SAL 28/2016

Syntactically Coded Corpus of Spoken Lithuanian: Developmental Issues and Pilot Studies

Received 12/2015
Accepted 11/2016

Syntactically Coded Corpus of Spoken Lithuanian: Developmental Issues and Pilot Studies

Sintaksiškai anotuotas Sakytinės lietuvių kalbos tekstynas: metodiniai aspektai ir žvalgomieji tyrimai

COMPUTATIONAL LINGUISTICS / KOMPIUTERINĖ LINGVISTIKA

Laura Kamandulytė-Merfeldienė, Ingrida Balčiūnienė
PhD, Vytautas Magnus University, Lithuania.

http://dx.doi.org/10.5755/j01.sal.0.28.15131

Abstract

The paper deals with the main methodological issues of development of the Corpus of Spoken Lithuanian with particular attention to its syntactic coding and applications for automatized language analysis. First, we consider a methodology of development of the Corpus as well as the principles of transcribing and coding Lithuanian speech data. The main concepts, such as "utterance" "sentence", etc. are discussed. Second, we present results of a pilot study in interrogatives that are typical for natural spontaneous spoken Lithuanian. Results of the automatized analysis of interrogatives revealed that a frequency and distribution of the *Wh-* and *yes/ no questions* is rather similar. Among the *Wh- questions*, the questions non-containing the interrogative particle seem to be dominant, while the questions containing the interrogative particle at the beginning ot at the end were much rarer. Among the different functional subtypes of *Wh- questions*, adverbial ones seem to be the most freequent; among the adverbial *Wh- questions*, the spatial ones were the most frequent. Certainly, the present study is rather pilot due to the novelty of automatized syntactic approach to the data of spoken Lithuanian, thus much more complex studies still await for future investigations. A use of interrogative sentences will be studied from the perspective of different genres (e.g., monologue vs dialogue), social characteristic of the speakers, and a situation of conversation (e.g., public vs private speech). Generally, we believe that future systematic corpus-based research of spontaneous spoken language will give more possibilities to identify, evaluate, and elaborate the development of the Lithuanian language.

KEY WORDS: corpus linguistics, syntax, syntactic coding, interrogatives, Lithuanian.

ktu 1922

Research Journal
Studies about Languages
No. 28/2016
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
pp. 92-101
DOI 10.5755/j01.sal.0.28.15131

Introduction

Systematic studies in natural spoken Lithuanian started in 2006 along with development of sufficient data basis collected by a group of researchers at Vytautas Magnus University[1] and called the *Corpus of Spontaneous Spoken Lithuanian* (Dabašinskienė, Kamandulytė, 2009). (Before 2006, some aspects of spoken TV and radio language in formal communication had been analyzed by several Lithuanian researchers (Vaicekauskienė, 2005; Girčienė, Tamaševičius 2012), but systematic morphological, syntactic or lexical features of spontaneous adult communication had not been investigated due to a lack of sufficient data basis. It should be particularly noted here that the collection and analysis of natural spontaneous language data is a complicated task requiring special preparation and adequate methods of data collection, transcribing and coding. Moreover, it requires considerable expenditure of time, financial means and personal efforts. This is the main reason why there had been no systematic research in Lithuanian spontaneous speech for such a long time). Later on, the Corpus was supplemented by a new data of spontaneous speech and expanded by a data of prepared speech, and thus renamed the *Corpus of Spoken Lithuanian*[2]. Now, the freely available *Corpus of Spoken Lithuanian* (http://donelaitis.vdu.lt/sakytines-kalbos-tekstynas) consists of almost 250000 grammatically annotated word forms. Finally, during the past years, the Corpus was syntactically annotated for automatized syntactic analysis[3].

Methodology of Development of the Corpus: The First Stage

Following McDaniel et al. (1996), like any other type of data collection, a corpus of spontaneous speech is useful only if the methods of data collection have been carefully planned. Thus, the key issues of recording, transcribing, and coding of the data were considered since the very beginning of the development of the Corpus.

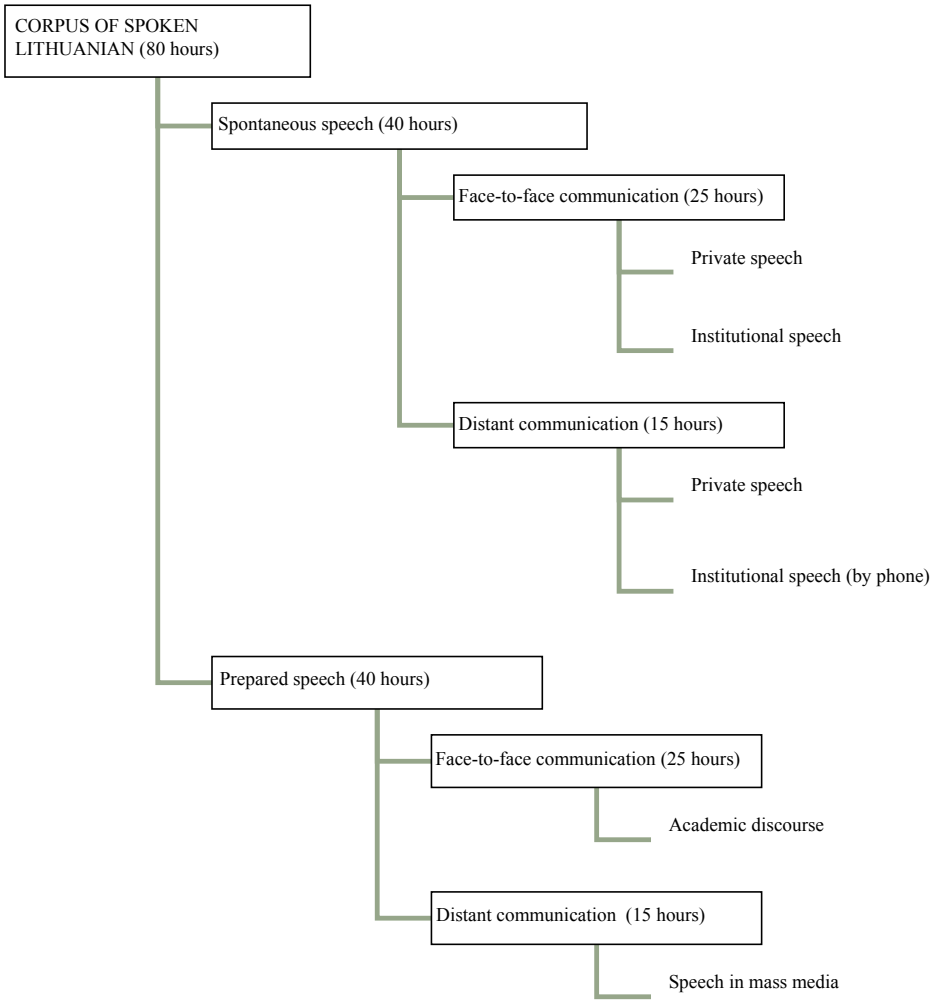
Collecting the data was based on the principles of *balance* and *naturality*. The *principle of balance* means that we aimed at developing a representative corpus of modern spoken Lithuanian that would be balanced from the perspective of a) different communication situations (such as institutional vs familiar conversations); b) different socio-economic status of the informants; and c) different situation and genre of conversation. Familiar interaction considered typical for private conversations, family members, or friends when speaking in an informal way. Institutional interaction, in contrary, take place in different institutional environments: at work, bank, school, shop, market, and other places where speakers usually keep a distance and resort to a more formal way of communication (Dabašinskienė, Kamandulytė, 2009). Certainly, specifics of spoken language depends not only on the situation and setting of communication, but also on the gender, age, education, or occupation of the speaker, e.g., adults addressing young children or old people tend to modify their language (Kamandulytė, 2006, 2007). Therefore the we aimed at collecting data with regard to different demographic criteria, such as gender, age, education, and place of residence (city/ town vs countryside). To develop even more extensive and multi-purpose data basis, different types of communication, i.e., face-to face and distant conversations (phone conversations, TV/ radio speech), were collected. Finally, the corpus data can be classified into following parts (see Figure 1).

1 The group of researchers was leaded by Prof. Ineta Dabašinskienė (former Savickienė) and included Dr. Laura Kamandulytė-Merfeldienė, Dr. Ingrida Balčiūnienė, Dr. Andrius Utka. Development of the *Corpus of Spontaneous Spoken Lithuanian* was funded by the Lithuanian State Science and Studies Foundation

2 The Corpus was developed in the framework of national project coordinated by Prof. Ineta Dabašinskienė (Vytautas Magnus University) and funded by the Lithuanian State Science and Studies Foundation. Researchers from Lithuanian academic institutions, namely, Vytautas Magnus University, Vilnius University, Institute of Lithuanian Language, Klaipėda University, and Šiauliai University, were involved for collecting the new data.

3 The syntactical annotation was completed in a framework of a national project supported by a grant No. LIT-9-11 from the Research Council of Lithuania. Also, we would like to thank our students Ieva Prameneckienė and Laura Simonavičienė for the initial syntactic annotation.

Figure 1
The structure of the Corpus



The principle of naturality was particularly respected when collecting the data. It was essential for our purposes that the speakers would not feel discomfort and could communicate naturally while recording their conversations. Therefore, it was decided to inform the speakers about recording only after the recording process ends.

Transcribing and coding the data

The recorded speech was transcribed according the CHAT (*Codes for the Human Analysis of Transcripts*) requirements of CHILDES (*Child Language Data Exchange System)* (MacWhinney, 2000). The main rules and processes of the transcription have been discussed in detail by Dabašinskienė and Kamandulytė (2009) but we still would like to emphasize an issue of speech segmentation we have faced with when transcribing the data. While a sentence is generally considered the main syntactical unit of written language, the main units of spoken language are still under discussion. Nowadays, an utterance seems to be considered the main unit while transcribing spoken data (MacWhinney, 2000), however some other units of segmentation are applied for specific purposes of the study (namely, a segmentation of a text into Communication units (Loban, 1976) are recommended for narrative analysis). In

our case, an utterance, i.e., a stretch of speech preceded and followed either by silence or by a change of speakers (Crystal, 2003), was agreed to be the main transcription unit. However, despite our previous experience in the child language transcription (Savickienė, 2003; Balčiūnienė, 2009; Kamandulytė, 2007) it was not that simple to distinguish one utterance from another in natural spontaneous adult speech. People usually speak very fast, they tend to interrupt and/ or overlap each other (Jefferson, 2004) and this cause difficulties in decision where one utterance ends and the other begins. Following Crystal (2003), we have been trying to identify an utterance by a pause or turn taking. The utterances were transcribed orthographically (phonetical transcription was not provided); contextual notes were inserted where necessary.

All the transcripts were annotated morphologically and double-checked. Morphological coding was completed following semi-automatized process. First, the transcribed data was coded automatically by searching the grammatically annotated lexicon. Then, disambiguation was completed manually and double-checked. Due to a high rate (up to 70 % of all word forms, see Rimkutė, 2003) of morphological ambiguity in Lithuanian language this stage of the Corpus development was extremely time consuming and required not only time and human resources but also special training. The main principles of morphological coding of the *Corpus of Spoken Lithuanian* and methodological discussions can be found in the papers of Kamandulytė and Savickienė (2008), Dabašinskienė and Kamandulytė (2009).

The development of the *Corpus of Spoken Lithuanian* has lead to a constant increase in studies on adult spontaneous communication. Various papers have dealt with a distribution of parts of speech in different registers and types of spoken language (Kamandulytė, Tuškevičiūtė, 2009), forms and functions of diminutives (Dabašinskienė, 2009a), distribution and use of different grammatical forms (Dabašinskienė, 2008), morphological features of spoken Lithuanian (Dabašinskienė, 2008, 2009b), variation of inflectional paradigms in the new morphologically integrated loanwords (Kamandulytė-Merfeldienė, 2010), forms and distribution of fillers (Kamandulytė-Merfeldienė, 2014).

Methodology of Development of the Corpus: The Second Stage

As it was mentioned before, syntactic analysis of spoken Lithuanian has long been limited due to a lack of sufficient data basis. In 2015, when the syntactic annotation of the Corpus started, we faced again a problem of speech segmentation. Since the data of the Corpus was segmented into utterances, the boundaries of the utterances did not necessarily match the boundaries of the sentence, e.g. (1):

(1) *INF1: *Žiūriu*.
"[I am] looking."
*INF1: *Ko jinai tą daikt--*.
"Why she that thing: UNFINISHED."
*INF1: *Ko jinai tą lėkštę čia neša*.
"Why she brings that plate."

Thus, before the syntactic annotation, the text first was re-segmented into syntactic units, as exemplified below (2):

(2) *INF1: *Žiūriu, ko jinai tą daikt-- ko jinai tą lėkštę čia neša.*
"[I am] looking, why she that thing:UNFINISHED, why she brings that plate."

During the syntactic annotation, a special syntactic line (%syn) as counterpart to the main text line was generated for each of the utterances, e.g. (3):

(3) *INF1: *Žiūriu, ko jinai tą daikt-- ko jinai tą lėkštę čia neša.*

"[I am] looking, why she that thing:UNFINISHED, why she brings that plate."

%syn: d:cs:com|kas=ko/ kas=ko

*INF1: *Žiūri į mano lėkštę, į viršų.*

"[She is] looking at my plate, upwards."

%syn: d:ss

*INF1: *Tą pasiėmė.*

"[She] took that."

%syn: d:ss

For the syntactic annotation, the following categories were suggested. Communicative type was identified as a) declarative, b) exclamative, c) imperative, or d) question. According the structure, the sentences was encoded as simple or composite ones, and the latter were further encoded as a) compound sentences, b) complex sentences, c) mixed-type (i.e., compound-complex) sentences, or d) asyndetic sentences. As for the complex sentences, their subordinated clause was encoded according its function, i.e., subject, object, attribute, or adverbial one. Now, the syntactically annotated corpus data enables for automatized approach to syntactic analysis of the spoken Lithuanian.

Methodology of a Corpus-based Analysis of Interrogatives in Lithuanian Spontaneous Speech

The present study migh be considered the first attempt to examine a variety of forms and functions of interrogative sentences in Lithuanian natural spontaneous speech. Due to a lack of similar studies, we could only hypothesize that a distribution of interrogative sentences might be different between written and spoken Lithuanian. Thus the main aim of the study was to compare theoretical models of interrogative taxonomy with a distribution of interrogative sentences in spontanous spoken Lithuanian.

According the Modern Lithuanian Grammar (2005), the Functional Lithuanian Grammar (Valeckienė, 1998), the Practical Lithuanian Grammar (Ramonienė, Pribušauskaitė, 2008), the Modern Lithuanian Syntax (Balkevičius, 1963), and the Lithuanian Syntax (Labutis, 2002), interrogatives were opposed to declaratives, imperatives, and exclamatives as one of communicative types of utterances. Then all the interrogatives were encoded as particular functional and structural types. Due to a lack of complex studies in Lithuanian syntax (especially in the syntax typical for the spoken Lithuanian), different papers have provided contradictory statements on the function of interrogatives. E.g., Balkevičius (1963) and Labutis (2002) have classified interrogatives into so called "clarification questions" (they match *yes/ no questions* according the English terminology) and "special questions" (*Wh- questions*, consequently). In the Modern Lithuanian Grammar (2005), the same classification has been applied but the term "special question" has been replaced by the "complementary question" and three more type of interrogatives, namely, the alternative questions, the rhetorical questions (that "do not require and answer" p.580) and the "title question" (that function "as an announcement of the following topic", p.580) have been added. Obviously, these papers (Balkevičius, 1963; Labutis, 2002; Ambrazas (ed.), 2005) have focused on the speaker's intention, e.g., to receive some information, to check his/ her own knowledge, to express his/ her feelings, or to announce a new topic of discourse. The Functional Lithuanian Grammar (Valeckienė, 1998) has emphasized not only intention of the speaker but also a way the listener should respond to the question. E.g., the "clarification questions" have been defined as those which require to confirm/ disconfirm an information provided by the speaker; the "content questions" (*Wh- questions* according the English terminology), consequently, require to give more additional information on the topic of

the question. In the Practical Lithuanian Grammar (Ramonienė, Pribušauskaitė, 2008), interrogatives have been classified into the "real questions" (that are further divided into "clarification questions" (*yes/ no questions* according the English terminology) and "concrete questions" (*Wh- questions* according the English terminology). The variety of terms for different functional types of interrogatives, one one hand, illustrates the main gaps in the theory of Lithuanian grammar but, on the other hand, highlights the main aspects (namely, communicative and pragmatic ones) of a methodological approach to the interrogative analysis. The structure and form of interrogatives in Lithuanian is not as complicated as the function. Despite various terms for the *Wh- questions*, they usually are classified into questions containing vs non-containing interrogative particle such as *ar* "so", *be* "probably", *bene* "probably", *gal* "maybe", *galgi* "maybe", *negi* "indeed", *nejaugi* "indeed". The tag questions are also mentioned as a form of interrogatives which explicitly encourages for responding (Valeckienė, 1998). Taking into account the given theoretical background, all the interrogatives found in the Corpus (4596 in total) were classified into *yes/ no questions* and *Wh- questions* (see Figure 2).

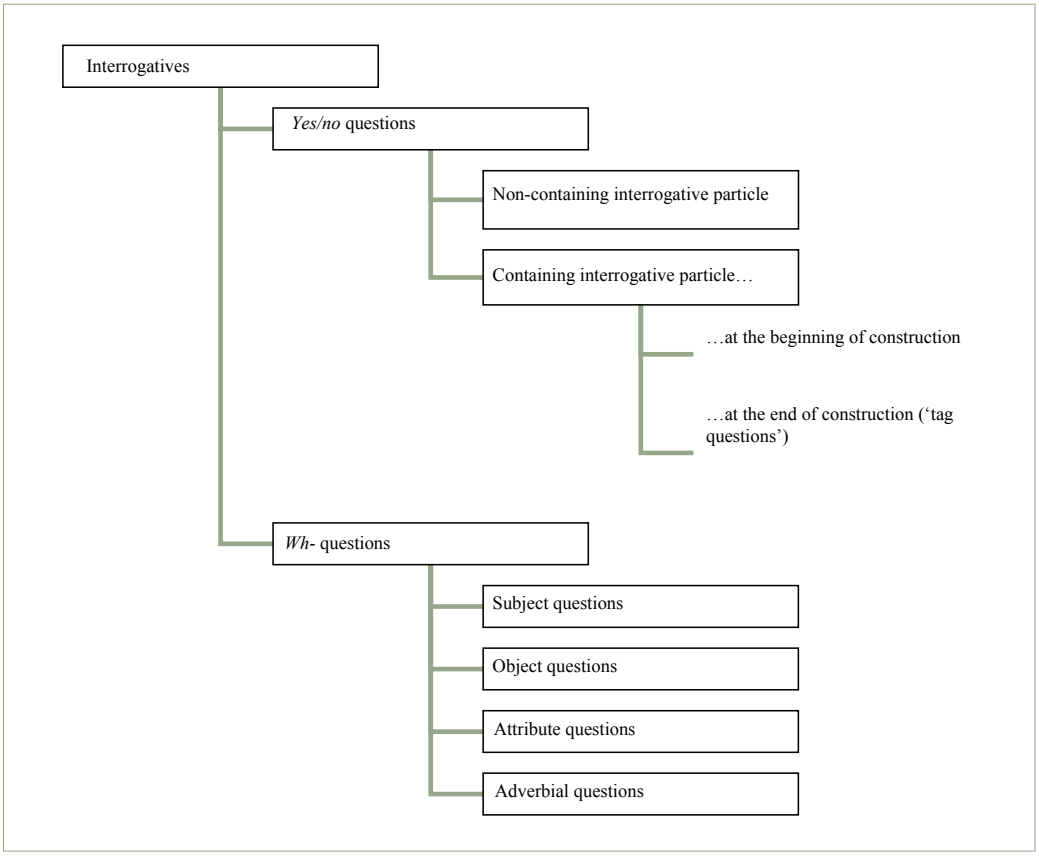


Figure 2
Classification of interrogatives

The *yes/ no questions* were further encoded as the questions (4) with vs (5) without an interrogative particle; among the questions with interrogative particles, a distinction between the questions with the particle (6) at the beginning vs (7) at the end of the question was considered. The *Wh-questions* were encoded as the (8) subject, (9) object, (10) attribute, and (11) adverbial questions.

(4) *Ar tai labai didelė būtų problema?*

"**So** would this be a big problem? "

(5) *Galėtum tepalus išpilti?*

"Could [you] remove the oil? "

(6) *O tai gal turi skaičiavimo mašinėlę?*

"So maybe you have a calculator?"

(7) *Įspūdžių daug, ar ne?*

"A lot of impressions, right?"

(8) *O kas čia duos pinigų?*

"And who [will] give [us] money here?"

(9) *Ką čia dedam(e)?*

"What [will we] place here?"

(10) *Kokia čia muzika?*

"What [kind of] music [is] here?"

(11) *O kada išpurškei?*

"And when [did you] spray [it]?"

During the automatized analysis, frequency and distribution of each of the (sub-) types of the interrogatives was measured and compared among different parts of the Corpus such as spontaneous vs prepared speech; private vs public speech, etc. Due to the limited size of the current paper, the interrogatives typical for only the spontaneous spoken Lithuanian will be discussed further.

Results of the Pilot Study

Automatized syntactic analysis evidenced that a frequency and distribution of the *Wh-* and *yes/ no questions* is rather similar. Namely, we found 578 *Wh- questions* and 455 *yes/ no questions* within the data of spontaneous speech. A slight dominance of the yes/ no questions may lead to a prediction that during the spontaneous conversation, interlocutors tend to clarify (to confirm or disconfirm) their own statements rather than to elicit a new information. In some cases, mixed-type questions, i.e., constructions including both *Wh-* and *yes/ no* pattern, were observed, e.g. (12):

(12a) *Kiek ten mililitrų, dešimt?*

"How many milliliters there, ten?"

(12b) *Čia dabar jisai perka kokius, stacionarinius?*

"What [computers] does he buy, the PCs?"

In such cases, the *Wh-* pattern usually preceded the *yes/ no* pattern.

Among the *Wh- questions*, the questions non-containing the interrogative particle seem to be dominant (449 occurrences), while the questions containing the interrogative particle (16, 17, 18) were much rarer (129 occurrences). Among the latter structures, the tag questions seem to be less frequent (45 occurrences) than the questions containing the interrogative particle at the beginning of the sentence (84 occurrences). A distribution of the interrogative particles was not significant: the particle *gal* "maybe" occurred at the beginning of 38 questions and the particle *ar* "so" occured at the beginning of 36 questions. Other interrogative particles such as *be* "probably", *bene* "probably", *galgi* "maybe", *negi* "indeed", *nejaugi* "indeed" were not observed within the data of spontaneous speech and thus should be considered as more typical for the written than for the spoken Lithuanian. The interrogatives tagged at the end of the question were more diverse: *ar ne* "isn't it", *ane* "isn't it", *ne* "not", and *taip* "yes".

Among the different functional subtypes of *Wh- questions*, adverbial ones (215 occurrences) seem to be the most freequent. The frequency of the adverbial questions might be partially explained by a need to receive more detailed and/ or additional information (time, place, cause, etc.) on a topic of conversation. Still, this sub-type is the most productive among other *Wh- questions*. The adverbial questions can be further divided into (13) causal, (14) temporal, (15) intentional, (16) spatial, (17) manner, and (18) quantity questions:

(13) *Tai kodėl vėluoja autobusas?*

"So why [is] the bus late?"

(14) *O kada eisi?*

"So when [will you] go?"

(15) *O tai kam tau jo reikia?*

"So why do you need this?"

(16) *Iš kur gavai?*

"Where [did you] get [it]?"

(17) *Kaip čia tariamas?*

"How to pronounce it?"

(18) *Kiek aš pinigų esu* [gavęs]*?*

"How much money have I [received]?"

Spatial and manner questions seem to be the most frequent (64 and 59 occurrences respectively) among the adverbial *Wh- questions*.

Object questions were in the second most frequent (144 occurrences) functional subtype of the *Wh- questions*. Among them, the direct object questions seem to be dominant (93 occurrences), while the indirect object questions were much rarer (51 occurrences).

Subject and attribute questions were the rarest (48 occurrences of each of the subtype) among the functional subtypes of the *Wh- questions*.

Methodology of Development of the Corpus: The Second Stage

The results of the study revealed almost equal number of the *yes/ no* and *Wh- questions* in the spoken Lithuanian. Among the *yes/ no questions*, those without the interrogative particles were dominant. Among the functional subtypes of the *Wh- questions*, the adverbial questions, especially, the spatial ones, were the most frequent. Certainly, the present study is rather pilot due to the novelty of automatized syntactic approach to the data of spoken Lithuanian, thus much more complex studies still await for future investigations. Namely, a use of interrogative sentences should be studied from the perspective of different genres (e.g., monologue vs dialogue), social characteristic of the speakers, and a situation of conversation (e.g., public vs private speech). Following previous studies based on English data (Tracy, Robles, 2013), we presume that interrogative sentences may be more numerous in women than in men conversations and that conversational discourse provokes more interrogatives than does narrative discourse. Research on spontaneous spoken language is inspiring and promising from at least a few points of view: first, it reflects the real situation of language usage and can inform about tendencies of its further development; second, its results can serve as a reliable source of authentic speech which can be used in translation studies, second language learning, etc.; finally, the data stored in the digital form ensures its availability for future studies. Thus, generally, we believe that future systematic corpus-based research of spontaneous spoken language will give more possibilities to identify, evaluate, and elaborate the development of the Lithuanian language.

References

1. Ambrazas, V. (ed.), 2005. Dabartinės lietuvių kalbos gramatika. Vilnius: Mokslo ir enciklopedijų leidybos institutas.
2. Balčiūnienė, I., 2009. Pokalbio struktūros analizė kalbos įsisavinimo požiūriu. Kaunas: Vytautas Magnus University.
3. Balkevičius, J., 1963. Dabartinė lietuvių kalbos sintaksė. Vilnius: Valstybinė politinės ir mokslinės literatūros leidykla.
4. Crystal, D., 2003. A Dictionary of Linguistics and Phonetics. Malden, MA: Blackwell Publishing.
5. Dabašinskienė, I., 2008.Trumpinimas ir dažnumo poveikis šnekamojoje kalboje. In: Darbai ir dienos, no 50, pp.109–117.
6. Dabašinskienė, I., 2009a. Intimacy, Familiarity and Formality: A Case of Diminutives in Modern Lithuanian. In: Lituanus, No 55.
7. Dabašinskienė, I., 2009b. Šnekamosios lietuvių kalbos morfologinės ypatybės. In: Acta Linguistica Lithuanica, no 60, pp. 1–15.
8. Dabašinskienė, I; Kamandulytė, L., 2009. Corpora of Spoken Lithuanian. In: Estonian Papers in Applied Linguistics, no 5, pp.67–77. http://dx.doi.org/10.5128/erya5.05
9. Girčienė, J.; Tamaševičius, G., 2012. Five Decades of Television: from Language Homophony to Polyphony. In: Lituanus. The Lithuanian Quarterly Journal of Arts and Sciences, no 58 (2), pp. 31–43.
10. Jefferson, G., 2004. A Sketch of Some Orderly Aspects of Overlap in Natural Conversation. In: Conversation Analysis. Studies from the First Generation. Lerner, G. H. (ed.). Amsterdam/ Philadelphia: John Benjamins Publishing Company. http://dx.doi.org/10.1075/pbns.125.05jef
11. Kamandulytė, L., 2006. Vaikiškosios kalbos ypatybės. In: Kalbos kultūra, no 79, pp.264–273.
12. Kamandulytė, L., 2007. Morphological Modifications in Lithuanian Child Directed Speech. In: Estonian Papers in Applied Linguistics, no 3, pp.155–166. http://dx.doi.org/10.5128/erya3.10
13. Kamandulytė-Merfeldienė, L., 2010. Daiktavardžio paradigmų produktyvumas: skolinių morfologinio įforminimo ir fleksijų varijavimo analizė. In: Lietuvių kalba, no 4.
14. Kamandulytė-Merfeldienė, L., 2014. Pertarų dažnumas ir įvairovė sakytinėje lietuvių kalboje. In: Bendrinė kalba, no 87.
15. Kamandulytė, L.; Tuškevičiūtė, M., 2009. Būdvardžio vartojimo skirtumai sakytinės kalbos registruose. In: Darbai ir dienos, no 50, pp.91–108.
16. Kamandulytė, L.; Savickienė, I., 2008. The Corpus of Spoken Lithuanian: Methodology and Development. In: Proceedings of the Third Baltic Conference on Human Language Technologies Čermak, F. et al. (eds.). Vilnius: Vytautas Magnus University, pp.127–135.
17. Labutis, V., 2002. Lietuvių kalbos sintaksė. Vilnius: Vilniaus universiteto leidykla.
18. Loban, W., 1976. Language Development: Kindergarten through Grade Twelve. National Council of Teachers of English, Urbana, Ill.
19. MacWhinney, B., 2000. The CHILDES Project: Tools for Analyzing Talk, vol. I: Transcription, Format and Programs. Mahwah, NJ: Lawrence Erlbaum Associates.
20. McDaniel, D.; Caims, H.; McKee, C., (Eds.), 1996. Methods for Assessing Children's Syntax. Cambridge, MA: MIT Press.
21. Ramonienė, M.; Pribušauskaitė, J., 2008. Praktinė lietuvių kalbos gramatika. Vilnius: Baltos lankos.
22. Rimkutė, E., 2003. Morfologinio daugiareikšmiškumo tipologija. In: Lituanistica, no 4 (56), pp.60–78.
23. Savickienė, I., 2003. The Acquisition of Lithuanian Noun Morphology. Wien: Verlag der Österreichischen Akademie der Wissenschaften.
24. Tracy, K.; Robles, J. S., 2013. Everyday Talk: Building and Reflecting Identities. NY, London: The Guilford Press.
25. Vaicekauskienė, L., 2005. Televizijos reklama – prarandamas lietuvių kalbos domenas? In: Bendrinė kalba ir visuomenė, pp.37–39.
26. Valeckienė, A., 1998. Funkcinė lietuvių kalbos gramatika. Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Santrauka

Laura Kamandulytė-Merfeldienė, Ingrida Balčiūnienė. Sintaksiškai anotuotas Sakytinės lietuvių kalbos tekstynas: metodiniai aspektai ir žvalgomieji tyrimai

Straipsnyje pristatoma *Sakytinės lietuvių kalbos tekstyno* (VDU) kūrimo ir sintaksinio anotavimo metodika, aptariamos automatizuotos sintaksinės analizės galimybės. Pirmojoje straipsnio dalyje supažindinama su *Tekstyno* kūrimo ir tobulinimo metodika bei etapais, aptariamos esminės sintaksinio kodavimo sąvokos. Antrojoje dalyje pristatomi vieno iš žvalgomųjų tyrimų, sutelkto į spontaninės sakytinės lietuvių kalbos klausiamųjų sakinių vartoseną, rezultatai. Atlikus automatizuotą *Tekstyne* užfiksuotų klausiamųjų sakinių analizę, paaiškėjo, tikrinamojo ir specialiojo klausimo sakiniai tekstyne pasiskirsto daugmaž tolygiai. Tarp tikrinamojo klausimo sakinių (angl. *yes/ no questions*) vyrauja klausimai be klausiamosios dalelytės, rečiau vartojami klausimai su klausiamąja dalelyte sakinio pradžioje ar pabaigoje. Tarp specialiojo klausimo sakinių (angl. *Wh-? questions*) vyrauja aplinkybės (ypač – vietos) klausimai. Suprantama, žvalgomasis tyrimas atskleidė tik esmines klausiamųjų sakinių, vartojamų spontaninėje lietuvių kalboje, ypatybes, tad ateityje planuojama šį tyrimą išplėsti tarpusavyje lyginant atskiras tekstyno dalis ir ieškant žanro (pvz., monologo *vs*. dialogo), kalbėtojo socialinių charakteristikų bei pokalbio situacijos (pvz., viešosios *vs*. privačios kalbos) poveikio klausiamųjų sakinių vartosenai. Sintaksiškai anotavus *Sakytinės lietuvių kalbos tekstyną*, atsivėrė galimybė atlikti automatizuotą sintaksinę šios duomenų bazės analizę, tad tikimasi ateityje išplėtoti kiekybinius natūralios sakytinės lietuvių kalbos sintaksės tyrimus.

About the Authors

Laura Kamandulytė-Merfeldienė

PhD, Vytautas Magnus University, Lithuania.

Academic interests

Corpus linguistics, first language acquisition, and grammar.

Address

Vytautas Magnus University, K. Donelaičio 58, 44248 Kaunas, Lithuania.

E-mail:

l.kamandulyte-merfeldiene@hmf.vdu.lt

Ingrida Balčiūnienė

PhD, Vytautas Magnus University, Lithuania.

Academic interests

First language acquisition, narrative analysis, and discourse analysis.

Address

Vytautas Magnus University, K. Donelaičio 58, 44248 Kaunas, Lithuania / Saint-Petersburg State Pediatric Medicine University, Litovskaya 2, 194100 Saint-Petersburg, Russian Federation.

E-mail:

i.balciuniene@hmf.vdu.lt