, the reason for undertaking this research in order to establish certain regularities of morphemic word structure in the Lithuanian language.

The aim of this research is to define and describe morphemic structural models of inflective parts of speech (i.e. nouns, adjectives, numerals, pronouns, and verbs) and regularities of their usage in contemporary Lithuanian. The main works on morphemics are presented in Chapter 2 of this paper. In Chapter 3 we describe the analysed data and present problems that have occurred during the compilation of this data (morpheme boundary detection, boundaries of certain words, and an uncertain line between synchronic and diachronic aspects). The main results of this research are presented in Chapter 4, where morphemic structures of nouns, adjectives, pronouns, numerals, and verbs are presented. In Chapter 5 we compare morphemic structures of different parts of speech.

One should consider that morphemic and derivational analysis do not correspond, e.g. the morphemic analysis of the noun *parašas* ‘signature’ splits the wordform into prefix+root+flexion (*pa*-ra-šas), while the derivational analysis into stem+flextion (*paraš-as*). The derivational analysis (simple and derivative words) is beyond the scope of this paper, and we will only briefly refer to some of its aspects in the paper.

Readers should note that this is one of the first works for the Lithuanian language, which presents such an extended analysis of morphemic structural models and their frequencies of usage that is based on empirical data. We compare morphemic structures of different parts of speech, we assess their level of complexity, regularity, and we highlight differences. The research is also useful for showing the complexity of Lithuanian words and multi-functionality of Lithuanian affixes. We also believe that these results will be useful in comparative studies.

### 2 Previous Research in Morphemics

The morphemic structure of Lithuanian words was analysed by Kuosienė (1986) and Keinys (2009), however, their research does not employ a quantitative data analysis. Typically, Lithuanian grammars (e.g. DLKG 1996) present only general information about morphemes. The morphemic structure of Lithuanian verbs was recently analysed by Rimkutė et al. (2011).

Other works related to Lithuanian morphemics deal with phonological structure of morphemes (Kruopienė 2000a, Kruopienė 2000b, Kruopienė 2001, Kruopienė 2002, Kruopienė 2005; Kaukienė 1994, Kaukienė 2002, Karosienė 2004; Akelaitienė 1996, Akelaitienė 2001, Akelaitienė 2008) but, as is the case regarding morphemic structure, none of them are based on a quantitative analysis. In view of this, the present research is important not only for Lithuanian scholars but also for researchers of other languages who analyse morphemic word structures. It should be noted that this research is based on rich electronic data of contemporary Lithuanian; therefore, it is believed that at least main regularities in morphemic word structure may be observed and reliable conclusions may be drawn concerning regularities of morphemic structure.

Comprehensive information about morphemic structure of other related languages, i.e. Latvian, Russian, and Czech, was not found. The existing research is limited to presenting what sort of morphemes may form words in the respective languages (see, e.g., Kalnača 2004; Sarkans 1996 and Levāne-Petrova et al. 2000 for Latvian; Polikarpov 2000 for Russian; Sedláček 2004 for Czech).

---

\(^6\) Morphotactics deals with regularities of morphemic composition, while morphemics deals with splitting of word forms into morphemes.
3 Data and Methods

The analysed data consist of samples from the Corpus of the Contemporary Lithuanian Language (http://tekstynas.vdu.lt), which covers a wide variety of texts: scientific publications, periodical texts, fiction, documents (laws, decisions, protocols, statutes). Each functional style is represented by samples that contain an equal number of words. Fragments from The Database of Spoken Language (http://donelaitis.vdu.lt/garsynas) have also been used. The size of the whole experimental corpus is 310,000 words. Texts have been semi-automatically morphologically annotated. The morphological tagger has identified the part of speech, gender, number, case, person, tense, etc.; it has also disambiguated morphologically ambiguous words (e.g. the word vidutiniškai ‘on the average’ may be an adverb or an adjective, while nužudyti ‘to kill’ may be an infinitive or a participle in the passive voice). However, it has ignored morphological multiword units and has not identified grammatical categories for unrecognized words (such as proper names, shortened word forms, new words). Some mistakes of the morphological tagger have been corrected, and missing information for unrecognized words have been added manually.

Fragments of natural written language (not dictionaries) have been chosen for the analysis of Lithuanian morphemic structure. There are several reasons that justify this choice. First of all, there is no good contemporary dictionary of standard Lithuanian. The content of the existing Modern Lithuanian Dictionary (DLKŽ 2006) does not match its name, as it includes many obsolete, dialectic, rare, and jargon words whose morphemic structure is often difficult to define.

The next argument concerns the very structure of dictionaries: they only present headwords in spite of the fact that the Lithuanian language has a very elaborate inflectional system. For example, the verb skaiti ‘read’ is presented by three main forms skait-y-ti, skait-o, skait-e, while verb forms that are derived from them (e.g. skait-ant-is (participle), skait-ant (adverbial participle), skait-y-dav-o (past frequentative tense), skait-y-dam-as (half participle) by adding inflectional suffixes are not included. In addition, dictionaries often do not include many derivational words that are very important for the morphemic analysis, such as prefix derivatives (te-be-skait-o ‘still reading’, ne-skait-o ‘not reading’) or diminutive words (e.g. stal-iuk-as ‘a little table’, sūn-el-is ‘a little son’).

Even though dictionary-based research could be interesting as well, it will not concern us here. Many problems have arisen while trying to establish boundaries of morphemes. In this study, theoretical principles for morpheme boundary detection are formulated following Urbutis (2009, 165-166). A part of a word is considered to be a morpheme if both (or all) parts of a word (potential morphemes) occur in other words (e.g. the root in nam-el-is ‘a little house’ also occurs in nam-as ‘a house’ and nam-in-is ‘domestic’, while the suffix is found in such words as stal-g-is ‘a little table’ or vaik-el-is ‘a little child’). Unfortunately, these assumptions do not solve all problems that may occur while analysing the large volume of empirical data. The problems include the following: an overlap of morphemes, the boundary detection according to synchronic or diachronic view, as well as boundary problems within international words.

Words are decomposed into as many morphemes (the smallest meaningful parts of a word) as it is allowed by the contemporary Lithuanian language. However, in some cases the analysis of certain words had to include the diachronic aspect. For example, adjectives gyvas ‘alive’, pilnas ‘full’ are semantically associated with the verbs gy-ti ‘to heal’ and pil-ti ‘to pour’, however contemporary speakers do not associate these words any more. Some researchers present these adjectives as suffix derivatives (Pakerys 1994, 356), while others as inflectional derivatives (DLKG 1996, 223). In the cases, when it is not clear whether morphemes can be separated according to the synchronic view, we present the boundaries in parentheses, e.g., gy(-)v-as ‘alive’.

Inasmuch as our research is synchronic, we treat pronominal inflections as one morpheme, although historically these forms were formed, while the pronouns jis, ji ‘he, she’ were attached to the simple inflection, e.g. gero+jo=gerojo ‘good’.

The boundaries of morphemes constituting words in the Lithuanian language are not always clear; various phonological processes take place at a morphemic juncture: either front or back consonants are
adjusted to the adjacent consonants of the other morpheme, sometimes morphemes overlap, then sounds are skipped or contracted. The reasons of such variation are phonological (degemination, dissimilation, contraction, when geminates that are not characteristic to the Lithuanian language are eliminated) and phonetic (metathesis and elimination, when sounds are switched or omitted to make pronunciation easier). The classification of these processes varies in Lithuanian linguistics (Kazlauskienė 2012). Some authors (DLKG 1996) classify these processes as contraction, elimination, and degemination. Other linguists do not classify these processes and identify them by the term “sound omission” (Urbutis 2009). The described individual cases have only marginally influenced our analysis, as in our work the biggest attention has been paid to the distribution of morphemes and their number and not to the problems of morpheme boundary detection, which would require a separate study.

It is a complicated matter to detect morpheme boundaries within international words, as commonly a borrowed word is being adapted to the Lithuanian language and it acquires grammatical features of the part of speech, e.g. the verb bomb-ard-uo-ti ‘to bombard’ is borrowed from the French word bombarder, however in the Lithuanian language there is also the noun bomb-a ‘a bomb’. If international words were broken into smaller parts, it would not contradict with regularities of morphosyntaxactics, and it would hardly have influence on the variability of models. However, the further analysis of morphemic structure and allomorphs would become more complicated, as the morphemic inventory would be supplemented with uncharacteristic morphemes for the Lithuanian language, as in the case of bomb-ard-uo-ti, we would need to distinguish the foreign suffix -ard-.

We understand that we still have left some open issues when establishing morphemic boundaries in our analysed data. The reason for this is the fact that due to the lack of theoretical discussion on morphemic boundaries and on the difference between morphemic and derivational analyses some of our choices may be debatable. Nevertheless, individual interpretations cannot change the main regularities of Lithuanian morphemic structure.

All the problematic cases mentioned above are treated consistently, and in the future they may be reviewed. Certainly not all difficulties of morphemic analysis have been mentioned, as first of all we would like to focus on regularities of morphemic composition in nominal and verbal words, and therefore we do not deal with other known problems of morphemic analysis, e.g. whether or not all words can be split into morphemes, whether or not some morphemes may be treated as such according to the synchronic point of view, and etc.

The initial marking of morphemic boundaries was performed manually by using special symbolic notifications (see examples below, e.g. the word rodydavosi ‘appeared’ was marked as ro9d!y-dav=o+si (where the exclamation mark separates the root rod- from the derivational suffix -y-, the hash separates the inflectional suffix -dav-, the equals sign separates the inflection, the number 9 denotes the acute (falling pitch) accent that falls onto the sound o, and the plus sign separates the reflexive affix).}

| nor!ė9-t=ų+si | laik!y9-k=itė+s | "euro*info*cen3tr=ų | *siaur#a*a3k=is |
|---|---|---|---|
| *si!p=n-ia9us=iq | me9kl!y9-k=kite+s | po9string!avl=im=ai | Su*ďe*ling-e4sn=io |
| vai!ši4n-dav=o+si | *už*dar/oji | at!@pa*la*da3v!lm=im=0 | ti<n-k-aml!au3 |
| il!š!ė9-k=itė+s | *#ą*syk | brolle3l=is | *#ą3*a3k=is |
| ty9c!l!o10-dav=o+si | *śim3t#ą*kart | dźiau3g!sl=ma=ą | *śim3t#ą*bal3l=is |
| terl!l!o9-t=umėmė+s | su*būr!i4m=o | iš+si!*dé9st!t!ym=ą | Su*būr!i4m=o |
| te@su4*geb=a | *t#ą3*syk | la9id!o!tuv=ėse | *t#ą3*syk |
| ap*žiūr!line9-dam=as | pa*da*ly9-t=omis | su+si*būrli4m=0 | *t#ą3*syk |
| *euro*info*cen3tr=ų | *śim3t#ą*kart | *siaur#a*a3k=is | *t#ą3*syk |

* In other examples 3 denotes the circumflex (rising pitch) accent, and 4 – the grave accent.
A special tool was developed, which can process a corpus with such notations\(^9\). Based on these notations the tool efficiently identifies different morphemic models and their quantitative distribution within the whole data set (see examples above).

*The Database of Lithuanian Morphemics Data*\(^{10}\) has been compiled and it is freely available online. Users can search for wordforms and morphemes in the database, where he/she can find wordforms with marked morphemic boundaries (the type of morphemes is not given yet). The database is important that it shows the big morphemic variation in the Lithuanian language.

The data in the *The Database of Lithuanian Morphemics Data* can be useful for improving applications, for instance the data may improve the automatic morphological analysis by using information about potential affixes of a given word (currently the Lithuanian morphological analyser expects many non-existing wordforms). The data could also be used for the lemmatisation of multi-word units (see Boizou et al. 2015). Besides, the affix information could be used in the language teaching, as learners could be taught how to form new words and which derivative or grammatical senses result from adding a certain affix.

### 4 Nominal and Verbal Models of Morphemic Structure

This chapter discusses the most typical examples and the most productive models of morphemic structure as found in Lithuanian nominal words (nouns, adjectives, pronouns, and numerals) and verbal items. The results are based on the 310,000 word corpus that has been described in the previous chapter. Although a corpus of different size and composition would give slightly different results, we do not think that the difference would be essential.

Figure 1 presents the quantitative information regarding structural models as identified for different parts of speech. It is obvious that parts of speech could be divided into three groups in terms of diversity of morphemic structures: verbal items belong to the *high diversity group*, nouns and adjectives to *medium diversity group*, and pronouns and numerals to *low diversity group*. Nouns and adjectives as well as pronouns and numerals have equal number of structural models in the present research.

![Figure 1. Number of Structural Models for Different Parts of Speech](image)

---

9 The tool was developed by Gailius Raškinis, Centre of Computational Linguistics, at the Vytautas Magnus University.

10 see http://tekstynas.vdu.lt/page.xhtml?id=morfema-db
4.1 Nouns

The analysed material comprises 142,086 noun tokens, covering 11,799 different word forms. All the nouns can be generalised as representing 63 structural models.

Nouns may have from one to nine morphemes. Bi-morphemic nouns are the most frequent while tri-morphemic and quadri-morphemic nouns are somewhat less frequent. Nouns with 2-4 morphemes make up 96% of all analysed nouns (see Figure 2).

![Figure 2. Percentage of Nouns with a Different Number of Morphemes](image)

**Uni-morphemic** nouns are uninflected international words, e.g. *meniu, ego, interviu* (R model11, 0.1% of all nouns).

**Bi-morphemic** nouns dominate and they make up almost half of all noun usage cases (46.5%). These are mostly nouns that have a root and an inflection forming the most frequent noun model RF, e.g. *darb-o* ‘work’ (Nmsg12), *žem-ės* ‘land’ (Nfsg). Endlingless Vocative case also belong to this group (the R$S_d$ model), e.g. *tėt-uk* ‘daddy’ (Nmsv), *mam-ut* ‘mummy’ (Nfsv). This model is very rare and it is only used in the spontaneous speech.

**Tri-morphemic** nouns make up one third of all nouns. The structure of this group of nouns can be generalised into 5 models. The RS$S_F$ represents the second most frequent noun model (here belong even 24.6% of all nouns, e.g. *sveik-at-os* ‘health’ (Nfsg), *gam-yb-os* ‘production’ (Nfsg)). The PRF model with the 7.5% frequency is the third most frequent noun model, e.g., *į-mon-ės* ‘enterprise’ (Nfsg/Nfpn13), *ap-saug-os* ‘security’ (Nfsg) while the RRF model (e.g. *darb-dav-ių* ‘employers’ (Nmpg/Nfpg), *čia-buv-io* ‘indigene’ (Nmsg)) make up 1.6%. Two less frequent models are RS$_F$ (0.9%, e.g., *vand-en-s* ‘water’ (Nmsg) and RS$_S_d$ (0.004%, e.g., *vaik-el-iuk* ‘a little child’, *graž-uol-iuk* ‘cute’ (both Nmsv)).

**Quadri-morphemic** nouns make up 14% of all analysed nouns; they can be generalised into 15 structural models. The RS$_S_F$ model constitutes half of all quadri-morphemic nouns (7.5%, the fourth most frequent noun model, e.g. *darb-uo-toj-ų* ‘employers’, *vart-o-toj-ų* ‘users’ (both Nmpg)). Other more frequent models are PRS$_F$ (4.2%, e.g. *j-stat-yms* ‘laws’ (Nmpg), *su-tar-ties* ‘agreement’ (Nfsg)), RcRF

---

11 See abbreviation list for description of abbreviations at the end of the paper.
12 Explanation of grammatical codes for nouns: 1st position (part of speech): N – noun; 2nd position (gender): m – masculine, f – feminine; 3rd position (number): s – singular, p – plural; 4th position (case): n – Nominative, g – Genitive, d – Dative, a – Accusative, i – Instrumental, l – Locative, v – Vocative).
13 Typically the forms in feminine gender that are in singular Nominative coincide with the forms of singular Instrumentative, and singular Genitive with plural Nominative. Similar instances of case syncretism could be found in other paradigms too. Such instances of case syncretism are separated by ‘/’ mark.
Morphemic Structure of Lithuanian Words

It should be noted that when the noun has two or more derivational suffixes, then only the last one (the one that precedes the inflection) is the derivational suffix of a given word, while the others belong to the stem, e.g. the noun *darb-u-o-toj-as* ‘employer’ that is derived from the verb *darb-u-o-tis* ‘working’, which is derived from the noun *darb-as* ‘work’.

Penta-morphemic nouns (4% of all analysed nouns) have the largest variety of models – 23. However, only three of them are used relatively frequently: PRS_{d}S_{d}F (1.2%, e.g. *ap-dor-oj-im-o* ‘processing’ (Nmsg), *par-ei-g-ūn-ai* ‘officers’ (Nmpn/Nmpv)), RS_{d}S_{d}F (1.1%, e.g. *reik-al-av-im-us* ‘requirements’ (Nmpa)), and PdRS_{d}F (0.6%, e.g. *pa-si-rink-im-o* ‘choice’ (Nmsg)).

Six-morphemic nouns make up only 0.6% (21 models), but their usage frequencies are low. In fact, many occurrences of this model are unique. Three of the most frequent are PdRS_{d}S_{d}F (0.2%, e.g. *pa-si-tik-ėj-im-o* ‘confidence’ (Nmsg)), PRS_{d}S_{d}F (0.2%, e.g. *ap-dov-an-oj-im-as* ‘award’ (Nmsn)), and RRS_{d}F (0.1%, e.g. *bendr-a-darb-iaj-um-o* ‘cooperation’ (Nmsn)).

Hepta-morphemic nouns fall into 9 models (0.05% of all nouns). The type of nouns in them are verbal abstracts derived from suffixal verbs, e.g. the PdRS_{d}S_{d}F model *at-si-stat-y-din-im-o* ‘resignation’ (Nmsg) and the PPdRS_{d}F model, *ne-pa-si-tik-ėj-im-as* ‘distrust’ (Nmsn).

Only two octa-morphemic nouns (0.02%) have been found in the data, the PdPRS_{d}S_{d}F model, as in *j-si-par-ei-g-oj-im-as* ‘commitment’ (Nmsn) and the PRS_{d}RS_{d}F model, e.g. *ne-pil( )n-a-vert-įšk-um-o* ‘inferiority’ (Nmsg).

One nona-morphemic word that has 7 roots was also found: *fibro-ezo-fago-gastro-duo-deno-skop-i-j-a* ‘fiber esophagogastroduodenoscopy’ (Nfsn/Nfsi) (the RRRRRRRS_{d}F model).

Simple nouns RF are the most numerous (approx. 47% of all nouns). However, they are not all simple, e.g. *gėr-is* is the inflectional derivative that is derived from the adjective *ger-as* ‘good’. A large part of the nouns have derivational suffixes (approx. 42%); the number of derivational suffixes in these nouns range from one to four. More than one tenth of nouns (14%) have prefixes (from one to three). These nouns may also contain many derivatives of inflectional morphemes, and thus the number of their models is quite big (39 models). Only 4% of all nouns have two roots. In spite of such small percentage, their morphemic structure is very diverse (38 models). Note that if a noun has more than two roots, it must be a borrowing.

Although our database contains 63 different morphemic structural models for nouns, the most productive 5 models account for 91% of all the noun forms. The models are the following: RF (46.8%; *darb-as* ‘work’), RS_{d}F (26.6%, *sveik-at-a* ‘health’), PRF (7.5%, *j-mon-ė* ‘enterprise’), RS_{d}S_{d}F (7.5%, *darb-u-o-toj-as* ‘worker’), PRS_{d}F (4.2%, *j-stat-ym-as* ‘law’). The generalised formula for all the most frequent nouns is $P_{0}R_{0}S_{0}F$.

### 4.2 Adjectives

The analysed material takes account of 30,097 adjectives (word tokens); 11,799 of them are different wordforms, which can be generalised as representing 63 structural models.

Adjectives may have from one to seven morphemes. Adjectives of 2-4 morphemes, just like nouns, make up even 95% of all cases. However, in the case of adjectives tri-morphemic wordforms are the most frequent ones (bi-morphemic for nouns) (see Figure 3).
Uni-morphemic adjectives as nouns are uninflected international words, e.g. *makro, mikro, mezo* (the R model, 0.02% of all adjectives).

Bi-morphemic adjectives are of two types: adjectives with a root and an inflection (the RF model is the second most frequent model of adjectives, e.g. *aisk-u‘clear*, *sunk-u‘heavy* (both neutral gender)).

Tri-morphemic adjectives are represented by the following models: the RS_F model (the most frequent one (38.2%), e.g. *did-el-is‘big* (Amsn^{14}), *social-in-ės‘social* (Afsn/Afpg); PRF (5.7%, e.g. *at-skir-u‘separate* (Ampg/Afsp), *pa-naš-us‘alike* (Amsn)); PS,F (5.6%, e.g. *svarb-iaus-ia‘the most important* (Afsn/Afspi/An), *did-esn-is‘bigger* (Amsnc)) and the very rare model RRF (0.3%, e.g. *pus-kieč-ų‘semisolid* (Ampg/Afsp)).

Quadri-morphemic adjectives make up one fifth of all analysed adjectives; they fall into 11 structural models. Almost half of them belong to the RS_SdF model (8.5%, e.g. *reik-al-ing-a‘necessary* (Afsn/Afspi/An)). The PRS_d,F model is rarer (5.9%, e.g. *pa-grind-in-io‘basic* (Amsn), *po-zem-in-io‘underground* (Amsg)). This group includes adjectives with two roots or two affixes: the RRS_dF model (2.5%, e.g. *geo-log-in-ių‘geologic* (Ampg/Afspg)); RS_S,F (1.4%, e.g. *auk-št-ėsn-ė‘higher* (Afsnc), *vis-ok-iaus-įų‘all sorts’ (Ampgs/Afspg)); RcRF (1.1%, e.g. *ilg-a-laik-įų‘long term‘ (Amsn)); RS,F_S,F (0.5%, e.g. *rud-en-in-ių‘autumnal‘ (Ampn)); PPRS (0.4%, e.g. *pra-naš-ėsn-įų‘superior’ (Amsn)); RRFS (0.2%, e.g. *ne-pa-jėg-ųs‘unable’ (Ampn)); RS,S,F (0.05%, e.g. *myl-ims-iaus-ia‘most loved’ (Amscs/Afscp)); RS,F (0.02%, e.g. *raud-on-pus-įų‘red-sided‘ (Amsn)); and PPRF (0.01%, e.g. *ne-to-lyg-aus‘uneven‘ (Amsg)).

Penta-morphemic adjectives comprise 4% of all analysed adjectives and are prominent because they consist of the largest variety of models (24 models). However, only three adjective models occur more frequently: the PRS_dS_d,F model (1.3%, e.g. *ne-reik-al-ing-as‘unnecessary‘ (Amsn/Afpa)) and the RRS_dS_dF model (1.1%, e.g. *gyv-yb-in-io‘vital‘ (Amsg)); and RsRS_d,F (0.7%, e.g. *ak-i-vaiz-d-ų‘obvious‘ (Ams)). The rest of adjectives of five morphemes make up only 1.2%.

Hexa-morphemic adjectives make up just 0.4% (15 models) and their usage frequencies are very low. In fact, many occurrences of this model are solitary. Somewhat more frequent are the following models: PRS_dS_dS_d,F (0.1%, e.g. *su-dė-t-ing-ėsn-įų‘more difficult‘ (Amsnc)); RRs_dS_dF (0.05%, e.g. *dv-i-pra-sm-išk-a‘equivocal‘ (Afsn/Afspi/An)); and PRS_d,S_dF (0.04%, e.g. *ap-skri-t-im-in-ė‘oblique‘ (Afsn)).

Only 8 hepta-morphemic adjectives have been found (0.03%): *ne-ak-i-vaiz-d-in-ėmis,‘extramural* (Afpi), *angl-ia-vand-en-il-in-ių‘hydrocarbon‘ (Amsi). All these comprise two roots, a connective vowel, two derivational suffixes, and either a prefix or an inflectional suffix.

---

14 Explanation of grammatical codes for adjectives: 1st position (part of speech): A – adjective; 2nd position (gender): m – masculine, f – feminine; 3rd position (number): s – singular, p – plural; 4th position (case): n – Nominative, g – Genitive, d – Dative, a – Accusative, i – Instrumental, l – Locative, v – Vocative; 5th position (degree): c – comparative, s – superlative.
In Lithuanian, adjectives are the most diverse part of speech in terms of inflectional forms (hence many morphemic structural models have been identified), as adjectives are inflected for number (e.g. *graž-us* [singular] – *graž-ūs* [plural]); gender (e.g. *graž-us* [masculine] – *graž-i* [feminine]), degree (*graž-us, graž-ėsn-is, graž-iaus-ias* ‘nice, nicer, nicest’) and definite forms (*graž-usis, graž-įjoji* ‘the nice one’). Only international adjectives are without grammatical inflections, e.g. *makro, mikro, mezo*.

The overall usage of adjectives is dominated by adjectives that have derivational suffixes (approx. 62%). More than one tenth of adjectives (14%) have prefixes. Compound adjectives are rare. Only 5% of all adjectives have two roots.

Summing up the adjective analysis, it can be observed that from 63 morphemic structural models for adjectives the most productive 6 models account for 89% of all adjective wordforms. The models are the following: RSdF (38.2%, *did-ėl-is* ‘big’), RF (24.9%, *aišk-us* ‘clear’), RSs,F (8.5%, *reik-al-ing-as* ‘necessary’), PRs,F (5.9%, *pa-grind-in-is* ‘main’), PRF (5.7%, *at-skir-as* ‘different’), RS,F (5.5%, *svarb-iaus-ias* ‘most important’). The first five models are identical to noun models and thus the generalised formula is also identical – $P_{0,1}$RS$_{0,2}$F.

### 4.3 Pronouns

There are 39,738 pronouns in the analysed material; they include 734 different word forms. All pronouns fall into 10 structural models. Pronouns may have from one to five morphemes. Most frequent are bi-morphemic pronouns (see Figure 4).

![Figure 4. Percentage of Pronouns with a Different Number of Morphemes](image)

The most frequent **uni-morphemic** pronoun is aš ‘I’ (P0sn15), which belongs to the R model.

**Bi-morphemic** pronouns have the RF model, e.g. *jis* ‘he’ (Pmsn), *kien-o* ‘whose’ (P00g).

**Tri-morphemic** pronouns belong to 3 models: RS,F (8.3%, e.g. *t-oks* (Pmsn), *k-oks* ‘what’ (Pmsn)), RCs (0.02%, e.g. *kel-ias-dešimt* ‘several tens’ (P)), and RRF (3.2%, e.g. *kiek-vien-as* ‘each’ (Pmsn/Pfpa)).

**Quadri-morphemic** pronouns also belong to 3 structural models: RCsRF (1.2%, e.g. *š-its-as* ‘this’ (Pmsn/Pfpa)), RRRF (0.9%, *ka(-)ž-kien-o* ‘whose’ (P00g)), and a single pronoun with a prefix *nė-mūs*–*išk-ė* ‘not ours’ (Pfns).

15 Explanation of grammatical codes for pronouns: 1st position (part of speech): P – pronoun; 2nd position (gender): m – masculine, f – feminine, n – neutrum; 3rd position (number): s – singular, p – plural; 4th position (case): n – Nominative, g – Genitive, d – Dative, a – Accusative, i – Instrumental, l – Locative, v – Vocative; 5th position (degree): c – comparative, s – superlative. The uninflected pronouns are marked by the single letter P. 0 can occur in all positions, when the category is non applicable.
The longest **penta-morphemic** pronouns belong to 2 structural models: $\text{RRS}_F \text{ ka(-)į-k-o} \text{-š-o} \text{k}$ (0.42% of all analysed pronouns) and $\text{i-š-t-o} \text{k-s} \text{ ‘such’}$ belongs to the $\text{RcRS}_F \text{ model}$ (0.09% of all analysed pronouns).

Two models ($\text{RF}$ and $\text{RS}_F$) make the core of pronoun models (more than 90%). The overall usage of pronouns is dominated by **primary** pronouns (approx. 87%). The diversity of structural models is not possible in this case. 8% of pronouns are suffixal derivatives and belong to the structural model $\text{RS}_F$; only 5% of all pronouns have **two roots**. Connective morphemes, inflections, and suffixes determine their diverse morphemic structure: $\text{RR, RRF, RRR, RR, RcRS}_F, \text{ RRRF}_F$. Pronouns with prefixes are very rare.

In sum, the most typical morphemic features of Lithuanian pronouns are as follows: they are primary, bi-morphemic, and belong to the $\text{RF}$ model.

### 4.4 Numerals

The analysed material contains 5,008 numerals (word tokens) of 513 different word forms. All numerals may be generalised into 10 structural models.

Numerals have from one to five morphemes. As in the case of pronouns, the most frequent are bi-morphemic numerals; however, tri-morphemic ones are also numerous (see Figure 5).

**Figure 5.** Percentage of Numerals with a Different Number of Morphemes

There are only two **uni-morphemic** numerals $\text{du ‘two’}$ and $\text{dešimt ‘ten’}$, but they are rather frequent (8.4% of all numerals).

Two thirds of all numerals are **bi-morphemic**. Most of bi-morphemic numerals are simple and definite, consisting of a root and an inflection, and belong to the $\text{RF}$ model (66.4%, e.g. $\text{dv-i ‘two’}$ (Mcfn17), $\text{tr-ys ‘three’}$ (Mcm0n)).

16 Lithuanian numerals are classified into 1) cardinal (main, collective, multiple, fractional), and 2) ordinal. Different numerals may have very different morphemic structures and sets of grammatical categories (e.g. some are not inflected (e.g. dešimt ‘ten’), the others are declined similarly do adjectives (e.g. antrasis ‘the second one’)). The section contains generalised information about morphemic structure of numerals without taking into account all the different types of numerals.

17 Explanation of grammatical codes for numerals: 1st position (part of speech): M – numeral; 2nd position (type): c – cardinal, o – ordinal; 3rd position (gender): m – masculine, f – feminine, n – neutrum; 4th position (number): s – singular, p – plural; 5th position (case): n – Nominative, g – Genitive, d – Dative, a – Accusative, i – Instrumental, l – Locative, v – Vocabative; position (degree): c – comparative, s – superlative. The uninflected numerals are marked by Mc. 0 can occur in all positions, when the category is non applicable.
Tri-morphemic numerals may belong to three structural models: RS_F (12%, e.g. tr-eč-ias ‘third’ (Momsn/Mofpa), dv-ej-us ‘two’ (Mcm0a)), RcR (6.8%, e.g. aštuon-ias dešimt ‘eighty’ (Mc)), and RRF (0.5%, e.g. pus-antr-o ‘one and a half’ (Mcmsg)).

Quadri-morphemic numerals can be classified into three models: RcRF (3.6%, e.g. penk-io-lik-a ‘fifteen’ (Mc00n/Mc00a/Mc00i)), RSdSdF (0.2%, e.g. tr-ej-et-ą ‘threesome’ (Mc00a)), and RRSdF (0.1%, e.g. pus-tr-eč-io ‘two and a half’ (Mcmsg)).

Penta-morphemic numerals are ordinal numerals showing the sequence RcRsF (2.1%, e.g. dv-y-lik-tą ‘twelfth’ (Momsa/Mofsa)).

The numeral usage is dominated by primary numerals (approx. 74%, two structural models, R and RF). 13% of the numerals are suffixal derivatives and belong to the structural models RSdF, RSiF, and RSdSdF. 13% of all numerals are compounds; they show the structural models RcR, RRF, RcRF, RRSdF, and RcRsF. Numerals do not have prefixes.

The data analysis has revealed that most typical numerals in the Lithuanian language are primary ones and suffixal derivatives with 1-3 morphemes. Out of 10 morphemic structural models for numerals, the most productive 3 models account for 87% of all wordforms. The three models are the following: RF (66.4%, e.g. vient-ąs ‘one’), RS_F (12%, e.g. tr-eč-ias ‘third’), R (8.4%, du ‘two’). The generalised formula for numerals is RS_d0 -1F0 -1. They differ from pronouns by the rather high frequency of occurrence of R model.

4.5 Verbs

The analysed material comprises 92,468 verbs (word tokens) which include 29,878 different word forms. Their morphemic structures are very diverse: all the verbs can be classified into 116 structural models. However, their frequencies of occurrence are immensely different (from just 1 occurrence up to 18 thousand).

Verbs may have from one (e.g. reik ‘must’) up to nine morphemes (e.g. ne-į-si-par-ei-g-o-k-ite ‘do not take commitment’). Uni-morphemic verbs and verbs longer than five morphemes are very rare. Thus the usage core (89%) is formed from verbs consisting of 2-4 morphemes (see Figure 6).

Uni-morphemic verbs are very infrequent. Just one occurrence has been found in the data: it is the shortened verb reik ‘must’ (main form, indicative mood, present tense, 3 person), which has been used in spontaneous speech (the R model, 0.02% of all verbs).

Bi-morphemic verbs make up one fourth of all analysed verbs. The diversity of structural models cannot be very big. The following three bi-morphemic models have been found: RF (19.9%, conjugated forms and participles, e.g. buv-o ‘was’ (main form, indicative mood, past tense, 3 person), RS (5.5%, e.g. bu-s ‘will be’ (main form, indicative mood, future tense, 3 person), ei-k ‘walk’ (main form, imperative mood, singular, 2 person), dirb-ti ‘to work’, es-ant ‘being’ (adverbial participle, present tense)), and PR (ne-reik ‘must not’ (main form, imperative mood, present tense, 3 person)). It should be noted that verbs do not have inflections only when their models contain inflectional suffix.

Tri-morphemic verbs (37% of all verbs) may be formed from various elements: roots, derivational suffixes, inflectional suffixes, inflections, and reflexive affixes. There are nine tri-morphemic models: PRF (13%, conjugated forms and participles, e.g. at-rod-o ‘appear’ (main form, indicative mood, present tense, 3 person), su-stij-ęs ‘related’ (participle, active voice, past tense, masculine, singular, Nominative)); RS_F (7.8%, forms with inflections, e.g. bū-tų, bū-dav-o, bū-dam-as ‘to be forms’ (the first form – main form, subjunctive mood, 3 person; the second form – main form, indicative mood, past frequentative tense, 3 person; the third form – half participle, masculine, singular)); PRS (5.3%, forms without inflections, e.g. at-ei-s ‘will come’ (main form, indicative mood, future tense, 3 person), at-lik-tı ‘to perform’, su-dar-ant ‘forming’ (adverbial participle, present tense)); RS_F (4.6%, conjugated forms and participles, e.g. gal-øj-o, gal-øj-ęs ‘could forms’ (the first form – main form, indicative mood, past tense, 3 person; the second form – participle, active voice, past tense, masculine, singular, Nominative)); RS_S (4.5%, mostly infinitives, e.g. dar-y-ti ‘to do’, but there are some participles and conjugated forms, e.g. ūn-o-k ‘know’ (main form, imperative

18 Abbreviations are not used due to the large variety of verb forms.
mood, singular, 2 person), reik-ė-s ‘will need’ (main form, indicative mood, future tense, 3 person), naud-oj-ant ‘using’ (adverbial participle, present tense)); RFd (1.4%, conjugated forms and several participles, e.g. skir-ia-si ‘differ’ (main form, indicative mood, present tense, 3 person), lank-ė-s-is ‘visiting’ (participle, active voice, past tense, masculine, singular, Nominative)); RSd (0.6%, infinitives, one conjugated verb and one participle, e.g. elg-ti-s ‘to behave’, baig-s-is ‘will finish’ (main form, indicative mood, future tense, 3 person), rem-iant-is ‘referring’ (adverbial participle, present tense)), and some standalone cases, such as RRS_s (kone-veik-ti ‘to scold’), and RS_S (bū-s-iant ‘be form’ (adverbial participle, future tense)).

Figure 6. Percentage of Verbs with a Different Number of Morphemes

Quadri-morphemic nouns make up 26% of all verbs (16 models). The four most frequent quadri-morphemic models are PRS_F (7.2%, e.g. ne-gal-im-a ‘not allowed’ (participle, passive voice, present tense, neutral)), RS_S_F (4.7%, mostly participles, naud-oj-am-i ‘used’ (participle, passive voice, present tense, masculine, plural, Nominative)), PRS_S_F (3.8%, mostly infinitives, e.g. nu-stat-y-ti ‘to define’), PdRF (3.2%, e.g. at-si-rad-o ‘appeared’ (main form, indicative mood, past tense, 3 person)), PRS_F (2.6%, e.g. ne-gal-ėj-o ‘could not’ (main form, indicative mood, past tense, 3 person)). Other models are not so frequent: PdRS (1.2%, e.g. pa-si-rink-ti ‘to choose’, at-si-žvelg-iant ‘considering’ (adverbial participle, present tense)), PPR_F (1.1%, e.g. ne-su-prat-au ‘misunderstood’ (main form, indicative mood, past tense, singular, 1 person)), RS_S_F (0.6%, e.g. dal-yv-auj-a ‘participate’ (main form, indicative mood, present tense, 3 person)), RS_Fd (0.5%, e.g. vadin-a-si ‘call’ (main form, indicative mood, present tense, 3 person)), RS_S_d (0.5%, e.g. mok-y-ti-s ‘to learn’), RS_S_d (0.3%, e.g. reik-al-au-ti ‘to require’), RS_Fd (0.3%, e.g. kreip-k-ité-s ‘address’ (main form, imperative mood, plural, 2 person), de-dav-o-si ‘pretended’ (main form, indicative mood, past frequentative tense, 3 person)), and PPRS_d (0.3%, e.g. ne-už-tek-s ‘will not be enough’ (main form, indicative mood, future tense, 3 person)). Three models among these 16 are very rare: verbs with two roots, e.g. RRS_S_F (0.01%, e.g. pus-ryč-iau-o ‘took breakfast’ (main form, indicative mood, past tense, 3 person)), RRS_S (0.01%, e.g. sen-berniau-ti ‘to be a bachelor’) and some active voice participles of the model RS_S (e.g. trūk-tel-ė-š ‘wrench’ (participle, active voice, past tense, masculine, singular, Nominative)).

Penta-morphemic verbs make up 9.5% of all analysed verbs and they exhibit 34 models. As a result of this, individual models are rather infrequent. Six most frequent models in this group make up two thirds (or 8%) of all verbs: PRS_S_F (4.4%, e.g. nu-stat-y-ta ‘set’ (participle, passive voice, past tense, feminine, singular, Nominative/Instrumental/neutral gender), at-stov-auj-ant-is ‘representing’ (participle, active voice, present tense, masculine, singular, Nominative)); PdRS_F (1%, e.g. su-si-rink-us-iems ‘gathering’ (participle, active voice, present tense, masculine, plural, Dative)); PdRS_S (0.8%, e.g. at-si-sak-y-ti ‘to refuse’); PPRS_F (0.7%, e.g. ne-ťis-veig-iam-as ‘unavoidable’ (participle, passive voice, present tense, masculine, singular,
Nominative/feminine, plural, Accusative)); PdRS,F (0.6%, e.g. su-si-rūp-in-o ‘concerned’ (main form, indicative mood, past tense, 3 person)); and RS₆,₅,F (0.5%, e.g. dal-yr-auj-anč-ių ‘participating’ (participle, active voice, present tense, masculine/feminine, plural, Genitive)).

**Hexa-morphemic** verbs have the largest number of models – 35, but they include only 1.5% of all verbs. The following three most frequent models make 1% of all verbs: PdRS,F (0.1%, e.g. pa-si-raš-y-t-as ‘signed’ (participle, passive voice, past tense, masculine, singular, Nominative/feminine, plural, Accusative)), PRS₆,F (0.03%, e.g. ne-į-tik-ė-tin-a ‘incredible’ (participle, necessity, feminine, singular, Nominative/Instrumental/neutral gender)), and PRS₆,₅,F (0.02%, e.g. ne-ab-iej-oj-am-a ‘unquestioning’ (participle, passive voice, past tense, feminine, singular, Nominative/Instrumental/neutral gender)).

In total, there are less than 100 (0.1%) **hepta-morphemic** words, but they fall into 12 models. The most frequent are conjugated forms and participles that have 1-2 prefixes, 2 suffixes; in addition, they are often reflexive: e.g. ne-pri-si-galv-o-dav-o ‘did not thought’ (main form, indicative mood, past tense, 3 person), at-si-pa-laid-uo-t-ų ‘relaxed’ (main form, subjunctive mood, 3 person), pri-si-stat-inė-dav-us-i ‘introducing herself’ (participle, active voice, past frequent tense, feminine, singular, Nominative). This group also includes a couple of verbs with two roots, e.g. ne-tuš-ia-žodž-iau-dam-as ‘not waffling’ (half participle, masculine, singular).

There are only ten **octa-morphemic** words (0.01%, 5 models), but words have occurred just 1-2 times. Such lengthy words are participles and conjugated verbs with two or three prefixes, as well as two or three suffixes; j-si-par-ëi-g-oj-us-iais ‘self-committed’ (participle, active voice, past tense, masculine, Instrumental).

Just one **nona-morphemic** verb has been found: ne-į-si-par-ëi-g-o-k-ite ‘do not take commitment’ (main form, imperative mood, plural, 2 person).

All models can be classified into three groups according to their frequencies: a) 10 most frequent models that have occurred more than 2 thousand times (they make up 75% of all analysed verbs; see Table 2); b) 13 models of medium frequency that have occurred more than 400 times (10% of all analysed verbs); c) the 90 remaining models are very rare (they make up 5% of all verbs). To summarize, it can be said that only about 20 morphemic models of verbs are most typical of the Lithuanian language, while the rest of the models are peripheral.

**Table 2. The Most Productive Verb Models**

| Model  | Word forms | Per cent | Typical examples |
|--------|------------|----------|------------------|
| RF     | 18,377     | 19.9     | yr-a ‘is’         |
| PRF    | 12,058     | 13.0     | n-ėr-a ‘is not’   |
| RS₆,F  | 7,236      | 7.8      | bū-t-ų ‘to be’ form |
| PRS₆,F | 6,687      | 7.2      | ne-bū-t-ų ‘was not’ |
| RS₅    | 5,039      | 5.5      | bū-ti ‘to be’    |
| PRS₅   | 4,865      | 5.3      | at-lik-ti ‘to do’ |
| RS₆,F  | 4,361      | 4.7      | tur-ė-t-ų ‘had’   |
| RS₆,₅  | 4,260      | 4.6      | tur-ėj-o ‘had’    |
| PRS₆,₅ | 4,158      | 4.5      | dar-ų-ti ‘to do’  |
| PRS₅,₅ | 4,052      | 4.4      | pa-mat-ų-s-i ‘you will see’ |

About one third of verbs are primary (approx. 33% of all verbs). These are verbs with a root, an inflection or (and) inflectional suffix: R (reik ‘need’ (main form, indicative mood, present tense, 3 person)), RF (20% of all analysed verbs, e.g. yr-a ‘is’ (main form, indicative mood, present tense, 3 person), būv-o ‘was’ (main form, indicative mood, past tense, 3 person), Žin-ai ‘you know’ (main form, indicative mood, present tense, singular, 2 person)), RS₅ (6%, e.g. bu-s ‘will be’ (main form, indicative mood, future tense, 3 person), et-k ‘go’ (main form, imperative mood, singular, 2 person), RS₆,F (8%, e.g. bū-t-ų (main form, subjunctive mood, 3
person), \( \text{bu-dav-o} \) (main form, indicative mood, past frequent tense, 3 person), \( \text{bu-s-i} \) (main form, indicative mood, future tense, singular, 2 person) "to be forms"), \( \text{RS}_{S_{1}} \) (e.g. \( \text{bu-s-iant} \) 'to be form' adverbial participle, future tense), and \( \text{RS}_{S_{1}} \) (very rare, e.g. \( \text{kabl-`e-t-a} \) 'spoken' (participle, passive voice, past tense, feminine, singular, Nominative/Instrumental/neutral gender)).

Two thirds of verbs in the analysed material have derivational affixes. No verbs with all four types of derivational morphemes (prefix, suffix, reflexive particle, and two roots) were found. Verbs with three different derivational morphemes are also very rare (approx. 2% of all verbs). Typically, it is a prefix, a suffix, and a reflexive particle, e.g. \( \text{su-si-rup-in-o} \) "concerned", or a prefix, suffix and two roots, e.g. \( \text{nu-foto-graf-uo-ti} \) "to take a picture". Verbs with two affixes of different types are relatively frequent (approx. 20% of all verbs). Characteristic combinations are as follows: a prefix and a suffix (e.g. \( \text{ne-gal-ej-o} \) "could not" (main form, indicative mood, past tense, 3 person)), a prefix and a reflexive particle (e.g. \( \text{at-si-rad-o} \) "occurred" (main form, indicative mood, past tense, 3 person)), and a suffix and a reflexive particle (e.g. \( \text{laik-y-ki-s} \) "hold on" (main form, imperative mood, singular, 2 person)). Most frequent are verbs with prefixes (approx. 28%) and suffixes (approx. 15%). Thus the most frequent types of derivational verbs contain a prefix (28%), a suffix (15%), a prefix and a suffix (12%), a prefix and a reflexive particle (6%), a reflexive particle (2%), a prefix, a suffix and a reflexive particle (2%), and, finally, a suffix and a reflexive particle (1%). Verbs with other combinations of derivational morphemes make up only about 1% of all verbs.

Almost a half (48%) of verbs in the analysed material have one prefix, 3% of verbs have two prefixes, and only 0.1% of verbs have three prefixes, e.g. \( \text{ne-`zin-`au} \) "I do not know" (main form, indicative mood, past tense, singular, 1 person), \( \text{ne-su-prat-`au} \) "I did not understand" (main form, indicative mood, past tense, singular, 1 person), \( \text{ne-pri-`pa-`zin-o} \) "did not acknowledge" (main form, indicative mood, past tense, 3 person).

About one third (31%) of all used verbs have derivational suffixes. Verbs with two derivational suffixes are quite rare (only 2% of all verbs), while verbs with three suffixes are very uncommon, e.g. \( \text{kabl-ej-o} \) "spoke" (main form, indicative mood, past tense, 3 person), \( \text{mok-y-toj-av-`au} \) "I used to be a teacher" (main form, indicative mood, past frequent tense, singular, 1 person).

12% of all verbs are reflexive, e.g. \( \text{at-si-rad-o} \) "appeared" (main form, indicative mood, past tense, 3 person).

Verbs with two roots are scarce in the Lithuanian language (about 0.5%), e.g. \( \text{bendr-a-darb-iau-ti} \) "to cooperate".

To summarise, typical verbs in the Lithuanian language are basic verbs consisting of 2-4 morphemes or verbs with one derivational affix. The generalised formula for a typical verb is \( P_{0} RS_{j} S_{0} F_{0}^{0} \).

5 Overview of Regularities of Morphemic Structures across Different Parts of Speech

Table 3 and Figure 7 show that contemporary Lithuanian is dominated by bi-morphemic words with a root and an inflection (RF). Pronouns, numerals, verbs and nouns have the largest proportion of bi-morphemic words. Adjectives are mostly derivatives with the model root + derivational suffix + inflection \( (RS_{F}) \). Such a model is quite frequent among nouns too. The Figure 7 also suggests that suffixes are quite typical to inflected parts of speech. This fact coincides with one of language universalities, that “suffixes are more prototypical affixes than prefixes” (Plungian 2003: 90). The prefixal derivation is more characteristic to verbs rather than to nominal words. It is also evident from the Table 6 that while each part of speech is dominated by one or two morphemic models, morphemic models of verbs are distributed more evenly.
The generalised morphemic structure formula of a Lithuanian word is the following:\(^1^9\): \(P_{0} R_{0,1} S_{0,2} F_{0} -1\).

Theoretically, one word may contain three prefixes, a reflexive affix, and two roots with a connective vowel, two suffixes, and an inflection; however, such words do not exist in the language. The maximum realisation of the formula is very rare, but possible, e.g. the noun – ne-su-der-in-um-as ‘incompatibility’ (Nmsn) (the PPRSdSiSdF model), the adjective – pa-vyzd-ing-iaus-iu ‘with the most exemplary’ (Amsis) (the PRSdSiF model), the verb – ne-i-si-sq-mon-in-t-as ‘not realised’ (participle, passive voice, past tense, masculine, singular, Nominative/feminine, plural, Accusative) (the PPdPRSsdSiF model).

There are differences between different parts of speech, for instance, numerals do not have prefixes, while pronouns can only have one prefix ne-. Lithuanian compound words have no more than two or at most three roots. Numerals and pronouns have only one or two derivational suffixes. Nouns do not have inflectional suffixes except those that are derived from verbs (see generalised morphemic structure in Table 4).

---

19 The formula does not include reflexive affixes (which can occur after a prefix or after an inflection) and connective vowels. These morphemes are not characteristic to all parts of speech, therefore they have been excluded from the generalised formula.
Table 4. Generalised Morphemic Structure of Nouns, Adjectives, Pronouns, Numerals, and Verbs

|      | P     | R     | Sd    | Si    | F     |
|------|-------|-------|-------|-------|-------|
| Nouns| P₀₋₃  | R₁₋₇  | Sd₀₋₄ | Si₀₋₂ | F₀₋₁  |
| Adjectives| P₀₋₂  | R₁₋₃  | Sd₀₋₄ | Si₀₋₂ | F₀₋₁  |
| Pronouns| P₀₋₁  | R₁₋₂  | Sd₀₋₁ | Si₀   | F₀₋₁  |
| Numerals| P₀    | R₁₋₂  | Sd₀₋₂ | Si₀   | F₀₋₁  |
| Verbs | P₀₋₃  | R₁₋₂  | Sd₀₋₄ | Si₀₋₂ | F₀₋₁  |

The Table 4 shows that nouns may have the maximum number of 3 prefixes. Statistics in the analysed data show that nouns most frequently have one prefix, less frequently two, while the rarest are nouns with three prefixes, e.g. ne-pri-pa-žin-im-as ‘disallowance’ (Nmsn) (the PPPRS₅F model). Derivational suffixes are more frequent, as there are quite a number of nouns that have two or even three derivational suffixes, e.g. reik-al-av-im-us ‘requirements’ (Nmpa) (the RSdS₀S₃S₃₅F model). Nouns with four derivational suffixes are unusual, mostly these are international words, e.g. centr-al-iz-av-im-as ‘centralisation’ (Nmsn) (the RS₀S₀₃S₃₅F model).

As in many other languages, the Lithuanian language is dominated by one-root nouns. Two-root nouns are also quite frequent. The largest number of roots (even 7) has the international noun fibro-ezo-fag-gastro-duo-deno-skop-ij-a ‘fibre esophagastroduodenoscopy’ (Nfsn/i) (the RRRRRRRS₅F model) in the analysed data. However, this individual case does not mean that multi-root words are typical to Lithuanian. By and large, international nouns have more roots than Lithuanian nouns. It should be noted, that sometimes two or more roots do not occur sequentially as they can be intervened by other morphemes (i.e. connective vowels), e.g. ei-lė-raš-tuk-as ‘a little poem’ (Nmsn) (the RSdcRSdS₃F model).

As it has already been mentioned, inflectional suffixes are not typical to nouns, as they occur only when a noun is derived from a verb, e.g. pri-klaus-om-yb-ės ‘dependence’ (Nfsa) (the noun is derived from the passive present tense participle pri-klaus-as ‘dependent’) or in the cases, when a suffix occurs in all cases of words (except singular Nominative) that have consonant endings of stems, e.g. vand-en-į ‘water’ (Nmsn) (where -en- by different researchers is treated as a part of root or as a suffix). The maximum number of inflectional suffixes is two, e.g. mir-št-am-um-as ‘mortality’ (Nmsn) (the RS₅S₅₃F model). However, one-suffix words are more frequent.

The prefixal derivation is not typical to adjectives and therefore they are not very frequent. Adjectives with two prefixes have been found in the analysed data, e.g. ne-pa-nas-us ‘not similar’ (Amsn) (the PPRF model). We have not found any adjective with more prefixes, however theoretically more prefixes could be used. Adjectives may have the maximum of three roots. Three-root adjectives may be of the international origin, e.g. mikro-bio-log-in-ės ‘microbiological’ (Afsn/Afpn) (the RRRRS₅F model), as well as of the Lithuanian origin, e.g. vien-uo-lk-meč-io ‘eleven years old’ (Amsg) (the RCRRF model). The most frequent adjectives are with one root, although two-root adjectives may also occur.

We have found adjectives with as much as 4 derivational suffixes, e.g. gy(-)v-en-įšk-us ‘true to life’ (Ampa) (the RS₅S₅₃S₅₃₅F model). The number of derivational suffixes is not large. Two-suffix adjectives are the most frequent ones, e.g. žin-om-iaus-iu ‘the most famous’ (Amsis) (the RS₃S₃F model; this adjective is of the verbal origin, which is shown by the suffix -om-, while the other suffix -iaus- is used to form the superlative degree).

Pronouns do not have many different morphemic structural models. They may have the maximum of two roots, e.g. kiek-vien-as ‘everyone’ (Pmsn/Pfpa) (the RRF model) and one derivational suffix, e.g. t-ok-s ‘such’ (Pmsn) (the RS₅F model). Inflectional suffixes and prefixes are not typical to pronouns, although we have found one exception in the analysed data – ne-mūs-įšk-ė ‘not our’ (Pfsn) (the PRS₅F model).

Numerals have a rather simple morphemic structure. They may have the maximum of two roots, e.g. aštuon-io-lk-t-ojo ‘the eighteenth’ (Momsg) (the RCRS₅F model). They also may have the maximum of two derivational suffixes, e.g. tr-aj-ėt-o ‘triplet’ (Mc00g) (the RS₅₃₅F model), while inflectional suffixes are not typical to numerals.
The prefixal derivation is typical to verbs, therefore a large part of verbs have one, two or even three prefixes. It is typical for verbs that prefixes are not necessarily distributed sequentially, but they are often intervened by a reflexive affix, e.g. ne-i-si-sq-mon-in-t-as ‘unrealised’ (participle, passive voice, past tense, masculine, singular, Nominative/feminine, plural, Accusative) (the PdPRS_S,F model).

As far as roots are concerned, the maximum number of roots in verbs is two, e.g. bad-mir-iau-o ‘starve to death’ (main form, indicative mood, past tense, 3 person) (the RRS_d,F model).

Verbs may have up to four derivational suffixes, e.g. dé-sty-toj-ai-ti ‘to lecture’ (the RS_d,S_d,S_d,S_d model) and up to two inflectional suffixes, e.g. bu-s-im-o-s ‘to be in future’ (participle, active voice, future tense, feminine, singular, Genitive/plural, Nominative) (the R,S,F model).

The largest and the smallest structural model is always a rare phenomenon in the language. So what is typical to all inflected parts of speech according to this analysis? The generalised model P_o,RS_d0,F is the most characteristic to nouns and adjectives. The root and the flexion are necessary morphemes for nouns and adjectives, if they have affixes, they are typically limited to 1 prefix and 2 derivational suffixes. The prefix is not a characteristic affix for numerals (RS_d0,F_o) and pronouns (RF), while flection-less numerals are common in the natural language. Verbs (P_o,RS_d0,F_o,F_o) differ from other inflected parts of speech that flection-less forms with inflectional suffix are quit common in the language.

Although our research is synchronic, but a couple of notes regarding historical aspects may be observed. The grammatical categories inherited from Proto-Indo-European are typically expressed by multi-sense flexions, e.g. gender: ger-as – ger-a ‘good’, number: ger-as – ger-i; case ger-as – ger-o, ger-am; person: dirb-u – dirb-i ‘work’; while more recent grammatical categories are expressed by inflectional suffixes, e.g. future and frequentative tense dirb-u – dirb-s-iu, dirb-dav-au.

6 Concluding remarks

Having analysed a 310 thousand-word Lithuanian corpus (where nouns make up 46% of all words, verbs – 30%, pronouns – 13%, adjectives – 10%, and numerals – 2%), some important conclusions may be drawn about the morphemic structure of the Lithuanian language.

Lithuanian inflective parts of speech may consist of up to 9 morphemes. Word forms of two, three, and four morphemes make up the largest part of the Lithuanian vocabulary (93% of all analysed words). The data contained 43% of bi-morphemic words, 33% of tri-morphemic, and 17% of quadri-morphemic words. The number of morphemes in a word is not limited. One can say that uni-morphemic words and words that are longer than 4 morphemes belong to the language periphery.

The range of morphemic structural models as found for different parts of speech is very diverse: numerals and pronouns have just 10 models, nouns and adjectives 63, while verbs – 116. In addition, frequencies of different morphemic structural models are quite different. The most frequent are simple words and words with one suffix or one prefix: RF (41% of all analysed words), RS_F (17%), and PRF (8%). The following three models are also among the most frequent: RS_d,F (6%), RS,F, and PRS,F (3% each). These six models make up even 78% of all inflective parts of speech in the analysed texts.

It may be concluded that the most typical words in the Lithuanian language are primary words with 2-4 morphemes.

The authors plan to continue the morphemics research, while further focusing on the statistical analysis of the data, e.g. describing regularities of the Lithuanian language according to Menzerath-Altmann’s law (cf. Köhler et al. 2005: 349), statistical regularities of morpheme and word length (cf. Meyer 1999), morphemic ambiguity, and frequency regularities of morphemic usage (cf. Krott 1999).

Acknowledgements: The research has been financed by the Lithuanian Council of Science (agreement No. MIP-06/2010). Special thanks to assoc. prof. dr. Violeta Kalėdaitė for her valuable comments and to assoc. prof. dr. Gailius Raškinis for the compilation and primary processing of the analysed data. We also thank reviewers for their worthy comments that have lead to the improvement of the paper and the revision of the The Database of Lithuanian Morphemics Data.
Abbreviations

c connecting vowel
d reflexive affix
F flection
P prefix
R root
Sd derivational suffix (derivational suffix is used to create new words, e.g. nouns from verbs or adjectives from nouns)
Si inflectional suffix (inflectional suffix is used to create new grammatical forms, which are not new words, e.g. degrees of adjectives and adverbs, past tense verbs, or participles)

References

Akelaitienė, Gražina. 1996. Morfonologinės balsių kaitos žodžių daryboje (Morphonological Vowel Alternation in Word Derivation). Vilnius: VPU leidykla.
Akelaitienė, Gražina. 2001. Šaknies balsių kaita: morfemų modeliai ir struktūriniai šaknies tipai (Root Vowel Alternation: Models of Morphonemes and Structural Root Types). Žmogus ir žodis. Didaktinė lingvistika 1 (3), pp. 3-9.
Akelaitienė, Gražina. 2008. Priešbalsines priesagų turinčių vardažodių šaknies balsių kaita (Root Vowel Change in Derivatives with Consonant Suffixes). Žmogus ir žodis. Didaktinė lingvistika 10 (1), pp. 5-9.
Božou, Loïc. Jolanta Kovalevskaitė, Erika Rimkutė, Automatic Lemmatisation of Lithuanian MWEs. Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015, pp.41-49.
DLKG. 1996. Dabartinės lietuvių kalbos gramatika (Contemporary Grammar of Lithuanian Language). Vilnius: Mokslo ir enciklopedijų leidykla.
DLKŽ. 2006. Dabartinės lietuvių kalbos žodynas, trečias elektroninis leidimas (Modern Lithuanian Dictionary, 3rd electronic edition). Vilnius: Lietuvių kalbos institutas. http://dz.lki.lt/ (27.04.2014).
Kalnača, Andra. 2004. Morfēmika un morfonoloģija (Morphemics and Morphology). Rīga: LU Akadēmiskais apgāds 2004.
Karosienė, Vida. 2004. Bendrinės lietuvių kalbos vardažodių šaknies struktūra (The Structure of Nominal Root in the Standard Lithuanian). Vilnius: Vilniaus universiteto leidykla.
Kaukienė, Audronė. 1994. Lietuvių kalbos veiksmažodžio istorija (History of Lithuanian Verb) (1st volume). Klaipėda: Klaipėdos universitetas.
Kaukienė, Audronė. 2002. Lietuvių kalbos veiksmažodžio istorija (History of Lithuanian Verb) (2nd volume). Klaipėda: Klaipėdos universitetas.
Kazlauskienė, Asta. 2012. Priešbalsiai veiksmažodžio morfemų sandūroje (Consonants at the Juncture of Verb Morphemes). Lietuvių kalba 6. http://www.letiuiukalba.lt/index.php?id=213 (15.03.2014).
Keinys, Stasys. 2009. Bendrinės lietuvių kalbos veiksmažodžio šaknies struktūra (The Structure of Verb Root in the Standard Lithuanian). PhD thesis. Vilnius: Vilniaus universiteto leidykla.
Köhler, Reinhard, Gabriel Altmann, Rajmund G. Piotrowski (eds.). 2005. Quantitative Linguistics. An International Handbook. Handbooks of Linguistics and Communication Science (HSK) 27. Berlin: De Gruyter Mouton.
Krott, Andrea. 1999. Influence of Morpheme Polysemy on Morpheme Frequency. Journal of Quantitative Linguistics 6 (1/1999), pp. 58-65. http://dx.doi.org/10.1076/jqul.6.1.58.4144 (27.04.2014).
Kruopienė, Irena (2000a), Bendrinės lietuvių kalbos veiksmažodžio šaknies struktūra (The Structure of Verb Root in the Standard Lithuanian). PhD thesis. Vilnius University.
Kruopienė, Irena (2000b), Stabiliojo centro veiksmažodžių šaknies struktūra (The Structure of Stable Nucleus Roots of Verbs). Kalbotyra 48 (1), pp. 71-82.
Kruopienė, Irena. 2001. Pradinių ir galinių priešbalsių derinimas veiksmažodžio šaknyse (The Combination of Initial and Final Consonants in Verb Roots). Kalbotyra 50 (1), pp. 67-82.
Kruopienė, Irena. 2002. Veiksmažodžio kintamojo centro šaknų struktūra (The Structure of Alternating Nucleus Roots of the Verb in Lithuanian). Kalbotyra 51 (1), pp. 89-97.
Kruopienė, Irena. 2005. Veiksmažodžio šaknies trinarės inicialės ypatumai (Peculiarities of three-member initial consonant clusters in the verbal root). Kalbotyra 54 (1), pp. 73-82.
Kuosienė, Monika. 1986. Morfeminė analizė (Morphemic Analysis). Šiaulių: Šiaulių K. Preiķšo pedagoginis institutas.
Levāne-Petrova, Kristine, Andrejs Spektors, Morphemic Analysis and Morphological Tagging of Latvian Corpus. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), pp. 1095-1098. http://gandalf.aksis.uib.no/non/lrec2000/pdf/107.pdf (15.03.2013).
Meyer, Peter. 1999. Relating Word Length to Morphemic Structure: A Morphologically Motivated Class of Discrete Probability Distributions. Journal of Quantitative Linguistics 6 (1), pp. 66-69. http://dx.doi.org/10.1076/jqul.6.1.66.4143
Pakerys, Antanas. 1994. Akcentologija, I dalis: daiktavardis ir būdvardis (Accentology, 1st Part: Noun and Adjective). Kaunas: Šviesa.

Plungian, Vladimir A. 2003. Общая морфология (General Morphology). Moscow: УРСС 2003.

Polikarpov, Anatoliy A. 2000. Chronological morphemic and word-formational dictionary of Russian: Some system regularities for morphemic structures and units. Linguistische Arbeits Berichte 5, pp. 201-212.

Rimkutė, Erika, Asta Kazlauskienė, Gailius Raškinis. 2011. Lietuvių kalbos veiksmažodžių morfeminė struktūra (The Morphemic Structure of Lithuanian Verbs). Acta Linguistica Lithuanica 64-65, pp. 87-105.

Sarkans, Ugis, Morphemic and morphological analysis of the Latvian language. Proceedings of the Forth conference on Computational Lexicography and Text Research 1996, pp. 219-225. http://www.ailab.lv/users/usarkans/complex.html (15.03.2013).

Sedlāček, Radek, The Core of the Czech Derivational Dictionary. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04). Lisboa: ELRA, 2004, pp. 1279-1282.

Urbutis, Vincas. 2009. Žodžių darybos teorija (Word Derivation Theory). Vilnius: Mokslo ir enciklopedijų leidybos institutas.