https://doi.org/10.18485/stijkk.2023.1.ch5
811.511.141'242'27-053.6
371.3:811.511.141'243

Boglárka Vermeki[1]
University of Belgrade – Faculty of Philology, Hungarology Department, visiting lecturer

FORMULAIC LANGUAGE IN TEENAGE SPEECH CORPORA

This study's primary focus is examining formulaic sequences in the spontaneous speech of teenagers. Previous research has demonstrated that proficient language users frequently employ fixed or semi-fixed lexical units in their discourse. These formulaic sequences play a crucial role in communication due to their coherence and fluency-enhancing properties. Acquisition of such formulaic language can also greatly benefit language learners, as it enables more natural speech production and fluency. Consequently, it is essential to thoroughly investigate and incorporate the most frequent formulaic sequences when composing educational materials for language teaching.

The methodology employed in this corpus-based research relies on the utilization of the KorSzak Child Language Corpus, a pedagogical spoken language corpus specifically designed for curriculum development. The corpus contains dialogues and monologues of children aged 11–15, discussing specified topics, which are the topics of the language teaching material in preparation. Within the scope of this study an extensive exploration of formulaic lexical units was conducted, leveraging a systematic approach to identify and categorize these linguistic constructs within the corpus under investigation.

The findings of this study provide valuable insights into the use of formulaic language by Hungarian children speaking on the recordings, thereby informing the development of effective language learning materials for this target population. During this phase of the research, nearly a thousand sequences were identified, which were categorized into the following groups based on their pragmatic functions: expressing doubt and uncertainty, conveying opinions (agreeing, disagreeing), emphasising, soliciting opinions, and repairing. The most frequent speech act with a relatively significant number of hits (275) was doubting and expression of uncertainty. However, valuable expressions can be

[1] Contact: vermekiboglarka@yahoo.com

utilised within each category to enrich teaching materials targeted at children learning Hungarian.

Key words: corpus linguistics, child language, spontaneous speech productions, formulaic language, fluency

Introduction

Pawley and Syder (1983) called the 'puzzle of native-like selection' the phenomenon in which competent language users select the most appropriate of all possible and grammatically acceptable formulations. The puzzle is why we choose that particular formulation and feel that the other possible structures do not sound good or are not English, Hungarian or Serbian enough. Due to the cognitive demands associated with spontaneous speech processing, conversations have a "limited and repetitive repertoire" (Biber et al., 1999: 1049) and depend on formulaic, "pre-constructed" lexical units (Wray, 2002). Corpus linguistic studies have shown that competent language users rely heavily on formulaic language during their speech production (O'Keefe et al., 2007) and that they do not make use of the limitless variation possibilities theoretically provided by grammar (Pawley–Syder, 1983). Several studies have shown a high proportion of formulaic lexical units in the discourse of native speakers (cf. Altenberg, 1998; Biber, Johnson, Leech, Conrad–Finegan et al., 1999; Pawley–Syder, 1998.). For instance, this is also proven by Altenberg's research, who, using the London-LundCorpus, a corpus of English spoken language, came to the conclusion that the structures that have become routine, partly or entirely prefabricated, have an impact on all levels of language organisation, from utterances to discourse. The structures found, which can best be described as more or less conventionalised expressions, are located along the border between fully lexicalised units and free constructions. He also estimated in his study that more than eighty per cent of the words in the corpus are part of word combinations (Altenberg, 1998). Erman-Warren (2000), in addition to the London-LundCorpus, examined a total of nineteen samples taken from the Lancaster-Oslo-Bergen written language corpus; based on their calculations, formulaic language makes up approximately fifty-five per cent of the speech and writing of competent language users (Erman–Warren, 2000). Another research showed that multiword speech made up 28% of the spoken and 20% of the written discourse studied, according to Biber et al. (1999). Lewis also concluded that "language consists of chunks which, when combined, produce continuous coherent text" (Lewis, 1997: 7). In her book, *Fluency in Native and Nonnative English Speech* (2013), Götz writes about the necessity for lexical units, which she calls *multiword sequences*. She concludes that these multiword sequences are learned in childhood as a routine and develop in adulthood. They do not

require much attention during speech. It does not require cognitive effort to use or extend each utterance. Research examining first language acquisition has also concluded that formulaic sequences are first memorised without analysis, and children only begin to analyse them later (Tomasello, 2003: 305–307).

Among the many types of research that deal with this topic, if we only take into account the ones listed above, we can already conclude that formulaic language use is indeed an essential part of communication. Competent language users know and actively use a significant amount of lexical units. So for language learners, using these units would also be important (Siyanova-Chanturia–Pellicer-Sánchez, 2019: 1). This idea is the basis of a larger project, of which the present research is also a part. The main aim of the project is to develop teaching material for children who are heritage language speakers of Hungarian, learning in the so-called Sunday or Weekend schools around the world.

To achieve the project's main aim, we are building a children's spontaneous speech corpus, which is needed to inform the teaching material. For several reasons, using corpora as a source in preparing teaching materials is important. In language teaching, corpora were usually used to create dictionaries for language learners. However, more and more teaching materials are created using them today because "in order to become a competent language user, it is essential that students become familiar with the expressions used by native speakers in everyday situations" (Szita–Pelcz, 2017: 263). Corpus-informed teaching materials reflect actual language use, and this way, they are effective tools for language learning. Therefore, we are basing the teaching material – among other sources – on the KorSzak Child Language Corpus and its examination. One research contributing to the realisation of the project's main aim is the present research on teenagers' formulaic language use.

Definition of formulaic sequences and their features

According to Wray (2019: 267), a formulaic sequence refers to any multiword string that is perceived by the agent (e.g., the language learner) to have an identity or usefulness as a single lexical unit. Formulaic language use in more detail is defined as follows:

"[Formulaic language is] a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray, 2002: 9).

Numerous terms can be found for this phenomenon in the literature. In her book, Wray (2002) collected more than fifty expressions that describe different aspects of formulaic language use. Among the fifty expressions, some emphasise the *meaning* (e.g., "idiom", "unit of meaning") and its structure (e.g. "sentence stem"), and others emphasise their *fixedness* (e.g., "multiword unit")

(Hunston, 2022: 102). Although there is a nuance of differences between them, they all try to describe the same phenomenon: they are *groups of words that often occur together in each other's company*.

In addition to formulaic sequences, the terms *lexical chunks*, *lexical bundles* and *multiword expressions* are mainly used in the English literature. According to Siyanova-Chanturia–Pellicer-Sánchez (2019: 2), it is best to use the term formulaic sequences or formulaic language use, because it can also include single-word items such as *exclamations* (e.g., hurrah) or various *speech formulas* (e.g., yeah), while if we use for example the notion multiword expressions, the above mentioned would be excluded from the definition, even though they function similarly to more extended multiword units.

The main characteristic of formulaic sequences is that they are holistically stored and recalled from memory, thus reducing the cognitive load (Wray, 2002). Besides the characteristic of being holistically stored, there are four other features of formulaic language: *frequency*, *familiarity* and *predictability*, *fixedness*, and *pragmatic function*, which are also important (Siyanova-Chanturia–Pellicer-Sánchez, 2019: 3).

Frequency is considered one of the most defining attributes, but research results show a low correlation between frequent formulaic sequences and the sequences learned by the studied group (cf. Siyanova-Chanturia –Pellicer-Sánchez, 2019: 4).

According to Siyanova-Chanturia-Pellicer-Sánchez, **familiarity**, convention, and **predictability** by the community are much more determining factors than that. An example of the latter is reading, during which the reader predicts the following words and even jumps in the text, thereby speeding up their reading speed (Siyanova-Chanturia–Pellicer-Sánchez, 2019: 4).

Based on their **fixedness**, these prefabricated or formulaic sequences can be fixed or semi-fixed, which means their parts are variable. For example, in Hungarian:

(1) őszintén szólva ('to be honest')

(2) jó emlékem van a *NOUN+-val/-vel* kapcsolatban ('I have good memories connected to NOUN')

Of the two examples above, the first is a fixed formulaic sequence, while the second is a partially fixed one. From a language learner's point of view, it is important to note that, in addition to making our speech more natural, such partially fixed formulaic sequences have a great communicative value since we only have to learn them once. However, they can be used in several ways and sometimes in several contexts.

Another key feature of formulaic sequences that makes them worth learning is that we often use them in **pragmatic functions** during our conversations.

(3) Ja, értem. ('Yeah, I see.')

(4) Én is egyetértek az XY-val/-vel. (’I also agree with Sb’)
(5) Nekem az lenne az első kérdésem, hogy… / Nekem az első kérdésem az lenne, hogy… (’ My first question would be that...’)

There are expressions that we use during interactions, for example, to *maintain communication* (3), to *express our agreement* (4) or to *organise our discourse* according to our communication goals (5) (Siyanova – Chanturia – Pellicer – Sánchez, 2019: 5).

Formulaic sequences in language acquisition

According to usage-based approaches to language acquisition (cf. Langacker, 1987; Bybee, 1985, 1995; Croft, 2001; Tomasello, 2003, 2009, 2015), a central element of first language acquisition is the use of formulaic language. Since the rise of corpus linguistics research, there is no doubt that formulaic language use is a significant part of children’s language and nursery language, i.e., child-directed speech. Several studies have been carried out on this topic, which shows that child-directed language is full of formulaic units (cf. Bannard & Matthews, 2008; Theakston - Lieven, 2017). These formulaic units are memorised by children after hearing them and are later rehearsed and used. In their experiment, Bannard and Matthews (2008) not only analysed the language directed towards children but also went further to investigate the use of formulaic language by children in more detail. Three-year-old children were asked to repeat certain words they knew and similar but unfamiliar words. The children were naturally more effective and faster at repeating items that they had heard several times in their everyday lives.

In addition to the research described above, the process of grammatisation (the term is also used in the form of grammaticalization cf. Dér, 2019), during which, in a general sense, a lexical item gradually acquires grammatical status, and which is in many ways the macro-level equivalent of the micro-level process during which children acquire the syntactic structure of language (Dörnyei, 2009: 210). According to Tomasello (2003: 5), during communication, words are interwoven into sequences, forming patterns of use that later solidify into grammatical structures. Ellis and Larsen-Freeman (2006: 567) argue that the process of grammatisation involves the automation of frequently occurring sequences of linguistic elements, which can lead to the emergence of syntactic constructions and then to changes in function as a result of frequent encounters. Frequency, however, plays an important role in language change and, at the micro-level, in language acquisition. According to researchers in usage-based theories, the pattern-finding function of children’s language processing systems, the inference of regularities from memorised constructions, is strongly dependent on the frequency of patterns (Ellis, 2002: 144). Ellis’s idea is confirmed by Bybee (2008), who writes that psycholinguistics, cognitive and functional linguistics,

computational linguistics, corpus-based analyses, and discourse analysis have all come to the realisation that linguistic knowledge is firmly based on linguistic experience, and that frequency of use is a fundamental determinant of the grammatical properties of language. Larsen-Freeman (2002: 281), however, points out that frequency alone cannot explain children's language acquisition (cf. Steinkrauss, 2017), something else is needed, which she calls probabilistic tendencies. This idea is based on Bob, Hay and Jannedy's book *Probabilistic Linguistics (2003)*, in which they explain that language learning should not be conceived as a minimal set of categorical rules or constraints, but as a set of gradient rules characterized by a statistical distribution (Dörnyei, 2009: 219). This implies that, in addition to frequency, the predictability of linguistic patterns and structures also plays a significant role, and that, contrary to previous theories, grammatical rules are associated with the probability of use. Thus, it is not the structure that is linguistically possible that is recorded, but the one that is thought to be linguistically probable, the one that has been encountered repeatedly in that context, in the company of already known linguistic elements.

The importance of formulaic language in language teaching

As mentioned earlier, formulaic language use is an essential part of the communication of competent language users, but it is also important for language learners. Numerous recent studies have been devoted to lexical units and formulaic language use. Formulaic language has been proven effective in developing the ability to communicate in a second language and contributes to speech fluency (Wray &Fitzpatrick, 2008). Segalowitz (2010: 126) states that "the ability to correctly use formulaic sequences contributes to the nativelike naturalness of speech, and to modulating the message-processing load to make communication easier and more efficient". Language learners who cannot use these lexical items properly will not be able to benefit from the use of such sequences, and this deficit may increase their speech processing burden, which may compromise their fluency. Hill (2000) also claims that learners who lack collocational competence employ longer, wordier phrases with grammatical faults and struggle with comprehension frequently. Using these formulaic sequences, by using collocations and expressions similar to those of competent language users, supports the fluency of speech and the proper use of the language (Kirk, 2014: 105). Lewis's Lexical Approach (1993) is based on these basic ideas. However, formulaic language has played a significant role in language teaching since the spread of the audio-lingual method (Tavakoli, 2020: 32). Even much earlier, Harold Palmer stated in 1925 about the English conversational language that the basic guideline for improving it may be to memorise groups of frequent, useful words (1925, 1999: 185). However, the active use of formulaic sequences improves language learners' receptive and productive skills. It is due to the fact that the recognition of frequently repeated

lexical units enables faster processing of linguistic input. It has been shown that language learners recognise words faster when they hear them in the company of words with which they normally appear.

Similarly, eye-tracking studies (Siyanova-Chanturia et al., 2011; Pellicer-Sánchez et al., 2022) have also shown that test subjects could read paragraphs that consisted of familiar lexical units faster. We can say that formulaic sequences are of fundamental importance from the point of view of language processing and language production. Their use allows our productive skills and our language use to be natural, fluent and accurate, and they also help with the development of receptive skills, reading and listening comprehension. In summary, formulaic language use is significant in the field of language learning and teaching because it:

- frequently occur in competent language user's language,
- aids fluency in speaking and writing,
- aids accuracy,
- makes learners use the target language more naturally,
- makes comprehension of texts easier, as it eases the load for the reader or listener.

Language corpora informing teaching material

As a result of the recognition of the importance of corpus use, more and more corpus-informed teaching materials are now appearing in language textbooks. Whatever the language, it can be argued that the use of corpora as a resource is important in the writing of curricula for several reasons (McCarthy - McCarten, 2022). One of the most important is that, in order to become competent language users, it is essential that learners become familiar with the expressions that native speakers use in everyday situations (Szita - Pelcz, 2017: 263). In turn, the natural language use of native speakers can be easily observed by relying on spoken and written language corpora (Conrad, 2000: 548). When the authors interpret their corpus research, McCarthy says that they have "three overarching goals in mind:

1. to identify authentic, motivating language,
2. to weave these findings into a carefully designed curriculum,
3. and to create textbooks that are familiar in structure and easy to use" (McCarthy, 2004: 15).

The use of corpora can support the work of curriculum developers in several areas (cf. McCarthy - McCarten, 2022). Using them can help avoid over-reliance on intuition and thus avoid misrepresented language use (McEnery - Xiao - Tono, 2006), help to create a levelled lexicogrammatical syllabus, and help to select appropriate texts and real-life situations in which language learners can practice the language elements they have learned (McCarten, 2010: 415;

McCarthy - McCarten, 2022). By examining the frequency of language forms, corpus linguistics can also help to make decisions about how the curriculum will reflect real language use (Conrad, 2000). In addition, the use of corpora can be very useful in other areas. These include the selection of useful vocabulary, the relationship of vocabulary to different text types (formal - informal style, written or spoken language), grammatical patterns (Kaltenböck - Mehlmauer - Larcher, 2005), and the observation of the use of discourse markers (McCarten, 2010). The use of corpora also facilitates the work of authors in areas such as determining the proportion of the phenomenon under study in a text or selecting the order of the linguistic forms to be studied (Meunier - Reppen, 2015: 501). In addition, Biber and Reppen suggest that authors of teaching materials should draw on data from frequency studies to increase the meaningful linguistic input available to learners (Biber - Reppen, 2002: 207). Authors should compare the vocabulary to be taught with the topic, the context in which competent speakers use it, and the grammatical patterns that often appear in their context, so that the texts in the corpora can serve as models for learners (Kaltenböck - Mehlmauer-Larcher, 2005: 72).

Crosthwaite (2019) mentions in his book that language teaching of children through the data-driven learning [DDL] method is not yet widely used and has only been studied by a few. The main reason for this is that language teachers find it too much of a task to introduce the computer programs required for data-driven language learning to younger learners. However, using corpora in language teaching would be beneficial to children, as well.

Not many pedagogical corpora were designed with the purpose of using it with children. Sinclair had planned to produce a language teaching corpus for primary school-age children in Scotland, but his death in 2007 meant he could not complete it. He wrote about the project, PhraseBox, in the local newspaper West Word:

"... it is like giving each pupil real-time access to a huge memory of all the different ways in which thousands of people have expressed themselves over several years, all instantly available in a highly organised presentation. Gradually, students are expected to internalise what they need of the resource and gather confidence in their ability to express themselves publicly; but the resource will always be available when it is needed" (Sinclair, 2006).

Later, others have also attempted to compile such corpora, mainly for teenagers. One such attempt was ELISA, the English Language Interview Corpus as a Second-Language Application (Braun, 2007), which consisted of 25 video interviews with native English speakers. Braun (2007: 322) tested the use of the corpus with twenty-six English language learners aged 14-15 for a month and found that the corpus-informed tasks created for it proved effective. A second pedagogical corpus for children, SACODEYL (System-Aided Compilation and Open Distribution of European Youth Language), was

developed between 2005 and 2008 as part of the European Union's Minerva project. The corpus was created to fill a gap in the corpus used in secondary school language teaching. The project website states, "SACODEYL sees itself as a pedagogical mediator in the language learning process of young Europeans, using ICT resources to provide opportunities for data-driven language learning based on a constructivist approach" (SACODEYL, 2008). Finally, a project in the pipeline on educational corpus for children exists. CorpusMate, developed by Peter Crosthwaite and Vít Baisa, provides simplified linguistic data analysis for children learning English as a second language in secondary schools. The project aims to integrate the best features of the available corpus tools into an easy-to-use digital environment. The corpus texts on the interface relate to secondary school and university subjects.

Methods and materials

About the corpus

The KorSzak Child Language Corpus, a dynamic corpus for pedagogical purposes, currently, in September 2022, contains seventy-three recordings of thirty-one children aged 11-15 (twelve boys and nineteen girls). The corpus is dynamic because it is constantly growing, and it is a corpus created for pedagogical purposes, which means the texts are annotated by predetermined topics. The main goal of corpus building is to create corpus-informed teaching material for children. The corpus contains more than 70,000 tokens and consists of sixty-four dialogues and nine monologues. The children-informants are classmates and friends learning in different schools in the countryside of Hungary. They are all monolingual Hungarian speakers, studying English or German as a foreign language in primary schools. They talk freely in pairs or small groups about a particular topic during the video and audio recordings. The task's duration was not defined, which is why there are recordings of 1-2 minutes and much longer ones, up to 45 minutes. The audio materials were transcribed using the speech recognition program called Alrite.[2] And after proofreading, they were uploaded to the Sketch Engine[3] interface.

Building the corpus began by defining the topics the children talked about during the recordings. A survey was conducted among 11-14-year-old children about what topics they like to talk about, what their favourite leisure activities are, and with whom they like to spend their free time. During the survey, in which 138 primary school children participated, they had to complete a mind map in pairs or small groups and answer open-ended questions. The topics were sorted according to age and frequency. Then we selected twelve of them (animals, dogs; trips; Harry Potter; hobbies: creative hobbies, sports;

[2] Website and more information: https://alrite.io/ai/hu/
[3] Website and more information: https://www.sketchengine.eu/

influencers; TV series; inventions, technical innovations; travel, talking about their own stories, food). For each topic, we created a mind map with short questions, and the purpose is that if the children get stuck during the recordings, they can get new ideas to continue by looking at the questions. (Baumann et al., 2021: 33-35). The process of recording the conversations started in the spring of 2020 and has been ongoing ever since.

Procedure

In order to identify formulaic sequences, corpus linguistic tools can be used. At the same time, the results cannot be obtained immediately using a single tool. When it comes to fixed lexical units, the so-called congram (i.e. Ngram with a variable end) searches can be used to find them. Another method is to find them manually from the concordance lines (Hunston, 2022: 102). In order to be able to establish that these word connections selected based on frequency are real formulaic sequences, specific criteria needed to be defined. Biber (2006), for example, looked for such lexical units (he refers to them as "lexical bundles") in university teaching materials and classroom language. The lexical units he found could be three or four words long, but he only used four-word units for his research. In addition, he also determined that the selected lexical units appear at least forty times in a corpus of one million words (Biber, 2006: 133–134). Since the corpus, I am using for the research is only a spontaneous speech corpus and since the Hungarian language has different features from English, it is necessary to define different criteria. Due to the agglutinating nature of the Hungarian language, I also list the two-word lexical units based on frequency in the first phase. Furthermore, I will examine the first five appearances due to the corpus size. Later, following the ratio defined by Biber (2006: 133–134), I plan to examine in detail only those lexical units that appear at least twenty-eight times in the corpus.

As for the formulaic language use, in the first current phase of the research, I only focus on lexical units consisting of several words and will analyse the one-word formulaic language use only later in the research. I examine the multiword units based on the frequency with the Sketch Engine N-grams search tool (cf. Biber, 2006), which generates frequency lists from token sequences. After I have selected the most frequently used formulaic language, I conduct concordance searches. Due to the word order characteristic of the Hungarian language, there may be expressions that the tool finds only with the second search or classifies them in separate groups. Here is an example of this issue:

(6) ezzel én is így vagyok

(7) én is így vagyok **ezzel/vele**

The above examples, which mean *'I think the same'*, clearly shows the challenges of working with Hungarian word order. The meaning of both sentences is the same, the children expressed their agreement on a particular

topic, but the search tool treated them as two separate N-grams during the first search due to the position of the pronoun 'ezzel', which means 'with this'. The pronoun 'ezzel' may be at the beginning or the end of this expression, while the other pronoun 'vele', which has a similar meaning ('with this or someone'), may be only at the end. Therefore, it is essential to examine the search results in detail and perform a second Key Word in Context (KWIC) concordance line analysis.

Table 1 Concordance line exact of 'én is így vagyok'

	Details	Left context	KWIC	Right context
1	doc#2	ɔn embertelen, vagy hogy mondjam. - Igen, ezzel	**én is így vagyok**	, hogyha már egyszer valamin ki kell próbálni, akk
2	doc#8	ıet kitalálni, mert nem ismered a személyiséget. -	**Én is így vagyok**	vele.</s><s>És milyen színű golden retriever lenn
3	doc#53	gymond bárkivel, akivel jól érzem magam, igen. -	**Én is így vagyok**	ezzel, mint az Amira, hogy én is bárkivel kimegyek
4	doc#53	yünk, akkor csak egyedül is szeretek sétálgatni. -	**Én is így vagyok**	ezzel a dologgal, hogyha családdal vagyunk, akkc
5	doc#60	ɣiket se nézem, én inkább az angol sorozatokat. -	**Én is így vagyok**	vele, mint az Eszter.</s><s>Igazából nem nagyon
6	doc#61	ɜm tudom abbahagyni, mivel izgalmas. - Hű ezzel	**én is így vagyok**	.</s><s>Én függővé válok és ez nagyon durva. - I
7	doc#68	n természeti látnivalók inkább jobban vonzanak. -	**Én is így vagyok**	vele, de például, amikor mentünk Firenzébe, akko

After the frequency analyses, as the last part of the process, I categorised the results obtained in this way according to the following speech acts: *expressing an opinion* (agreeing, disagreeing), *asking for an opinion*, *emphasising*, *repairing*, *doubting and expressing uncertainty.*

Results

During the present research, I found a total number of 1434 different multiword expressions, 945 expressions after lemmatization. Multiword lexical items are claimed to range from two to six words. However, this is only true for languages like English. In the initial stages, due to the agglutinating nature of the Hungarian language, the search tool found only a few 6-word, some 5 and 4-word units. However, most of the results were 3-2-word units.

Rank	Multiword expressions	Occurances
1	2-words	861
2	3-words	73
3	4-words	5
4	5-words	3
5	6-words	3

Table 2 Number of multiword expressions found in the corpus

Among the small number of 4-6-word findings, there are some useful semi-fixed (examples 8 and 9) and fixed expressions (example 10). Although these examples are not very frequent in the corpus, they are practical, and therefore it would be worthwhile to deal with them in more detail in a later phase of the research.

Example of a 6-word expression:

(8) jó emlékem van a sporttal kapcsolatban ('I have a good memory connected to sport/or another noun')

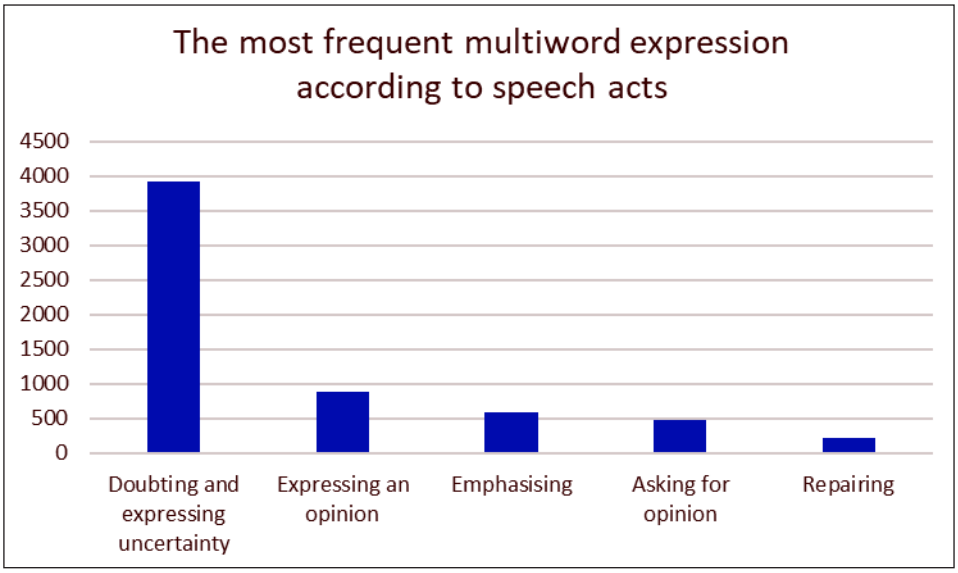
Example of a 5-word expression:

(9) általában a családommal szoktam kirándulni ('I usually go for a trip with my family / or another noun')

Example of a 4-word expression:

(10) semmi értelme nem volt ('it did not make any sense')

When examining the corpus, I found the most expressions belonging to the speech act *doubting and expressing uncertainty*.



Graph 1 The most frequent multiword expression according to speech acts

There are relatively few types of lexical units expressing this speech act, but they are found in the most significant number in the corpus. It is not only characteristic of children's speech but also of spontaneous speech itself since, unlike written communication, spontaneous speech is "real-time, improvised, and affects the capacity of the short-term memory of the listener and the speaker" (Andó, 2003: 98). For this reason, different lexical units are sometimes needed to spare time for the informants, so phrases such as *'I do not know'* can be used as time fillers. In the table below, different lexical units can be seen to

express uncertainty. The last two (4 and 5) are variations of the first two (1 and 2) completed with the word 'vagy', meaning 'or', which means they are also included in the first two findings.

Rank	Lexical unit	Translation	Hits	Frequency (Number of hits per million tokens)
1	nem (is) tudom	'I do not know'	275	3914.98
2	hogy (is) mondjam	'how should I say'	31	441.32
3	mit tudom én	'how should I know'	29	412.85
4	vagy nem (is) tudom	'or I do not know'	27	284.38
5	vagy hogy mondjam	'or how should I say'	18	256.25

Table 3 Multiword expressions showing 'doubt and uncertainty'

The second largest category was the *expression of opinion* due to the nature of the task. During the recordings of the conversations, children had to share their opinions on various topics, which is why the corpus contains many expressions related to this speech act. It is important to remember that children informants are friends and classmates. Consequently, they usually agree as they think similarly on specific topics. Even if they disagree, they only use lexical units expressing soft disagreement. We must be aware of this limitation of the research, but at the same time, it must also be noted that the circumstances did not allow for other types of recordings to be made for the time being.

Rank	Lexical unit	Translation	Hits	Frequency (Number of hits per million tokens)
1	ez / az igaz	'this/that is true'	23	327.43
2	hát én/nekem is	'well me too'	14	142.36
3	én is egyetértek	'I also agree'	8	113.89
4	(ezzel) én is így vagyok ezzel / vele	'I think the same'	7	99.65
5	igen, (az/ azok) abszolút	'yes, absolutely'	7	99.65
6	(én is) egyet tudok érteni	'I can agree, too'	6	85.42
7	én is inkább	'I would rather, too'	6	85.42
8	nekem is hasonló a véleményem / a helyzet / mint…-nél	'I have the same opinion'	4	42.71

Table 4 Multiword expressions showing agreement (the subcategory of expressing an opinion)

	Details	Left context	KWIC	Right context
1	doc#28	az utálódással kapcsolatban sajnos van, amikor	**egyet tudok érteni**	.</s><s>Nekem is van az osztályból, sőt a másik
2	doc#39	resszt. - Hát én ezzel a kapcsolattárssal teljesen	**egyet tudok érteni**	, meg mondjuk például, hogyha lehullott igazából
3	doc#39	rt kreatívabbak a gyerekek. - Ezzel én is teljesen	**egyet tudok érteni**	.</s><s>Szerintem is a felnőtteknek kicsit kevese
4	doc#39	át igen, foglalkozott vele. - Én is mindkettőtökkel	**egyet tudok érteni**	.</s><s>Szerintem is sokkal nagyobb értéke van
5	doc#60	an, másik világnak lehetek a része. - Ezzel én is	**egyet tudok érteni**	, szerintem jó kikapcsolódás.</s><s>És ha pihen
6	doc#60	vagy ilyesmi, ami egyébként meg jót tesz. - Én is	**egyet tudok érteni**	mindkettőtökkel, hogy ha nagyon sokat nézzük, a

Table 5 One of the multiword expressions to express agreement

Rank	Lexical unit	Translation	Hits	Frequency (Number of hits per million tokens)
1	(így) annyira nem	'not so much'	39	555.22
2	de szerintem ez/az	'but I think this'	10	142.36
3	nem azt mondom, hogy	'I am not saying that'	8	113.89
4	én / nekem nem nagyon	'not so much for me'	7	99.65
5	de szerintem nem	'but I do not think so'	3	42.71
6	hát nekem nem	'well not for me'	3	42.71

Table 6 Multiword expressions showing (soft) disagreement (the subcategory of expressing an opinion)

Closely related to this speech act is the category of opinion requests. The children informants did not use many variations of this speech act, but the following examples appeared several times.

Rank	Lexical unit	Translation	Hits	Frequency (Number of hits per million tokens)
1	Szerinted az fontos / lényeges, (hogy)…?	'do you think that is important'	34	484.03
2	Mit gondolsz azokról / arról, hogy…?	'what do you think about'	8	113.89
3	És szerinted…?	'and what do you think'	9	128.13
4	Neked mi a kedvenc…?	'what is your favourite'	2	28.47
5	Neked mi a véleményed…?	'what is your opinion'	2	28.47

Table 7 Multiword expressions of the speech act ,asking of an opinion'

Emphasis is expressed in many different ways, but the most typical are *'de még akkor is'*, *'és igenis'* and *'semmiképp(en) sem'*. The children used the following lexical units in the corpus:

Rank	**Lexical unit**	**Translation**	**Hits**	**Frequency (Number of hits per million tokens)**
1	az nem + Object/ Adjective/ Possession	'that is not'	41	583.69
2	(de még) akkor is	'but even then'	36	512.51
3	mennyire + Adjective	'how…'	26	370.14
4	pont a(z) …-ról/-ről (már) beszéltünk	'that is what we were talking'	5	71.18
5	és igenis	'igenis' is used for expressing the contrary of something	4	56.95
6	semmiképp se(m)	'by no means'	3	42.71
7	mindegyik ugyanolyan fontos / aranyos	'they all are equally important/nice…'	2	28.47

Table 8 Multiword expressions of the speech act 'emphasis'

Just as *doubting and expressing uncertainty* is characteristic of spontaneous speech, so is *'repairing'*, which is why this category also brought a relatively significant number of hits. The children used the following multiword units.

Rank	**Lexical unit**	**Translation**	**Hits**	**Frequency (Number of hits per million tokens)**
1	tehát, hogy	'so that'	16	227.78
2	tehát nem	'so not'	8	113.89
3	nem feltétlenül, (hanem)	'not necessarily'	7	99.65
4	mármint, hogy (nem)	'I mean that'	5	71.18

Table 9 Multiword expressions of the speech act 'doubting and expressing uncertainty'

Suppose we collect the lexical units categorised according to the different speech acts. In that case, we can see that overall, the most frequently used ten multiword lexical units in the KorSzak Child Language Corpus are the followings:

Rank	Lexical unit	Translation	Hits	Frequency (Number of hits per million tokens)	Speech act
1	nem (is) tudom	'I do not know'	275	3914.98	doubting and expressing uncertainty
2	az nem + Object/ Adjective/ Possession	'that is not'	41	583.69	emphasising
3	(így) annyira nem	'not so much'	39	555.22	expressing an opinion (disagreement)
4	(de még) akkor is	'but even then'	36	512.51	emphasising
5	Szerinted az fontos / lényeges, (hogy)…?	'do you think that is important'	34	484.03	asking for an opinion
6	hogy (is) mondjam	'how should I say'	31	441.32	doubting and expressing uncertainty
7	mit tudom én	'how should I know'	29	412.85	doubting and expressing uncertainty
8	mennyire + Adjective	'how…'	26	370.14	emphasising
9	ez / az igaz	'this/that is true'	23	327.43	expressing an opinion (agreement)
10	tehát, hogy	'so that'	16	227.78	repairing

Table 10 The most frequent multiword lexical units in the KorSzak Child Language Corpus

Discussion and conclusion

The primary purpose of the current study on the KorSzak Child Language Corpus, which is a dynamic, spontaneous speech corpus with a pedagogical purpose, was to learn and categorise the multiword lexical units most frequently used by the informants according to speech acts. The research was necessary

to show the first directions for preparing a Hungarian teaching material being edited for children who are heritage language speakers of Hungarian. Using the research results for the teaching material is necessary because corpus-informed teaching materials based on corpus linguistic investigations are more effective since they reflect natural language use. Just as it is for adult students, it is also essential for children to know the language use of their L1 Hungarian-speaking peers to use their acquired knowledge in a target language environment easily. The examination of formulaic language contributes to the knowledge of natural language use because it is known from several studies (cf. Altenberg, 1998; Erman - Warren, 2000) that these lexical units appear in a large proportion in the speech of competent language users. Research shows that children learn these formula-like sequences all at once without analysing them (Tomasello, 2003: 305–307). Formulaic sequences can be similarly mastered by second or heritage language learners, as learning them brings many benefits. Among others, the two most important benefits of learning formulaic sequences are the following:

1) It develops productive skills (speaking and writing comprehension) because the time and energy freed up by recalling the holistic stored units from memory can be utilised in other areas. Language learners become more fluent by using these sequences.

2) It develops receptive skills (reading, listening comprehension). Since the recognition of frequently repeated lexical units enables faster processing of linguistic input, if we know a collection of formulaic sequences during reading and listening, we can create predictions by which we can understand texts faster and more successfully. Similar to competent language users.

In the first phase of the research, multiword lexical units were examined. Due to the agglutinating nature of the Hungarian language, most of these formulaic sequences consist of two or three words. However, we can also find examples of multiword expressions, including 4-6 words. The present research shows that five speech acts appear most often in the KorSzak Child Language Corpus. Based on their order of frequency, they are the following:

1. doubting and expressing uncertainty,
2. expressing an opinion (agreeing, disagreeing),
3. emphasising,
4. asking for an opinion,
5. repairing.

The most frequent speech act with a relatively significant number of hits (275) is *doubting and expression of uncertainty*. However, it should be noted that children often use multiword lexical units in this category as time fillers while thinking during a conversation. This aspect of the expressions used should be examined in more detail. The second most frequent category is *expressing an opinion*, and within it, the subcategories of *agreement and disagreement*.

Considering the expressions belonging to the disagreement subcategory, it should be mentioned that the relationship of the informants to each other is friendship, which influences the results since mostly soft disagreement expressions can be found among them.

Formulaic sequences selected on the basis of frequency are used in several ways in the teaching material. I will only mention a few because this could be the subject of a separate study. On the one hand, they are taken into account in the didacticization of spoken language texts, the frequent elements are not taken out of the texts but are presented several times in several contexts so that learners notice, practice and later use the expressions themselves. In speaking tasks, these expressions are used as a model, facilitating learners' linguistic production, as they do not have to compose their whole utterance themselves but can select parts of it from a database of frequently used formulaic language. This speeds up their speaking and comprehension and allows them to concentrate on the relevant information in their conversations, which leads to developing their fluency.

Limitations of the study

While this study contributes valuable insights into the role of formulaic sequences in teenagers' speech productions, it is important to acknowledge certain limitations that may have influenced the findings and implications of the research.

Firstly, the recordings for this study were collected from two counties in Hungary. As a result, the sample may not be fully representative of the entire population of Hungarian-speaking children. Hungary's regional and cultural variations might affect formulaic sequences' frequency and usage patterns. Therefore, caution should be exercised when generalizing the results to the broader Hungarian-speaking population.

Secondly, the recordings in the corpus are not balanced, with certain topics having a higher number of recordings (e.g., Animals) compared to others. This imbalance in the data might introduce a bias in the analysis. Future research should aim to collect a more balanced and diverse dataset to mitigate this limitation.

Another limitation of the study relates to the inclusion of monologues in the corpus. While monologues can provide valuable insights into individual language production, they may differ in terms of formulaic sequence usage compared to dialogues. Including monologues alongside dialogues might introduce data variability and could affect the overall analysis and interpretation of formulaic sequences. Future studies could consider separating monologues and dialogues for more focused analysis or explore the specific characteristics and functions of formulaic sequences in each type of speech.

Lastly, the categorization of formulaic sequences in this study was conducted by a single researcher. Although efforts were made to ensure reliability and accuracy, the categorization process could benefit from the involvement of multiple researchers to enhance inter-rater reliability.

The obtained results are only partial results of the first phase, but by getting to know the preliminary results and limitations of the research, we got closer to realising the project's main goal. Further investigations are needed in the next period and the following phases. It is necessary to examine the expressions that appear in the categories of individual speech acts in more detail, conduct further frequency studies, and analyse formulaic language use, taking into account the characteristics of the Hungarian language.

References

Altenberg, B.: 1998, On the phraseology of spoken English: the evidence of recurrent word-combinations, In A.O. Cowie (Ed.), *Phraseology: theory, analysis and application*, Oxford: Oxford University Press, 101–122.

Andó É.: 2003, Beszélt nyelvi narratívumok szerkezeti összetevőinek és beszédtempójának összefüggése. In Tóth Szergej (Ed.): *Nyelvek és kultúrák találkozása. A XII. Magyar Alkalmazott Nyelvészeti Kongresszus kiadványai III.,* Szeged, 98–103. web: https://manye.hu/wp-content/uploads/2020/12/Toth-Szergej-szerk.-NYELVEK-ES-KULTURAK-TALALKOZASA.pdf (26.09.2022)

Bannard, C., Matthews, D.: 2008, Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19(3), 241–248.

Baumann T., Majoros J., Pelcz K., Schmidt I., Szita Sz., Vermeki B.: 2020, Bemutatkozik a Korpusznyelvészeti és Szakmódszertani Munkacsoport. *Hungarológiai Évkönyv* 21: 1-2, 32-41.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E.: 1999, *Longman Grammar of Spoken and Written English*. Harlow: Longman

Biber, D., Reppen, R.: 2002, What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, 24, 199–208.

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

Bod R. Hay J., Jannedy S.: 2003, *Probabilistic linguistics*. MIT Press.

Braun, S.: 2007, Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, 19 (3), 307–328.

Bybee, J.: 2008, Usage-Based Grammar and Second Language Acquisition. In: Robinson P. J., Ellis, N. C. (Eds.). *Handbook of cognitive linguistics and second language acquisition*. Routledge.

Bybee, J.: 1985, *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J.: 1995, Regular morphology and the lexicon. *Language and Cognitive Processes*, 10 (5), 425–455.

Conrad, S.: 2000, Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, 34, 548–560.

CorpusMate: 2023, *Information*. Web: https://corpusmate.baisa.cz/info (Utoljára megtekintve: 2023.04.11).

Croft, W.: 2001, *Radical Construction Grammar: Syntactic Theory In Typological Perspective*. Oxford: Oxford University Press.

Crosthwaite, P. (Ed.): 2019, *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners*, Routledge.

Dér Cs. I.: 2019, *Grammatikalizáció*. Budapest: Akadémiai Kiadó.

Dörnyei, Z.: 2009, *The Psychology of Second Language Acquisition - Oxford Applied Linguistics*. Oxford University Press.

Ellis, N. C., Larsen-Freeman, D.: 2006, Language Emergence: Implications for Applied Linguistics--Introduction to the Special Issue. *Applied Linguistics*, 27(4), 558–589.

Ellis, N. C.: 2002, Frequency effects in language processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, 24 (2), 143–188.

Erman, B., Warren, B.: 2000, The idiom principle and the open choice principle, *Text & Talk*, 20(1), De Gruyter Mouton, 29–62.

Hill, J.: 2000, Revising priorities: from grammatical failure to collocational success. In Lewis, M. (Ed.) *Teaching Collocation: Further Developments in the Lexical Approach*. London: Heinle Cengage Learning, 47-67.

Hunston, S.: 2022, *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kaltenböck, G., Mehlmauer-Larcher, B.: 2005, Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. *ReCALL*, 171, 65–84.

Kirk, S.: 2014, Addressing Spoken Fluency in the Classroom. In Muller, T. (2014). *Exploring EFL fluency in Asia*. 101–119.

Langacker R. W.: 1987, *Foundations of cognitive grammar*. Stanford University Press.

Larsen-Freeman, D.: 2006, The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics*, 27 (4), 590–619.

Lewis M.: 1993, *The lexical approach: the state of elt and a way forward*. Language Teaching Publications.

Lewis, M.: 1997, *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

McCarthy M.: 2004, *Touchstone–From Corpus to Course Book*, Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J.: 2022, Writing corpus-informed materials. In: Norton, J., Buchanan, H. (Eds.). *The Routledge Handbook of Materials Development for Language Teaching*. Routledge. 170–183.

McEnery, T., Xiao, R., Tono, Y.: 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meunier, F., Reppen, R.: 2015, Corpus versus non-corpus-informed pedagogical materials: grammar as the focus In: *The Cambridge Handbook of English Corpus Linguistics*, Biber, D., Reppen, R. (Eds.). Cambridge: Cambridge University Press.

Nattinger, J.R., DeCarrico, J.S.: 1992, *Lexical Phrases and Language Teaching*, Oxford: Oxford University Press.

Palmer, H.: 1925, Conversation: The fundamental guiding principle for the student of conversation, In: Smith, R.C. (1999) *The Writings of Harold E. Palmer: An Overview*, Tokyo: Hon-no-Tomosha. web: http://homepages.warwick.ac.uk/~elsdr/WritingsofH.E.Palmer.pdf (12.06.2022)

Pawley, A., Syder, F.H.: 1983, Two puzzles for linguistic theory: Native-like selection and native-like fluency. In: Richards, J., Schmidt, R. (Eds), *Language and Communication*, London: Longman. 191–226.

Pellicer-Sánchez, A., Siyanova-Chanturia, A., Parente, F.: 2022, The effect of frequency of exposure on the processing and learning of collocations: A comparison of first and second language readers' eye movements. *Applied Psycholinguistics*, 43(3), 727-756. doi:10.1017/S014271642200011X

SACODEYL: 2004-2008, http://webapps.ael.uni-tuebingen.de/backbone-search/faces/search.jsp

Segalowitz, N.: 2010, *Cognitive bases of second language fluency*. New York, NY: Routledge.

Sinclair, J.: 2006, A language landscape. *West Word*. Web: www.westword.org.uk/jan2006.html (Utoljára megtekintve: 2022.08.02.)

Siyanova-Chanturia, A., Pellicer-Sanchez, A.: 2019, *Understanding formulaic language: A second language acquisition perspective*. New York, NY: Routledge.

Siyanova-Chanturia, A., Conklin, K., Schmitt, N.: 2011, Adding more fuel to the fire: An eye-tracking study of idiom processing by native and nonnative speakers, *Second Language Research*, 27 (2), 1-22.

Steinkrauss, R.: 2017, 5. L1 acquisition beyond input frequency. In: Evers-Vermeul, J., Tribushinina, E. (Eds.). *Usage-Based Approaches to Language Acquisition and Language Teaching.* Berlin, Boston: De Gruyter Mouton. 117–142.

Szita Sz., Pelcz K.: 2017, Modellalapú nyelvoktatás – Természetes nyelvhasználat a tanteremben és a tantermen kívül. *THL: A magyar nyelv és kultúra tanításának szakfolyóirata,* 1–2., 262–269.

Tavakoli P., Wright C.: 2020, *Second language speech fluency: from research to practice.* Cambridge: Cambridge University Press.

Theakston, A., Lieven, E.: 2017, Multiunit Sequences in First Language Acquisition, *Topics in Cognitive Science*, 9, 588–603.

Tomasello, M., Dweck, C. S., Silk, J. B., Skyrms, B., Spelke, E. S.: 2009, *Why we cooperate.* MIT Press.

Tomasello, M.: 2003, *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M.: 2015, The usage-based theory of language acquisition. In: Bavin, E., Naigles, L. (Eds.). *The Cambridge Handbook of Child Language*. Cambridge: Cambridge University Press. 89–106.

Wray, A., Fitzpatrick, T.: 2008, Why can't you just leave it alone? Deviations from memorised language as a gauge of nativelike competence. In: F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching*, Amsterdam: John Benjamins. 123–148.

Wray, A.: 2002, *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press.

Wray, A.: 2019, Concluding Question, Why Don't Second Language Learners More Proactively Target Formulaic Sequences? In: Siyanova-Chanturia, A., Pellicer-Sanchez, A. (Eds.) *Understanding formulaic language: A second language acquisition perspective*, New York, NY: Routledge. 248 – 269.

FORMULAIČNI JEZIK U GOVORNOM KORPUSU TINEJDŽERA

S a ž e t a k

Ovaj istraživački rad se bavi nedostatkom udžbenika na mađarskom jeziku za decu na osnovu istraživanja korpusa i nedostatkom nedavnih istraživanja upotrebe jezika kod dece, posebno u vezi sa formulačkim jezikom. Primarni fokus ove studije je ispitivanje formalnih sekvenci u spontanom govoru tinejdžera. Prethodna istraživanja su pokazala da iskusni korisnici jezika često koriste fiksne ili polufiksne leksičke jedinice u svom diskursu. Ove formulaične sekvence igraju ključnu ulogu u komunikaciji zbog svoje koherentnosti i svojstava koja poboljšavaju tečnost. Usvajanje takvog formulačnog jezika takođe može od velike koristi učenicima koji uče jezik, jer omogućava prirodniju proizvodnju i tečnost govora. Shodno tome, neophodno je temeljno istražiti i uključiti ove formulne sekvence prilikom sastavljanja obrazovnih materijala za nastavu jezika.

Metodologija korišćena u ovom istraživanju zasnovanom na korpusu oslanja se na korišćenje korpusa dečijeg jezika KorSzak, pedagoškog korpusa govornog jezika posebno dizajniranog za razvoj kurikuluma. Korpus sadrži dijaloge i monologe dece uzrasta 11–15 godina, koji govore o određenim temama.

Nalazi ove studije pružaju vredan uvid u upotrebu formulačnog jezika od strane mađarske dece, čime se informišu o razvoju delotvornih materijala za učenje jezika za ovu ciljnu populaciju. Tokom ove faze istraživanja identifikovali smo skoro hiljadu sekvenci, koje su kategorisane u sledeće grupe na osnovu njihovih pragmatičnih funkcija: izražavanje sumnje i neizvesnosti, prenošenje mišljenja (slaganje, neslaganje), isticanje, traženje mišljenja i popravljanje. Najčešći govorni čin sa relativno značajnim brojem pogodaka (275) bila je sumnja i izražavanje nesigurnosti. Međutim, vredni izrazi se mogu koristiti u svakoj kategoriji kako bi se obogatili nastavni materijali namenjeni deci koja uče mađarski.

Ključne reči: korpusna lingvistika, dečji jezik, spontana govorna produkcija, formulački jezik, tečnost jezika