EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 5, 67–77

CORPORA OF SPOKEN LITHUANIAN

Ineta Dabašinskienė, Laura Kamandulytė

Abstract. The paper discusses the development of spoken Lithuanian corpora. In the analytical part longitudinal child language data as well as adult conversations are discussed in view of the issues that occurred during the period of data collection, transcription and coding. The data are transcribed and coded according to the requirements of CHILDES.

The second part of the paper presents a corpus based analysis and provides preliminary results. The data of adult-directed speech, child-directed speech and child speech are analysed to reveal the frequency distribution of parts of speech. Spoken language is compared to written language in order to observe the tendencies of usage. The main differences and similarities within the spoken language registers are discussed as well.

Keywords: corpus of spoken language, grammatical annotation, grammatical disambiguation, lexicon, adult-directed speech (ADS), child-directed speech (CDS), child speech (CS), Lithuanian

1. Introduction

Spoken language research requires special preparation, which, first of all, involves the development of a corpus.

Systematic research of spoken Lithuanian is closely related to the development of child language corpora. The main aim in this field was a comparative cross-linguistic investigation of the first phases of morphology acquisition. For this purpose in 1993 we began working on a project for theory-guided research which included developing a parallel longitudinal data collection of children from different languages from about age 1;4 to at least 3;0 as well as applying identical methods of transcription, morphological coding and analysis within nearly two dozens of languages. The research was commenced by taking part in the international project "Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition" (supervised by Wolfgang U. Dressler).

The recorded speech was transcribed according to the requirements of CHILDES (MacWhinney, Snow 1990, MacWhinney 2000).[1] The transcripts were coded for morphological analysis and double-checked. Adult utterances were transcribed orthographically; children's utterances were transcribed both orthographically and phonetically. Contextual notes were inserted where necessary.

The CHILDES program consists of two tools for analysing talk: the CHAT format is used for transcription and coding of the data, and the CLAN programs are used for data analysis, such as MLU (mean length of utterance), frequencies of different linguistic elements, collocations, etc. Thus, after having starting the development of a one-child language corpus, the work has been extended not only to develop a more intensive child language longitudinal data collection, but also to embark on adult language research. The so-called spoken Lithuanian corpora[2] today consist of a morphologically coded corpus of child and child-directed speech of about 200 hours of conversations (Savickienė 1998, 2002, 2003, 2006; Balčiūnienė 2005, 2006, 2007; Kamandulytė 2005, 2006, 2007), a corpus of adult speech called "Corpus of Spoken Lithuanian" (about 80 hours of talk) (Kamandulytė, Savickienė 2008) and a small corpus of foreign talk (about 12 hours of talk) (Čubajevaitė 2006a, 2006b). All the data are transcribed and coded according to CHILDES which is now adapted to the Lithuanian language.

However, until the end of 2006 there was no corpus of Lithuanian adult speech to provide for spontaneous adult speech analysis. Some aspects of spoken TV and radio language in formal communication had been analyzed by several Lithuanian researchers (Girčienė 2004, Vaicekauskienė 2005), but systematic morphological, syntactic or lexical features of spontaneous adult-directed speech (ADS) had not been investigated until 2006.

Development of the Corpus of Spoken Lithuanian (freely available for public on the Internet)[3] started in 2006 and was funded by the Lithuanian State Science and Studies Foundation. At present a freely available corpus of morphologically coded spoken language consists of almost 50 000 grammatically annotated word forms, and at the beginning of 2009 it will be expanded up to 250 000 word forms.

2. Corpora of spoken Lithuanian

2.1. Development of corpora

Like any other type of data collection, a corpus of spontaneous speech is useful only if methods of data collection are carefully planned (McDaniel et al. 1996: 7). Therefore, the issue of methodologies applied in recording, transcribing and coding of the data was considered since the very beginning of the development of the spoken Lithuanian database. It is very important to understand that the role of the researchers involved in the project, their theoretical and methodological approach is crucial: they have to decide what to record, how to transcribe and what to mark or code in the process of corpus development. Consequently, the issue of subjectivity and particular interests of the researchers does exist.

1 Child Language Data Exchanges System, see http://childes.psy.cmu.edu/ (23.01.2009).

2 The discussion of spoken Lithuanian corpora is related mainly to the work of Vytautas Magnus University (Kaunas) team that includes Ineta Dabašinskienė (former Savickienė), Ingrida Balčiūnienė and Laura Kamandulytė. Some researchers from Vilnius University were involved in the data collection of adult spoken speech as well.

3 See http://www.vdu.lt/LTcourses/ (see MOKSLAS 'education') (23.01.2009).

2.1.1. Recording spoken language

2.1.1.1. Corpus of child language

As mentioned above, the development of a child language corpus started in 1993. Child language data were recorded by tape recorders, which later were substituted by digital recorders. Parents were instructed to record conversations in different settings and different communication situations, e.g., while bathing, cooking, eating, playing outdoors, and visiting other people.

The age range of the children and the length of the recording process were determined on the basis of linguistic purposes. If the corpus is being created for language acquisition analysis it is advisable to begin recordings with children under 2 years of age in order to catch the transition stages (McDaniel et al. 1998). The child language data were collected starting from 1;6 (with some children later) and continued until the children acquired all grammatical categories and reached a modular morphology stage (at 3–4 years) (Savickienė 2003).

While compiling a corpus it is advisable to collect more data than is actually needed to ensure that at least a certain number of relevant utterances (i.e., utterances containing constructions of a certain grammatical type) are included in every recording session. Therefore, it was decided to record conversations with children, which would be at least 15–30 minutes long, three to four times per week. Intensive and systematic recordings are crucial especially during the initial stages of language acquisition; later intervals between the recording sessions can be reduced.

The summary of the relevant information on the Lithuanian child language corpus collection is presented in Table 1.

Table 1. The structure of Lithuanian child language corpus

	Gender of child	Age range	Duration of recordings (in hours)	Number of words	Participants (number of words)
1	female	1;7–2;5	34	155 414	child (91 646) mother (58 763) father (2220) other (2785)
2	female	1;8–2;8	27	122 114	child (30 439) mother (87 847) father (3007) other (821)
3	male	2;1–4;3	14	45 902	child (15 011) mother (13 365) father (17 526)
4	male	1;6–2;7	20	71 728	child (27 610) mother (40 234) father (1364)

2.1.1.2. Corpus of spoken Lithuanian

While developing a corpus of spoken adult Lithuanian it was decided to follow the principle of balance and to record conversations, which take place in different communicative situations and settings. It was planned to compile a corpus consisting of two main parts: spontaneous speech and prepared public speech. To develop a more extensive and multi-purpose corpus, different types of communication, i.e., direct and indirect conversations, were recorded. The recordings of direct spontaneous interactions include *private* and *institutional* conversations (see Figure 1). Familiar interactions are typical for private conversations, family members or friends when speaking in an informal way. Institutional interactions are related to conversations taking place in different institutional environments: at a working place, bank, school, shop, market, etc., where speakers usually keep a distance and resort to a more formal way of communication.

Indirect or direct communication can take place in both private and institutional conversations. Indirect institutional conversations were divided into phone conversations and media speech (TV, radio). Private indirect communication is possible only while talking on the phone.

Prepared public conversations were divided into direct and indirect interactions; TV recordings were classified as indirect prepared speech, while academic discourse was regarded as direct prepared speech.

Specific features of spoken language depend not only on the situation and setting of communication, but also on the gender, age, education and occupation of the speaker. For example, adults addressing young children or old people tend to modify their language (Savickienė 2006, Kamandulytė 2006, 2007). Therefore the

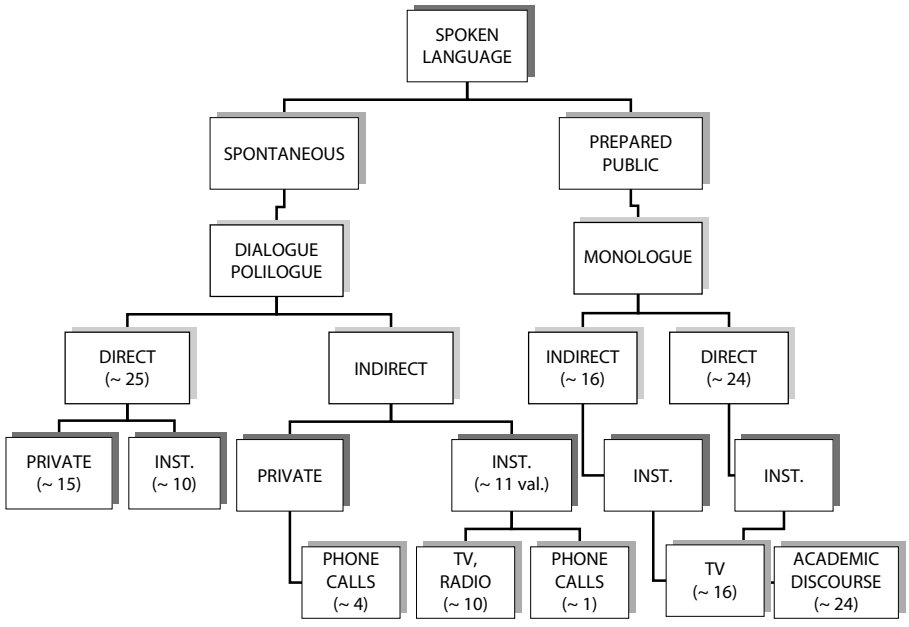

Figure 1. The structure of the Corpus of spoken adult Lithuanian.

main aim was to record speech samples according to different demographic criteria, such as gender, age, education, living place (town or countryside).

2.1.2. Transcribing and coding of spoken language

During the process of transcription and coding while using the CHILDES program some problems have emerged. These are presented and discussed below:

a) The problem of orthography. When transcribing spoken language it is possible to use either standard orthography or a phonetic representation of sounds. Phonetic representation of sounds is usually relevant for carrying out specific research, such as dialectal or phonetic analysis. For the development of spontaneous Lithuanian corpora standard orthography was chosen; in addition, phonetic transcriptions were introduced for representing child speech. Annotated grammatical forms of standard Lithuanian were presented in standard orthography because the CHILDES program can find a word in the corpus automatically and code it if that word is included in the lexicon, which incorporates words written in standard orthography. The program cannot find and code words that are not included in the lexicon.

   The creation of a lexicon and the coding process of spontaneous language are very complicated issues due to the fact that colloquial speech includes very many specific lexical and morphological features, such as shortened forms, non-standard pronunciation of certain words, jargon, slang words, etc. which may all occur during speech production. Therefore, it was decided to add the standard form of a differently pronounced word while transcribing the data, e.g. *mazas* [: mažas]; *mazas* is an incorrectly pronounced child language form. The correct form *mažas* 'small' is given in the brackets (it is the form of standard Lithuanian). Another example is a shortened form of the locative case *klasėj* [: klasėje] 'in the classroom'. Non-standard word forms were provided with standard versions put in brackets. This 'explanation' enables CHILDES to automatically find and code the form, which is written in brackets. Non-standard forms were additionally marked in order to show the real production.

b) The problem of a transcription unit. A sentence is the main syntactical unit of written language, whereas the main unit of spoken language is an utterance. An utterance is also considered to be the main unit while transcribing spoken data.

   It is not complicated to transcribe child speech because the utterances of child and child-directed speech are very short; in addition, participants of conversations usually speak slowly. However, it is not that simple to distinguish one utterance from another in spontaneous adult speech. People speak very fast, they often interrupt each other and this creates problems for distinguishing between utterances: it is difficult to decide where one utterance ends and the other begins.

   According to Crystal (2003), an utterance is a stretch of speech preceded and followed either by silence or by a change of speakers. We followed Crystal's definition and tried to identify an utterance by a pause or change of speaker, for example:

PERSON 1: *Einam* 'Let's go'
PERSON 2: *Kur?* 'Where?'
PERSON 1: *Į universitetą* 'To the university'.

The above example of Lithuanian speech consists of three utterances.

c) The third problem is morphological disambiguation. The CHILDES program (command *mor*) codes the transcribed data automatically by using the grammatically annotated lexicon that consists of the 65 000 most frequently used Lithuanian word forms. The main problems we were faced with were related to morphological disambiguation. A number of Lithuanian word forms are ambiguous and the program cannot choose the correct one from those given in the lexicon. Therefore, disambiguation should be done manually. It is not difficult to choose the correct noun or verb form, but to choose the correct version for some prepositions, particles, conjunctions and interjections is rather problematic as the meaning of such words depends on the context. In addition, morphological description of these words differs across different dictionaries.

In order to make the task of identification of some words easier the following criteria were followed:

- meaning in the context (Rimkutė 2006);
- relations with other words (Paulauskienė 1994);
- function (for example, a particle modifies word meaning, a conjunction links words, a conjunction links sentence elements, an interjection marks emotions).

2.2. Corpora-based morphological analysis

2.2.1. Parts of speech

The study is based on the analysis of spoken Lithuanian. The data of adult-directed speech (ADS), child-directed speech (CDS) and child speech (CS) are analysed to reveal the frequency distribution of parts of speech. Spoken language is compared with written language (the data provided by Rimkutė 2006 is based on the analysis of a written language corpus)[4] in order to observe the tendencies in usage (see Table 2).

Table 2. Distribution of parts of speech in spoken and written corpora

Parts of speech	CS (%)	CDS (%)	ADS (%)	Written (%)
Noun	15.9	14.6	16.2	39.4
Adjective	2.5	2.6	2.8	7.3
Pronoun	15.8	19.0	16.9	8.7
Numeral	0.2	1.4	1.9	1.0
Verb	28.1	20.3	22.8	20.5
Adverb	15.1	15.8	10.2	6.7
Preposition	3.1	4.5	4.2	4.6
Particle	10.5	13.4	12.0	3.0
Conjunction	6.2	7.0	8.2	7.6
Interjection	2.6	1.4	4.8	0.2
Total	**100**	**100**	**100**	**100**

[4] See http://donelaitis.vdu.lt (23.01.2009).

Table 2 shows that all words were classified as particular parts of speech. Ambiguous words were disambiguated according to their meaning and function in the context (see the criteria above). Unheard or unfinished and unintelligible words were excluded from the analysis.

First, we will discuss the usage of main parts of speech. The most frequent words used in spoken language (in all three registers) are verbs, whereas nouns appear more often in written language. Nevertheless, the frequency of verb tokens in spoken and written language is almost the same (around 20%).[5] This is because spoken language is much more expressive and a frequent use of verbs emphasizes this feature. The use of pronouns in spoken language shows high frequency, exceeding that of nouns. Moreover, in written language pronouns appear almost twice less frequently than in spoken language, where pronouns tend to replace nouns. Differently from written language, adverbs, particles and interjections are used much more frequently in spoken language. These words are not incidental, being related to the expressiveness, spontaneity and emotionality typical of spoken discourse. Numerals, conjunctions and prepositions show similar frequencies. Another major difference is associated with the use of adjectives: in spoken language these words are not so frequent as in the written form.

A comparison of the three registers of spoken language (ADS, CDS, CS) yields the following results: child speech is marked by a greater usage of verbs (28.1% vs. 20.3%, 22.8%), child-directed speech has a higher number of pronouns (19% vs. 15.8%, 16.9%), whereas particles, conjunctions and interjections are used more frequently in adult-directed speech. In both CDS and CS registers, a dominance of adverbs is noticed (15.1%, 15.8% vs. 10.2%).

To sum up, the usage of parts of speech has shown that differences are observed only between spoken and written forms, whereas within the spoken language register differences occurred in CDS and CD as opposed to ADS.

CDS corpus analysis revealed that this register differs a lot from ADS in terms of phonology, morphology, syntax and pragmatics (Ferguson 1977, Kempe, Brooks 2001; for Lithuanian see Savickienė 2003, Kamandulytė 2007, Wójcik 1994). Therefore, we hypothesize that a deeper investigation of one category, at least on the level of morphology, might reveal some differences in usage.

2.2.2. Case

The starting point of our analysis is the classification of cases proposed by Kuryłowicz (1964, 1977). We will thus analyse grammatical cases from the point of view of their syntactic functions, whereas concrete cases will be discussed with respect to the semantic functions they usually perform. Our hypothesis is that from a statistical point of view the frequency of occurrence of a certain case is inversely proportional to the degree of its functional markedness (Laskowksi 1989). The frequency of occurrence (in percentages) of all cases found in ADS, CDS and CS is presented in Table 3.

[5] The data are presented only in percentages because absolute numbers differ greatly.

Table 3. The frequency of case forms (in percentages)[6]: ADS vs. CDS vs. CS

Corpus	NOM	ACC	GEN	DAT	INS	LOC
ADS	33%	19%	29%	10%	5%	4%
CDS	48%	21%	19%	5%	4%	3%
CS	59%	18%	15%	4%	2%	2%

The above data clearly indicate that grammatical cases, i.e., the nominative, accusative, and genitive are much more frequent than the concrete ones, i.e., those of the dative, instrumental and locative. Thus it can be concluded that the sub-system of concrete cases, which is functionally marked, is characterised by a low frequency of occurrence. The frequency of grammatical cases differs greatly from that of concrete ones. For example, the frequency of the genitive case alone is higher than that of all concrete cases taken together. The most frequent case, then, is the unmarked nominative case, whereas the locative and the instrumental represent the cases with the lowest frequency of occurrence. Likewise, the locative case, due to its lowest frequency of occurrence, should be considered the most marked member in the case system.

Spoken language research into Slavic languages provides similar results. Thus in Russian the frequency of occurrence of cases is as follows: NOM 32.6%, ACC 25.3%, GEN 22%, DAT 4.1%, LOC 10.1%, INS 5% (Zemskaja 1979: 74). In Polish the respective numbers are as follows: NOM 34.2%, ACC 29.8%, GEN 19.2%, DAT 4.8%, INS 4.4%, LOC 7.6% (Laskowski 1989: 212).

The results obtained from the analysis of the data demonstrate that similar tendencies prevail in both CDS and CS but differ in ADS. Therefore, our further discussion will be based on the general use of cases in ADS, CDS and CS.

Despite the fact that the differences exist, some similarities can be observed, i.e., the nominative case is the most frequent among all three registers. A comparison of the nominative usage shows that the occurrence of this form in CS is 10% higher than in CDS and almost 27% higher than in ADS. The frequent use of the nominative in CDS and CS is a specific feature of these registers in the early period of language acquisition (Savickienė 2003, Voeikova, Savickienė 2001), because it is related to the process of teaching and learning. The occurrences of other case forms differ. In ADS the genitive is more frequent than the accusative, but the accusative appears more often in CDS and CS. The dative case forms are almost twice as frequent in ADS, but the usage of the instrumental and the locative seems similar in all three registers.

3. Discussion and conclusion

Spoken language research requires special preparation. In order to analyse spontaneous speech, a corpus requiring enormous human, technical and financial resources has to be developed. The process is long, time-consuming and requires accuracy, discipline and devotion on the part of the researcher, because preliminary results can be obtained only after a few years of intensive work. There is a great difference in developing spoken or written, adult-directed or child speech corpus, and first of all it is related to human efforts. The most difficult work awaits those

[6] Kuryłovicz (1964) did not separate the vocative as a discrete case, therefore we do not include the vocative either.

who are involved in child language research, which requires lots of manual work and time. Therefore even representatives of so-called big languages do not have longitudinal child language data of very many children; usually data obtained from one child are used.

Research on spontaneous language is interesting from many points of view: first, it shows the real situation of language usage and can inform about tendencies of further development; second, it creates an important source of authentic speech which can be used in translation studies, second language learning etc.; third, if the data are stored in an electronic form it ensures its availability for future studies.

From a linguistic point of view this preliminary study based on a corpus approach has shown some differences between spoken and written language, especially in the distribution of parts of speech. A deeper analysis of the category of case was carried out in order to reveal some peculiarities of different registers of spoken language. The frequency of occurrence of different case forms of specific words reflects the nature of the category of case in Lithuanian, i.e., the degree of markedness of each case. On the other hand, noun semantics is a basic factor that influences the frequency of its case forms (Laskowski 1989, Savickienė 2003). Therefore, further analysis related to the semantic analysis of the category of case is necessary to show the relation between form and meaning.

We believe that systematic, corpus-based research of spontaneous language will give more possibilities to identify, evaluate, and change the development of the Lithuanian language.

References

Balčiūnienė, Ingrida 2005. Parodomųjų įvardžių įsisavinimas. – Lituanistica, 4, 45–54.

Balčiūnienė, Ingrida 2006. Do mothers imitate their children? – Prace Bałtystyczne, 3, 19–28.

Balčiūnienė, Ingrida 2007. Kodėl tėvai kartoja vaikų pasakymus. – Gimtasis žodis, 8, 2–5.

CHILDES = Child Language Data Exchanges System. http://childes.psy.cmu.edu/ (23.01.2009).

Crystal, David 2003. A Dictionary of Linguistics and Phonetics. Malden, MA: Blackwell Publishing. doi:10.1002/9781444302776

Čubajevaitė, Laura 2006a. Verbal behaviour of Japanese students in conversational Lithuanian. – Regioninės studijos, 1, 190–199.

Čubajevaitė, Laura 2006b. Lithuanian as a foreign language. – Kalbotyra, 56 (3), 33–38.

Ferguson, Charles A. 1977. Baby talk as simplified register. – C. A. Snow, Ch. A. Ferguson (Eds.). Talking to Children: Language Input and Acquisition. Cambridge: Cambridge University Press, 209–235.

Girčienė, Jurgita 2004. Naujųjų skolinių ir jų atitikmenų konkurencija sakytinėje vartosenoje. Skoliniai ir bendrinė kalba, 120–146.

Kamandulytė, Laura 2005. Vaikiškosios kalbos registras. – Gimtoji kalba, 7, 12–16.

Kamandulytė, Laura 2006. Vaikiškosios kalbos ypatybės. – Kalbos kultūra, 79, 264–273.

Kamandulytė, Laura 2007. Morphological modifications in Lithuanian child directed speech. – Estonian Papers in Applied Linguistics, 3, 155–166.

Kamandulytė, Laura; Savickienė, Ineta 2008. The Corpus of Spoken Lithuanian: Methodology and development. – František Čermak, Rūta Marcinkevičienė, Erika Rimkutė, Jolanta Zabarskaitė (Eds.). Proceedings of the Third Baltic Conference on Human Language Technologies, Vilnius: Vytautas Magnus University, 127–135.

Kempe, Vera; Brooks, Patricia 2001. The role of diminutives in Russian gender learning: Can child-directed speech facilitate the acquisition of inflectional morphology? – Language Learning, 51, 221–256. doi:10.1111/1467-9922.00154
Kuryłowicz, Jerzy 1964. The Inflectional Categories of Indo-European. Heidelberg: Universitätsverlag.
Kuryłowicz, Jerzy 1977. Problèmes de linguistique indoeuropèenne. Wroclaw: Zakład Narodowy im. Ossolińskich.
Laskowski, Roman 1989. Markedness and the category of case in Polish. – O. M. Tomiã (Ed.). Markedness in Synchrony and Diachrony. Berlin etc.: Mouton de Gruyter, 207–226.
MacWhinney, Brian; Snow, Catherine 1990. The child language data exchange system: An update. – Journal of Child Language, 17, 457–472. doi:10.1017/S0305000900013866
MacWhinney, Brian 2000. The CHILDES Project: Tools for Analyzing Talk. Vol. I: Transcription, Format and Programs. Mahwah, NJ: Lawrence Erlbaum Associates.
McDaniel, Dana; McKee, Cecile; Smith, H. Cairns 1998. Methods for Assessing Children's Syntax. Cambridge, Massachusetts, London: The MIT Press.
Paulauskienė, Aldona 1994. Lietuvių kalbos morfologija. Vilnius: Mokslo ir enciklopedijų leidybos institutas.
Rimkutė, Erika 2006. Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne. PhD Thesis. Kaunas: VDU.
Savickienė, Ineta 1998. The acquisition of diminutives in Lithuanian. Studies in the acquisition of number and diminutive marking. – Antwerpen Papers in Linguistic, 95, 115–135.
Savickienė, Ineta 2002. The emergence of case distinctions in Lithuanian. – M. D. Voeikova, W. U. Dressler (Eds.). Pre- and Protomorphology: Early Phases of Morphological Development in Nouns and Verbs. Lincom Studies in Theoretical Linguistics 9. München: Lincom, 105–115.
Savickienė, Ineta 2003. The Acquisition of Lithuanian Noun Morphology. Wien: Verlag der Österreichischen Akademie der Wissenschaften.
Savickienė, Ineta 2006. Komunikacinė pragmatika ir kalbėjimo situacijos tikslas: deminutyvų vartojimo atvejis. – Kalbos kultūra, 79, 258–265.
Savickienė, Ineta; Dressler, Wolfgang U. 2007. The Acquisition of Diminutives. A Cross-Linguistic Perspective. Amsterdam: Benjamins.
Vaicekauskienė, Loreta 2005. Televizijos reklama – prarandamas lietuvių kalbos domenas? – Tarptautinė Jono Jablonskio konferencija: Bendrinė kalba ir visuomenė. Pranešimų tezės. 37–39.
Voeikova, Maria; Savickienė, Ineta 2001. The acquisition of the first case oppositions by a Lithuanian and a Russian child. – Wiener Linguistische Gazette, 67–69, 165–188.
Wójcik, Paweł 1994. Some characteristic features of Lithuanian baby talk. – Linguistica Baltica, 3, 71–86.
Zemskaja, Elena A. 1979. Russkaja razgovornaja retš'. Moskva: Nauka.

Ineta Dabašinskienė (Regional Studies Department at Vytautas Magnus University, Kaunas, Lithuania). Her research interests cover interdisciplinary areas such as socio- and psycholinguistics, especially first and second language acquisition, normal and impaired language development, language use and variation.
i.dabasinskiene@pmdf.vdu.lt

Laura Kamandulyte (Department of the Lithuanian Language at Vytautas Magnus University). Her research interests are corpus linguistics, psycholinguistics, first language acquisition, second language acquisition, language impairment.
l.kamandulyte@hmf.vdu.lt

LEEDU SUULISE KEELE KORPUSED

Ineta Dabašinskienė, Laura Kamandulytė
Vytautas Magnuse Ülikool

Artiklis käsitletakse leedu suulise keele korpuste arendamist. Arutletakse probleemide üle, mis kerkisid longitudinaalse lapsekeele uuringu ning täiskasvanute vestluste andmete kogumisel, transkribeerimisel ja kodeerimisel. Andmed on transkribeeritud ja kodeeritud CHILDES-i nõudmisi järgides.

Artikli teises osas tutvustatakse korpuspõhise analüüsi esialgseid tulemusi. Otsitakse sõnaliikide ja käänete sageduserinevusi kolmes suulise keele registris: täiskasvanule suunatud kõnes (ingl ADS), lapsele suunatud kõnes (ingl CDS) ja lapsekõnes (ingl CS). Tulemused näitavad olulist erinevust kirjaliku ja suulise keele vahel, mitte aga suulise keele registrite vahel. Lapsekõnes on küll rohkem verbe (28,1% – vrd 20,3% lapsele suunatud kõnes ja 22,8% täiskasvanule suunatud kõnes) ning lapsele suunatud kõnes rohkem asesõnu (19% – vrd vastavalt 15,8% ja 16,9%), samas leidub enim partikleid, side- ja hüüdsõnu täiskasvanule suunatud kõnes. Nii lapsele suunatud kui ka lapsekõnes leidub aga rohkem määrsõnu (vastavalt 15,1% ja 15,8%) kui täiskasvanule suunatud kõnes (10,2%).

Käändekasutuses ilmneb sarnasusi lapsekõne ja lapsele suunatud kõne vahel, mis eristavad neid täiskasvanule suunatud kõnest, kus on oluliselt vähem nominatiivi ja mõnevõrra rohkem genitiivi kui kahes eelmises registris. Lapsekõnes on nominatiivi koguni 27% ja lapsele suunatud kõnes 16% rohkem kui täiskasvanule suunatud kõnes. Rohke nominatiivikasutus ilmestabki neid kaht registrit keele omandamise varasel perioodil (Savickienė 2003, Voeikova, Savickienė 2001), olles seotud õpetamise ja õppimisega. Ka ühendab lapsele suunatud kõnet väike akusatiivi ülekaal genitiivi suhtes, samas kui täiskasvanule suunatud kõnes esineb genitiivi tervelt 10% rohkem kui akusatiivi. Oluliselt haruldasemat daativi leidub täiskasvanule suunatud kõnes pea kaks korda rohkem kui ülejäänud registris. Instrumentaali ega lokatiivi kasutuses erinevusi ei ilmnenud.

Võtmesõnad: suulise keele korpus, grammatiline märgendus, grammatiline ühestamine, leksikon, täiskasvanule suunatud kõne (ingl ADS), lapsele suunatud kõne (ingl CDS), lapsekõne (ingl CS), leedu keel