World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:11, No:4, 2017

Grammatically Coded Corpus of Spoken Lithuanian: Methodology and Development

L. Kamandulytė-Merfeldienė

Open Science Index, Cognitive and Language Sciences Vol:11, No:4, 2017 publications.waset.org/10006833.pdf

Abstract**—**The paper deals with the main issues of methodology of the *Corpus of Spoken Lithuanian* which was started to be developed in 2006. At present, the corpus consists of 300,000 grammatically annotated word forms. The creation of the corpus consists of three main stages: collecting the data, the transcription of the recorded data, and the grammatical annotation. Collecting the data was based on the principles of balance and naturality. The recorded speech was transcribed according to the CHAT requirements of CHILDES. The transcripts were double-checked and annotated grammatically using CHILDES. The development of the Corpus of Spoken Lithuanian has led to the constant increase in studies on spontaneous communication, and various papers have dealt with a distribution of parts of speech, use of different grammatical forms, variation of inflectional paradigms, distribution of fillers, syntactic functions of adjectives, the mean length of utterances.

Keywords**—**CHILDES, Corpus of Spoken Lithuanian, grammatical annotation, grammatical disambiguation, lexicon, Lithuanian.

I. Introduction

OVER the past two decades, spoken-language corpora have been increasingly used to gain more direct insights into human communication, since they reflect the actual use of language in everyday situations [1]. Spoken language has been the object of many previously published studies [2]-[8]. Recently, due to a rapid development of technology, investigations into spoken language have improved in quality and diversity. Improved computers and newly-created programs allow researchers to store speech databases and corpora of different sizes, to apply different analysis tools, and to conduct various researches quickly and efficiently. Taking advantage of such possibilities, many studies have used speech databases and corpora as representatives of language reflecting features of language in the best way. However, it should be noted that most of the studies are restricted to the analysis of the speech of specially selected informants recorded under laboratory conditions or the analysis of public discourse (lectures, public speeches, TV or radio speech) leaving aside spontaneous, natural speech. This might be explained by the fact that the collection and processing of the data of spontaneous speech, which is usually done manually, are extremely time consuming and very complicated. Thus, even today, all over the world, the investigation of spontaneous speech is technologically and methodologically challenged.

In order to conduct representative research into spontaneous language, it is necessary to have a database of sufficient scope, which would include digitized sound recordings of spontaneous speech processed with special programs. The creation of such a database is obviously a tremendous and time-consuming work requiring great financial and human resources.

The grammatically coded Corpus of Spoken Lithuanian [9] was started to be developed in 2006 and used the CHAT transcription format and CLAN programs of the database CHILDES (Child Language Data Exchange System, [10], [11]). In 2006, Dabašinskienė, Utka and Kamandulytė-Merfeldienė adapted the program CHILDES for the Lithuanian language and used it to develop the first grammatically annotated corpus of spoken adult speech [12]. The corpus was developed in the framework of national projects coordinated by Vytautas Magnus University (Kaunas, Lithuania) and supported by the Lithuanian State Science and Studies Foundation (grant No. L-12/2008) and by the Research Council of Lithuania (grants No. LIT-9-11, No. LIP-085/2016). Currently, the corpus consists of 300,000 grammatically annotated word forms. There are two primary aims of this study: to present the Corpus of Spoken Lithuanian and to explain the development of its methodology.

II. Methodology and Development of a Grammatically Coded Corpus

This section describes the procedure and the main stages in the development of the Corpus of Spoken Lithuanian: the collection of the data, the transcription of the recorded data, and the grammatical annotation.

A. Procedure

The data of the Corpus of Spoken Lithuanian were collected in several stages covering the following two periods: 2006-2010 and 2015-2017. A number of linguists from different universities of Lithuania belonging to different regions of Lithuania (Klaipėda University, and Šiauliai University from Western Lithuania, Vytautas Magnus University from Middle Lithuania, Vilnius University, and Institute of Lithuanian Language from Eastern Lithuania) participated in the collection of the data for the corpus. Trained linguists (local coordinators) were responsible for groups of several people who were asked to collect samples of spoken language including spontaneous speech and prepared public speech. Most of the collectors of speech samples were students from different fields as well as the family members and friends of the students and local coordinators. Recordings were made in different regions of Lithuania, i.e. in cities/towns and

L. K. M. is with the Faculty of Humanities, Vytautas Magnus University, Kaunas, Lithuania (e-mail: laura.kamandulyte@vdu.lt).

countryside, using high-quality dictaphones. However, during the collection of samples of spoken language, dialects, which differ greatly from the standard language, were outside the focus of attention. The sessions of recordings differ in their length ranging from several minutes to half an hour. Recordings include conversations which take place in different settings, such as universities, shops, hospitals, schools, churches, a hairdresser's shop, as well as the home, and other institutions and enterprises. A part of the recordings is made of TV talk shows. Currently, the Corpus of Spoken Lithuanian consists of 232 conversations involving 998 speakers (552 women, 446 men). The age of the speakers range from 3 years to 81 years, while the largest group of informants belong to the age range from 25 years to 50 years.

B. The Collection of the Data

The collection of the data was based on the principles of ***naturality*** and ***balance***. *The principle of naturality* was particularly respected when collecting the data [13]. 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 the recording only after the recording process ends (see more [13]).

In order to create a *balanced* and multi-purpose database, it was decided to include conversations taking into consideration the following aspects: a) different communication situations (such as public speech vs. private speech; institutional conversations vs. familiar conversations; formal speech vs. informal speech); b) different demographic criteria and socio-economic status of the informants [13].

1) **Different communication situations.** Spoken language includes various communicative situations. According to these situations, spoken language can be divided into two varieties: spontaneous private speech (informal communication) and prepared public speech (formal communication). As opposed to the prepared public speech, spontaneous speech shows a number of differences in different language levels, such as lexis, morphology, and syntax [14]–[16] etc. It is significant to distinguish spontaneous conversations from public speeches not only for linguistic, but also for psycholinguistic and sociolinguistic purposes. When speaking publicly, people always try to control their speech, i.e. they try to use the standard language, to speak correctly, and to respect the requirements of the polite conversation. Therefore, recordings of public speeches can reveal the status of the standard language and help to determine the most typical violations of standard language rules. Nonetheless, it should be remembered that even though public speech is prepared in advance, it does not loose spontaneity, since a thought, even if it is contemplated, is formulated at the moment of speech delivery and its expression is unique.

When communicating in a private environment with familiar people, such as friends, relatives, family members, and colleagues, people tend to control their language less, they communicate more freely and boldly, and they do not avoid using dialects, professional jargon, or slang. Therefore, recordings of spontaneous conversations are valuable for the study of the interaction between the standard language, dialects and sociolects, the change of codes and social roles as well as the structure of natural conversation.

Thus, when creating a balanced corpus, it was decided to collect the data of spontaneous private speech and prepared public speech, since the analysis of such data is informative and revealing not only from the linguistic but also from the sociolinguistic and psycholinguistic perspective. The recorded conversations also include institutional and familiar conversations. Familiar interaction is typical of private conversations between family members or friends speaking in an informal way, whereas institutional interaction takes place in different institutional environments, such as work places, banks, schools, shops, markets, and other places where speakers usually tend to keep a distance and resort to a more formal way of communication. For the database to be even more extensive and multi-purpose, different types of conversations, i.e. face-to-face and distant conversations (phone conversations, TV/radio speech), were collected. Finally, the corpus data can be classified into spontaneous conversations and prepared conversations; face-to-face communication and distant communication [13] (see Fig. 1).

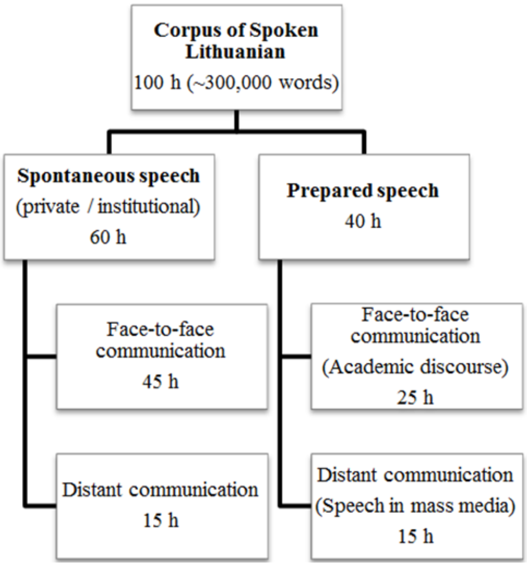


Fig. 1 The structure of the Corpus of Spoken Lithuanian

2) **Different demographic criteria and socio-economic status of the informants.** Certainly, specific features of spoken language depend not only on the situation and setting of communication but also on the gender, age, education, or occupation of the speaker. For example, adults addressing young children or old people tend to modify their language [17]. For this reason, the data were collected taking into consideration different demographic criteria, such as gender, age, education, and place of residence (city/town vs. countryside) [12].

Most features of language are determined not only by the type of conversation, the gender, age, or occupation of the speaker but also by the social role of the situational speaker. As individuals switch between their roles, their speech and code change, too. For example, the role of an office worker or a director is mainly associated with a formal subject-specific style of language, whereas usual means of careless spoken language are more attributable to the roles of co-workers or buyers. The social role of a friend, in its turn, is characterized by a familiar and colloquial speech. Accordingly, it was decided that conversations collected for the corpus should encompass socially different levels, i.e. conversations should be recorded in different settings and different situations.

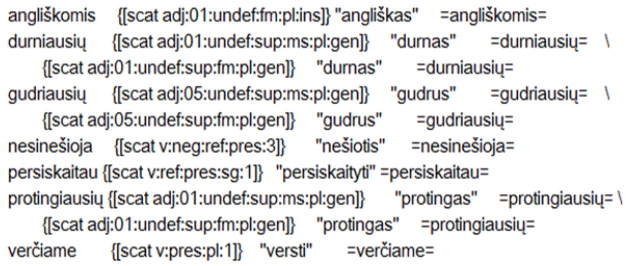
C. Transcription of the Recorded Data

The recorded speech was transcribed in accordance with the CHAT (*Codes for the Human Analysis of Transcripts*) requirements of CHILDES [10]. The transcription was accomplished by researchers-linguists participating in the projects intended for the development of the corpus. Each transcript was double-checked by other experienced researchers responsible for the quality of the Corpus of Spoken Lithuanian. The main rules and processes of the transcription have been discussed in detail by [12], so only the major methodical problem of the transcription of recordings related to the segmentation of speech will be presented herein.

While a sentence is generally considered to be the main syntactical unit of written language, the main units of spoken language are still under discussion [13]. However, nowadays, when transcribing and analyzing spoken data, it is the utterance that seems to be regarded as the main unit of spoken language. Although only a few authors discuss differences between a sentence and an utterance, it is obvious that the usage of both terms is related not only to different modes of expression (written or spoken). Linguists observe that spoken language is characterized by incomplete sentences, pauses, repetitions, corrections, jumps of thought, and interruptions [18], [19], which condition ambiguous boundaries of utterances (sentences). As a result, it becomes complicated to determine the boundaries of utterances even by taking into consideration the context, for example, if the speaker speaks very quickly, expresses several thoughts nonstop, or if speakers interrupt each other not allowing to finish the thought. In such cases, the definition suggested by reference [20] can be employed. Reference [20] defines an utterance as a stretch of speech preceded and followed either by pause (silence) or by a change of speakers. This definition is also used by Lithuanian linguists who work with specialized corpora (e.g. a corpus of children speech, which is characterized by short utterances) [12], [13]. However, such segmentation is not suitable for the transcription of adult speech, since if an utterance is considered to be a stretch until a pause or a change of speakers, it is often difficult to understand the meaning of an utterance (if an utterance breaks or is interrupted without being completed), to determine functions of an utterance, and to define its structure. During the annotation of the Corpus of Spoken Lithuanian, an *utterance* is considered to be a stretch of speech which is marked by a completed intonation and relatively completed thought. In this research, the main features of an utterance include predication, typical formal structural scheme, a clear communicative function, and a completed intonation. In case the pace of the speaker is very fast and few pauses are made, utterances are distinguished from the speech flow taking into consideration the above-mentioned features. If a thought of one speaker is incomplete and the speech is interrupted by a pause or another speaker's words, an interrupted stretch of speech is regarded as a part of the utterance, which is connected with the following stretch of speech; both these stretches constitute a single utterance. Admittedly, even applying the above-mentioned criteria, it is not always possible to identify the exact boundaries of an utterance and very often it is the intuition of a transcriber that is relied upon. However, this problem is faced not only by the creators of the Corpora of Spoken Lithuanian, but also by those researchers who develop speech corpus of other languages.

D. Grammatical Annotation

Grammatical annotation of a corpus is one of the most important stages in the development of the Corpus of Spoken Lithuanian. For the grammatical annotation of the corpus, the program CLAN of the database CHILDES was used. The command MOR of the program automatically annotates words according to the lexicon, i.e. a list of word forms with morphological tags (see more [12]). The lexicon of the Lithuanian language included 65,000 word forms most frequently used in written language (the lexicon was compiled by A. Utka who used the corpus of written language "Corpus of Contemporary Lithuanian"). Later, the lexicon was expanded and now it consists of 90,000 word forms. In the lexicon used for morphological annotation each word form is marked with all grammatical categories, for example, if it is a noun, its gender, number, case, and paradigm are indicated. Besides, certain information about the derivation of words (e.g. diminutives, compounds) as well as certain semantic information is given (the fragment of the lexicon is presented in Fig. 2).



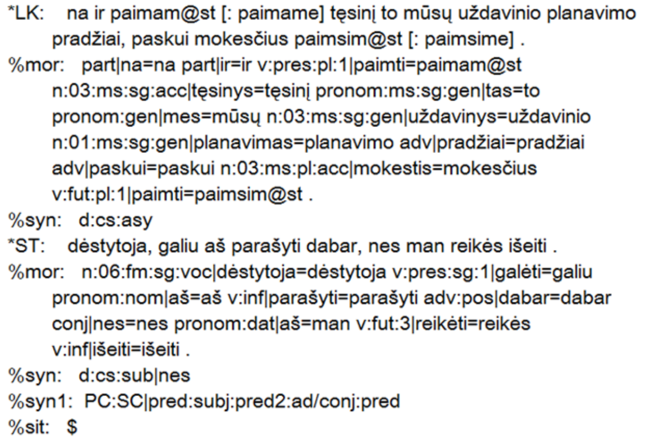
```
angliškomis    {[scat adj:01:undef:fm:pl:ins]} "angliškas"    =angliškomis=
durniausių     {[scat adj:01:undef:sup:ms:pl:gen]}    "durnas"      =durniausių=   \
       {[scat adj:01:undef:sup:fm:pl:gen]}    "durnas"      =durniausių=
gudriausių     {[scat adj:05:undef:sup:ms:pl:gen]}    "gudrus"      =gudriausių=   \
       {[scat adj:05:undef:sup:fm:pl:gen]}    "gudrus"      =gudriausių=
nesinešioja    {[scat v:neg:ref:pres:3]}      "nešiotis"     =nesinešioja=
persiskaitau {[scat v:ref:pres:sg:1]}   "persiskaityti" =persiskaitau=
protingiausių {[scat adj:01:undef:sup:ms:pl:gen]}      "protingas"    =protingiausių= \
       {[scat adj:01:undef:sup:fm:pl:gen]}    "protingas"    =protingiausių=
verčiame       {[scat v:pres:pl:1]}   "versti"      =verčiame=
```

Fig. 2 Fragment of the Lexicon

It should be noted that even though the collected lexicon is large enough, the coding process of spontaneous language is a very complicated task, since colloquial speech includes a lot of specific lexical and morphological features, such as shortened forms, non-standard pronunciation of certain words, jargon, slang words, etc., which may all occur during the production of speech [12]. Presently, by using the lexicon, 80% of words forms are automatically annotated. Unrecognized word forms are included manually, which adds to the expansion of the lexicon.

After having annotated the transcribed data, the problem of ambiguity is faced. A number of Lithuanian word forms are ambiguous; therefore, the program cannot choose the correct form from those given in the lexicon. For this reason, disambiguation should be done manually [12]. It is not difficult to choose the correct noun or verb form, but to choose the correct version of some prepositions, particles, conjunctions and interjections is rather problematic as the meaning of such words depends on the context. Moreover, different dictionaries provide different morphological description of these words. In order to facilitate the process of the identification of some words, the following criteria were taken into consideration: a) meaning in the context; b) relations with other words; c) function (for example, a particle modifies the meaning of a word, a conjunction links elements in a sentence, an interjection marks emotions) [12].

A disambiguated corpus can be expanded by various tags, and it can also be additionally annotated manually or semi-automatically. Having accomplished the morphological annotation and disambiguation of the Corpus of Spoken Lithuanian, particular syntactic tags were added to it. At the present moment, the following information is marked in the Corpus: functional types of utterances (declarative, interrogative, imperative, exclamatory), the structure of utterances (simple, compound, asyndetic, subordinate, coordinate, and compound-complex) as well as the word order of some combinations of words (attributive utterances, relative clauses). A disambiguated and both, morphologically and syntactically coded text is presented in Fig. 3.



```
*LK:    na ir paimam@st [: paimame] tęsinį to mūsų uždavinio planavimo
        pradžiai, paskui mokesčius paimsim@st [: paimsime] .
%mor:   part|na=na part|ir=ir v:pres:pl:1|paimti=paimam@st
        n:03:ms:sg:acc|tęsinys=tęsinį pronom:ms:sg:gen|tas=to
        pronom:gen|mes=mūsų n:03:ms:sg:gen|uždavinys=uždavinio
        n:01:ms:sg:gen|planavimas=planavimo adv|pradžiai=pradžiai
        adv|paskui=paskui n:03:ms:pl:acc|mokestis=mokesčius
        v:fut:pl:1|paimti=paimsim@st .
%syn:   d:cs:asy
*ST:    dėstytoja, galiu aš parašyti dabar, nes man reikės išeiti .
%mor:   n:06:fm:sg:voc|dėstytoja=dėstytoja v:pres:sg:1|galėti=galiu
        pronom:nom|aš=aš v:inf|parašyti=parašyti adv:pos|dabar=dabar
        conj|nes=nes pronom:dat|aš=man v:fut:3|reikėti=reikės
        v:inf|išeiti=išeiti .
%syn:   d:cs:sub|nes
%syn1:  PC:SC|pred:subj:pred2:ad/conj:pred
%sit:   $
```

Fig. 3 Fragment of the grammatically coded text

When using grammatically annotated corpus together with the program CHILDES, it is possible to do the analysis of different lexemes, coded morphological categories, and syntactic features by analyzing all conversations together or each separately, which helps to focus on the features of speech of each speaker or a group of speakers (according to their age, gender, profession and other defined data). Currently, the grammatically annotated Corpus of Spoken Lithuanian provide Internet users who do not apply the program CHILDES with a possibility to search for a word or a word form and get statistical information and the context of usage. The results obtained offer not only a concordance, but also a grammatical annotation and statistical data about all grammatical forms of the word being searched for. In addition, Internet users have the possibility to search for collocations and see information about each speaker and situation of conversation.

III. Possible Applications

The Corpus of Spoken Lithuanian is the first publicly available representative database of the Lithuanian language. First of all, the corpus serves as a rich database for linguists who are interested in the investigation of spoken language. As mentioned above, at present, while using the corpus, which includes mainly conversational and spontaneous informal speech (as well as prepared and more formal speech), it is possible to conduct different morphological and in certain respects, syntactic and semantic analysis of spoken language. Recently, a number of researches on spoken Lithuanian language have been conducted. Some of the studies focus on morphological and lexical features of speech [14], [21] etc., others aim at the description of certain syntactic features of spoken language [15], [16] etc. It should be noted that these investigations view spoken language as a solid form of expression and distinguish only differences between prepared public speech (media, academic discourse) and spontaneous colloquial speech. However, by using the corpus, the analysis of different registers can also be made taking into consideration the function and situation of speech as well as social roles, age, and gender of speakers. At the moment, the transcribed data is not linked to the audio files, but, after accomplishing this task in the future, it will be possible to conduct studies in the field of phonetics. Presently, the Corpus of Spoken Lithuanian is suitable for synchronic analysis but, if expanded constantly, it could be used for diachronic research.

The Corpus of Spoken Lithuanian could be useful and interesting not only to linguists, but also to a wider public, as it gives a possibility to have a different look at ordinary everyday speech.

References

[1] J. Kuvač Kraljević, and G. Hržica, "Croatian adult spoken language corpus (HrAL)," *FLUMINENSIA: Journal for Philological Research*, vol. 28, no. 2, 2017, pp. 87–102.

[2] D. Biber, "Investigation language use through corpus-based analyses of association patterns," *International Journal of Corpus Linguistics*, vol. 1, no. 2, 1996, pp. 171–198.

[3] D. Biber, *University Language: A Corpus-based Study of Spoken and Written Registers.* Amsterdam: John Benjamins, 2006.

[4] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *LREC-Eighth international conference on Language Resources and Evaluation,* Turkey, 2012.

[5] R. Simpson, and D. Mendis, "A Corpus-Based Study of Idioms in Academic Speech", *Tesol Quarterly*, vol. 37, iss. 3, 2003, pp. 419–441.

[6] R. Reppen, "English language teaching and corpus linguistics: Lessons

from the American National Corpus," in *Contemporary Corpus Linguistics*, P. Baker, Ed. London: Continuum, 2012, pp. 204–213.

[7] R. Carter, and M. McCarthy, *Exploring Spoken English*. Cambridge: Cambridge University Press, 1997.

[8] M. McCarthy, and M. Handford, "Invisible to us: A preliminary corpus-based study of spoken business English," in *Discourse in the Professions: Perspectives form Corpus Linguistics*, U. Connor, T. Upton, Eds. Amsterdam: John Benjamins, 2004, pp.167–201.

[9] Corpus of Spoken Lithuanian, http://donelaitis.vdu.lt/sakytines-kalbos-tekstynas/ Accessed on 20/03/2017.

[10] Child Language Data Exchange System, https://childes.psy.cmu.edu/ Accessed on 20/03/2017.

[11] B. MacWhinney, "The TalkBank Project," in *Creating and Digitizing Language Corpora: Synchronic Databases,* vol. 1, J. C. Beal, K. P. Corrigan & H. L. Moisl, Eds. Houndmills: Palgrave-Macmillan, 2007, pp. 163–180.

[12] I. Dabašinskienė, and L. Kamandulytė, "Corpora of Spoken Lithuanian," *Estonian papers in applied linguistics,* no. 5, 2009, pp. 67–77.

[13] L. Kamandulytė-Merfeldienė, and I. Balčiūnienė, "Syntactically Coded Corpus of Spoken Lithuanian: Developmental Issues and Pilot Studies," *Studies about Languages*, no. 28, 2016, pp. 92–101,

[14] L. Kamandulytė-Merfeldienė, "Pertarų dažnumas ir įvairovė sakytinėje kalboje (The Frequency and Variety of Fillers in Spoken Lithuanian Language)," *Bendrinėkalba*, no. 87, 2014, pp. 1–10.

[15] L. Kamandulytė-Merfeldienė, and I. Balčiūnienė, "Funkciniai pasakymų tipai sakytinėje kalboje (Types of Sentences and their Functions in Spoken Lithuanian)," *Thought elaboration: linguistics, literature, media expression*: *coolection of scientific papers*, 2016, pp. 11–29.

[16] L. Kamandulytė-Merfeldienė, and I. Balčiūnienė, "Atributinių ir predikatinių junginių su būdvardžiais dažnumas ir struktūra sakytinėje kalboje (Frequency and structure of attributive and predicative utterances in spoken Lithuanian)", *Lituanistica*, vol. 62, no. 2, 2016, pp. 127–137.

[17] L. Kamandulytė-Merfeldienė, "Morphological modifications in Lithuanian child directed speech", *Estonian Papers in Applied Linguistics*, no. 3, 2007, pp. 155–166.

[18] A. J. Liddicoat, *An Introduction to Conversion Analysis*, London: Continuum, 2007.

[19] G. Brown, and G. Yule, *Discourse analysis,* Cambridge University Press, 2001.

[20] D. Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell Reference, 2003.

[21] I. Dabašinskienė, "Šnekamosios lietuvių kalbos morfologinės ypatybės (The Morphological Features of Spoken Lithuanian)", *Acta Linguistica Lithuanica*, no. 60, 2009, pp. 1–15.