Morphemic structure of the Lithuanian prefixes
Lietuviešu valodas priedėkļu morfēmiskā struktūra

Daiva Šveikauskienė
Institute of the Lithuanian Language
Vileišio str. 5, LT-10308 Vilnius
E-mail: *daiva.sveikauskiene@lki.lt*

The article describes the morphemic structure of words. The structures of the word in various languages are reviewed. The morphemic dictionaries are discussed too. The insufficiency of information in the morphemic dictionary of the Lithuanian language is pointed out. The generalized structure of the Lithuanian word is presented. The results of the comprehensive analysis of the Lithuanian prefixes are described.

Keywords: Morphology, morphemic structure, prefix, computational linguistics, natural language processing.

1. Introduction

According to the data of the META-NET project, Lithuanian belongs to the group of the worst computerized languages of the Europe (Vaišnienė, Zabarskaitė 2012, 35). In many areas the Latvian language is more computerized than Lithuanian. For example, the inverse dictionary of Latvian by Emilija Soida (Soida, Kļaviņa 1970) is computerised. The reverse dictionary of the Lithuanian language by Juozas Korsakas (Korsakas 1991) is not computerized yet.

Morphological analysis is one of the main parts of the language processing. It is very important to have information about the morphemic structure of a word. The described structures of the word differ in various languages, because of the difference in structure of language and because of the difference in the goal of representing information in the structure of the word.

2. Word structure in various languages

Various languages represent their word structure in different ways. Czech word is depicted as consisting of three sections. Finnish and Latvian words have four parts with different representation of information. Russian linguists describe the word as consisting of two parts. English word has three parts as in Czech, but the kind of information represented in it is different.

2.1. Word structure in English

Structurally, English word consists of three parts. It is made up of prefix, base form and suffix. Figure 1 shows the formula of the English word (Adedimeji 2005, 10).

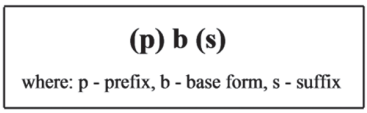

Figure 1. Formula of the English word (Adedimeji 2005, 10)

Base form is obligatory, whereas prefix and suffix are optional. The brackets indicate that their contents are optional. Possible structures of an English word may be the following: *b, pb, bs, pbs*. Many English words have more than one *p* element and more than one *s* element. Compound words have more than one *b* element. The formula of the English word can be expanded and Figure 2 shows it (Adedimeji 2005, 11).

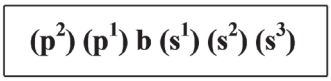

Figure 2. Expanded structure of the English word (Adedimeji 2005, 11)

The author doesn't note if the structure with two prefixes and three suffixes is most possible extension in the English language.

2.2. Word structure in Czech

Sedlaček (2004) describes the detailed morphemic structure of Czech words. Each Czech word is divided into three fields: root segment, preroot segment, and postroot segment. A tool for research on Czech derivation morphology was created. Figure 3 shows an example of printout.

```
       od /běr/ a tel
mal  o od /běr/ a tel          compound
mal  o od /běr/ a tel k  a     derivative
mal  o od /běr/ a tel sk ý     derivative
mal  o od /běr/ a tel sk y     derivative
velk o od /běr/ a tel          compound
velk o od /běr/ a tel k  a     derivative
velk o od /běr/ a tel sk ý     derivative
velk o od /běr/ a tel sk y     derivative
       od /běr/ a tel k  a     derivative
       od /běr/ a tel sk ý     derivative
```

Figure 3. Example of the morphemic structure of the Czech word family (Sedlaček 2004, 1281)

The words are organized into derivational families. The root is marked with slash. The preroot segment includes all prefixes and also the first element of compounds and the postroot segment contains all suffixes and the flexional ending if any (Sedlaček 2004, 1280).

2.3. Word structure in Finnish

The word structure in Finnish consists of four parts: stem, sign, ending, and suffix (Website 1). Figure 4 shows the generalized structure of the Finnish verbs and nominals.

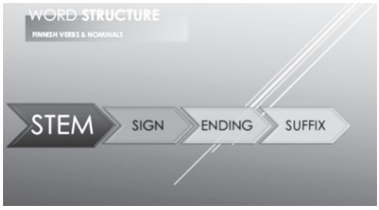

Figure 4. Word structure of the Finnish words (Website 1)

Stem indicates the root. Sign can indicate the mood, for example, conditional, or number. Ending gives the grammatical information, for example, person, case. Suffix can indicate possessive feature or signify yes/no question.

2.4. Word structure in Latvian

The structure of the Latvian word is represented as consisting of four parts too, only the information represented is different. Figure 5 shows the general format for single-rooted words (Levāne, Spektors 2000, 1095).

<prefix>* <root> <suffix>* <[ending]>

Figure 5. Word structure of the single-rooted Latvian words (Levane, Spektors 2000, 1095)

The possibility of zero or more occurrences of element is denoted by * and optional elements are included in square brackets. Figure 6 shows an example of the „Derivational Dictionary of Latvian" (Metuzāle-Kangere, 1985, 92).

GRUP	GRUP A	GROUP
	GRUP Ē T (IE S)	TO GROUP
	GRUP Ē J UM S	GROUPING
	GRUP Ē ŠAN A	GROUPING
	GRUP Ē ŠAN ĀS	GROUPING
	NO GRUP Ē T IE S	TO GROUP
	NO GRUP Ē J UM S	GROUPING
	PĀR GRUP Ē T (IE S)	TO RE-GROUP
	PĀR GRUP Ē ŠAN A	RE-GROUPING
	PĀR GRUP Ē ŠAN ĀS	RE-GROUPING
	PĀR GRUP Ē J UM S	RE-GROUPING
	SA GRUP Ē T (IE S)	TO GROUP
	SA GRUP Ē ŠAN A	GROUPING
	AROD GRUP A	TRADE-UNION GROUP
	ASIN S GRUP A	BLOOD GROUP
	HIDR OKS IL GRUP A	HYDROXIDE GROUP
	TRIEC IEN GRUP A	ASSAULT GROUP

Figure 6. Example of the „Derivational Dictionary of Latvian" (Metuzale-Kangere, 1985, 92)

The words are represented in word family. Each family is introduced with underlined root. The English equivalent is given for the each word in the family. Roots are placed in the same column. Prefixes, suffixes and endings are separated with spacing.

2.5. Word structure in Russian

Each flexible word in the Russian language has two parts: stem and inflectional formant. Stem contains the root, optional prefixes and suffixes. Inflectional formant consists of ending and optional postfix, for example, reflexive particle (Website 2). Single-rooted words can consist of one to eight morphemes (see Figure 7).

Figure 7. Structure of the Russian word with eight morphemes (Website 2)

The morphemic dictionary of the Russian language gives exhaustive information about the type of each morpheme in the word. Figure 8 shows two examples of the word representation (Website 3).

Figure 8. Examples of the word representation in the Russian morphemic dictionary (Website 3)

Roots, prefixes, suffixes and endings are represented in different manner.

3. Works created in morphemic of Lithuanian

The works made in Lithuanian on computerisation of morphology are following: automatic morphologic analysis, created at the Institute of Mathematics and Informatics (Zinkevičiu[illegible]00) is used in morphologic analyser of the Vytautas Magnus university in Kaunas. The dictionaries of morphemic, both alphabetic and reverse, are created at the Centre of computational linguistics in Vytautas Magnus university in Kaunas. Derivational morphemic database is created at the Institute of Lithuanian Language. These works are very important, however, they have some negatives, and must be improved.

3.1. Morphemic dictionary

Alphabetical morphemic dictionary of the Lithuanian language is corpus-based (Rimkutė, Kazlauskienė, Raškinis 2011, 4). It contains information about the frequency of the word, that is, times of its occurrence in the corpus, the word

divided in morphemes, and grammatical information, that is, lemma, part of speech, tense, person, case, number, gender etc.

The negative of this dictionary is absence of any information about the type of morpheme. For example, two words *viršsvoris* and *viršūnėmis* are presented as in Figure 9 (Rimkutė, Kazlauskienė, Raškinienė, 2011, 777).

2	virš-ūn-ėmis	viršūnė; dkt. mot. g. dgs. įnag.
1	virš-svor-is	viršsvoris; dkt. vyr. g. vns. vard.

Figure 9. Examples of the words *viršūnėmis* and *viršsvoris* (Rimkutė, Kazlauskienė, Raškinienė, 2011, 777)

Both words have three morphemes and begin with the same morpheme *virš*. There is no information in the dictionary that in the word *viršsvoris* the first morpheme (*virš*) is the prefix and in the word *viršūnėmis* the first morpheme (*virš*) is the root.

The same lack of information can be observed in compounds. Figure 10 shows the representation of two words, one of which is compound (*laikrodis*) and the second is simple (*laikmenoje*). The structure of the both words is identical, that is, both words consist of three morphemes separated with hyphens and the first morpheme is *laik-* (Rimkutė, Kazlauskienė, Raškinienė, 2011, 331–332).

9	laik-rod-is	laikrodis; dkt. vyr. g. vns. vard.
3	laik-men-oje	laikmena; dkt. mot. g. vns. viet.

Figure 10. Examples of the words *laikrodis* and *laikmenoje* (Rimkutė, Kazlauskienė, Raškinienė, 2011, 331–332)

The type of the second morpheme in both words differs. There is no information indicating, that the second morpheme *men* in the word *laikmenoje* is the suffix, and the second morpheme *rod* in the word *laikrodis* is the second root.

3.2. Derivational and morphemic database of Lithuanian

The first linguistic database in Lithuania was created in the last decade of the 20th century at the Institute of Mathematics and Informatics. It contains ca. 75 thousand word entries of the Dictionary of Contemporary Lithuanian (Murmulaitytė 2012, 97). Morphemic information is following: the root is represented with bold font, upper case is used for the derivational suffix, endings have normal font, not derivational suffixes are depicted with italic case and infixes are double underlined. The underlying words are given for compounds and derivatives. Figure 11 shows

two fragments of the database: derivative *taisyklingas* and compound *toliaregis* (Murmulaitytė 2012, 98).

tais *ykl ING* as	taisyklė dkt.
tol ia **reg** is	toli prv. + regėti vksm.

Figure 11. Examples of the first database: derivative *taisyklingas* and compound *toliaregis* (Murmulaitytė 2012, 98)

The database was created before the Morphemic Dictionary (see Section 3.1.), however, it contains more information about the morpheme type.

4. Word structure in Lithuanian

The structure of the Lithuanian words is not sufficiently analysed. We have no publications about the generalized structure of Lithuanian word. Rimkutė (2010, 95) describes the frequency characteristics of word patterns in Contemporary Lithuanian text corpus. The largest word found in corpus consists of 9 morphemes. The type of morphemes is not presented. It is planned at the second stage of the project LIEPA to create the grammatical database with exhaustive information about the words used in the project. To achieve this goal, we need a comprehensive research of the structure of Lithuanian words.

4.1. Generalized structure of the word

The publications describing structure of the Lithuanian word give word patterns. No general structure is presented. It was decided to create common structure reflecting all Lithuanian words. We prefer to treat the Lithuanian word as consisting of three parts. However, the information belonging to each part is different from the English and Czech languages, which have the same number of parts in the word structure (see Section 2.1. and Section 2.2.). English word has prefix, base form and suffix. There is no ending depicted in it, because endings are very rare in English. The endings are very important for the Lithuanian language, because they have almost all syntactic information. Thus the method of representation of the word structure used in Czech is more suitable for the Lithuanian language. The only difference is in the composition of word segments. The postroot segment contains suffix and ending. It is very useful for Lithuanian with the large number of endings. It is not very meaningful to make the ending a separate part of the word, like in the case of Latvian (see Section 2.4.), because the ending in a Lithuanian word is always an only one. Thus, the better choice is to treat the ending as a component of the postroot segment.

The representation of information in root segment and preroot segment in the Lithuanian differ from Czech. Preroot segment contains prefixes and optional reflexive particle. Root segment is common for the single rooted words and compounds. It contains one or more roots, infixes, connective vowel between

the roots. So we have generalized word structure which is able to represent any Lithuanian word (see Figure 12).

Figure 12. Generalized structure of the Lithuanian word

Each part must be detailed. The additional research is needed in the structure of Lithuanian words in the aspect essential for the computer processing. The first part of the structure of the Lithuanian word was particularly analysed in its composition. Next section describes the results of the research.

4.2. Structure of the prefixes in Lithuanian

The publications represent some patterns of Lithuanian word. Figure 13 shows most frequent pattern of Lithuanian verbs (Murmulaitytė 2012, 99).

$$\mathbf{pd_{1.1}^{d} + š_1 + ps_{1.1} + ps_{1.2}}$$

Figure 13. Most frequent word pattern of Lithuanian verbs

It consists of one prefix, one root and two suffixes. Only the infinitives were taken into account.

Tree most typical models of Lithuanian verbs are following (Rimkutė 2010, 88):

1. prefix + root + ending,
2. prefix + root + derivational suffix + ending,
3. prefix + root + derivational suffix

The largest number of prefixes in given patterns is two, for example, PPRF (prefix, prefix, rot, flexion), PPRS (prefix, prefix, root, suffix), PPRSF, PPRSSF (Rimkutė 2010, 95). The patterns of Russian words have the largest number of prefixes – three (see Section 2.5.).

It is important to know how many positions in the database are needed for each part of the words. The preroot part of the Lithuanian word was analysed. All possible combinations of prefixes were made. The combination was deleted if the word as example for that combination of prefixes was not found. The goal of research was to find most possible number of morphemes bevor the root, that is, it was more technical than linguistic research. The number positions is important, not the number of prefixes. Some Lithuanian linguists treat the reflexive particle before the root as a prefix, however, others state that the reflexive particle is not prefix. The number of morpheme combinations before the root reaches about 600. The largest number of morphemes before the root in the word is 6, for example, *tenebeprisipažįsta* (let he do not confess it more). It was decided to make the

first part of the Lithuanian word structure consisting of eight positions with hope that the words with more than eight morphemes will not arise in the Lithuanian language.

The list of all possible combinations of morpheme was made by representing each type of prefixes with different colour. Figure 14 shows a fragment of the list.

467. → **ne-be-si-a-**¶
468. → **ne-be-si-ant-**¶
469. → **ne-be-ap-si-**¶
470. → **ne-be-at-si-**¶
471. → **ne-be-si-dis-**¶
472. → **ne-be-si-eks-**¶
473. → **ne-be-si-hiper-**¶
474. → **ne-be-į-si-**¶
475. → **ne-be-si-inter-**¶
476. → **ne-be-iš-si-**¶
477. → **ne-be-si-kon-**¶
478. → **ne-be-si-ko-**¶
479. → **ne-be-nu-si-**¶
480. → **ne-be-pa-si-**¶
481. → **ne-be-par-si-**¶
482. → **ne-be-per-si-**¶

Figure 14. Fragment of the list of the morpheme combinations in the preroot part of the Lithuanian word

It can be observed, that the reflexive particle *si* is allocated before the international prefix and after the Lithuanian prefix. The particles *te*, *be*, *ne* are located before the prefixes, that is, at the beginning of the word. This information can be useful for the linguistic research of the regularity of location of certain morphemes in preroot part of the word.

Conclusions

1. The morphemic structure of the word in various languages is reviewed.
2. Works created in morphemic of Lithuanian words are described.
3. Generalized morphemic structure of Lithuanian words is presented.
4. The results of the comprehensive research of Lithuanian prefixes are described.

Sources

1. Website 1: http://www.slideshare.net/riortamm/word-structure-42667712

2. Website 2: http://www.langust.ru/rus_gram/rus_gr03.shtml
3. Website 3: http://russkiy-na-5.ru/dictionary/morphemics/архитектурный
   http://russkiy-na-5.ru/dictionary/morphemics/беспамятный

References

1. Adedimeji, Mahfouz A. 2005. *Word Structure in English*. https://www.unilorin.edu.ng/publications/ADEDIMEJI/WORD%20STRUCTURE%20IN%20ENGLISH.pdf
2. Korsakas, Juozas. 1991. *Lietuvių kalbos inversinis žodynas*. Kaunas: Šviesa.
3. Levāne, Kristīne, Spektors, Andrejs. 2000. Morphemic Analysis and Morphological Tagging of Latvian Corpus. *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, May 31–June 2, 2000. V. 2. 1095–1098. Available: *http://www.lrec-conf.org/proceedings/lrec2000/pdf/107.pdf*
4. Murmulaitytė, Daiva. 2012. Lietuvių kalbos morfemikos ir žodžių darybos tyrimų perspektyvos. *Žmogus ir žodis*. Nr. 1. Vol. 14. 96–102.
5. Rimkutė, Erika. 2010. Lietuvių kalbos veiksmažodžių morfeminė struktūra. *Acta linguistica Lituanica*. LXIV–LXV. 87–105.
6. Rimkutė, Erika, Kazlauskienė, Asta, Raškinis, Gailius. 2011. *Abėcėlinis lietuvių kalbos morfemikos žodynas*. I dalis. Kaunas: Vytauto Didžiojo universitetas.
7. Sedlaček, Radek. 2004. *The Core of the Czech Derivational Dictionary*. Available: *http://www.lrec-conf.org/proceedings/lrec2004/pdf/696.pdf*
8. Soida, Emīlija, Kļaviņa, Sarma. 1970. *Latviešu valodas inversā vārdnīca*. Rīga, Latvijas valsts universitāte.
9. Vaišnienė, Daiva, Zabarskaitė, Jolanta. 2012. The Lithuanian Language in the digital age. *META-NET White paper series*. Rehm, Georg, Uszkoreit, Hans (eds.). Berlin, Heidelberg: Springer-Verlag.
10. Zinkevičius, Vytautas. 2000. Lemuoklis – morfologinei analizei. *Darbai ir dienos 24*. Kaunas: Vytauto Didžiojo universitetas, 245–273.

Kopsavilkums

Rakstā īpaša uzmanība pievērsta vārdu struktūrai. Tiek aplūkotas dažādu valodu vārdu struktūra, kā arī morfēmikas vārdnīcas. Iesniegtajā pētījumā apgalvots, ka lietuviešu valodas morfēmikas vārdnīcā nav pietiekamas informācijas. Tiek apkopota lietuviešu valodas vārda struktūra, aprakstīti rūpīgi analizētu lietuviešu valodas priedēkļu rezultāti.