Analysis of the Use of the Same Frequent Verb Collocations in Three Different Corpora

Zsófia Antal
International Studies Center, University of Pécs Medical School
(<antal.zsofia@pte.hu>)

Abstract

The present study analyzes the use of the frequent verb collocations in the *KorSzak tanulói korpusz* (*KorSzak* learner corpus) and compares them with the frequent verb collocations of the *MagyarOK nyílt pedagógiai korpusz* (*MagyarOK* open pedagogical corpus), and the *huTenTen12 native language corpus*. It investigates whether the foreign language learners use in their written performances the same verb collocations as native speakers, whether they use the right collocations, whether they overuse or underuse them. The study reveals and compares the lexicogrammatical characteristics of the frequent verbs in the above-mentioned corpora and describes the formal and contentual characteristics of these results.

Keywords

corpus-based language analysis; learner corpus; native language corpus; pedagogical corpus; verb collocation

1. Introduction

Corpora can play a very important role in the study of foreign language acquisition. Available corpora and text analysis software allow us to systematically study how learners acquire a new language based on a large linguistic sample. The primary advantage of corpus-based language analysis is that it gives a picture of the language actually used. It can investigate almost any language patterns. Statistical indicators in corpus research provide information about the frequency with which items are used and the context in which they typically occur. They may be aimed at exploring aspects of grammar (word

 | SFU – Studi Finno-Ugrici, n.s. 2 (2022), 1-27 | ISSN 1826-753X
https://doi.org/10.6093/1826-753X/9859

order, use of verbs, verb conjugations, etc.) or lexis (collocations, frequently used language elements and their context), but also at pragmatic issues or at exploring register features, the linguistic devices of a text type. The study briefly introduces the *KorSzak tanulói korpusz* (*KorSzak* learner corpus) (Antal *et al.* 2020), and then reveals and compares the lexicogrammatical characteristics of the frequent verbs in the *KorSzak learner corpus* and compares them with the frequent verb collocations of the *MagyarOK nyílt pedagógiai korpusz* (*MagyarOK* open pedagogical corpus) (Szita, Pelcz 2020-), and the *huTenTen12 native language corpus* (Jakubíček *et al*. 2012), and describes the formal and contentual characteristics of these results. The study also seeks to answer the question whether the foreign language learners use in their written performances the same verb collocations as native speakers, whether they use the right collocations, whether they overuse or underuse them.

2. About the learner corpora

Learner corpus research is a relatively young but highly dynamic branch of corpus linguistics, which began to emerge as a discipline in its own right in the late 1980s and early 1990s. Learner corpora are electronic collections of texts of written and oral expressions collected from the products of foreign language learners (Granger 2004, 124, Szirmai 2005, 34). Nesselhauf (2005, 40) also adds the notion of systematicity to the definition of a learner corpus, by which she means that the texts in the corpus are selected on the basis of specified criteria (such as the language learner's first language, level, etc.). Nesselhauf in the same work also points out that the emphasis in the learner corpus is on spontaneous linguistic expression. In her view, the least controlled corpus-compatible texts are essays (where only the topic is given) and oral interviews. The computer's ability to store and process language provides tools for the study of language learners' linguistic products that were not possible before. These carefully compiled databases can prove to be a very useful resource for anyone who wants to know how foreign language learners learn a language and how the language learning process can be made even better and more efficient, how the needs of learners can be better addressed in language teaching. The main aim of compiling learner corpus is to collect objective linguistic data that can help to describe the learners' language. Learner corpora are also important because they show differences in language

use from that of native speakers of a given language, and they have inspired a great deal of research and academic work, as well as providing ongoing support for measurement and evaluation and curriculum development in foreign language teaching. Corpus analysis can also provide satisfactory answers to many questions that could only be answered incompletely by relying on linguistic intuitions, textbooks, or monolingual dictionaries.

English, the language with the largest number of learners, has the largest and most extensive learner corpus, and the number is growing steadily as more and more people are starting to build learner corpora, even for their own use, thanks to the spread of corpus research methods and the recognition of the practical importance of corpora. However, it is encouraging that such corpora also exist in many other languages; of which the error annotated *Fehlerannotiertes Lernerkorpus (Falko)*[1] in German built at Humboldt University in Berlin is perhaps one of the best known. The corpus contains German language learners' essays, letters, fiction, academic writings, journal articles and book reviews from beginner to advanced level and includes a number of native sub-corpora collected from native German-speaking students at the university (Reznicek *et al*. 2012).

Another error tagged and widely known corpus is *The Learner Corpus of Czech as a Second Language (CzeSL)*[2]. It is the first of its kind in the Czech Republic to contain spoken and written texts and several sub-corpora (e.g., Russian, Romani and Vietnamese L1) for multiple language levels that examines an inflected language and uses a multi-layered error-coding method (Hana *et al*. 2010; Rosen 2016).

The *International Corpus of Learner Finnish (ICLFI)*[3] is also worth mentioning, which is one of the largest learner corpora representing Balto-Finnic languages. The corpus of almost one million words is a collection of essays, narratives, and diary entries from language learners with 22 different native languages at beginner, intermediate and advanced levels. In addition to the texts, the corpus also contains a set of metatextual information about each variable; for example, the age and mother tongue of the learners, the

[1] <https://korpling.german.hu-berlin.de/falko-suche/> (10/2022).
[2] <http://utkl.ff.cuni.cz/dokuwiki/doku.php?id=czesl:czesl> (10/2022).
[3] <https://korp.csc.fi/korp/#?cqp=%5B%5D&corpus=iclfi&stats_reduce=word> (10/2022).

mother tongue of the language teacher collecting the data, and the genre of the text (Jantunen 2011, Jantunen, Brunni 2013, 237)[4].

3. The *KorSzak* learner corpus

Hungarian as a foreign language learner corpora are not really abundant. In 2012, a paper was published in which two researchers from the University of Indiana reported on their corpus linguistic studies (Dickinson, Ledbetter 2012). Their mini corpus contains 10-15-line diary entries (with topics chosen by the students themselves) of 14 (9 beginners, 1 intermediate and 4 advanced) Hungarian language learners. Another published corpus of learners for the analysis of Hungarian as a foreign language is *HunLearner*, a project of researchers at the University of Szeged. The first two sub-corpora of the corpus contain the written submissions of 35 students majoring in Hungarian at the University of Zagreb. According to the latest publication data, the corpus contains 1,427 sentences and about 22,000 tokens (Durst *et al*. 2013, 2014).

In February 2020 our working group on learner corpus (Baumann *et al*. 2020, 35-37) started to build the dynamic *KorSzak learner corpus* (Antal 2021) with the aim of creating a searchable, public database of significant size, which is constantly expanding along defined principles, and which can be a rich resource for research and study by linguists, hungarologists and Hungarian as a Foreign Language teaching specialists. The basic criterion for the building of the *KorSzak learner corpus* is that the corpus contributors learn from the *MagyarOK textbook* family (Szita, Pelcz 2013-2022) in different educational formats and at different language levels, in order for the *MagyarOK open pedagogical corpus* and the *KorSzak learner corpus* to be comparable.

The *KorSzak learner corpus* is divided into two main parts: written and oral language productions. The constantly growing language database currently contains the language production of 257 language learners at A1-B2 level from 57 L1 backgrounds. At the time of writing, the written corpus is the more significant corpus, with 2,338 written texts, 373,664 tokens and 288,308 words.

The sub-corpus of written texts in the corpus is composed of written submissions by students of Hungarian as a foreign language. Some of the

[4] The list of learner corpora is not intended to be exhaustive.

texts are handwritten and typed in such a way that they remain faithful to the original version in content and, as far as possible, in form, so that they can be examined with regard to both content and form from a research point of view. The other part of the student texts was submitted electronically as a consequence of the introduction of digital education. In order to avoid the potential for errors that may occur during digitalization (typos, omissions in typing, etc.), returning to face-to-face teaching we continue to provide online platforms for students to allow them to submit their work electronically. Texts are stored in two formats: original texts are retained and made available without annotation, and among the texts featured in the open corpus only the ones that inhibit searching are corrected. Correction is done on the word stem, but nothing else is corrected. The fact that a correction has been made is indicated in the text after the concerned lexeme, but not in detail. In order to preserve anonymity, we use uniform fictitious names instead of names that indicate vowel mismatches. Other personal information is omitted if it would allow the respondent to be identified.

Although the corpus is not yet available to everyone, we aim to make it available to the general public in the near future. We have chosen *Sketch Engine*[5] as a repository because it provides researchers, language teachers and language learners with a number of easy-to-use tools to search the corpus according to different criteria. For example, the *Word Sketch* tool can be used to query multi-element language units (collocations), *Concordancer*'s concordance lines provide a number of examples of the use of the language element (morpheme, word or phrase) being searched for, and *Wordlist* can be used to generate frequency lists by word type.

4. Corpus-based analysis on learner, pedagogical, and native corpus

Digital corpora that can be subjected to multi-criteria searches also provide an excellent opportunity among others to research collocations, i.e., word combinations that occur with statistically detectable frequency. Every lexical unit has collocational partners, and these structures, so characteristic of natural language use, should also be central to language teaching. Without them, we can formulate what we say, but our linguistic product is likely to

[5] <https://www.sketchengine.eu/> (10/2022).

be far removed from the way native speakers express themselves (O'Keeffee, McCarthy, Carter 2007, 62). Corpus-based search allows the words searched for to be listed together with their multi-word context, analyzed, and used to create a systematic set of examples. One of the most obvious areas of investigation is the comparison of language use between language learners and native speakers using Hungarian language corpora (e.g., *Hungarian National Corpus*, *Hunglish*, *huTenTen12*, and the *MagyarOK open pedagogical corpus*).

4.1 The most frequent verbs in the corpora studied

In what follows, I will use three different corpora (*KorSzak*'s sub-corpus of written texts, *MagyarOK*, *huTenTen12*) to map which verbs are most frequently used by learners of Hungarian as a foreign language and which by Hungarians; and then I will examine the most frequent collocate of one of the most frequent verbs in each of the three corpora, by corpus.

I searched the first 50 most frequent verbs in the *KorSzak learner corpus* (Figure 1).

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	van	11,693	14	olvas	553	27	ír	349	40	találkozik	266
2	szeret	2,547	15	lakik	507	28	fél	348	41	ismer	266
3	tanul	2,358	16	akar	483	29	tölt	347	42	érdekel	261
4	megy	1,777	17	főz	464	30	sétál	344	43	tetszik	258
5	tud	1,433	18	segít	438	31	hallgat	338	44	fcg	251
6	él	1,421	19	jár	428	32	beszélget	322	45	mond	245
7	eszik	1,058	20	csinál	409	33	pihen	313	46	veszik	223
8	beszél	832	21	utazik	396	34	hiányzik	296	47	lát	212
9	nincs	793	22	alszik	382	35	reggelizik	296	48	ker	201
10	kell	766	23	lesz	371	36	vesz	292	49	szokik	199
11	dolgozik	692	24	vásárol	366	37	sportol	291	50	talál	199
12	iszik	654	25	tart	360	38	elég	290			
13	néz	568	26	használ	351	39	vacsorázik	273			

Figure 1 – 50 most frequent verbs in the *KorSzak learner corpus*[6].

[6] 1. be, 2. like/love, 3. study, 4. go, 5. can/know, 6. live, 7. eat, 8. speak, 9. not have/there is no 10. must/have to, 11. work, 12. drink, 13. look/watch, 14. read, 15. live, 16. want, 17. cook, 18. help, 19. walk/go/attend, 20. do, 21. travel, 22. sleep. 23. will be, 24. purchase, 25. keep/hold, 26. use, 27. write, 28. be afraid, 29. spend, 30. walk, 31. listen, 32. talk, 33. rest, 34. miss, 35. have breakfast, 36. buy/take, 37. do sport, 38. be enough, 39. have dinner/supper, 40. meet, 41. know, 42. be interested, 43. like, 44. catch/grab/will/be going to, 45. say, 46. be lost, 47. see, 48. ask, 49. get used to, 50. find.

The 5 most frequent verbs in the *KorSzak learner corpus*, ranked by frequency index, are *van* 'be', *szeret* 'like/love', *tanul* 'study', *megy* 'walk/go/attend' and *tud* 'can/know'. There are

— 11,693 occurrences of *van* (3,129% of the total corpus),
— 2,547 of *szeret* (0,6816% of the total corpus),
— 2,358 of *tanul* (0,6310% of the total corpus),
— 1,777 of *megy* (0,4756% of the total corpus),
— 1,433 of *tud* (0,3835% of the total corpus).

The *MagyarOK open pedagogical corpus* was created for the *MagyarOK textbook* family, with the primary aim of making language teaching and learning more effective. It reflects natural language usage and is a collection of adapted and authentic texts. It consists of two sub-corpora, the first of which is a collection of the full text of the textbooks, and the second consists of semi-authentic narrative texts on the textbook topics from native speakers. It is also available from the *Sketch Engine* interface, with 144,832 words and 201,079 tokens (Szita 2020, 174-175, Szita 2021, 74-75). For the *MagyarOK open pedagogical corpus* the frequency list of 50 filtered by verbs are as follows (Figure 2):

	Lemma	Frequency ? ↓		Lemma	Frequency ? ↓		Lemma	Frequency ? ↓		Lemma	Frequency ? ↓		Lemma	Frequency ? ↓
1	van	4,357	11	akar	302	21	ír	210	31	szokik	154	41	iszik	120
2	tud	1,046	12	él	286	22	segít	187	32	jön	148	42	gondol	120
3	szeret	950	13	mond	257	23	fog	185	33	tölt	145	43	talál	119
4	megy	615	14	csinál	254	24	lát	181	34	tesz	143	44	ért	115
5	kell	559	15	jár	240	25	néz	173	35	ad	138	45	hoz	114
6	tanul	456	16	vesz	231	26	köszön	165	36	találkozik	137	46	kap	113
7	dolgozik	355	17	olvas	230	27	kér	161	37	keres	130	47	elmegy	112
8	beszél	343	18	lakik	227	28	ismer	160	38	vár	126	48	érdekel	112
9	nincs	341	19	eszik	223	29	beszélget	159	39	tetszik	126	49	játszik	103
10	lesz	323	20	tart	220	30	főz	157	40	használ	123	50	utazik	102

Figure 2 – 50 most frequent verbs in the *MagyarOK open pedagogical corpus*[7].

[7] 1. be, 2. can/know, 3. like/love, 4. go, 5. must/have to, 6. study, 7. work, 8. speak, 9. not have/there is no 10. will be, 11. want, 12. live, 13. say, 14. do, 15. walk/go/attend, 16. buy, 17. read, 18. live, 19. eat, 20. keep/hold, 21. write, 22. help, 23. catch/grab/will/be going to, 24. see, 25. look/watch, 26. say hello, 27. ask, 28. know, 29. talk, 30. cook, 31. get used to, 32. come, 33. spend, 34. do/put, 35. give, 36. meet, 37. search, 38. wait, 39. like, 40. use, 41. drink, 42. think, 43. find, 44. understand, 45. bring, 46. get/receive. 47. go away/leave, 48. be interested, 49. play, 50. travel.

The top 5 most frequent verbs in the *MagyarOK open pedagogical corpus*, ranked by frequency index, are *van* 'be', *tud* 'can/know', *szeret* 'like/love', *megy* 'walk/go/attend' and *kell* 'must/have to'. There are

— 4,357 instances of *van* (2,167% of the total corpus),
— 1,046 of *tud* (0,5202% of the total corpus),
— 950 of *szeret* (0,4725% of the total corpus),
— 615 of *megy* (0,3058% of the total corpus),
— 559 of *kell* (0,2780% of the total corpus).

Finally, I searched for the 50 most frequent verbs in the *huTenTen12*, a giant Hungarian native speaker corpus available from the *Sketch Engine*. The corpus is representative of the language use on the Internet and in the written vernacular, consisting of texts published on the Internet until 2012, and contains 2,572,620,694 words and 3,161,920,362 tokens. The 50 most frequently used verbs in the collection are the followings (Figure 3):

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	van	49,067,995	11	akar	3,024,986	21	kap	2,125,871	31	történik	1,439,083	41	dolgozik	1,185,360
2	kell	9,961,946	12	vesz	2,821,985	22	jelent	1,879,193	32	jut	1,355,538	42	szól	1,139,091
3	tud	9,142,019	13	megy	2,723,229	23	néz	1,837,588	33	beszél	1,331,544	43	játszik	1,095,373
4	lesz	6,669,952	14	ad	2,697,447	24	kezd	1,713,455	34	válik	1,306,079	44	ért	1,089,696
5	mond	5,293,753	15	kerül	2,559,533	25	hisz	1,660,818	35	használ	1,293,971	45	mutat	1,045,263
6	tesz	4,065,887	16	tart	2,525,031	26	él	1,640,815	36	kér	1,291,821	46	választ	1,021,860
7	lát	3,679,625	17	jön	2,355,900	27	talál	1,640,585	37	sikerül	1,255,722	47	ismer	1,014,583
8	fog	3,350,463	18	ír	2,256,527	28	jár	1,597,749	38	marad	1,231,495	48	olvas	1,006,999
9	szeret	3,145,155	19	áll	2,234,011	29	hoz	1,553,475	39	segít	1,214,134	49	hagy	980,044
10	nincs	3,054,168	20	gondol	2,204,515	30	vár	1,521,333	40	csinál	1,191,689	50	érik	962,000

Figure 3 – 50 most frequent verbs in the *huTenTen12 native language corpus*[8].

The top 5 most frequent verbs in the *huTenTen12 native language corpus*, ranked by frequency index, are *van* 'be', *kell* 'must/have to', *tud* 'can/know', *lesz* 'will be', and *mond* 'tell'. There are

[8] 1. be, 2. must/have to, 3. can/know, 4. will be, 5. say, 6. do/put, 7. see, 8. catch/grab/will/be going to, 9. like, 10. not have/there is no, 11. want, 12. buy/take, 13. go, 14. give, 15. cost, 16. keep/hold, 17. come, 18. write, 19. stand, 20. think, 21. get/receive, 22. report/mean, 23. look/watch, 24. start, 25. believe, 26. live, 27. find, 28. walk/go/attend, 29. bring, 30. wait, 31. happen, 32. get, 33. speak, 34. become, 35. use, 36. ask, 37. succeed, 38. stay, 39. help, 40. do, 41. work, 42. tell, 43. play, 44. understand, 45. show, 46. choose, 47. know, 48. read, 49. leave, 50. mature/ripen.

— 49,067,995 occurrences of *van* (0.2209% of the total corpus),
— 9,961,946 of *kell* (0.3151% of the total corpus),
— 9,142,019 of *tud* (0.1936% of the total corpus),
— 6,669,952 of *lesz* (0.1870% of the total corpus),
— 5,293,753 of *mond* (0.1031% of the total corpus).

We can thus observe that the three corpora of different sizes, with different data sources and different purposes, show a strong similarity in the frequency of verbs (Table 1). In all three corpora, the first verb in the frequency index is the substantive verb *van* 'be'. The verb *szeret* 'like/love' is ranked second in the learner corpus and third in the pedagogical corpus, and although it is not in the top five in the native corpus, it is also ranked ninth with 3,145,155 occurrences (0.09947% of the total corpus). The verb *tanul* 'study' is ranked third in the learner corpus, sixth in the pedagogical corpus with 456 instances (0.2268% of the total corpus) and ninetieth in the native corpus with 599,084 instances (0.01895% of the total corpus). The verb *megy* 'walk/go/attend' is ranked fourth in both the learner and the pedagogical corpus, and thirteenth in the native corpus with 2,723,229 instances (0.08613% of the total corpus). The verb *tud* 'can/know', like *van* 'be', is in the top five verbs in all three corpora. The verb *kell* 'must/have to' is ranked second in the native corpus and fifth in the pedagogical corpus, while in the learner corpus it is ranked tenth with 766 instances (0.2050% of the total corpus). The verb *lesz* 'will be' is ranked fourth in the native corpus, twenty-third in the learner corpus with 371 instances (0.09929% of the total corpus) and tenth in the pedagogical corpus with 323 instances (0.1606% of the total corpus). The verb *mond* 'tell' is ranked fifth in the native corpus, thirteenth in the pedagogical corpus with 257 instances (0.1278% of the total corpus) and forty-fifth in the learner corpus with 245 instances (0.06557% of the total corpus). The only significant difference is observed for the verb *tanul* 'study', which is not even among the top fifty verbs in the native corpus. However, the over-representation of the verb *tanul* 'study' in the learner corpus and the pedagogical corpus is not surprising, since the learner corpus is composed of written texts by language learners actively involved in the learning process, and the pedagogical corpus is aligned with the themes of the *Common European Framework of Reference for Languages*, in which the theme of (language) learning is prominent at all language levels.

order of frequency	*KorSzak*	*MagyarOK*	*huTenTen12*
1.	*van* 'be'	*van* 'be'	*van* 'be'
2.	*szeret* 'like/love'	*tud* 'can/know'	*kell* 'must/have to'
3.	*tanul* 'study'	*szeret* 'like/love'	*tud* 'can/know'
4.	*megy* 'walk/go/attend'	*megy* 'walk/go/attend'	*lesz* 'will be'
5.	*tud* 'can/know'	*kell* 'must/have to'	*mond* 'tell'

Table 1 – Five most frequent verbs in the corpora studied.

I have therefore found two verbs which are in one of the top five places in all three corpora in terms of frequency. Despite the fact that the substantive verb *van* 'be' is in the first place in all three corpora, in the 4.2 subsection I will examine the verb *tud* 'can/know', which is also present in all three corpora. The reason for this is that *A magyar nyelv értelmező szótára* (Explanatory Dictionary of the Hungarian Language)[9] distinguishes a total of 65 meanings and several nuances of the substantive verb *van* 'be' within nine major meaning groups. Kiefer (1968) in a summary of his study considers the substantive verb *van* 'be' to have eleven ambiguous meanings, taking into account both syntactic and semantic aspects. De Groot (1989) describes 7 types of structures with *van* 'be' from the perspective of functional syntactic theory, and the *Magyar Grammatika* (Hungarian Grammar) (2000) also lists 6 classes of substantive verbs. For reasons of space, I will not therefore attempt to analyze and explore the collocational profile of the substantive verb in this paper.

4.2 The most frequent collocation partners of the verb tud in the corpora studied – Analysis of the *tud-jól* collocation

Next, using the *Word Sketch* tool, I queried the most frequent collocation partners of the verb *tud* 'can/know' in all three corpora. The tables below show the typical collocation partners in the corpora, sorted by frequency (Table 2, Table 3, Table 4).

[9] <https://www.arcanum.com/hu/online-kiadvanyok/Lexikonok-a-magyar-nyelv-ertelmezo-szotara-1BE8B/> (10/2022).

	item	frequency
1.	*jól* 'well'	161
2.	*az* 'that'	44
3.	*nyelv* 'language'	27
4.	*sem* 'neither/nor'	25
5.	*akkor* 'then'	23

Table 2 – Top 5 collocates of *tud* 'can/know' in the *KorSzak learner corpus*.

	item	frequency
1.	*jól* 'well'	77
2.	*ahogy* 'as/like'	24
3.	*úgy* 'like/such as'	18
4.	*én* 'I'	18
5.	*az* 'that'	17

Table 3 – Top 5 collocates of *tud* 'can/know' in the *MagyarOK open pedagogical corpus*.

	item	frequency
1.	*sem* 'neither/nor'	380,008
2.	*hogy* 'that'	182,466
3.	*jól* 'well'	166,109
4.	*már* 'already'	142,991
5.	*csak* 'only'	126,604

Table 4 – Top 5 collocates of *tud* 'can/know' in the *huTenTen12 native language corpus*.

By comparing the tables, we can observe that it is the adverb *jól* 'well' that appears in all three corpora (first in the learner and pedagogical corpus and third in the native corpus), providing a perfect basis for comparing the usage characteristics of the collocations of *tud* 'can/know' and *jól* 'well' in the three corpora.

Using the *Concordancer* tool, I arrived at the results by systematically analyzing the concordance lines. Due to the scope of the study, I will only illustrate 20 concordance lines for each of the three corpora in the following figures, which are, however, representative despite their relatively small number, thanks to the *Sketch Engine* software.

Figure 4 – Examples of the use of the *tud-jól* collocation in the *KorSzak learner corpus*.

By analysing the concordance lines of the *KorSzak learner corpus* (Figure 4), we can observe that:

— The collocates *tud* 'can/know' and *jól* 'well' are side by side, with only four instances of another lexical item wedged between them: the adverbs *elég* 'quite', *igazán* 'really', *annyira* 'so much' (*tudok elég jól* 'I can/know quite well', *tud elég jól* 'he/she can/knows quite well', *tudtam igazán jól* 'I could/knew really well', *tudtuk annyira jól* 'we could/knew so well').
— The lexeme *jól* comes before *tud* more often (*jól tudok* 'I can/know well', *jól tud* 'he/she can/knows well', *jól tudott* 'he/she could/knew well') than vice versa (*tud jól* 'he/she can/knows well', *tudtok jól* 'you [2PL] can/know well', *tudtam jól* 'I could/knew well').
— The word order *tud jól* occurs when the collocation is preceded by an adverb, interrogative and/or negative (*reggel tud jól* 'he/she can do well in the morning', *mikor tudsz jól* 'when can you do well', *nem tudtuk jól* 'we couldn't do/didn't know well').
— The verb *tud* is typically present tense, declarative, indefinite, first-person singular (*Este nagyon **jól tudok** tanulni.* 'I can study very well in

the evening.' *A kollégáim szerint jól kommunikálok szóban és írásban, és* ***jól tudok*** *csapatban dolgozni.* 'According to my colleagues I communicate well both orally and in writing, and I can work well in a team.' ***Jól tudok*** *tanulni, dolgozni és szerintem is a hosszú távú memóriám ilyenkor működik a legjobban.* 'I can learn and work well and I also think my long-term memory works best at this time.'). The second-person plural indefinite and the first-, second- and third-person plural definite infinitive forms are not in the corpus in the present tense at all.

— The past tense form of *tud* occurs only 16 times in the corpus. In first-person singular six times, in third-person singular eight times, in first-person plural twice (*De csak egy szemeszterig tanultam ott, mert a koronavírus idején nem* ***tudtam*** [1SG] ***jól*** *online tanulni.* 'But I only studied there for one semester, because I couldn't study well online during the time of the coronavirus.' *Például egy történész volt a középiskolában, és* ***jól tudott*** [3SG] *motiválni és nagyon érdekesen magyarázott.* 'For example, there was a historian in high school, and he was good at motivating, and he explained things in a very interesting way.' *Nem* ***tudtuk*** [1PL] ***jól*** *bejárni az országot, mert covid volt.* 'We couldn't travel around the country well because there was covid.'). The past definite conjugated form of *tud* is only used once (*Nem* ***tudtam*** [1SG] ***jól****, amit akartam.* 'I did not know well what I wanted.').

— The collocation is most often in the intrasentential position (*A matematikaóra a tanárom* ***jól tudott*** *motiválni és érdekesen magyarázott.* 'In my math class, my teacher was good at motivating and explained things in an interesting way.' *Este nagyon* ***jól tudok*** *tanulni.* 'I can study very well in the evening.' *Hakal nagyon* ***jól tud*** *főzni.* 'Hakal can cook very well.'), but it is also sometimes found in the sentence initial position (***Jól tudja*** *Javával, C++vel és Javascripttel programozni a programjait.* 'He/She is good at programming in Java, C++ and JavaScript.' ***Jól tud*** *úszni és nagyon szeret sportolni.* 'He/She can swim well and loves sports.' ***Jól tudom****, hogy vietnámiok nagyon jók sakkban és sportlövészetben.* 'I know very well that the Vietnamese are very good at chess and shooting sports.'). In the sentence-final position, it appears only once in a sentence with incorrect word order (*Szerintem szép nyelv a magyar, de a kiejtés nehéz és olvasás soha nem* ***tudok jól****.* 'I

think Hungarian is a beautiful language, but the pronunciation is difficult, and I can never read well.').

— It frequently occurs in negative sentences (*Sajnos* ***nem tud jól*** *főzni.* 'Unfortunately, he/she is not good at cooking.' *Mindenkinek más ötlete volt és* ***nem tudtuk*** *annyire* ***jól*** *megbeszélni a dolgokat.* 'Everyone had different ideas and we couldn't discuss things that well.' ***Nem tudtam jól****, amit akartam.* 'I didn't know well what I wanted.').
— It can occur in any clause of complex sentences (*A szobámban világos és csendes,* ***jól tudok*** *dolgozni.* 'My room is bright and quiet; I can work well.' ***Jól tudom****, hogy vietnámiok nagyon jók sakkban* és *sportlövészetben.* 'I know very well that the Vietnamese are very good at chess and shooting sports.').
— The comparative form of the collocate *jól* is often encountered (*Ezért fekvő testhelyzetben* ***tudunk jobban*** *koncentrálni.* 'Therefore, we can concentrate better in a lying position.' *Ha eleget alszol,* ***jobban tudsz*** *koncentrálni.* 'If you get enough sleep, you can concentrate better.' *Egyetlen dolog van, amitől* ***jobban tudok*** *összpontosítani: a csend.* 'The only thing that helps me concentrate better is silence.'). The superlative form of *jól* is also found in the corpus, but only in a negligible number (*Akkor* ***tudok legjobban*** *koncentrálni, ha jól aludtam és korán kelek fel.* 'I can concentrate best when I sleep well, and get up early.').
— The collocation is often followed by an infinitive (*Nem* ***tudok jól főzni*** [INF] *és nem is szeretek.* 'I can't cook well and I don't like to do it either.' *Nagyon* ***jól tud táncolni*** [INF]. 'He/She is very good at dancing.' *Van olyan étel, amelyet valamelyik családtagja különösen* ***jól tud elkészíteni*** [INF]? 'Is there a dish that a member of your family is particularly good at preparing?').

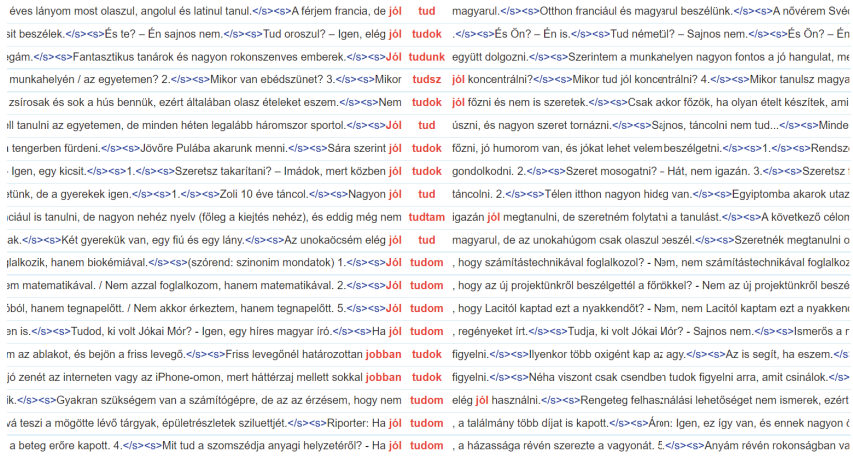

Figure 5 – Examples of the use of the *tud-jól* collocation in the *MagyarOK open pedagogical corpus*.

By analyzing the concordance lines of the *MagyarOK open pedagogical corpus* (Figure 5), we can observe that:

— The collocates *tud* 'can/know' and *jól* 'well' stand side by side, with only three instances of another lexical item wedged between them: the adverbs *elég* 'quite', *igazán* 'really' and *is* 'too, also' (*tudtam igazán jól* 'I could/knew really well', *tudom elég jól* 'I know quite well', *jól is tudtam* 'I knew it well').
— The lexeme *jól* comes before *tud* more often (*jól tud* 'he/she can/knows well', *jól tudok* 'I can/know well', *jól tud* 'he/she can/knows well', *jól tudnak* 'they can/know well'), than vice versa (*tudok jól* 'I can/know well', *tudsz jól* 'you can/know well', *tud jól* 'he/she can/knows well').
— The word order *tud jól* occurs when the collocation is preceded by an interrogative and/or negative (*mikor tudsz jól* 'when can you do well', *nem tudok jól* 'I can't do well').
— The verb *tud* is typically present tense, declarative, indefinite, first-person singular (*Igen, elég* ***jól tudok****.* 'Yes, I can do it quite well.' ***Jól tudok*** *úszni, és nagyon szeretek tornázni.* 'I can swim well, and I really like gymnastics.' *Imádok, mert közben* ***jól tudok*** *gondolkodni.* 'I

love it because I can think well while I'm doing it.'). Second-person plural indefinite and first-, second- and third-person plural definite forms are not found at all in this corpus, as in the learner corpus, and there is no second-person plural indefinite form.

— The past tense form of *tud* occurs only three times in the corpus. Twice in first-person singular, once in second-person plural. In all three cases as indefinite conjugation (*Többször elkezdtem franciául is tanulni, de nagyon nehéz nyelv (főleg a kiejtés nehéz), és eddig még nem* ***tudtam*** [1SG] *igazán* ***jól*** *megtanulni, de szeretném folytatni a tanulást.* 'I've started learning French several times, but it's a very difficult language (especially the pronunciation) and I haven't been able to learn it very well yet, but I'd like to keep learning.' *Nagyon szerettem spanyolul beszélni, és* ***jól*** *is* ***tudtam*** [1SG] *spanyolul, vagy legalábbis azt hittem, hogy jól tudok.* 'I really liked speaking Spanish, and I could also speak it well, or at least I thought I could.' ***Jól tudtatok*** [2PL] *franciául, amikor kimentetek?* 'Did you speak French well when you went abroad?').

— The collocation is most often in the intra-sentential position (*Ők* ***is jól tudnak*** *már németül.* 'They also speak German well already.' *A feleségem nagyon* ***jól tud*** *dánul.* 'My wife speaks Danish very well.' *Mikor* ***tud jól*** *koncentrálni?* 'When can he/she concentrate well?'), but it is sometimes found in the sentence initial position (***Jól tudok*** *veled dolgozni.* 'I can work well with you.' ***Jól tudunk*** *együtt dolgozni.* 'We can work well together' ***Jól tud*** *úszni, és nagyon szeret tornázni.* 'He/She's a good swimmer, and he/she really likes gymnastics.') and also in the sentence final position (*Végzettségét tekintve építész, ha* ***jól tudom.*** 'He is an architect by qualification, as far as I know.' *Nagyon szerettem spanyolul beszélni, és jól is tudtam spanyolul, vagy legalábbis azt hittem, hogy* ***jól tudok.*** 'I was very fond of speaking Spanish and I spoke it well, or at least I thought I did.' *Szerintem nem* ***jól tudod.*** 'I don't think you know it well.').

— It occurs in negative sentences (***Nem tudok jól*** *főzni és nem is szeretek.* 'I'm not a good cook and I don't like to cook either.' *Sajnos* ***nem tud jól*** *úszni, de imád kajakozni és vitorlázni.* 'Unfortunately, he/she can't swim well, but he/she loves kayaking and sailing.' *Gyakran szükségem*

van a számítógépre, de az az érzésem, hogy ***nem tudom*** *elég* ***jól*** *használni.* 'I often need the computer, but I have the feeling that I can't use it well enough.').

— It can occur in any clause of a compound sentence (*Otthon franciául beszélünk, de a férjem nagyon* ***jól tud*** *angolul is.* 'We speak French at home, but my husband also speaks English very well.' ***Jól tudom****, hogy számítástechnikával foglalkozol?* 'Am I right in thinking that you work in computer science?').

— The comparative form of the collocate *jól* of the collocative is often encountered (*Friss levegőnél határozottan* ***jobban tudok*** *figyelni.* 'I can definitely pay more attention in fresh air.' *Kiderült, hogy fekve* ***jobban tudunk*** *gondolkodni, mint ülve.* 'It turns out that we can think better lying down than sitting up.' *Esetleg keresnék valami jó zenét az interneten vagy az iPhone-omon, mert háttérzaj mellett sokkal* ***jobban tudnék*** *figyelni.* 'I might look for some good music on the internet or on my iPhone, because I could focus much more with background noise.'). The superlative form of *jól* is also found in the corpus, but only in a negligible number (*Akkor* ***tudok legjobban*** *koncentrálni, ha jól aludtam és korán kelek fel.* 'I can concentrate best when I sleep well and get up early.').

— The collocation is often followed by an infinitive (*Zsigmond szerint a friss levegőn* ***jobban tudunk koncentrálni*** [INF]. 'According to Zsigmond we can concentrate better in fresh air.' *Csilla barátja, Philippe, ráadásul francia, és ő is nagyon* ***jól tud főzni*** [INF]. 'Moreover, Csilla's friend, Philippe is French, and he can also cook very well.' ***Jobban tudsz figyelni*** [INF]. 'You can concentrate better.').

— Although in the present study, for reasons of scope, I will only examine the two compound collocations, I think it is very important to point out that the occurrence of the constructions *ha jól tudom* 'as far as I know' and *jól tudom, hogy* 'Am I right in thinking that' can be observed as a typical pattern (***Jól tudom, hogy*** *számítástechnikával foglalkozol?* 'Am I right in thinking that you're in computer science?' ***Jól tudom, hogy*** *az új projektünkről beszélgettél a főnökkel?* 'Am I right in thingking that you've been talking to the boss about our new project?' ***Jól tudom, hogy*** *Lacitól kaptad ezt a nyakkendőt?* 'Am I right in thinking that you

got this tie from Laci?' ***Ha jól tudom***, *a találmány több díjat is kapott.* 'As far as I know, the invention has won several awards.' ***Ha jól tudom***, *a házassága révén szerezte a vagyonát.* 'As far as I know, he made his fortune through his marriage.' ***Ha jól tudom***, *a múzeum hétfőn zárva van.*) 'As far as I know, the museum is closed on Mondays.'.

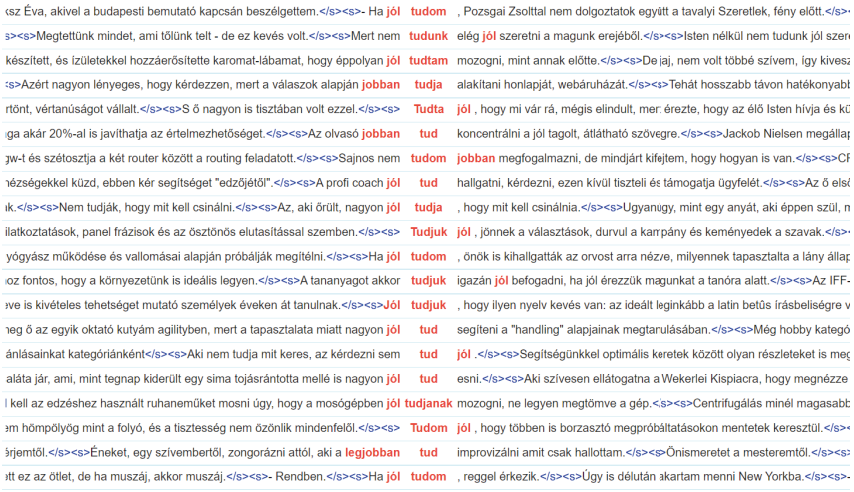

Figure 6 – Examples of the use of the *tud-jól* collocation in the *huTenTen12 native language corpus*.

By analyzing the concordance lines of the *huTenTen12 native corpus* (Figure 6), we can observe that:

— The collocates *tud* 'can/know' and *jól* 'well' are adjacent, but adverbs can sometimes be wedged between them (*tudunk elég jól* 'we can/know quite well', *tud csak jól* 'only he/she can/knows well', *tudom nagyon jól* 'I know very well', *tudok igazán jól* 'I can really well', *tudom annyira jól* 'I know so well').
— The lexeme *jól* is more often placed before *tud* (*jól tudom* 'I know well', *jól tudja* 'he/she knows well', *jól tudják* 'they know well') than vice versa (*tudok jól* 'I can/know well', *tudta jól* 'he/she knew well', *tudjuk jól* 'we know well').

— The word order *tud jól* occurs when the collocation is preceded by an adverb, interrogative and/or negative (*akkor tudunk jól* 'we can/know well if', *nem tudom jól* 'I don't know well', *ki tud jól* 'who can/knows well').
— The frequency indicators show that the present tense, definite conjugation, first-person singular, third-person singular and third-person plural forms of *tud* have the highest frequency (*Ha **jól tudom*** [1SG] *szerettetek volna az első négyben végezni.* 'As far as I know you wanted to finish in the top four.' *Nagyon **jól tudja*** [3SG], *milyen károkat okozunk neki, és hogy kell erre reagálnia.* 'He/She is well aware what damage we are doing to him/her and how he/she should react.' *Dolgozóink **jól tudják*** [3PL], *hogy partnereink elégedettsége és környezetünk rendje vezet a sikerességhez.* 'Our employees are well aware that the satisfaction of our partners and order in our environment lead to success.').
— Based on the frequency indicators, past tense, definite conjugation, first-person singular, third-person singular and third-person plural forms of *tud* have the highest frequency (*Ráadásul a wobblerfestésen kívül sok más területen is **jól tudtam*** [1SG] *már használni.* 'In addition, I have been able to use it well in many other areas besides wobbler painting.' *Pista **jól tudta*** [3SG], *hogy a vége felé jár a klasszikus népművelés ideje.* 'Pista was well aware that the time for classical folk education was coming to an end.' *A kérdésekre a gyerekek **jól tudták*** [3PL] *a választ.* 'The children knew the answers to the questions well.').
— The collocation is most often in intrasentential position (*Ezáltal az emberi szervezet különösen **jól tudja** felvenni és hasznosítani a tengeri összetevőket.* 'This allows the human body to absorb and utilise marine ingredients particularly well.' *Később **jól tudnak** jönni a fejlesztőkártyák.* 'Later, the development cards can come in handy.' *Otthonról mindenki nagyon **jól tudja** mit kellene csinálni.* 'Everyone at home knows very well what should be done.'), but it can also be found in the sentence initial position (***Jól tudom**, hogy a szívizom sejtek nem cserélődnek le?* 'Am I correct that myocardial cells are not replaced?' ***Jól tudjuk**, hogy az utazás, utaztatás bizalmi műfaj, pláne a mi fő tevékenységünk, a kulturális körutazások területén.* 'We are well aware that travel and travel management are a trusted art, especially in our main business,

cultural cruises.' ***Jól tudják*** *ezt a Suzuki mérnökei is, akik nagyon is nehéz feladatot vettek a nyakukba.* 'The engineers at Suzuki, who have taken on a very difficult task, are also well aware of this.'), and also in the sentence final position (Ő *már meghalt,* ***tudod jól****.* 'He is already dead, you know that.' *Aki nem tudja mit keres, az kérdezni sem* ***tud jól****.* 'If you don't know what you're looking for, you can't ask questions well either.' *Előttünk az élet, és álmokat szőni nem törvénytelen ha* ***jól tudom****.* 'Life is ahead of us, and dreaming is not against the law, as far as I know.').

— It occurs in negative sentences (***Nem tudunk jól*** *rágni, nem merünk mosolyogni, „félünk" a kivehető pótlásoktól.* 'We can't chew well, we don't dare to smile, we are "afraid" of removable replacements.' *Ami még azoknak is fel fog tűnni, akik egyébként* ***nem tudnak jól*** *angolul.* 'Even those who don't speak English well anyway will notice.' *De lehet* ***nem tudom jól****.* 'But I could be wrong.').

— It can occur in any clause of a compound sentence (*Bizonyára* ***jól tudja****, hogy megfelelő marketing stratégia nélkül ez igen nehezen menne.* 'You must be well aware that without a proper marketing strategy, this would be very difficult.' *Frau Eva-Maria azt mondja, nem* ***tudok jól*** *németül.* 'Frau Eva-Maria says I don't speak German well.').

— Most often, we find the comparative form of the collocate *jól* of the collocative (*Te tényleg okos vagy, hogy ez a Mazsola meg a gorillák* ***jobban tudják****, hogyan kell bocsánatot kérni, mint sok ember.* 'You are really smart that this Mazsola and the gorillas know how to apologize better than many people.' *Az olvasó* ***jobban tud*** *koncentrálni a jól tagolt, átlátható szövegre.* 'The reader can concentrate better on well-articulated, transparent text.' *Társas helyzetben is* ***jobban tudnak*** *teljesíteni.* 'They can also perform better in social situations.'), and the superlative form of the *jól* collocates is also found in the corpus (*Lehetőségeinket a rendezvényre szánt büdzsé után ez a tényező* ***tudja legjobban*** *szűkíteni.* 'This is the factor that can best limit our options after the budget allocated to the event.' *Nem a legerősebb éli túl, hanem aki* ***legjobban tud*** *alkalmazkodni az új helyzethez.* 'It is not the strongest that survive, but those who can best adapt to the new situation.' *A realisták* ***tudták legjobban*** *a regényt.* 'The realists knew the novel best.').

— The collocation is often followed by an infinitive (*Mára ezt nálamnál már **jobban tudja művelni*** [INF]*!* 'Today, he/she can do it better than I can!' *Sajnos nem **tudom jobban megfogalmazni*** [INF], *de mindjárt kifejtem, hogy hogyan is van.* 'Unfortunately, I can't put it better, but I'll explain how it is immediately.' *A profi coach **jól tud hallgatni*** [INF], *kérdezni, ezen kívül tiszteli* és *támogatja ügyfelét.* 'A professional coach is good at listening, asking questions, in addition, respects and supports his client.').
— A typical pattern is the frequent use of the constructions *ha jól tudom* 'as far as I know'; *jól tudom/tudják* 'I/they know it well', *hogy; tudom/tudjuk jól, hogy* 'I am/we are well aware that' (***Ha jól tudom**, ezután Szlovákiába mentél.* 'As far as I know, you then went to Slovakia.' És *nagyon **jól tudom hogy** akkor teljesen összeomlana!* 'And I know very well that he/she would completely collapse!' ***Tudom jól, hogy** többen is borzasztó megpróbáltatásokon mentetek keresztül*. 'I am well aware that many of you have been through terrible ordeals.').

4.3 Analysis of errors in the learner corpus

Only 14 instances of incorrect use of a collocation are found in the corpus. I have classified the errors into 4 groups according to their type and I illustrate each type with an example below:

1. Overuse/redundant use of collocation:
   **Dolgozom sok, mert tudok jól beszélni magyarul.* For a native speaker, the sentence *Sokat dolgozom, mert jól beszélek magyarul.* 'I work a lot because I speak Hungarian well.' sounds natural, the collocation *tud jól* + infinitive structure is inappropriate, sounding "not like Hungarian".
2. Incorrect word order:
   **Tudok elég jól beszélni japánul.* The focal point of the sentence is on the degree of proficiency (*elég jól* 'quite well.'), so it is placed before the verb according to the rules of Hungarian: *Elég jól tudok beszélni japánul.* 'I can speak Japanese quite well.'.
3. Lack of collocation + INF:

**Remélem tudok jobban beszél*. The collocation is always followed by the infinitive of the verb: *Remélem, jobban tudok beszélni.* 'I hope I can speak better.'.

4. Conjugation error:
   **De nem tud jól játszani, mert nem vagyok zenei tehetség*. The collocation is correct, but the verb in the first clause is not in agreement with the verb in the second clause. The first clause uses the third-person singular form of the verb *tud*, while the second clause uses the first-person singular form. The sentence *De nem tudok jól játszani, mert nem vagyok zenei tehetség.* 'But I can't play well because I'm not musically talented.' would therefore be correct according to the matching rules.

5. Summary

We can therefore observe that the collocates of the collocation are typically juxtaposed in all three corpora, with the exception of adverbs wedged between them. In terms of frequency indicators, the two most frequently wedged adverbs are *elég* 'quite' and *igazán* 'really'. The lexeme *jól* is more frequently placed before the verb *tud*. The word order *tud jól* occurs when the collocation is preceded by an interrogative and/or a negative. In contrast to the other two corpora, in the native corpus the verb *tud* occurs in all tenses (present, past, future), in all numbers (singular, plural), and in all persons, in the definite and indefinite forms, but the frequency indicators show that the first-person singular form of the present tense predominates everywhere. In the native corpus there is also a significant presence of the third-person singular (*tud*/*tudja*) and the third-person plural (*tudnak*/*tudják*) forms, but their absence or under-representation in the learner (and pedagogical) corpus is not (necessarily) due to the fact that the language learners have not acquired this form, but rather for the nature of the texts in the corpus: while language learners generally write about themselves, in the native corpus, which includes press materials, we observe opinions about others and interpretation of others' ideas, as is specific to journalistic language. The position of the collocation is the same across the corpora, no significant differences are found in terms of its position in the sentence (sentence initial, final, or intrasentential), and it occurs in the same pro-

portion in any of the constituent clauses of compound clauses in all three corpora. Negative constructions are found in all three sources, as well as the comparative and superlative forms of the *jól* collocate in comparative sentences. Also the fact the the collocation is typically followed by an infinitive is observed on all the three corpora, but while in the learner and pedagogical corpus the infinitive verbs that are following the collocations are generally related to language learning, language proficiency, and learning processes, in the native corpus the topics are diverse and cannot be delimited. The corpus-based volumes of the *MagyarOK textbook* family reflect natural language use and also meet the requirements of the Common European Framework of Reference for Languages, which includes language learning at all language levels.

The overlaps between the native corpus, the pedagogical corpus and the learner corpus are therefore very clear, but we can also see that the content of the native corpus does not always match the content of the pedagogical and learner corpus, since, while native corpora are intended to represent the natural language use of native speakers, pedagogical corpora are intended to illustrate elements of natural language use in a way that is comprehensible to language learners according to their language level (a condition that does not exclude the representativeness of the corpus). The data contained in the learners' corpora show the extent to which language learners have adapted and transformed the given material.

Despite the fact that we found less than 10% of incorrect collocations in the material studied, we were able to identify four major types of errors which are not at all negligible and which we have to focus more on eliminating in language teaching.

In our case, in addition to the analysis of the errors, one shortcoming will provide a relevant conclusion for practical teaching, namely, the absence of some typical patterns in the learner corpus. What is striking in the comparison of the corpora is the under-representation, almost total absence, of the structures *ha jól tudom*; *jól tudom/tudják*, *hogy*; *tudom/tudjuk jól, hogy*. (Only once does the structure *ha jól tudom* appear in the corpus: *Nagyon megváltozott,* ***ha jól tudom****, ő alacsony és sovány volt, de most magas és jóképű egyetemista.* 'He's changed a lot, as far as I know he was short and skinny, but now he's a tall and handsome university student.').

In language teaching, our primary goal is to make the language learner a competent language user who, over time, will participate in the communicative process as an equal partner with the native speaker (cf. Szita, Pelcz 2017). The proper use of collocations is now widely regarded as one of the most important prerequisites for authentic, natural language use (cf. Cowie 1998, Lewis 1993, Schmitt 2000, Sinclair 1991, Wray 2002). In addition to specific and correctable errors, it is essential to incorporate into the language teaching process collocational elements that make the students' communication smoother. Corpus-based tasks can greatly help us to achieve our goal, in this case, to memorize the missing patterns listed above. We can also encourage language learners to use corpora, so that they can explore these patterns themselves and incorporate them into their own texts (Kennedy, Miceli 2010, 2017, Szita 2020), or we can use or create task sets using the corpora studied above.

The results of the corpus studies can point out the shortcomings of language learners, which can be eliminated to make language teaching and learning more effective. Not only theoretical researchers and practising language teachers, but also independent language learners can benefit from the research findings. And the wider dissemination of the corpus-based approach can bring about a qualitative change in the teaching of Hungarian as a foreign language.

Works cited

Antal, Zsófia. «A *KorSzak* tanulói korpusz bemutatása és a magyar nemzetiséget jelölő melléknév vizsgálata» [The *KorSzak* learner corpus, and a corpus analysis of the adjective 'magyar' denoting Hungarian nationality]. *Hungarológiai Évkönyv* vol. 22, nn. 1-2 (2021): 7-21. URL: <https://epa.oszk.hu/02200/02287/00022/pdf/>.

Antal, Zsófia *et al*. 2020-. *KorSzak tanulói korpusz* [*KorSzak* learner corpus].

Baumann, Tímea *et al*. «Bemutatkozik a Korpusznyelvészeti és Szakmódszertani Munkacsoport» [Introduction of the Work Group for Corpus Linguistics and Didactics (*KorSzak*)]. *Hungarológiai Évkönyv* vol. 21, nn. 1-2 (2020): 23-31. URL: <https://epa.oszk.hu/02200/02287/00021/pdf/>.

Balogh, Judit *et al*. 2000. *Magyar grammatika* [Hungarian Grammar]. Budapest: Nemzeti Tankönyvkiadó.

Cowie, Anthony Paul. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

De Groot, Caspar. 1989. *Predicate Structure in a Functional Grammar of Hungarian*. Dordrecht: Foris Publications. DOI: <https://doi.org/10.1515/9783110250480>.

Dickinson, Markus, Scott Ledbetter. 2012. «Annotating Errors in a Hungarian Learner Corpus». In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey, May 21-27, 2012, Istanbul Lüfti Kirdar Convention & Exhibition Centre, Istanbul, Turkey*, edited by Nicoletta Calzolari *et al.*, 1659-1664. URL: <http://www.lrec-conf.org/proceedings/lrec2012/index.html> (*open access*).

Durst, Péter *et al*. «A „HunLearner" magyar tanulói korpusz fejlesztése és várható hozadékai» [The development and expected results of the "HunLearner" Hungarian learner corpus]. *THL2* nn. 1-2 (2013): 28-41. URL: <https://epa.oszk.hu/01400/01467/00010/pdf/>.

—. «Using Automatic Morphological Tools to Process Data from a Learner Corpus of Hungarian». *APPLES: Journal of Applied Language Studies* vol. 8, n. 3 (2014): 39-54. URL: <https://apples.journal.fi/article/view/97871>.

Granger, Sylviane. 2004. «Computer learner corpus research: current status and future prospects». In *Applied Corpus Linguistics: A Multidimensional Perspective*, edited by Ulla Connor and Thomas A. Upton, 123-145. Amsterdam-Atlanta: Rodopi.

Hana, Jirka *et al*. 2010. «Error-tagged learner corpus of Czech». In *Fourth Linguistic Annotation Workshop. Proceedings of the Workshop. 15-16 July 2010, Uppsala University, Uppsala, Sweden*, edited by Nianwen Xue and Massimo Poesio, 11-19. Uppsala: The Association for Computational Linguistics. URL: <https://aclanthology.org/W10-1802.pdf> (*open access*).

Jakubíček, Miloš *et al*. 2012. *Hungarian Web 2012 (huTenTen12)*. URL: <https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhutenten12_hp2>.

Jantunen, Jarmo Harri. «Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi» [International Corpus of Learner Finnish (ICLFI): typology, variables, and annotation]. *Lähivõrdlusi. Lähivertailuja* vol. 21 (2011): 86-105. DOI: <http://dx.doi.org/10.5128/LV21.04>.

Jantunen, Jarmo Harri, Sisko Brunni. 2013. «Morphology, lexical priming and second language acquisition: a corpus-study on learner Finnish». In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Corpora and Language in Use 1*, edited by Sylviane Granger, Gaetanelle Gilquin, and Fanny Meunier, 235-246. Louvain-La-Neuve: Presses Universitaires de Louvain.

Kennedy, Claire, Tiziana Miceli. «Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource». *Language Learning and Technology* vol. 14, n. 1 (2010): 28-44. DOI: <http://dx.doi.org/10125/44201> (*open access*).

—. «Cultivating effective corpus use by language learners». *Computer Assisted Language Learning* vol. 30, nn. 1-2 (2017): 1-24. DOI: <https://doi.org/10.1080/09588221.2016.1264427>.

Kiefer, Ferenc. 1968. «A transformational approach to the verb van ('to be') in Hungarian». In *The Verb 'Be' and its Synonyms. Part 3. Philosophical and Grammatical Studies*, edited by John W. M. Verhaar, 53-85. Foundations of Language/ Supplementary Series/Volume 8. Dordrecht: D. Reidel Publishing Company.

Lewis, Michael. 1993. *The Lexical Approach. The State of ELT and the Way Forward*. Hove: Language Teaching Publications.

Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins. DOI: <https://doi.org/10.1075/scl.14>.

O'Keeffe, Anne, Michael McCarthy, Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1075/ijcl.13.4.09rep>.

Reznicek, Marc *et al*. 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01* [The Falko Handbook. Corpus Structure and Annotations Version 2.01]. Berlin: Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik – Korpuslinguistik.

Rosen, Aleksandr. 2016. «Building and using corpora of non-native Czech». In *Proceedings of the 16th ITAT Conference Information Technologies – Applications and Theory. Tatranské Matliare, Slovakia, September 15-19, 2016*, edited by Broňa Brejová, vol. 1649 of *CEUR Workshop Proceedings*, 80-87. Bratislava: Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics. URL: <https://ceur-ws.org/Vol-1649/80.pdf>.

Schmitt, Norbert. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Szirmai, Mónika. 2005. *Bevezetés a korpusznyelvészetbe: A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában* [Introduction to corpus linguistics: the application of corpus linguistics to the learning and teaching of mother tongue and foreign language]. Budapest: Tinta Kiadó.

Szita, Szilvia. «Korpuszépítés és korpuszhasználat alacsonyabb nyelvtudási szinteken» [Corpus building and the use of corpora at lower levels of language learning]. *Hungarológiai Évkönyv* vol. 21, nn. 1-2 (2020): 173-179. URL: <https://epa.oszk.hu/02200/02287/00021/pdf/>.

—. «A *MagyarOK* nyílt korpusz használatáról» [On the use of the *MagyarOK* open pedagogical corpus]. *Hungarológiai Évkönyv* vol. 22, nn. 1-2 (2021): 72-88. URL: <https://epa.oszk.hu/02200/02287/00022/pdf/>.

Szita, Szilvia, Katalin Pelcz. 2020-. *MagyarOK nyílt pedagógiai korpusz* [*MagyarOK* open pedagogical corpus]. Online as «*MagyarOK* Teaching Materials for Hungarian, Levels A1 to B2» (old version). URL: <https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fmagyarok> (*open access*).

—. «Modellalapú nyelvoktatás, természetes nyelvhasználat» [Model-based language teaching, natural language use]. *THL2* nn. 1-2 (2017): 262-269. URL: <https://epa.oszk.hu/01400/01467/00015/pdf/>.

—. 2013-2019. *MagyarOK A1+, A2+, B1+, B2+ (Magyar nyelvk*önyv + *Nyelvtani munkafüzet* [*MagyarOk* Hungarian coursebook and workbook, CEFR level A1+, A2+, B1+, B2+]. Pécs: University of Pécs.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. New York: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511519772>.