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

doi:10.3233/FAIA200627

Quantitative Analysis of Language Competence *vs.* Performance in Russian- and Lithuanian-Speaking 6 Year-Olds

Ingrida BALČIŪNIENĖ [a,b1] and Aleksandr N. KORNEV [b]
[a] *Vytautas Magnus University, Lithuania*
[b] *Saint-Petersburg State Pediatric Medical University, Russia*

Abstract. The paper deals with a comparative analysis of the *Part-of-Speech Profile* between different languages and discourse genres in 6-year-old typically developing Russian- *vs.* Lithuanian-speaking children. Results of the study inspire a discussion on a possibility to evaluate both language competence and language performance of the same subject on the basis of his/her distribution of parts of speech in the discourse.

Keywords. Corpus linguistics, language acquisition, part of speech, language competence, language performance

1. Introduction

Among numerous studies in child language, the development of separate linguistic patterns such as morphology and morphosyntax of some parts of speech as well as morphological derivation and compounding has been well described both in Russian and Lithuanian. In Russian, the essential longitudinal studies [1, 2, 3, 4, 5, 6, 7] have focused on the acquisition of nouns, adjectives, pronouns, and verbs. In Lithuanian, longitudinal studies have been devoted to the acquisition of nouns [8], verbs [9], and adjectives [10]. Some comparative studies [11, 12] should also be noted. Much less is still known about relationships between different parts of speech (PoSs) along the developmental course and their role in the acquisition of discourse skills. [13] discussed the effect of grammatical, lexical, and pragmatic categories on the mean length of utterance (MLU) rate. Authors proposed that during the 2nd-3rd years of life, function words play the central role in the syntactic development. Verbs appeared to be especially important for the late syntactic development [14, 15, 16], whereas nouns are important for the nominal function development. The acquisition of PoSs means "…knowing how to use the word in the language. The grammatical category of a word determines (1) the position it is allowed to occupy in the clause /…/; (2) the range of syntactic functions it can occupy /…/; (3) the types of words with which it co-occurs /…/; (4) the types of morphemes it requires or accepts /…/" [17: 434].

[1] Corresponding Author: Ingrida Balčiūnienė; Department of Lithuanian Studies, Vytautas Magnus University, V. Putvinskio st. 23-206, LT-44243 Kaunas, Lithuania; E-mail: ingrida.balciuniene@vdu.lt.

As Slobin proposed, "…child acquires more than a system of grammatical forms and semantic/communicative functions. In acquiring the grammar of a particular language, the child comes to adopt a particular framework for schematizing experience" [18: 7]. In different languages, categorical/syntactic function of word plays different roles in entity assignment. On the other hand, to become a proficient native speaker, a child has to learn language-specific rhetorical style which, in turn, influences lexical and morphosyntactic features of discourse [19]. Speakers of two different languages will organize the same reality in slightly different ways and, thus, they will employ the PoSs in some different proportions.

The **aim** of our study was to compare PoS distribution in the discourse of Russian-speaking children and their Lithuanian-speaking peers. The point of our interest was to analyze quantitively the PoS distribution from both static (language knowledge/competence) and dynamic (language behavior) perspectives in different genres. It was hypothesized that *lemma distribution* to more extent reflects the *language competence* of a subject, while *word token distribution* is more sensitive to *language behavior* demands in the given discourse context.

Among various quantitative approaches to corpus data, the distributive PoS-analysis should reveal some syntactic pattern information [20, 21, 22]. Following Lyashevskaya, the "grammatical behavior" of language units in corpus data manifests in the item distribution in a context. This is relevant to PoS Profile (PoSP), i.e. the distribution of word types [23: 7].

2. Methodology

For this comparative quantitative study, we accessed two corpora of child language. The *Corpus of Lithuanian Children Language* has been developed at Vytautas Magnus University and comprises morphologically annotated longitudinal and semi-experimental data (~106 hours) of the Lithuanian L1 development [24]. The *Corpus of Russian Children Language* has been compiled at Saint-Petersburg State Pediatric Medical University and comprises morphologically annotated semi-experimental data (~65 hours) of Russian L1 development [25].

For this study, we selected 24 typically developing (TD) 6-year-olds and analyzed their PoSs in different discourses (Table 1).

Table 1. The data

Subjects	Russian TDs (n = 12)	Lithuanian TDs (n = 12)
Morphologically annotated transcripts:		
Fictional narratives	2466 word tokens	2975 word tokens
Conversational dialogues	3074 word tokens	13279 word tokens

Namely, we selected (1) narratives told by the subjects according to the picture sequence and (2) conversational dialogues. As for narratives, Lithuanian children told stories according the *Cat Story* picture sequence developed by [26]; Russian children told stories according the sequence slightly modified in the framework of the COST Action IS0804 (http://bi-sli.org). Conversational dialogues were elicited in a slightly different way: Russian children were asked to answer 10 comprehensions about the story they told. Lithuanian children were not controlled for story comprehension. Their conversational dialogues were based on brief semi-structured interviews about the daily activities at the kindergarten.

Word tokens included only words and excluded punctuation marks, symbols, and acronyms. Linguistic disfluencies, such as hesitations, incomplete/revised words, were also excluded from the analysis (on linguistic disfluencies, see [27]). Morphological multiwords (such as Lithuanian *vos ne vos* 'hardly', *iš tikrųjų* 'in fact' or Russian *kak budto* 'as it were' *vse ravno* 'even so') were analyzed as entire units (on morphological multiword units, see [28]). All word tokens were lemmatized by means of the CLAN [29]. During the analysis, all the children (a) word tokens and (b) lemmas were classified into PoSs. The distribution of them was compared from the perspective of the language (Lithuanian *vs.* Russian) and the genre (narrative *vs.* dialogue).

3. Results

3.1. PoSP in Different Genres in Russian-speaking Children

The between-genre comparison of *word token distribution* in Russian-speaking children revealed multiple distinctions. The majority of PoSs (with the exception of adjectives, participles and prepositions) significantly discriminated narratives from conversations (Figure 1).

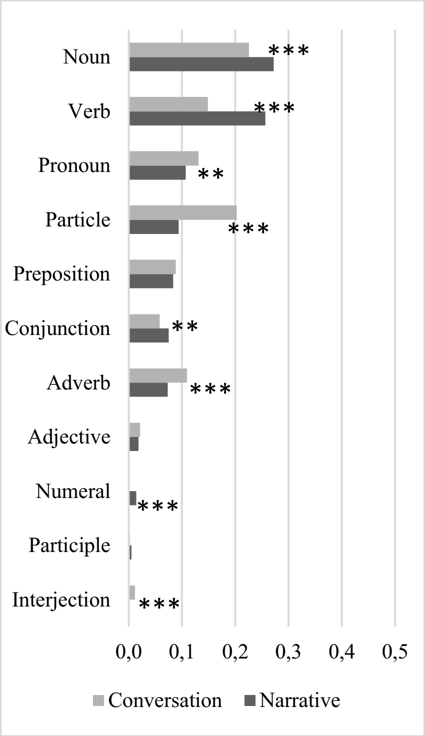

Figure 1. Distribution of *PoSs (word tokens)* in Russian-speaking discourse

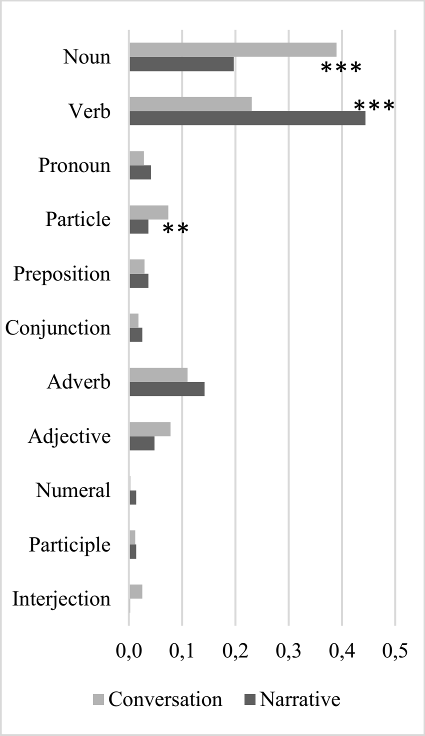

Figure 2. Distribution of *PoSs (lemmas)* in Russian-speaking discourse

Notes: *** means $p \leq 0.001$; ** means $p \leq 0.01$; * means $p \leq 0.05$

The directions of these distinctions were different: in conversations, significantly less verbs, numerals, and conjunctions, but significantly more pronouns, particles, interjections, and adverbs were produced. However, in *lemma distribution,* only three PoSs (verbs, nouns and particles) discriminated the genres (Figure 2).

3.2. PoSP in Different Genres in Lithuanian-speaking Children

Lithuanian-speaking children demonstrated partially similar PoS distribution as their Russian-speaking peers. In conversations, more adverb, pronoun, adjective, numeral, and, especially, particle *word tokens* were produced, while in narratives, nouns, verbs, and conjunctions were significantly more frequent (Figure 3).

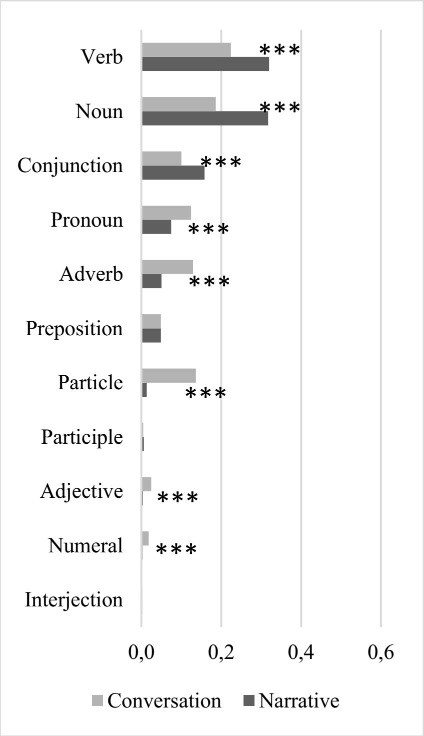

Figure 3. Distribution of *PoSs (word tokens)* in Lithuanian-speaking discourse

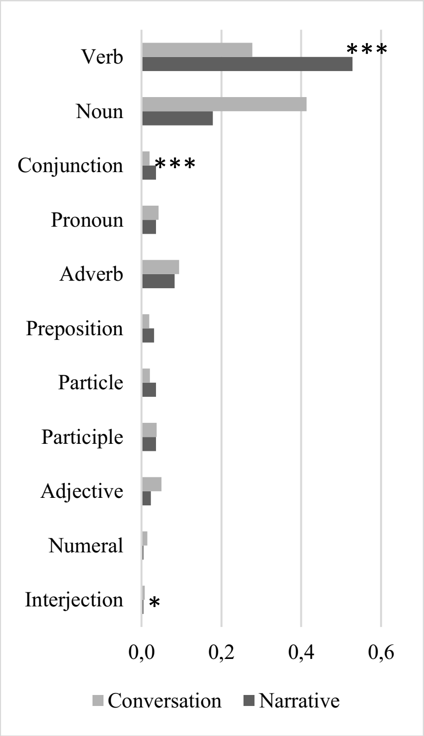

Figure 4. Distribution of *PoSs (lemmas)* in Lithuanian-speaking discourse

In *lemma distribution* (Figure 4), only interjections were more frequent in conversations, while verbs and conjunctions were more frequent in narratives.

3.3. Between-group Analysis of the PoSP

Many between-group distinctions in *word token distribution* were revealed in both genres.

In *narratives*, Lithuanian-speaking children produced more nouns, verbs, and conjunctions, whereas Russian-speaking peers produced more pronouns, prepositions, adjectives, numerals, and, especially, particles (Figure 5).

In *conversations*, the main patterns of PoS distribution were similar to narratives, but nouns were more frequent in the Russian data, while adverbs and numerals were more frequent in the Lithuanian one (Figure 6).

Between-group comparative analysis of the *lemma distribution* revealed only two distinctions in narratives (Figure 7) where adverbs and particles were more frequent in the Russian data; slightly more differences were revealed in conversations (Figure 8) where verbs were more frequent in the Lithuanian data, while particles, adjectives and interjections were more frequent in the Russian one.

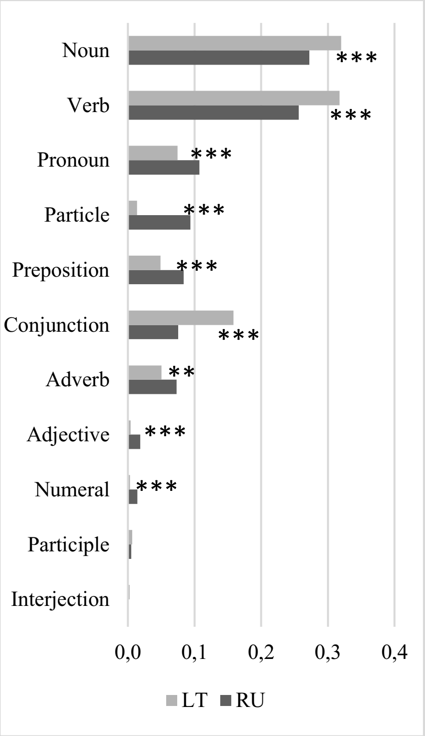

Figure 5. Distribution of *PoSs (word tokens)* in narratives

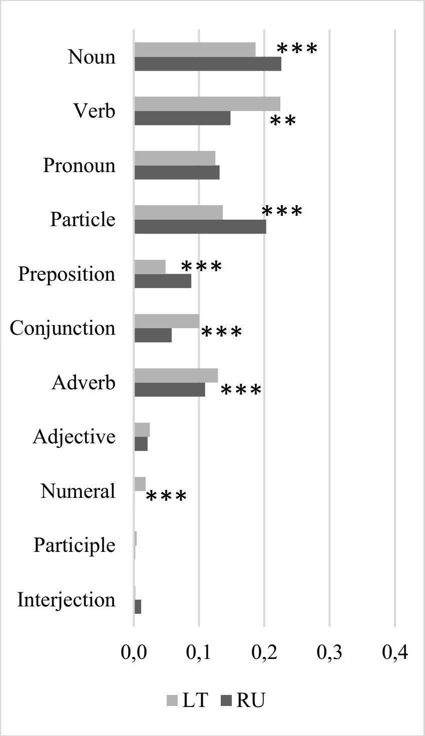

Figure 6. Distribution of *PoSs (word tokens)* in conversations

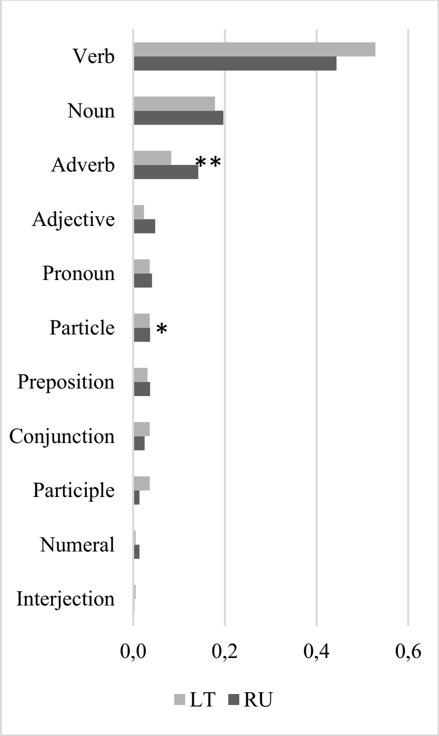

Figure 7. Distribution of *PoSs (lemmas)* in narratives

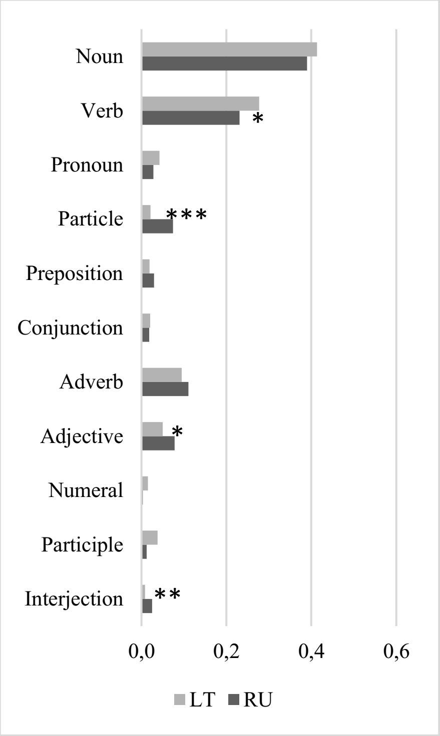

Figure 8. Distribution of *PoSs (lemmas)* in conversations

4. Conclusions and Discussion

Results of the PoS distribution analysis in Russian- and Lithuanian-speaking TD children in different discourses evidenced that PoSP was a rather sensitive measure that discriminated both genres and languages. The genre had a strong influence on the distribution of several PoS (especially, in word token items) in both languages. This probably means that genre demands govern the PoS distribution in language behavior. On the other hand, the PoS distribution in the sample of lemmas revealed much less between-genre distinctions. In other words, within the variety of different lemmas used in the data, only a few PoS discriminated narratives and conversations. It seems reasonable to consider this measure relevant to *language competence* (the variety of acquired lemmas). As for the between-group comparison, new data were obtained about discourse language distinctions. Despite the very similar PoS distribution between contemporary Russian [30] and Lithuanian [31], our study evidenced several between-group distinctions (especially, in the tokens sample). In the narratives, Lithuanian-speaking children produced significantly more nouns, verbs, conjunctions and adverbs, while Russian-speaking children were significantly more productive in pronouns, prepositions, adjectives, numerals, and particles. As for nouns, verbs, and conjunctions, this difference was close to the distinctions between the Russian [32] and Lithuanian national corpora [33]. Hence, despite a rather similar language competence, Russian and

Lithuanian children tended to recruit some PoSs in slightly different ways in different genres.

In addition to the between-group distinctions in language behavior, we found some minor distinctions in language competence. In the narratives, Lithuanian-speaking children used more verbs, while Russian-speaking children used more adverbs. In the conversations, Lithuanian-speaking children used more verbs, conjunctions, and numerals, whereas Russian-speaking peers used more particles and interjections.

In the lemma's PoSP, only two PoSs (verb and particle) discriminated the languages. To sum up, it should be concluded that genres of discourse govern the PoS distribution in both languages and this manifests in *language behavior* measures much stronger than in *language competence* measures.

In multiple publications related to child discourse development, many age-related features have been described. However, a syntactic role of the lexicon in different genre patterns still remains the least analyzed. In some quantitative studies, lexical diversity (e.g. type/token ratio) has been discussed as a language competence measure. However, the lexical (PoSs) richness and diversity (i.e. language competence) and word production (i.e. language behavior) have almost never been disentangled in the same discourse. Our results inspire an assumption that PoS variety in the mental lexicon of the narrator is not the same as the PoSs variety he/she produces in discourse. Also, distinctions between languages and the related pattern of using PoS in child discourse should be considered.

Acknowledgements

The study was supported by a research grant No. 18-18-00114 from the Russian Science Foundation. Our sincere thanks are due to the anonymous reviewers, for their helpful suggestions and comments on previous versions of this paper.

References

[1] Gvozdev AN. Formirovanie u rebenka grammatičeskogo stroja russkogo jazyka. Moskva: Akad. pedag. nauk RSFSR; 1949.

[2] Voeikova MD. Kvalitativnye semantičeskie kompleksy i ih vyraženie v sovremennom russkom literaturnom jazyke i v detskoj reči. SPb.: RGPU; 2004.

[3] Voeikova MD. Rannie ètapy usvoenija det'mi imennoj morfologii russkogo jazyka. Moskva: Znak; 2011.

[4] Dobrova GR. Ontogenez personal'nogo dejksisa (ličnye mestoimenija i terminy rodstva). SPb.: RGPU im. A.I.Gercena; 2005.

[5] Gagarina NV. Stanovlenie grammatičeskih kategorij russkogo glagola v detskoj reči. SPb.: Nauka; 2008.

[6] Ceytlin SN. Očerki po slovoobrazovaniju i formoobrazovaniju v detskoj reči. Moskva: Znak; 2009

[7] Eliseeva MB. Stanovlenie individual'noj jazykovoj sistemy rebenka. Rannie ètapy. Moskva: JaSK; 2014.

[8] Savickienė I. 2003: The Acquisition of Lithuanian Noun Morphology. Wien: Verlag der Österreichischen Akademie der Wissenschaften

[9] Wójcik P. The Acquisition of Lithuanian Verb Morphology: A Case Study. Kraków: Quartis, 2000.

[10] Kamandulytė L.Lietuvių kalbos būdvardžio įsisavinimas: leksinės ir morfosintaksinės ypatybės. Kaunas: VDU; 2009.

[11] Voeikova MD. The acquisition of case in typologically different languages. In: MD Voeikova, WU Dressler (eds) Pre- and protomorphology. Early phases of morphological development in nouns and verbs. Vienna: University of Vienna, 2007; p.

[12] Dabašinskienė I, Voeikova M. Diminutives in Spoken Lithuanian and Russian: Pragmatic functions and structural properties. In: P Arkadiev, A Holvoet, B Wiemer (eds) Contemporary approaches to Baltic linguistics. Moscow: RAS, 2015; p.203-234.

[13] Le Normand MTh, Moreno-Torres I, Parisse C, Dellatolas G. How do children acquire early grammar and build multiword utterances? A corpus study of French children aged 2 to 4. Child Development 2013;84(2):647-661.

[14] Lieven EVM, Pine JM, Baldwin G. Lexically-based learning and early grammatical development. Journal of Child Language 1997;24(1):187-219.

[15] Tomasello M. A usage-based approach to child language acquisition. Annual Meeting of the Berkeley Linguistics Society 2000;6(1):305-319.

[16] Tomasello M. Constructing a language: A usage-based approach to child language acquisition. Cambridge, MA: Harvard University Press, 2003.

[17] Labelle M. The acquisition of grammatical categories: The state of the art. In: Cohen H, Lefebvre C, editors, Handbook of Categorization in Cognitive Science. Amsterdam: Elsevier; 2005. p. 433-458.

[18] Slobin DI. Learning to think for speaking: Native language, cognition, and rhetorical style. Pragmatics 1991;1:7-26.

[19] Berman R, Slobin DI. Relating Events in Narrative: A Crosslinguistic Developmental Study. Hillsdale, NJ: Erlbaum, 1994.

[20] Rayson P. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Lancaster: Lancaster University; 2002.

[21] Rayson P, Garside R. Comparing corpora using frequency profiling. Paper presented at the Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000); 2000.

[22] Rayson P, Berridge D, Francis B. Extending the Cochran rule for the comparison of word frequencies between corpora. Paper presented at the 7es Journées internationales d'Analyse statistique des Données Textuelles; 2004.

[23] Liashevskaya ON. Korpusnye instrumenty v grammatičeskih issledovanijah russkogo jazyka. Moskva: IRYA im. V.V.Vinogradova RAN; 2014.

[24] Balčiūnienė I, Kamandulytė-Merfeldienė L. The Corpus of Lithuanian Children Language: Development and application for modern studies in language acquisition. Kalbotyra. 2018;(71):7-25.

[25] Balčiūnienė I, Kornev AN. Osobennosti ustnogo diskursa u detej 4-5 let: aprobacija novogo metoda polučenija korpusnych dannych. In: 8 meždizciplinarnyj seminar «Analiz razgovornoj russkoj reči 2019»; 2019; SpB: SpBSU; p. 31-38.

[26] Hickmann M. Children's Discourse: Person, Time and Space across Languages. Cambridge: Cambridge University Press, 2003.

[27] Balčiūnienė I, Kornev AN. Linguistic disfluency in children discourse: Language limitations or executive strategy? In: Computational Linguistic and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016»; 2016; Online resource. http://www.dialog-21.ru/media/3381/bal%C4%8Di%C5%ABnien%C4%97ikornevan.pdf.

[28] Homola P, Rimkutė E, Jarašiūnaitė G. Morfologinių samplaikų atpažinimas ir klasifikavimas. Lituanistica. 2005;(2): 58-75.

[29] MacWhinney B. The CHILDES Project: Tool for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[30] Sičinava DV. Časti reči [Parts-of-speech]. Russkaja korpusnaja grammatika. 2011; Online resource. http://rusgram.ru/

[31] Utka A. Dažninis rašytinis lietuvių kalbos žodynas 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu. Kaunas: VDU; 2009.

[32] Bogdanova-Beglarian NV, Šerstinova TYu, Baeva EM, Blinova OV, Martynenko GYa, Ermolova OB, Ryko AI. et al. Russkij jazyk povsednevnego obŝenija: osobennosti funkcionirovanija v raznyh social'nyh gruppah. SPb.: LAIKA; 2016.

[33] Dabašinskienė I. Šnekamosios lietuvių kalbos morfologinės ypatybės. Acta Linguistica Lithuanica 2009; LX:1-15.