Vilnius University Press

Vilnius University Press

Kalbotyra
2018 (71) 7-25

ISSN 1392-1517 eISSN 2029-8315
DOI: https://doi.org/10.15388/Kalbotyra.2018.1

Papers

The Corpus of Lithuanian Children Language: Development and application for modern studies in language acquisition

Ingrida Balčiūnienė

Department of Lithuanian Studies
Vytautas Magnus University
V. Putvinskio g. 23–218
LT-44234 Kaunas, Lithuania
Email: ingrida.balciuniene@vdu.lt

Laura Kamandulytė-Merfeldienė

Department of Lithuanian Studies
Vytautas Magnus University
V. Putvinskio g. 23–218
LT-44234 Kaunas, Lithuania
Email: laura.kamandulyte@vdu.lt

Abstract. This paper describes *The Corpus of Lithuanian Children's Language* and its possible applications for modern studies on the first language acquisition. First of all, the procedure of data collection for the Corpus is discussed. Furthermore, the main methodological principles of longitudinal and experimental data compilation and transciption are decribed. Finally, different studies in developmental psycholinguistics which have been carried out so far and which demonstrate possible ways of the application of the Corpus data for different scientific purposes are introduced.

The Corpus of Lithuanian Children's Language developed at Vytautas Magnus University comprises typical and atypical, longitudinal and experimental data of the Lithuanian language development. The Corpus was compiled using different methodological approaches, such as natural observation and semi-experiment. The longitudinal data (conversations between the target children and their caretakers) compiled according to the requirement of natural observation includes transcribed and morphologically

annotated speech of two typically-developing children, one late talker, one early talker, one child from a low SES family, and a pair of twins. The data was collected during the period of 1993–2017 and and it can be divided into three cohorts. The semi-experimental data (~ 124 hours) comes from numerous studies in narratives and spontaneous dialogues elicited from typically-developing and language-impaired monolingual and bilingual (pre-) school age children.

From the very beginning of data collection for the *The Corpus of Lithuanian Children's Language*, studies in the develomental changes of typical child language have been carried out. Over the past decade, these studies have been supplemented by statistical analysis of elicited semi-experimental data; the majority of these studies deal with typical vs. atypical (delayed or impaired) language acquisition and with differences between acquision of Lithuanian in a monolingual vs. bi-/polylingual settings.

The paper provides an overview of data of *The Corpus of Lithuanian Children's Language*, which have been collected from 1993 but still needed to be structurized according to the employed methodology of data compilation and possible applications for different scientific purposes.

Keywords: corpus linguistics, language acquisition, child language, Lithuanian

1 Introduction

Systematic psycholinguistic studies on Lithuanian-speaking children's language started in 1993 and initially were based on longitudinal data of two Lithuanian children (Savickienė 1999; Wójcik 2000). Later on, along with the development of *The Corpus of Spoken Lithuanian*[1] and the adaptation of the CHAT (MacWhinney 2017a) software for the Lithuanian language (Dabašinskienė & Kamandulytė 2009), a great amount of data of Lithuanian-speaking children's language has been collected and prepared for an automatized linguistic analysis. Now, *The Corpus of Lithuanian Children's Language* comprises typical and atypical, longitudinal and experimental data of the Lithuanian language development. Longitudinal data (conversations between the target children and their caregivers) includes transcribed and morphologically annotated speech of: a) two typically-developing (TD) children; b) one late talker; c) one early talker; d) one child from a low SES family; and e) a pair of twins. The data was collected during the period

[1] *The Corpus of Spoken Lithuanian* was developed at Vytautas Magnus University (under a supervision of Ineta Dabašinskienė) in cooperation with Vilnius University, Klaipėda University, Šiauliai University, and The Institute of Lithuanian Language. The work was supported by the Lithuanian State Science and Studies Foundation (No. L-12/2008) and the Research Council of Lithuania (No. LIT-9-11, No. LIP-085/2016). The Corpus (http://sakytinistekstynas.vdu.lt/) comprises 256 conversations (~320,000 words) of 1,086 adult.

of 1993–2017 and it can be divided into three cohorts. Experimental data (~ 124 hours) comes from numerous studies in narratives and spontaneous dialogues elicited from TD and language-impaired (LI) monolingual and bilingual (pre-) school age children.

The aim of this paper is to give a structurized description of *The Corpus of Lithuanian Children's Language* with the main focus on methodological approaches to data compilation and on the most prominent ways of its application for the modern studies on language acquisition.

2 Data compilation and transcription methods

2.1 Data compilation methods

2.1.1 Longitudinal data

Following the universally agreed methodology of naturalistic studies (Voeikova & Dressler 2002), seven Lithuanian monolingual children were investigated from the target child's onset of speech (1;6-2;5) until the age of about 4 years. All the children (with exception of the child from a low SES family) come from middle-class families; one or both of the parents are highly educated professionals. Conversations of 15-20 minutes in length between the target child and his/her caretakers were recorded 2-3 times per week in a common (mainly, home) environment. The parents were instructed to record as many different situations as possible: games, cooking, eating, communication with guests, bathing, preparation for sleep, etc. The conversations were recorded at different times of the day, which mostly depended on the target child's willingness to communicate. Most of the recordings were conversations between the target child and his/her parents (usually, the mother) but there were also some dialogues between the child and his/her grandmother and some polylogues between the child, his/her mother/father, grandmother, other caretakers, and other children. The collected data was grouped by months trying to maintain similar size and length within each month's material (see Table 1).

Target child*	**Age range**	**Hours of recordings**	**The total number of words**
TD-1	1;7-2;5	~ 31	155,414
TD-2	1;8-3;8	~ 23	194,296
Early talker	1;6-2;6	~ 16	71,728
Late talker	2;5-4;3	~ 11	45,905
Low SES child	1;7-3;6	~ 8	18,900
Twin pair	2;5-3;6	~ 17	74,442
In total		~ 106	560,685

Table 1. The structure and size of the longitudinal sub-corpus

The data of the TD-1 was compiled by her mother during the period of 1993–1994 under the supervision of Ineta Dabašinskienė (former Savickienė) in the framework of her PhD research. The data of the TD-2 was compiled by her mother Ingrida Balčiūnienė during the period of 2000–2002 in the framework of her PhD research. The data of the early talker was compiled by his mother during the period of 2007–2008 under the supervision of Laura Kamandulytė-Merfeldienė in the framework of her PhD research. The data of the late talker was compiled by his father during the period of 2005–2007 under the supervision of Laura Kamandulytė-Merfeldienė in the framework of her PhD research. The data of the low SES child was compiled by his mother during the period of 2008–2010 under the supervision of Ingrida Balčiūnienė[2]. The data of the twins was collected during the period of 2015–2017 by their parents under the supervision of Ingrida Balčiūnienė[3].

The methodology of the longitudinal Lithuanian data collection has been described in detail by Savickienė (1997, 2003), Balčiūnienė (2009), Kamandulytė (2009), and Dabašinskienė & Kalėdaitė (2012). Despite universally agreed and clearly stated requirements for the naturalistic observation (Voeikova & Dressler 2002), some challenges that occurred during the data compilation could be mentioned. First of all, individual differences in child's language acquisiton should be taken into account. All the children (with the exception of the low SES child) were supposed to be typically developing (since they did not have any documented developmental disorders) and, thus, were selected as representers of the typical Lithuanian language acquisition. However, in the process of data compilation, two of them, i.e. the late talker and the early talker, turned out to be representers of the atypical (although not impaired) language acquisition. One more child selected for the very first attempt to collect Lithuanian data (1993) turned out to be phonologically-impaired and had to be excluded from the study. Secondly, the parents (although they were instructed to behave in the most natural manner), obviosly tried to talk to their children accurately, politely, and perfectly, at least, during the initial period of data compilation. Finally, some unplanned time lags occurred between the recording sessions because of summer holidays, the child's ilness, family trips, etc. All these challenges prevented from keeping the balance of the corpus material from the perspective of its monthly size and length.

2.1.2 Experimental data

Experimental psycholinguistic studies in the Lithuanian language acquisition started in 2006, when Lithuania was involved in the international projects COST Action

2 The work was supported by the Research Council of Lithuania, grant No. LIT-6-13.
3 The work was supported by the Research Council of Lithuania, grant No. LIP-020/2016.

A33 – *Cross-linguistically Robust Stages of Children's Linguistic Performance*[4] and CLAD – *Crosslinguistic Language Diagnosis*[5] and started developing and/or adapting internationally standardized language diagnostic tools. In the framework of the COST Action IS0804 – *Language Impairment in a Multilingual Society: Linguistic Patterns and the Road to Assessment*[6], the LITMUS-MAIN (*Language Impairment Testing in Multilingual Settings: Multilingual Assessment Instrument for Narratives*) was developed and piloted in 28 languages, including Lithuanian (Gagarina et al. 2012, 2015). At the same time, a great number of narratives[7] and elicited spontaneous dialogues[8] of (pre-) school age children was collected in the framework of Lithuanian national scientific projects and individual scientific research.

2.1.2.1 Narrative data

Following numerous previous studies (Botting 2002; McCabe & Bliss 2003), narrative analysis might serve as an efficient and ecologically valid diagnostic tool for the distinction between typically-developing and language-impaired children; moreover, oral narrative skills can serve as an informative predictor for written language acquisition and literacy development (Westerveld et al. 2008). Thus, during the past decades, narratives have been considered an informative data for studies on child's language.

Narrative acquisition might be divided into several stages. Following Hedberg & Stoel-Gammon (1986), children start with 1) heaps (at the age of 2 years) and 2) sequences, then they acquire 3) primitive narratives and 4) unfocused chains, and, finally, reach the stage of 5) focused chains and 6) true narratives (at the age of 5-6 years). One can presume that a child first familiarizes himself/herself with a narrative genre via fairytales and family narratives. This experience enables him/her to master the basic scheme to tell about his/her own life events and to connect them into a storyline. Later on, a child starts creating his/her own fictional stories (tales, scary stories, etc.) in everyday social communication. Then, after having started school, a child starts applying his/her experience to produce written stories and to comprehend printed narrative texts.

---

4 http://www.cost.eu/COST_Actions/isch/A33.

5 http://www.cladproject.eu/.

6 http://www.cost.eu/COST_Actions/isch/IS0804.

7 The data of monolingual children was collected by Ingrida Balčiūnienė in the framework of her Post-Doc research funded by the Research Council of Lithuania. The data of Lithuanian- and English-speaking bilinguals was collected by Agnė Blažienė (former Kalninytė) in the framework of her PhD research.

8 The data was collected with a financial support from the Research Council of Lithuania, grant No. LIT-1-18.

Usually, narratives are classified into scripts, personal narratives, and fictional stories (Hughes et al. 1997). In Lithuanian narrative studies, data of fictional stories, i.e. stories relating past, present, or future events that are not real and focus on someone or something attempting to carry out a goal, (see Hedberg & Westby (1993), was collected. Due to the lack of previous experience in narrative data collection and analysis, we started with preschoolers of 4-5-years and tried to cover all age groups up to 18-19 years of age. However, the main attention was paid to children from the pre-primary education group (6-7 years of age), since this age was presumed to be critical for the transition from oral to written communication (Hayward & Schneider 2000), which, consequently, appears to be crucial for the later development of literacy and academic attainment.

The sequence of coloured wordless pictures the *Baby-Birds*, originally developed by Hickmann (2003) and modified by Gagarina et al. (2012, 2015), was applied for narrative elicitation. The sequence is given in Figure 1.

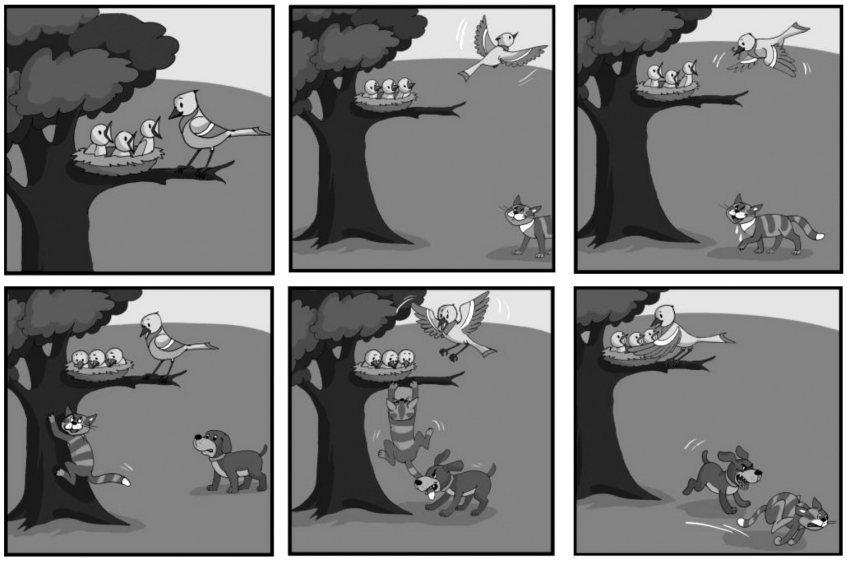

Figure 1. *The Baby-Birds* picture sequence (according to Hickmann 2003 and Gagarina et al. 2012, 2015)

In total, 626 subjects were assessed (see Table 2), i.e. 626 narratives were elicited.[9]

[9] A new narrative data elicited from primarily language-impaired (PLI) preschool age children will be added to the Corpus within the next two years. The work is supported by the Research Council of Lithuania, grant No. LIP-18-36.

Lingualism	**The number of children**	**Age range**	**Hours of recordings**	**The total number of words**
TD-MO	76	4;0-6;0	~ 12	3,076
TD-MO	120	6;0-7;0	~ 19	11,088
TD-MO	96	7;0-9;0	~ 16	12,084
TD-MO	96	9;0-11;0	~ 16	19,948
TD-MO	18	11;0-13;0	~ 3	1,060
TD-MO	12	13;0-15;0	~ 1.5	760
TD-MO	46	15;0-18;0	~ 7	1,993
LI-MO	12	4;0-7;0	~ 1.5	567
TD-BI-L1	100	4;0-10;0	~ 16	5,952
TD-BI-L2	50	6;0-7;0	~ 8	1,446
In total	626		~ 100	57,974

Table 2. The structure and size of the narrative sub-corpus. *TD-MO – typically-developing monolinguals; LI-MO – language-impaired monolinguals; TD-BI-L1 – typically-developing bilinguals speaking Lithuanian as L1; TD-BI-L2 – typically developing bilinguals speaking Lithuanian as L2

Each of the subjects was tested individually, in a quiet room at the kindergarten or at school. First, for a warming-up, the child was asked whether he/she likes fairy-tales and stories, who tells stories to him/her, and then the experimenter said: "Today I would like you to tell me a story." The experimenter took the pictures and continued: "These pictures illustrates a particular story. First of all, I will show you all the pictures, and then you will look at each picture carefully and tell me the story you see." Then the experimenter placed the pictures in the correct sequence in a single, horizontal row in front of the child, without saying anything except, "The story starts like this…". The child was allowed to look at the pictures for a few minutes to get the gist of the story. Then the experimenter said: "Now I want you to tell the story. This is the beginning of the story. Look at the pictures and try to tell the best story you can." No questions such as "What is he/she doing here?"; "What is this?"; "Who is coming?", etc. were used in order not to disrupt or influence the child's narration. Allowable prompts, if the child was hesitant to continue, were, "Tell me a story about what happens in this picture" or "Tell me what happened".

In total, ~100 hours of narratives were recorded and prepared for the analysis. More on the methodology of the narrative data compilation, see Gagarina et al. (2012, 2015). As for Lithuanian, most of the challenges were related to the pictorial content, especially in the older children. While young (4-6 years of age) children found the story-telling task attractive and funny, the older ones (starting with 7 years of age) referred to the pictures as "too childish", "too simple"; thus, it was much more difficult to engage the older

children into the task. For future studies, some more complex picture (and maybe photo) sequences or videostimuli would be advised for narrative elicitation.

Our attempt to elicit personal narratives in preschool children should also be mentioned as an illustration of unsuccessful experience. Initially, we tried to ask children to tell about their group trip to the city castle; a shared experience was supposed to serve for a better comparability among the narrative texts. However, presumably due to different interests and low engagement into the task, the narratives turned out to be too short for the linguistic analysis. For future studies, *The Conversational Map Procedure* (McCabe & Rollins 1994) would be advised for personal narrative elicitation, since children (and even adults) are much more likely to share their own experiences if the experimenter does the same first.

2.1.2.2 Dialogue data

Usually, the development of communicative competence starts with the acquisition of a dialogue because of its relatively transparent structure and rules. Following a great number of previous studies, such as those by Snow (1977), McTear (1985), Clark (2009), a one-year-age child is already able to construct some simple dialogue structures; later on, a child becomes a more and more skilled participant in various extended discourses (Pan & Snow 1999). However, much less is still known about conversations between older children. Thus, we aimed at the compilation of spontaneous dialogical data of preschoolers of 6-7 years of age and students of 8-11 years of age.

In total, 288 monolingual typically-developing subjects were assessed (see Table 3).

Lingualism	**The number of children**	**Age range**	**Hours of recordings**	**The total number of words**
TD-MO	96	6;0-7;11	~8	35,820
TD-MO	96	8;0-9;11	~8	33,612
TD-MO	96	10;0-11;11	~8	27,468
In total	288		~24	96,900

Table 3. The structure and size of the dialogue sub-corpus

The dialogues were elicited following the method of a joint activity (Balčiūnienė & Ančlauskaitė 2011, Balčiūnienė & Miklovytė 2011). Children from the same group/class were paired for a joint task. Each pair was invited to a specially equipped room with a table, two chairs, and a hidden digital recorder. A simple black-and-white sketch and a set of six colouring pencils were given to the target pair and the children were asked to color the sketch within the next 10 minutes. The experimenter told that, due to the time limit, the children should try to cooperate with each other and decide about the

joint activity. Then, the experimenter left the room. Five minutes later, the experimenter entered the room to ensure that the children were busy with colouring and to remind that only five minutes left for the completion of the task. After the next five minutes, the experimenter entered the room again to stop the assessment.

In total, ~24 hours of dialogues were recorded and prepared for the analysis.

The joint task approach is still relatively new in the compilation of dialogal data, nevertheless, the procedure turned out to be sucessful. The children were provoked to share the pencils, to divide the sketch, to decide on the the final results and, thus, had to verbalize their suggestions, ideas, and requests. Usually, the conversations started with a discussion about the task but later new and new topics appeared. For researchers who intend to apply this kind of dialogue elicitation, we would advice not to pair the children but to suggest them to choose the partner: our data evidenced that a pair of close friends produced much longer and elaborate dialogues in comparison to a pair of less familiar classmates.

2.2 Data transcription

All the recordings (longitudinal and experimental) were transcribed orthographically by professional linguists (in some longitudinal cases – by the mothers) and checked independently by two experts (professional linguists). The transcripts were annotated for a multipurpose (lexical, morphological, and partially discourse) automatic linguistic analysis using the CHAT (*Codes for the Human Analysis of Transcripts*) software (MacWhinney 2017a) and checked independently by two experts (professional linguists). An example of the annotated transcripts is provided in Figure 2.

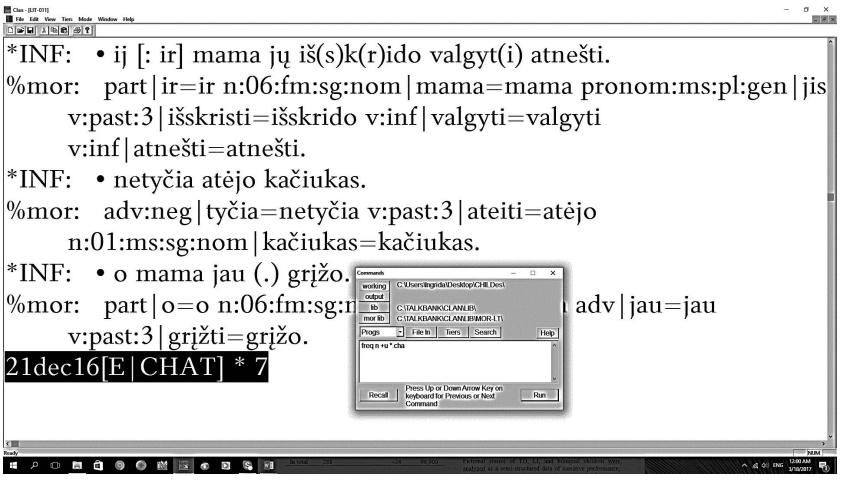

Figure 2. An excerpt from the CHAT (MacWhinney 2017a) window

In order to allow automated morphological analysis using CLAN tools (MacWhinney 2017b), the transcribers coded each word with morphological information, including the base form of a word and a set of tags expressing Lithuanian morphological characteristics (Lounela 2005).

The main problems related to the transcription and morphological annotation of Lithuanian speech data were discussed by Dabašinskienė and Kamandulytė (2009). To sum up, three difficulties faced at this stage could be mentioned: 1) the problem of orthography (the distinction between the standard orthography and the phonetic representation of sounds); 2) the problem of a transcription unit (the distinction between a sentence and an utterance); and 3) the problem of morphological disambiguation that prevents from an automatized morphological annotation and requires to annotate a vast number of word tokens manually.

3 The Corpus as a data-base for the modern studies on language acquisition

The first psycholinguistic studies based on the longitudinal data of Lithuanian children addressed the acquisition of the morphology of nouns (Savickienė 1999) and verbs (Wójcik 2000). Later, the acquisition of adjectives (Kamandulytė 2010) and conversation structure (Balčiūnienė 2009) were taken into consideration. The studies helped to shed a light on the developmental changes in the Lithuanian language acquisition. Namely, the main characteristics of different stages of morphology acquisition (premorphology, protomorphology, and modular morphology) were identified; the main patterns in the acquisition of noun case, gender, and number forms were higlighted; the role of diminutive forms in both child's and child-directed speech was evaluated. Besides pure morphological characteristics of adjectives, their syntactic functions were analyzed, too. Finally, the development of the conversation structure (turn taking, topic maintainence, and repair of conversation breakdowns) were discussed. To sum up, the first stage of Lithuanian studies related to developmental psycholinguistics helped to build a systematic picture of the acquisition of the Lithuanian language.

In 1993, longitudinal data of Lithuanian children was included in the international *Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition* supervised by W. U. Dressler (Austrian Academy of Sciences and Vienna University, Austria) and, since then, a lot of comparative studies have been carried out. Examples, worth mentioning include three monographs, namely, by Savickienė and Dressler (2007), Stephany and Voeikova (2009), and Tribushinina et al. (2015), and some peer-reviewed papers (e.g., Kilani-Schoch et al. (2009), Kazakovskaya, Balčiūnienė (2012),Tribushinina et al. (2013), and Dabašinskienė and Voeikova (2015). The first book is devoted to the role of diminutives in child's and child-directed speech in different languages; Lithuanian-speaking children and their parents (together with Italian- and Russian-speaking subjects), turned out to be extremely productive in the use of diminutive forms. Diminutives are suggested to be

some kind of morphological bootstrapping (Pinker 1984) elements that simplify the noun declension system and, thus, help a child with its acquisition (Savickienė et al. 2009). The second book highlights the main patterns of the acquisition of nominal words in different languages. The third one is devoted to different aspects of the acquisition of adjectives; as for Lithuanian language, the main lexical and morphological characteristics of child's adjectives (Kamandulytė-Merfeldienė 2015) and specific parental reactions to child's adjectives (Kazakovskaya, Balčiūnienė 2015) are presented.

The previous studies based on longitudinal data might be classified into two groups. Studies belonging to the first group address linguistic distinctions between typically- vs. atypically-developing children, while the other group of studies compares three cohorts of the longitudinal observation. A comparative analysis of typically- and atypically-developing children (such as the late talker, the early talker, the low SES child, and the pair of twins) evidence a number of lexical and grammatical distinctions. In the following two sub-sections, these distinctions will be discussed.

So far, experimental corpus data has been employed for the comparison of the acquisition Lithuanian language in monolingual vs. bi-/polylingual settings. An overview of these studies is given in the sub-section 3.3.

3.1 The application of the Corpus for a comparative analysis of typically- vs. atypically-developing children

A comparative analysis between the typically- and atypically-developing Lithuanian children (see Table 1) addressed the main lexical and grammatical measures, such as a productivity, lexical diversity, and (morpho-) syntactic complexity. The F-test two-sample for variances evidenced that the children did not differ from each other in respect of the productivity (the total number of words and utterances). During the analyzed period (2;5-3;5), their mean length of turn (MLT) varied from 1.4 to 5.8 utterances but the differencies were not statistically signifficant ($P \geq 0.05$). However, the low SES child demonstrated significantly lower mean length of utterance (MLU) rate than the TD child and the twin girl and boy (repectively, $P = 0.023$; $P = 0.028$; and $P = 0.040$, see Table 4). From this perspective, the productivity of the low SES child might be equaled to the productivity of the late talker.

	TD child	**Late talker**	**Low SES child**	**Twin girl**	**Twin boy**
TD child	1	0.132	**0.023***	0.783	0.734
Late talker		1	0.526	0.153	0.159
Low SES child			1	**0.028***	**0.040***
Twin girl				1	0.939
Twin boy					1

Table 4. Differences in the MLU rate among the children (2;5-3;5). * – $p \leq 0.05$

Morphosyntactic complexity, in contrast, was the worst developed in the twin pair. For instance, as illustrated in Table 5, the twins produced a great number of errors in adjective agreement in comparison with the rest of the children.

	Erroneous gender form	**Erroneous number form**	**Erroneous case form**
TD child	3.6	0.9	0.6
Late talker	0.7	2.5	2.5
Twin girl	3.7	1.9	6.5
Twin boy	3.9	2.7	7.8

Table 5. The percentage of erroneous adjectives (from all the adjective tokens) among the children (2;5-3;5)

The percentage of errors in adjective agreement seems to be extremely high in the twin boy's speech; this might be explained by a relatively slower general grammatical development that manifested as so called 'secret language of twins' (Thorpe et al. 2001; Thorpe 2006; Hayashi et al. 2013) in his speech.

3.2 The application of the Corpus for a comparative analysis of different cohorts

During the past decades, Lithuania, as many other Eastern European countries, has undergone intense changes in (socio-) cultural and (socio-) economic areas. One can presume that, along with dramatic changes in the so-called 'work culture', attitudes to parenting (including strategies and styles of communication with young children) have also changed. And, thus, language acquisition is now influenced by completely different features of a child-directed speech (CDS) than decades ago. In order to test this presumption, a comparative linguistic analysis of different longitudinal cases that fall into three cohorts was carried out. The first cohort was representend by a TD child whose data was collected in 1993–1994; the second cohort was represented by a TD child whose data was collected in 2000–2002; and the third cohort was represented by a TD child whose data was collected in 2008–2010. The main measures (for both children's (CS) and child-directed (CDS) speech) were as following: the MLU rate, the type/token ratio (TTR) of the content words; and the distribution of noun diminutives at the age of 1;8, 2;0, and 2;4. Although the F-test two-sample for variances did not reveal signifficant differences between the cohorts, one can observe a slightly higher difference between the first and the third cohort in comparison with a difference between the first and the second or between the second and the third one (see Table 6).

	Cohort 1		Cohort 2		Cohort 3	
The MLU rate	CS	CDS	CS	CDS	CS	CDS
1;8	1.100	2.752	1.221	2.987	0.334	1.635
2;0	1.812	3.021	1.531	2.825	1.697	3.684
2;4	2.624	3.592	1.682	2.721	1.114	3.364
Noun TTR	CS	CDS	CS	CDS	CS	CDS
1;8	0.588	0.439	0.242	0.293	0.071	0.800
2;0	0.428	0.417	0.553	0.316	0.407	0.673
2;4	0.409	0.517	0.475	0.899	0.375	0.586
Verb TTR	CS	CDS	CS	CDS	CS	CDS
1;8	0.727	0.359	0.457	0.642	0.214	0.567
2;0	0.390	0.467	0.407	0.322	0.478	0.569
2;4	0.483	0.551	0.307	0.827	0.600	0.738
Adjective TTR	CS	CDS	CS	CDS	CS	CDS
1;8	0.750	0.472	0.610	1.000	-	-
2;0	0.379	0.355	0.550	0.622	-	0.833
2;4	0.469	0.503	0.447	0.494	-	-
A percentage of diminutives among all noun tokens	CS	CDS	CS	CDS	CS	CDS
1;8	0.22	0.54	0.1	0.34	0	0.27
2;0	0.52	0.58	0.47	0.43	0.9	0.19
2;4	0.47	0.46	0.45	0.51	0	0

Table 6. Results of the comparative analysis among the cohorts

This presumption is particularly valid for the distribution of diminutives. In the first cohort, the percentage of diminutives among all noun tokens varied from 0.22 to 0.52 in the child's speech and from 0.46 to 0.58 in the CDS; while, in the third cohort, the percentage of diminutives among all noun tokens reached only 0.9 in the child's speech and only 0.27 in the CDS. The third cohort was also distinguished by the fact that there were almost no adjectives in its corpus. Taking into account relatively limited resources (one child per cohort), it is refrained from concluding the decrease in the quality of CDS, but the data of the Corpus enables for more detailed future studies.

Generally, longitudinal studies on the acquisition of the Lithuanian language now cover almost all fields of linguistics, i.e. phonetics/phonology, morphology, morphosyntax, and discourse; only vocabulary acquisition still lacks naturalistic longitudinal studies.

Besides pure scientific purposes, the longitudinal data might serve as a great basis for the development of language diagnostic (Kamandulytė et al. 2010) and therapy tools.

3.3 The application of the Corpus for a comparative analysis of monolingual vs. bilingual children

Experimental data of the acquisition of the Lithuanian language not only supplement the longitudinal naturalistic data but also enable for the quantitative analysis of oral discourse. Since 2012, fictional stories of TD and PLI monolingual and bilingual children have been analyzed as a semi-structured data of narrative performance; the main macro- and microstructural measures (such as story structure, episode completeness, general productivity, lexical diversity, and syntactic complexity) have been evaluated in each of the populations (Balčiūnienė 2012, 2013; Balčiūnienė, Kalninytė 2014; Blažienė 2016). For instance, a comparative analysis of narrative production in children acquiring Lithuanian as the first vs. a heritage language evidenced that monolinguals demonstrated greater ($p < 0.05$) lexical diversity and used a wider range ($p < 0.05$) of syntactic devices to create story cohesion than the heritage speakers, although a general story length (both in words and utterances) was higher ($p < 0.05$) in the heritage group (Balčiūnienė et al. 2017). Additionally, such features as a simplified language declension system, numerous errors in modifier agreement, various difficulties using transitive verbs, prepositional phrases, and verbs with prefixes, and broadened and/or narrowed meanings of words were observed in bilingual (pre-) schoolers speaking Lithuanian as the first and English as the second language (Blažienė 2016).

Finally, a part of narrative data was submitted for the statistical cross-linguistic comparative analysis (Gagarina et al. 2013, Balčiūnienė & Kornev 2016). Among the papers based on the experimental corpus data of Lithuanian children, *The Tool for Narrative Analysis in Bilingual Children* (Balčiūnienė & Dabašinskienė 2012) and a doctoral thesis (Blažienė 2016) should be particularly mentioned not only as scientific studies but also as the methodological basis for language intervention.

4 Conclusions

Developmental psycholinguistics was introduced in Lithuania only two decades ago and started with two longitudinal case-studies. Nevertheless, over the past decades, relatively small individual studies on the acquisition of the Lithuanian language have grown and transformed into systematic multiple research. As stated by Dabašinskienė and Kalėdaitė (2012), "while systematic research on the acquisition of Lithuanian in the 20th century was virtually non-existent, the beginning of the new millennium has witnessed Lithuanian emerging as one of the languages where acquisition of morphology is best documented and analyzed" (2012, 153). During this period, a number of papers dealing

with the acquisition of Lithuanian phonetics/phonology, morphosyntax, and discourse have been published.

One of the major problems of naturalistic studies on language acquisition (not only in Lithuania but also worldwide) is the fact that the lack of research control might lead to incomparable samples and make it difficult to study low-frequency phenomena (Eisenbeiss 2010). Experimental data, on the other hand, does not provide representative naturalistic speech samples (Eisenbeiss 2010). However, the combination of these methodological approaches might provide a full picture of language acquisition and production. Thus, the development of complex corpora comprising both longitudinal and experimental data should be particularly encouraged in the field of developmental psycholinguistics.

In this paper, only a few scientific studies based on the Corpus data and devoted to the most important problems in child's language studies (i.e. language acquisition in monolingual vs. bilingual settings; an impact of the cohort on various lexical and grammatical measures of child's language; individual linguistic differences among typically- and atypically-developing children) were discussed. In future studies, more specific and urgent problems, such as Lithuanian primary language impairment, will be analyzed by means of corpus linguistics.

Abbreviations

CDS child-directed speech
CHAT *Codes for Human Analysis of Transcripts* (MacWhinney 2017a)
CLAN *Computerized Language Analysis* (MacWhinney 2017b)
CS child speech
LI language impairment
MLT the mean lenght of turn
MLU the mean lenght of utterance
PLI primary language impairment
SES socio-economic status
TD typically-developing (child)
TTR type/token ratio

References

Balčiūnienė, Ingrida. 2009. *Pokalbio struktūros analizė kalbos įsisavinimo požiūriu.* [Analysis of conversational structure from the perspective of language acquisition]. Humanitarinių mokslų daktaro disertacija. [PhD dissertation]. Kaunas: Vytautas Magnus University.

Balčiūnienė, Ingrida. 2012. Lithuanian narrative language at preschool age. *Estonian Papers in Applied Linguistics* 8, 21–36.

Balčiūnienė, Ingrida. 2013. Linguistic disfluency in narrative speech: Evidence from story-telling in 6-year olds. *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association*, *August 25–29, Lyon, France, Proceedings*. 2143–2146.

Balčiūnienė, Ingrida, Jolita Ančlauskaitė. 2011. Nurodymų raiška priešmokyklinio amžiaus vaikų kalboje. [Directives among Lithuanian preschoolers]. *Filologija* 16, 5–18.

Balčiūnienė, Ingrida, Ineta Dabašinskienė. 2012. DVRPAM – Dvikalbių vaikų rišliojo pasakojimo analizės metodika. [A tool for narrative analysis in bilingual children]. *MAIN: Multilingual Assessment Instrument for Narratives*. ZASPil Nr. 56 – December 2012. Berlin: ZAS. http://www.zas.gwz-berlin.de/zaspil.html?&L=1.

Balčiūnienė, Ingrida, Ineta Dabašinskienė, Agnė Blažienė. 2017. *Narrative production of children acquiring Lithuanian as a heritage language*. Paper presented at the 14th International Congress for the Study of Child Language. Lyon: IASCL, University Lyon 2.

Balčiūnienė, Ingrida, Agnė Kalninytė. 2014. Priešmokyklinio amžiaus vaikų rišliojo pasakojimo ypatybės. [The main characteristics of narrative in preschoolers]. *Kalbos kultūra* 86, 212–236.

Balčiūnienė, Ingrida, Aleksandr N. Kornev. 2016. Linguistic disfluency in story-telling: Evidence from Lithuanian- and Russian-speaking preschoolers. *Pediatr* 7 (1), 142–147.

Balčiūnienė, Ingrida, Inga Miklovytė. 2011. Lietuvių vaikų ir suaugusiųjų kalbos vidutinis pasakymo ilgis. [MLU in Lithuanian Children and Adult Speech]. *Res Humanitariae* 2 (10), 302–313.

Bishop, Dorothy V. M., Sonia J. Bishop. 1998. "Twin language": A risk factor for language impairment? *Journal of Speech Language and Hearing Research* 41, 150–160.

Blažienė, Agnė. 2016. *Lietuvių vaikų leksikos ir gramatikos raida anglakalbėje aplinkoje.* [Lexical and grammatical development of Lithuanian children in an English-speaking environment]. Humanitarinių mokslų daktaro disertacija [PhD dissertation]. Kaunas: Vytautas Magnus University.

Botting, Nicola. 2002. Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy* 18 (1), 1–21.

Clark, Eve V. 2009. *First language acquisition*. Cambridge: Cambridge University Press.

Dabašinskienė, Ineta, Violeta Kalėdaitė. 2012. Child language acquisition research in the Baltic area. *Journal of Baltic Studies* 43 (2), 151–160.

Dabašinskienė, Ineta, Laura Kamandulytė. 2009. Corpora of spoken Lithuanian. *Estonian Papers in Applied Linguistics* 5, 67–77.

Dabašinskienė, Ineta, Maria D. Voeikova. 2015. Diminutives in spoken Lithuanian and Russian: Pragmatic functions and structural properties. In *Contemporary*

approaches to Baltic linguistics. Peter Arkadiev, Axel Holvoet & Björn Wiemer, eds. Berlin: Mouton de Gruyter. 203–234.

Eisenbeiss, Sonja. 2010. Production methods in language acquisition research. In *Experimental methods in language acquisition research*. Elma Blom & Sharon Unsworth, eds. Amsterdam: John Benjamins Publishing Company. 11–34.

Gagarina, Natalia, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ingrida Balčiūnienė, Uta Bohnacker, Joel Walters. 2012. *MAIN: Multilingual assessment instrument for narratives*, ZAS Papers in Linguistics 56. Berlin: ZAS.

Gagarina, Natalia, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ingrida Balčiūnienė, Uta Bohnacker, Joel Walters. 2015. Assessment of narrative abilities in bilingual children. In *Assessing multilingual children*. Sharon Armon-Lotem, Jan de Jong & Natalia Meir, eds. Bristol: Multilingual Matters. 243–269.

Hayashi, Chisato, Hirosi Mikami, Reiko Nishihara, Chihi Maeda, Kazuo Hayakawa. 2013. The relationship between twin language, twins' close ties, and social competence. *Twin Research and Human Genetics* 17 (1), 27–37.

Hayward, Denyse, Phyllis Schneider. 2000. Effectiveness of teaching story grammar knowledge to pre-school children with language impairment: An exploratory study. *Child Language Teaching and Therapy* 16 (3), 255–284.

Hedberg, Natalie L., Carol Stoel-Gammon. 1986. Narrative analysis: Clinical procedures. *Topics in Language Disorders* 7 (1), 58–69.

Hedberg, Natalie L., Carol E. Westby. 1993. *Analyzing storytelling skills. Theory to practice*. Tucson, AZ: Communication Skill Builders.

Hickmann, Maya. 2003. *Children's discourse: Person, space and time across languages*. Cambridge: Cambridge University Press.

Hughes, Diana, LaRae McGillivray, Mark Schmidek. 1997. *Guide to narrative language: Procedures for assessment.* PRO-ED.

Kamandulytė, Laura. 2009. *Lietuvių kalbos būdvardžio įsisavinimas: leksinės ir morfosintaksinės ypatybės.* [Acquisition of Lithuanian adjective: lexical and morphosyntactic features]. Humanitarinių mokslų daktaro disertacija. [PhD dissertation]. Kaunas: Vytautas Magnus University.

Kamandulytė, Laura, Bettina Fürst, Wolfgang U. Dressler, Ulli Sauerland. 2010. On the acquisition of adjective gradation in Lithuanian and German. *Darbai ir dienos* 54, 267–276.

Kamandulytė-Merfeldienė, Laura. 2015. The acquisition of Lithuanian adjectives: Lexical ad morphological features. In *Semantics and morphology of early adjectives in first language acquisition.* Elena Tribushinina, Maria D. Voeikova & Sabrina Noccetti, eds. Cambridge: Cambridge Scholars Publishing. 313–346.

Kazakovskaya, Victoria V., Ingrida Balčiūnienė. 2012. Interrogatives in Russian and Lithuanian child-directed speech: Do we communicate with our children in the same way? *Journal of Baltic Studies* 43 (2), 197–218.

Kazakovskaya, Victoria V., Ingrida Balčiūnienė. 2015. Adult contribution towards early adjective acquisition: Evidence from Russian and Lithuanian longitudinal data. In *Semantics and morphology of early adjectives in first language acquisition.* Elena Tribushinina, Maria D. Voeikova & Sabrina Noccetti, eds. Cambridge: Cambridge Scholars Publishing. 243–312.

Kilani-Schoch, Marianne, Ingrida Balčiūnienė, Katharina Korecky-Kröll, Sabine Laaha, Wolfgang U. Dressler, 2009. On the role of pragmatics in child-directed speech for the acquisition of verb morphology. *Journal of Pragmatics* 41, 219–239.

Lounela, Mikko. 2005. Exploring morphologically analysed text material. In *Inquiries into words, constraints and contexts*. Antti Arppe Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, Anssi Yli-Jyrä, eds. Stanford: Stanford University, 259–267.

MacWhinney, Brian. 2017a. *Tools for analyzing talk. Part 1: The CHAT program.* Available at: http://talkbank.org/manuals/CHAT.pdf. Accessed: 19 August 2017.

MacWhinney, Brian. 2017b. *Tools for analyzing talk. Part 2: The CLAN program.* Available at: http://talkbank.org/manuals/CLAN.pdf. Accessed 19 August 2017.

McCabe, Allyssa, Pamela Rosenthal Rollins. 1994. Assessment of preschool narrative skills. *American Journal of Speech-Language Pathology* 3, 45–56.

McCabe, Allyssa, Lynn S. Bliss. 2003. *Patterns of narrative discourse*. Boston: Allyn & Bacon.

McTear, Michael. 1985. *Children's conversation*. Oxford: Basil Blackwell.

Pan, Barbara A., Catherine E. Snow. 1999. The development of conversational and discourse skills. In *The development of language*. M. Barrett, ed. Psychology Press.

Pinker, Steven. 1984. *Language learnability and language development*. Harvard University Press.

Savickienė, Ineta. 1999. *Lietuvio vaiko daiktavardžio morfologija.* [The acquisition of Lithuanian noun morphology]. Humanitarinių mokslų daktaro disertacija. [PhD dissertation]. Kaunas: Vytautas Magnus University.

Savickienė, Ineta. 2003. *The acquisition of Lithuanian noun morphology*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Savickienė, Ineta, Wolfgang U. Dressler, eds. 2007. *The acquisition of diminutives: a cross-linguistic perspective*. Amsterdam: John Benjamins Publishing Company.

Savickienė, Ineta, Vera Kempe, Patricia J. Brooks. 2009. Acquisition of gender agreement in Lithuanian: exploring the effect of diminutive usage in an elicited production task. *Journal of Child Language* 36, 477–494.

Snow, Catherine E. 1977. The development of conversation between mothers and babies. *Journal of Child Language* 4, 1–22.

Stephany, Ursula, Maria D. Voeikova, eds. 2009. *Development of nominal inflection in first language acquisition: a cross-linguistic perspective*. Berlin: Mouton de Gruyter.

Thorpe, Karen. 2006. Twin children's language development. *Early Human Development* 82, 387–395.

Thorpe, Karen, Rosemary Greenwood, Areana Eivers, Michael Rutter. 2001. Prevalence and developmental course of 'secret language'. *International Journal of Language and Communication Disorders* 36 (1), 43–62.

Tribushinina, Elena, Huub van den Bergh, Marianne Kilani-Schoch, Ayhan Aksu-Koç, Ineta Dabašinskienė, Gordana Hrzica, Katharina Korecky-Kröll, Sabrina Noccetti, Wolfgang U. Dressler. 2013. The role of explicit contrast in adjective acquisition. *First Language* 33 (6), 594–616.

Tribushinina, Elena, Maria D. Voeikova, Sabrina Noccetti, eds. 2015. *Semantics and morphology of early adjectives in first language acquisition*. Cambridge: Cambridge Scholars Publishing.

Westerveld, Marleen F., Gail T. Gillon, Catherine Moran. 2008. A longitudinal investigation of oral narrative skills in children with mixed reading disability. *International Journal of Speech-Language Pathology* 10 (3), 132–145.

Voeikova, Maria D., Wolfgang U. Dressler, eds. 2002. *Pre- and protomorphology: early phases of morphological development in nouns and verbs*. München: Lincom.

Wójcik, Pawel. 2000. *The acquisition of Lithuanian verb morphology: a case study*. Kraków: Quartis.

Submitted 14 December 2017
Accepted 18 September 2018