Hungarológiai Közlemények 2023/4. Bölcsészettudományi Kar, Újvidék
Papers of Hungarian Studies 2023/4. Faculty of Philosophy, Novi Sad

ETO: 81'42
81'276.3-053.5
DOI: 10.19090/hk.2023.4.106-121

ORIGINAL SCIENTIFIC PAPER

VERMEKI Boglárka

Belgrádi Egyetem, Filológiai Kar
Hungarológia Tanszék
Belgrád, Szerbia
vermekiboglarka@yahoo.com

A GYERMEKEK SPONTÁN BESZÉDÉNEK JELLEMZŐI

Szógyakorisági és kulcsszóvizsgálatok a KorSzak Gyermeknyelvi Korpuszon

Characteristics of children's spontaneous speech

Word frequency and keyword analysis in the KorSzak Child Language Corpus

Karakteristike spontanog govora dece

Analiza učestalosti reči i ključnih reči u korpusu dečijeg govora KorSzak

A tanulmány célja, hogy korpusznyelvészeti vizsgálatok – szógyakoriság- és kulcsszóelemzések – segítségével tárja fel a kutatásban részt vevő gyermekek nyelvhasználatának tulajdonságait, különös tekintettel szókincsük összetételére. A kutatás alapjául szolgáló KorSzak Gyermeknyelvi Korpusz (Baumann et al. 2020) pedagógiai célú, dinamikus korpusz, amely jelenleg huszonkilenc adatközlő, 11–15 éves gyermek 78 felvételéből áll. A videó- és hangfelvételek során a gyermekek egy-egy adott témáról (például állatok, szabadidős tevékenységek) párban vagy kiscsoportban beszélgetnek szabadon. A jelen tanulmányban bemutatott eredmények egy PhD-képzés során készült kutatás részét képezik, amelynek célja a formulaszerű nyelvhasználat és lexikogrammatikai mintázatoknak a korpuszban történő megfigyelése, valamint az eredmények felhasználásával egy korpuszinformált gyermekek számára szerkesztett magyar mint idegen nyelvi tananyag készítése.
Kulcsszavak: korpusznyelvészet, gyermeknyelv, spontán beszéd, nyelvtanítás

Bevezetés

Az 1970-es évektől a korpusznyelvészet fejlődése fontos változásokat hozott az alkalmazott nyelvészetben, nagy hatást gyakorolt a nyelvek tanulmányozására (Hunston 2022, 1). A számítógép és az internet térhódításával pedig egyre

elterjedtebbé vált, és ma már a nyelvtanítás során többféleképpen is alkalmazzák a kutatások eredményeit. Ilyen terület például a tananyagkészítés.

A jelen tanulmányban bemutatott szógyakorisági és kulcsszóvizsgálatok egy doktori kutatás részét képezik, amelynek célja a KorSzak Gyermeknyelvi Korpusz adatközlő gyermekeinek nyelvhasználatában megfigyelhető formulaszerű elemek, valamint lexikogrammatikai mintázatok feltérképezése, hogy a kapott eredmények felhasználásával egy gyermekek számára szerkesztett korpuszinformált magyar mint idegen nyelvi tananyag jöhessen létre. A tanulmányban a KorSzak Gyermeknyelvi Korpusz céljának, készítési folyamatának és felépítésének bemutatása után, az első eredmények közlésére, valamint azok rövid elemzésére kerül sor.

A Korpusznyelvészeti és Szakmódszertani Munkacsoport

A KorSzak Gyermeknyelvi Korpusz a Korpusznyelvészeti és Módszertani Munkacsoport (KorSzak) által szerkesztett korpusz része, amely 2020 februárjában jött létre azzal a céllal, hogy áthidalja a korpusznyelvészeti kutatások és a magyar mint idegen nyelv oktatásmódszertana közötti szakadékot (Baumann et al. 2020, 32). Az alapító tagok felismerték, hogy a korpuszalapú empirikus kutatási eredmények gyakran nem épülnek be a nyelvtanítás módszertanába, amely pedig elengedhetetlen a megalapozott módszertani fejlesztéshez és a sikeres nyelvoktatáshoz. A korpuszkutatások eredményeinek hasznosítása ugyanis számos előnnyel jár a nyelvoktatásban és a tananyagkészítésben, hiszen betekintést nyerhetünk a nyelvi elemek használatának gyakoriságába, használatuk jellemző kontextusaiba és a hozzájuk kapcsolódó grammatikai jellemzőkbe (vö. Hoey 2005). Így ezen adatok segítségével megalapozottabb döntéseket tudunk hozni a természetes nyelvhasználatról, ahelyett, hogy kizárólag a saját intuícióinkra hagyatkoznánk (Hoey 2005; Hunston 2002; O'Keeffe et al. 2007). A Korpusz Munkacsoport megbízható, bizonyítékokon alapuló információkkal kíván hozzájárulni a nyelvoktatók, a tanulók és a tananyagtervezők munkájához. E célok elérése érdekében több almunkacsoport alakult, amelyek többféle korpuszt hoznak létre (Baumann et al. 2020, 32–33), így például:

1. élőnyelvi korpusz Pelcz Katalin vezetésével,
2. tanulói korpusz Baumann Tímea vezetésével és
3. gyermeknyelvi korpusz Vermeki Boglárka vezetésével.

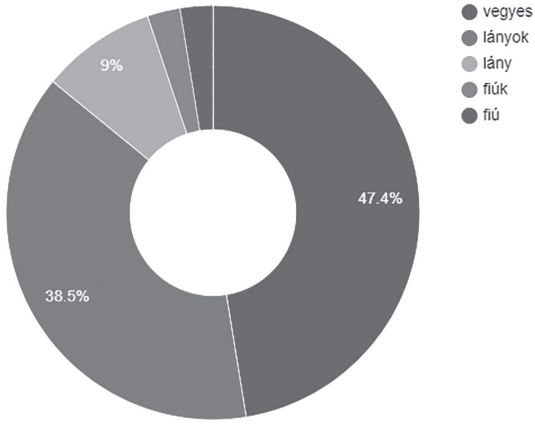
A KorSzak munkacsoport 2020 februárjában, a Pécsi Tudományegyetem Nemzetközi Oktatási Központjában megtartott alakuló ülése után, szinte azonnal

elkezdte a fent említett korpuszok létrehozását. Azóta rengeteg adatmennyiség gyűlt össze, a korpuszok folyamatosan bővülnek. A munkacsoport taglétszáma pedig az alakulás óta több mint duplájára emelkedett.

A KorSzak Gyermeknyelvi Korpusz

A KorSzak Gyermeknyelvi almunkacsoport célja olyan pedagógiai célú korpusz felépítése, amely különböző, a gyermekek számára is releváns és érdekes témákkal foglalkozik. Elsődleges célkitűzése, hogy biztosítsa azt a korpuszt, amelynek vizsgálata során nyert adatok felhasználásával a magyart idegen- vagy származásnyelvként tanuló gyermekek számára tananyag készülhet (Baumann et al. 2020, 33).

Mivel a korpuszépítés egyik fontos célkitűzése a korpuszinformált tananyagok készítése, a KorSzak Gyermeknyelvi Korpusz a mintavétel módja szerint dinamikus, a felhasználási módja szerint pedig egy speciális, pedagógiai célú korpusz. A korpusz dinamikus, mert folyamatosan növekszik. Jelenleg, 2022 októberében hetvennyolc felvételt tartalmaz huszonkilenc 11–15 éves gyermektől (kilenc fiú és húsz lány).



1. ábra. A lányok és fiúk aránya a KorSzak Gyermeknyelvi Korpuszban (2022. október)

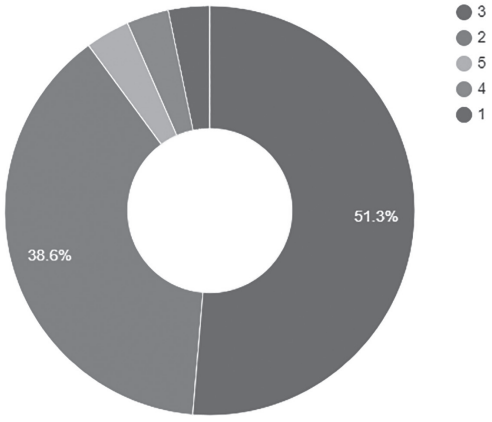
A diagramon (1. ábra) látható a rögzített felvételeken szereplő gyermekek nemek szerinti megoszlása. A vegyes kifejezés arra utal, hogy az adott felvételeken lányok és fiúk is hallhatóak, a lányok és fiúk jelentése pedig, hogy a rögzített felvételeken több gyermek is hallható, amíg az egyes számban szereplő lány és fiú kategória a monológokban szereplőkre utal.

A KorSzak Gyermeknyelvi Korpusz jelenleg, 2022 októberében több mint 71 000 tokent tartalmaz. A videó- és hangfelvételek során a gyermekek párokban vagy kiscsoportokban szabadon beszélgetnek egy-egy témáról. A feladat időtartama nem volt meghatározva, ezért vannak 1-2 perces és ennél jóval hosszabb, akár 45 perces felvételek is, a már említett 2020-ban Zoom videókonferencia program segítségével készült felvételek. A korpusz hatvannyolc dialógusból és tíz monológból áll.

Típus	Felvételek száma	Tokenek száma
Dialógus	68	69 101
Monológ	10	2445

1. táblázat. A KorSzak Gyermeknyelvi Korpuszban lévő szövegek típusai

A monológok a 2021-es évben először a járvány miatt keletkeztek, majd később a személyleírás témakör bevezetése miatt további felvételekkel gyarapodott a korpusz. A dialógusok során 2–5 adatközlő beszélget egymással az adott témáról. A felvételek nagy arányában két vagy három gyermek beszélget egymással (2. ábra). Mindösszesen 3,3 százalékukban szerepel négy, és 3,5 százalékukban öt beszélő.



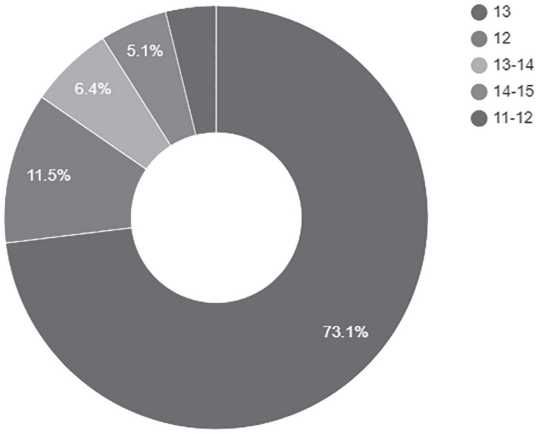
2. ábra. Beszélők számának felvételenkénti eloszlása a KorSzak Gyermeknyelvi Korpuszban

A huszonkilenc adatközlő gyermek magyarországi vidéki iskolákban tanuló osztálytársak és barátok, vagy csak barátok. A felvételek 64,1%-a Pest megyében, 35,9%-a Baranya megyében készült. Az első rögzítési fázisban, 2020-ban főleg Baranya megyei iskolások vettek részt. Ahogy már említettem, ekkor

készültek a hosszabb felvételek, ezért, ha a tokenek számának lefedettségét nézzük, a korpusz 84,1%-a Baranya megyéből származik.

A legtöbb gyermek egynyelvű magyar anyanyelvű, két gyermeknek (5. és 16. adatközlő) az édesapja nem magyar származású (holland, olasz), de otthon magyar nyelven kommunikálnak. Az adatközlő gyermekek az általános iskola felső tagozatán angol (83%) vagy német (17%) nyelvet tanulnak idegen nyelvként.

A felvételek nagy része a 13 éves korosztállyal készült, de a korpuszban szereplő gyermekek életkori megoszlása 11–15 éves korig terjed. Az egyes felvételeken az osztálytársak életkorában lehet életérés, valamint vannak olyan adatközlők, akik többször is adtak mintát a három év során, így az adatbázisban több életkorral is szerepelnek.



3. ábra. A KorSzak Gyermeknyelvi Korpusz felvételein szereplő gyermekek életkori megoszlása

Módszertan

A jelen tanulmányban bemutatott kutatáshoz az Adam Kilgarriff által fejlesztett Sketch Engine[1] elnevezésű korpusznyelvészeti eszközöket tartalmazó online felületet használtam, amelyre nemcsak azért esett a választásom, mert felhasználóbarát, hanem azért is, mert modern, sokféle funkcióval rendelkezik, és meglehetősen gyorsan fejlődik. A vizsgálatok elvégzéséhez a Worldlist és a Keywords elnevezésű korpuszeszközt alkalmaztam, amelyeket a következőkben röviden ismertetek.

[1] https://www.sketchengine.eu/

Gyakorisági listák készítése

A korpusznyelvészeti kutatások fő kvantitatív vizsgálata a gyakorisági vizsgálat, amely azt mutatja meg, hogy milyen gyakran fordul elő egy lemma, szó vagy szókapcsolat az adott korpuszban. A vizsgálatok során készült listákon láthatjuk a keresett elem összes előfordulási típusát az előfordulási számukkal együtt. A gyakorisági vizsgálatokkal kapcsolatban két lényeges aspektusra kell figyelni. Egyrészt a vizsgálat során kapott gyakorisági mutatóknak összehasonlíthatóknak kell lenniük. Másrészt pedig fontos megjegyezni, hogy az ilyen elemzések során kapott számadatok önmagukban nem informatívak, kontextusba helyezve értelmezhetőek megfelelően (Hunston 2022, 82–84).

A Wordlist, vagyis szólista funkció, a Sketch Engine által kínált korpuszeszközök egyike, amely lehetővé teszi a felhasználók számára, hogy gyakorisági mutatók alapján szólistákat hozzanak létre. Ennél a funkciónál kétféle, úgynevezett egyszerű vagy összetett keresési lehetőségek állnak rendelkezésünkre. Az egyszerű keresések szófaj, szavak, lemmák és olyan kritériumok, mint a kezdő- vagy záróbetű/frázis alapján végezhetők. Az összetett keresések testre szabhatók címkékkel, lempos-okkal (a lemma és a szófaj angol megfelelőjének rövidítése), szövegtípus vagy alkorpusz meghatározással és speciális kereső karakterekkel vagy szimbólumokkal [Regular Expressions – RegEx]. Az összetett keresésben lehetőség van olyan szólisták beillesztésére is, amelyeknek a korpuszban való gyakoriságára vagyunk kíváncsiak. Ezenkívül meghatározhatjuk, hogy mit zárjon ki és mit foglaljon bele a listába a program, a kis- és nagybetűk használatát, a gyakorisági minimumot és maximumot, valamint, hogy mely attribútuma alapján rendeződjön a lista (Sketch Engine é. n. a).

A Wordlist eszköz a tokenek szintjén működik, és összesen hat gyakorisági mutatót lehet megismerni a használata által. Megmutatja

1. a listában szereplő elemek gyakoriságát a korpuszban,
2. az elemek gyakoriságát per egymillió szó,
3. a dokumentumgyakoriságot [Document Frequency – DOCF], amely kimutatja, hogy hány feltöltött dokumentumban jelennek meg a listában szereplő elemek,
4. a relatív dokumentumgyakoriságot [Relative Document Frequency], amelyből megtudhatjuk, hogy a dokumentumok hány százalékában jelennek meg a listában szereplő elemek,
5. a redukált gyakoriságok átlagát [Average Reduced Frequency – ARF], amely nem veszi figyelembe a listán szereplő elemek közeli előfordulásait, hanem a teljes korpuszban eloszló megjelenésükből készít egy átlagot,

6. valamint az átlagos logaritmikus távolsági gyakoriságot [Average Logarithmic Distance Frequency – ALDF], amelynek figyelembevétele megakadályozza, hogy az eredményeket túlzottan befolyásolja egy token nagy koncentrációja a korpusz egy vagy több kis részében. Ha a token egyenletesen oszlik el a korpuszban, az ALDF és az abszolút gyakoriság hasonló vagy azonos értékeket kap (Sketch Engine é. n. a).

Kulcsszavak és kifejezések keresése

A másik korpuszeszköz, amelyet a kutatás során alkalmaztam, az a Keywords, vagyis kulcsszóvizsgáló eszköz. A Keywords funkció segítségével megállapíthatjuk, hogy az általunk vizsgált korpuszban (fókuszkorpusz), vagy alkorpuszban mely szavak – vagy lemmák, morfémák, n-gramok – szerepelnek gyakrabban, mint a referenciaként szolgáló másik korpuszban vagy alkorpuszban, tehát melyek az adott korpusz kulcsszavai, kulcskifejezései (Hunston 2022, 88). Referenciakorpuszként az alapbeállítás szerint a Sketch Engine felületén található legnagyobb azonos nyelvű korpusz használható (Sketch Engine é. n. b). A magyar nyelv esetében a Hungarian Web 2012 (huTenTen12). Ugyanakkor ez a beállítás módosítható, választhatunk másik készen hozzáférhető korpuszt, vagy akár egy saját korpuszunkat is.

A kulcsszókeresésnél csak az összetett keresés funkciót választhatjuk, az egyszerű keresés fülnél csak definíciókat mutat a felület (32. ábra), majd a go gombra kattintva átirányít bennünket a többszavas lexikai egységek [multiword expressions] keresési lehetőségére.

Az összetett keresésnél tehát beállíthatjuk a fókuszkorpuszunkat, a referenciakorpuszunkat. Kiválaszthatjuk, illetve egy skálán bejelölhetjük, hogy a keresés során mennyire koncentráljon a program a mindkét korpuszban megjelenő közös elemekre, és mennyire az eltérőkre. Az előzőleg bemutatott korpuszeszközökhöz hasonlóan beállíthatjuk a kis- és nagybetűk közötti különbséget, hogy az álszavakat, vagyis olyan szavakat, amelyek nem betűvel kezdődnek (például a három dimenzió rövidítése, a *3D*), megjelenítse-e, kizárhatunk az eredmények közül bizonyos elemeket, meghatározhatjuk, hogy mely szövegtípusokat használja a kereső, valamint egy szólistát is beilleszthetünk, ha meg szeretnénk tudni, hogy a benne szereplő elemek kulcsszavai-e a fókuszkorpuszunknak. Az összetett keresés alsó részén, a kulcsszóbeállításoknál [Keywords setting] megadhatunk attribútumokat (lemma, szó, szófajok, címkék, lempo-sok), amelyeket finomíthatunk jelekkel [RegEx]. A Keywords funkció a fókuszkorpuszban a fordításban használható terminusokat is megtalálja, ez a második

beállítási lehetőség [Identify terms] az összetett keresés alsó részén. Ezek olyan többszavas kifejezések, amelyek gyakrabban fordulnak elő a fókuszkorpuszban mint a referenciakorpuszban, és megfelelnek a terminusgrammatikának [term grammar], amely a lexikai struktúrák meghatározására készült, CQL nyelven íródott szabálykészlet. Az utolsó keresési lehetőség pedig a fókuszkorpuszra jellemző, tipikus n-gramok kilistázásának lehetősége, amelynél szintén mód van a szokásos szűrési feltételek (lemmára, szóra, szófajra, lempos-ra, címkére való keresés), valamint jelek használatára (Sketch Engine é. n. b).

A kulcsszókereső algoritmus a következő képletet használja az alkalmazása közben:

$$\frac{fpm_{focus} + n}{fpm_{ref} + n}$$

A számlálóban található fpm_{focus} a szó vagy kifejezés gyakorisága per egymillió szó a fókuszkorpuszban, amíg a számlálóban található fpm_{ref} a szó vagy kifejezés gyakorisága per egymillió szó a referenciakorpuszban. Az n matematikai paraméter értéke az alapbeállítások szerint egy (Kilgarriff 2015, 3).

A fent bemutatott két korpuszeszközzel létrehozott és kapott eredmények letöltése CSV, XLSX, XML és PDF formátumban történhet. A lementhető gyakorisági listák itemszáma, valamint a kulcsszavak maximális száma a korpusz típusától függ; a nyilvánosan elérhető korpuszok esetében ez egy 1000 itemből álló gyakorisági lista és 100 kulcsszó, míg saját készítésű korpuszok esetében nincs megadva (Sketch Engine é. n. d). Ez a korlátozás adatvédelmi megfontolásokból adódik.

A szógyakorisági vizsgálatok eredményei

A KorSzak Gyermeknyelvi Korpuszon történő szógyakorisági vizsgálatokkal – amelyeket a fentebb említett Sketch Engine Wordlist eszközzel végeztem el – az volt a célom, hogy feltérképezzem a korpuszban található szövegekben lévő leggyakrabban előforduló lemmákat, illetve szófajonként osztályozva: a főneveket, mellékneveket és igéket. Ennek az elemzésnek az elvégzése két okból volt szükséges. Egyrészt jó kiindulópontja volt a formulaszerű lexikai egységek kiválasztásának, valamint ezek későbbi részletesebb analizálásához nyújtott információt. Másrészt pedig a készülő korpuszinformált tananyag lexikájának kijelöléséhez biztosított adatokat. A következőkben közölt eredmények a KorSzak Gyermeknyelvi Korpusz 2022. októberi állapotára vonatkoznak. A vizsgálat referenciakorpuszaként használt Hungarian Web 2012, vagyis a

huTenTen12 egy magyar nyelvű, a világhálóról összeállított szövegkorpusz, amely a TenTen korpuszcsalád része, és amelynek célkitűzése egy olyan több mint 10 milliárd szóból álló korpusz létrehozása, amelynek a különböző nyelveken gyűjtött részei azonos módszertant alkalmazva épülnek fel. A Sketch Engine jelenleg több mint 30 nyelven kínál hozzáférést ezekhez a korpuszokhoz. A magyar változat adatait 2012 júniusában gyűjtötték össze a SpiderLing szoftver segítségével, így egy körülbelül 2,5 milliárd szavas korpuszt kaptak (Sketch Engine é. n. c). Mivel a referenciakorpusz írott nyelvi szövegeket, elsősorban cikkeket tartalmaz, nem vonhatunk le lényeges következtetéseket a fókuszkorpusszal való összevetése után, ugyanakkor arra megfelelő eszköz, hogy a segítségével kizárjuk azokat az elemeket, amelyekkel nem érdemes foglalkozni a kutatás során.

Az első vizsgálat alatt a korpuszban található leggyakoribb lemmákat kerestem meg a Wordlist funkció segítségével. A beállítások során minden gyakorisági mutatót láthatóvá tettem: az alapbeállításokhoz tartozó abszolút gyakoriságot, a gyakoriság per egymillió szót, a dokumentumgyakoriságot (DOKgy), vagyis hány dokumentumban jelenik meg az adott lemma, a relatív dokumentumgyakoriságot (Relatív DOKgy), amely azt mutatja meg, hogy a dokumentumok hány százalékában jelenik meg az adott lemma, a redukált gyakoriságok átlagát (RGyÁ), amely nem veszi figyelembe az egymáshoz közel eső elemeket, valamint az átlagos logaritmikus távolsági gyakoriságot (ÁLTGy) is, amelyre a későbbiekben vizsgált elemek kiválasztásához volt szükségem. Az így elkészült táblázat eredményeit az abszolút gyakorisági mutató alapján csökkenő sorrendbe rendeztem. Az alábbiakban látható a korpuszban található leggyakoribb húsz lemma.

Sz.	Lemma	Gyak.	Gyakoriság/ millió	DOKgy	Relatív DOKgy %	ARF	ALDF
1.	A	2,903	40,575.29	77	98.72	1,826.60	1,898.84
2.	VAN	2,309	32,272.94	74	94.87	1,468.05	1,530.72
3.	AZ	2,142	29,938.78	76	97.44	1,324.68	1,378.00
4.	ÉS	1,687	23,579.24	77	98.72	1,044.83	1,068.86
5.	NEM	1,412	19,735.55	66	84.62	852.99	856.72
6.	AHOGY	1,325	18,519.55	68	87.18	796.06	808.35
7.	IS	1,218	17,024.01	69	88.46	762.61	799.75
8.	ÉN	1,132	15,821.99	68	87.18	617.51	577.03
9.	MEG	863	12,062.17	55	70.51	513.81	513.82

Sz.	Lemma	Gyak.	Gyakoriság/ millió	DOKgy	Relatív DOKgy %	ARF	ALDF
10.	EZ	827	11,559.00	63	80.77	483.53	481.80
11.	EGY	812	11,349.34	67	85.90	492.12	501.54
12.	DE	769	10,748.33	69	88.46	496.96	515.95
13.	TUD	650	9,085.06	53	67.95	380.80	382.91
14.	ILYEN	597	8,344.28	52	66.67	331.77	328.82
15.	AKKOR	592	8,274.40	60	76.92	357.37	368.04
16.	JÓ	583	8,148.60	58	74.36	330.29	331.45
17.	IGEN	561	7,841.11	49	62.82	303.16	273.04
18.	HÁT	559	7,813.16	52	66.67	343.18	335.13
19.	NAGYON	556	7,771.22	70	89.74	326.92	327.81
20.	Ő	515	7,198.17	55	70.51	267.34	252.48

2. táblázat. A húsz leggyakoribb lemma a KorSzak Gyermeknyelvi Korpuszban

Az elkészült listát (2. táblázat) az elemzés során három szakaszra bontottam. Az első határt az első nyolc, vagyis az ezer fölötti abszolút gyakorisági mutatóval rendelkező lemma után húztam meg. A másodikat pedig az első húsz leggyakoribb lemma alatt. Az első szakaszt ezután összehasonlítottam a huTenTen12 első nyolc lemmájával. Két lemma kivételével az eredmények ennél a szakasznál még nagyon hasonlóak, ahogy várható is volt. A különbség annyi, hogy a KorSzak Gyermeknyelvi Korpusz első nyolc lemmája között szerepel az ahogy, valamint az én. Ezzel szemben a referenciakorpuszban az ahogy a kilencedik helyen, míg az én csak jóval később a huszonkettedik helyen jelenik meg. A személyes névmás korai megjelenése minden bizonnyal a beszélt nyelv egyik sajátosságának tudható be. Ezenkívül még egy apró, de érdekes különbséget megfigyelhetünk: a fókuszkorpuszban a van lemma az a és az névelők közé ékelődött be.

A következő szakaszban (9–20. hely) már több eltérés figyelhető meg, bár a legnagyobb eltéréseket majd csak a harmadik szakaszban (20–100. hely) találjuk. A második szakaszban két lemmát is találunk (hát, igen), amely egyáltalán nem szerepel a referenciakorpusz első száz lemmájának listáján, amely szintén a beszélt és az írott nyelv különbségének következménye. Ezenkívül a fókuszkorpuszban szerepel három lemma, amelyeknek jóval nagyobb gyakorisági mutatójuk (gyakoriság per egymillió szó) van, mint a referenciakorpuszban lévő párjuknak. A tud 71%-kal, az ilyen 88%-kal, a nagyon pedig 80%-kal magasabb értéket mutat.

A harmadik meghatározott szakaszban (20–100. hely) több olyan elem is szerepel, amely a referenciakorpusz első száz lemmája között nem jelenik meg. Ezek az alábbi táblázatban (3. táblázat) láthatók. Ezenkívül megfigyelhetünk egy duplumot is, az 56. helyen lévő énszerint elemet, amely a 29. helyen lévő szerint lemmához tartozó találatok egy részének külön közlése, és amelyet emiatt ki kell zárni a gyakorisági listából.

Sz.	Lemma	Gyak.	Gyakoriság/ millió szó	DOKgy	Relatív DOKgy %	ARF	ALDF
39.	UGYE	231	3,228.69	41	52.56	123.24	114.19
40.	SZOKIK	226	3,158.81	44	56.41	81.74	56.44
49.	IGAZÁBÓL	170	2,376.09	39	50.00	94.19	92.33
59.	INKÁBB	138	1,928.83	33	42.31	67.98	61.82
80.	TÖK	88	1,229.98	19	24.36	39.59	32.48
84.	ÚGYHOGY	81	1,132.14	26	33.33	47.35	48.79
87.	TÉNYLEG	78	1,090.21	25	32.05	40.50	40.16
90.	GONDOL	76	1,062.25	29	37.18	41.30	40.72
91.	EGYSZER	73	1,020.32	32	41.03	42.61	43.11
92.	NA	73	1,020.32	26	33.33	36.18	35.69
93.	HISZ	73	1,020.32	24	30.77	39.14	38.99
94.	SZÓVAL	73	1,020.32	23	29.49	42.16	41.28
95.	AMÚGY	72	1,006.35	21	26.92	37.63	35.49
99.	AZTÁN	69	964.41	24	30.77	33.63	32.79
100.	BESZÉL	68	950.44	17	21.79	30.91	28.67

3. táblázat. A fókuszkorpuszban megjelenő, a referenciakorpusz első száz lemmája között nem szereplő elemek

A korábbi szakaszokból felsorolt, a fenti táblázatban (3. táblázat) lévő, valamint a nagyobb gyakorisági mutatóval rendelkező elemek – amelyek ebben, a harmadik szakaszban a szeret (79%-kal magasabb), olyan (63%-kal magasabb) így (76%-kal magasabb), mond (76%-kal magasabb), szerint (63%-kal magasabb), megy (72%-kal magasabb) – használatát érdemes lesz a későbbiekben megvizsgálni.

Szófajonkénti (főnév, melléknév, ige) eredmények

A gyakorisági vizsgálatok szófajonkénti eredményeit röviden szeretném bemutatni, csak a legfontosabb jelenségekre koncentrálva a figyelmet. A főne-

vek tekintetében a legnagyobb gyakorisági mutatóval rendelkező lemmák a beszélgetések témaköreit reprezentálják (például a kutya, állat, sorozat, sport, film, barát, család, videó, hobbi, vagy egy kutya neve: Bubu). Ezeket kizárva, valamint az első ötven elemet áttekintve az ész és az érzés lemmák, amelyek egyáltalán nem szerepelnek a huTenTen12 első száz találata között, valamint a némileg eltérő gyakorisági mutatókkal rendelkező rész, baj, kérdés lemmák használati mintázatát célszerű lesz a későbbiekben részletesebben is megvizsgálni.

A melléknevek közül a KorSzak Gyermeknyelvi Korpusz leggyakrabban megjelenő lemmája a jó, amely a referenciakorpuszban csak a harmadik helyen szerepel. A melléknevek gyakorisági listáiban található a legnagyobb különbség a fókusz- és a referenciakorpusz között. Az adatközlő gyermekek által használt melléknevek nagy része nem szerepel a huTenTen12 első száz lemmája között. Ilyenek az aranyos, boldog, büdös, cuki, csodálatos, csodás, csúnya, fura, híres, ideális, idős, izgalmas, kedvenc, kreatív, menő, normális, pici, szerencsétlen, szörnyű, vicces.

Ezenkívül a huTenTen12 korpuszban szereplő elemek közül a fókuszkorpuszban lévők eltérő, legtöbbször sokkal alacsonyabb gyakorisági mutatóval rendelkeznek. Ilyenek például az új (79%-kal kevesebb) vagy az utolsó (60%-kal kevesebb).

Az igék kategóriájában a főnevekhez hasonlóan – de nem olyan nagy mértékben – megjelennek azok a lemmák, amelyek a témakörökhöz szorosan kapcsolódnak. Ilyenek például a szabadidős tevékenységekhez köthetőek közül a sportol, rajzol, sétál, élvez, amelyek nem szerepelnek a referenciakorpusz leggyakoribb száz igéje között. A referenciakorpuszban egyáltalán nem jelennek meg olyan igekötős igék, mint az elmegy, elkezd, megnéz vagy a kimegy, valamint nincs benne a felsorolásban a fókuszkorpusz 41. helyén álló eszik sem. Erre a kategóriára a melléknevekkel ellentétben inkább az jellemző, hogy azok az igék, amelyek mindkét korpuszban megjelennek, és nem azonos helyen állnak a listában, a fókuszkorpuszban nagyobb gyakorisági mutatóval rendelkeznek. Ilyen a már említett megy (72%-kal magasabb), a csinál (81%-kal magasabb), valamint az emlékszik (78%-kal magasabb) igék.

A kulcsszóvizsgálatok eredményei

A kulcsszavak vizsgálatánál az első szakaszban csak a lemmák alapján történő keresést alkalmaztam. Ez az eszköz szintén a huTenTen12 korpuszt használta referenciaként. Az alábbi táblázatban (4. táblázat) láthatjuk a korpusz állatok alkorpuszának 2022. októberi állapot szerinti első huszonöt kulcsszavát.

A találati eredmények között nem szerepelnek a gyermekek háziállatainak a nevei, a megemlített települések nevei, valamint a duplumok.

Sz.		Gyakoriság/millió szó		Relatív DOKgy %	
	Lemma	Fókuszk.	Referenciak.	Fókuszk.	Referenciak.
1.	MADÁRLES	369.44	0.12	30.00	< 0.01
2.	HÁZIÁLLAT	1,436.72	4.24	80.00	0.17
3.	LABRADOR	451.54	1.30	40.00	0.04
4.	RETRIVER	205.25	0.08	30.00	< 0.01
5.	CSIVAVA	287.34	0.67	40.00	0.02
6.	GIDA	287.34	0.88	10.00	0.02
7.	VIZSLA	574.69		30.00	0.07
8.	SPÁNIEL	287.34	1.02	10.00	0.03
9.	BICHON	164.20	0.17	10.00	< 0.01
10.	FÖLNEVEL	164.20	0.25	10.00	0.01
11.	HÖRCSÖG	410.49	2.14	40.00	0.06
12.	KUTYATULAJDONSÁG	123.15	< 0.01	30.00	< 0.01
13.	KUTYAPANZIÓ	123.15	0.08	10.00	< 0.01
14.	KUTYAFAJTA	246.30	1.17	40.00	0.04
15.	VADÁSZLES	123.15	0.12	10.00	< 0.01
16.	VÉNÜL	123.15	0.15	20.00	< 0.01
17.	ORVVADÁSZ	164.20	0.57	10.00	0.02
18.	BORDER	205.25	1.01	20.00	0.03
19.	EFFEKTÍVE	164.20	0.65	20.00	0.03
20.	COLLIE	164.20	0.66	20.00	0.02
21.	ÁLLATMENTŐ	123.15	0.30	10.00	0.01
22.	BORZ	287.34	2.03	10.00	0.07
23.	KUTYAISKOLA	164.20	0.82	20.00	0.03
24.	CUKI	574.69	5.86	40.00	0.23
25.	TIKTOKON	82.10	0.00	10.00	0.00

4. táblázat. Kulcsszavak a KorSzak Gyermeknyelvi Korpusz állatok alkorpuszában

Ahogy a szógyakorisági vizsgálatoknál, itt is érdemes megfigyelni a relatív dokumentumgyakoriságot ahhoz, hogy teljes képet kapjunk a korpuszban szereplő kulcsszavakról. Ebből láthatjuk, hogy az állatok alkorpuszban a

kutyafajták (labrador retriver, csivava, vizsla, spániel, bichon, border collie), valamint a madárles (30%), háziállat (80%), hörcsög (40%), kutyatulajdonság (30%) lemmák, valamint például a cuki (40%) jelző igen gyakran megjelenik az alkorpuszban.

Az eredmények alkalmazása a nyelvoktatásban

Ezek az első eredmények iránymutatásként, valamint kiegészítő információként szolgálhatnak mind a kutatás, mind pedig a tananyagkészítés további szakaszai során. A korpuszokból kapott eredmények számos módon hozzájárulhatnak a tananyagfejlesztők munkájához. Lehetővé teheti számukra, hogy ne kizárólag az intuícióra hagyatkozzanak, így biztosítva a nyelvhasználat pontos ábrázolását (McEnery–Xiao–Tono 2006). Emellett segíthet a megfelelő szintű lexikogrammatikai tananyag kidolgozásában, valamint a releváns szövegek és valós életből vett szituációk kiválasztásában, amelyekben a nyelvtanulók képesek alkalmazni a tanult nyelvi elemeket (McCarten 2010, 415). Ezen túlmenően a korpusznyelvészet befolyásolhatja azokat a döntéseket, amelyek arra vonatkoznak, hogy a tanterv hogyan tükrözi a valós nyelvhasználatot (Conrad 2000).

A szerzők támaszkodhatnak gyakorisági vizsgálatokra, hogy meghatározzák egy adott jelenség arányát egy szövegben, vagy kiválaszthatják, hogy mely nyelvi formákat vizsgálják meg először (Meunier–Reppen 2015, 501). Hasonlóképpen Biber és Reppen azt javasolják, hogy a tananyagok szerzői használják fel a gyakorisági vizsgálatokból származó adatokat, hogy a tanulók megfelelő mennyiségű, értelmes nyelvi inputot kapjanak (Biber–Reppen 2002, 207). A szerzőknek össze kell hasonlítaniuk a tanított szókincset a témával és a kontextussal, amelyekben azt kompetens beszélők használják, erre alkalmasok a kulcsszó- és kulcskifejezés-vizsgálatok. Emellett a szerzőknek figyelembe kell venniük a közös nyelvtani mintákat, hogy a korpuszokban szereplő szövegek modellként szolgálhassanak a tanulók számára (Kaltenböck–Mehlmauer-Larcher 2005, 72).

Összefoglalva kijelenthetjük tehát, hogy a korpuszokat forrásként használó tananyagok pontosabban tükrözik a valós nyelvhasználatot, amely a korábbi tananyagok esetében nem mindig mondható el, mert azok főként az intuícióra és a hagyományos megközelítésekre támaszkodtak. Valamint azt is, hogy a korpuszok használata nagyban megkönnyíti a tananyagok készítőinek munkáját.

Irodalom

Baumann Tímea – Majoros Judit – Pelcz Katalin – Schmidt Ildikó – Szita Szilvia – Vermeki Boglárka. 2020. Bemutatkozik a Korpusznyelvészeti és Szakmódszertani Munkacsoport. *Hungarológiai Évkönyv* 21 (1–2): 32–41.

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

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

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. Abingdon, England: Routledge.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

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

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

Kilgarriff, Adam. 2015. Statistics used in Sketch Engine. https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf (2023. jan. 31.)

KorSzak Gyermeknyelvi Korpusz. 2020. Sketch Engine.

McCarten, Jeanne. 2010. Corpus-informed course book design. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. 413–427. London: Routledge.

McEnery, Tony – Xiao, Richard – Tono, Yukio. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meunier, Fanny – Reppen, Randi. 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.

O'Keefe, Anne – McCarthy, Michael – Carter, Ronald. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Sketch Engine é. n. a. Wordlist – frequency lists and linguistic datbases. https://www.sketchengine.eu/guide/wordlist-frequency-lists/ (2023. jan. 31.)

Sketch Engine é. n. b. Keywords and term extraction – identifying typical words. https://www.sketchengine.eu/guide/keywords-and-term-extraction/ (2023. jan. 31.)

Sketch Engine é. n. c. huTenTen: Corpus of the Hungarian Web. https://www.sketchengine.eu/hutenten-hungarian-corpus/ (2023. febr. 4.)

Sketch Engine é. n. d. Trial and paid account limitations. https://www.sketchengine.eu/guide/account-limitations/ (2023. febr. 4.)

Szita Szilvia – Pelcz Katalin. é. n. MagyarOK teaching materials for Hungarian, levels A1 to B2. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fmagyarok_hp2 (2023. febr. 4.)

CHARACTERISTICS OF CHILDREN'S SPONTANEOUS SPEECH

Word frequency and keyword analysis in the KorSzak Child Language Corpus

The purpose of this study is to reveal the characteristics of children's language usage, with particular attention to the composition of their vocabulary, with the help of corpus linguistic investigations, such as word frequency and keyword analyses. The KorSzak Children's Language Corpus, which is the basis of the present research, is a dynamic corpus for pedagogical purposes currently consisting of 73 recordings of twenty-seven children aged 11-15. During the video and audio recordings, the children-informants talk freely about particular topics (e.g. animals, leisure activities) in pairs or small groups. The current research presents, in detail, the informants' most frequently used words, classifying them into word classes and lexical units associated with them and examining their lexico-grammatical patterns. The application of the investigation's findings in language education will be discussed as an outlook of the presentation.

Keywords: corpus linguistics, child language, spontatneous speech, language teaching

KARAKTERISTIKE SPONTANOG GOVORA DECE

Analiza učestalosti reči i ključnih reči u korpusu dečijeg govora KorSzak

Cilj rada je da se pomoću korpusne lingvistike – analize učestalosti reči i ključnih reči – prikažu osobenosti govora dece koja su učestvovala u istraživanju, sa posebnim osvrtom na strukturu njihove leksike. Osnovu za istraživanje čini dinamički korpus dečijeg govora nastao sa pedagoškim ciljem, pod nazivom *KorSzak* (Bauman et al. 2020), a koji sadrži 78 snimaka 29 ispitanika od 11 do 15 godina. Tokom zvučnih i video zapisa deca u paru ili u malim grupama slobodno međusobno razgovaraju o pojedinim temama (npr. o životinjama, aktivnostima u slobodno vreme). Rezultati predstavljeni u radu deo su istraživanja koje je sprovedeno u okviru doktorskih studija, sa ciljem da se sagleda šablonska upotreba jezika i leksičko-gramatički modeli u okviru korpusa, kao i da se upotrebom rezultata sačini nastavno sredstvo za učenje mađarskog jezika kao stranog.

Ključne reči: korpusna lingvistika, dečiji govor, spontani govor, nastava jezika

A kézirat beérkezésének ideje: 2023. márc. 9. Közlésre elfogadva: 2023. máj. 10.