DE GRUYTER
OPEN

Acta Univ. Sapientiae, Philologica, **9**, **3,** (2017) 99–116

DOI: 10.1515/ausp-2017-0032

Translation Studies and Corpus Linguistics: Introducing the Pannonia Corpus

Edina ROBIN, Andrea GÖTZ, Éva PATAKY,
Henriette SZEGH
ELTE (Budapest, Hungary)
Department of Translation Studies
pannonia.corpus@gmail.com

Abstract. The tools of corpus linguistics have become indispensable for research in descriptive translation studies (DTS), which aims to describe the characteristics of the translation process, and translational texts. Machine-readable corpora of translated texts are crucially important since they can yield statistically significant results that underpin the findings of empirical studies. Baker's (1993) seminal paper gave new impetus to translation research as it has re-calibrated the goals of DTS to study and uncover the particular properties of the so-called "third code" (Frawley 1984), i.e. the language of translated texts, with the help of computerized corpora. The present study, after providing a brief overview of international and Hungarian corpus linguistic research, introduces the Pannonia Corpus Project developed by Eötvös Loránd University's *Translation Studies Doctoral Programme*, which was created to make a Hungarian translation corpus, containing millions of words, available for translation researchers. The Pannonia Corpus (PC) is a multi-modal corpus: it contains translated, interpreted, and audiovisual texts. It represents a diverse array of texts of specialized and literary genres, reflecting modern language use and the current state of the translation industry. The PC provides researchers with a vital opportunity as its multi-modality, diverse textual make-up, and substantial size are unparalleled in the Hungarian context. Until now, there were no large corpora available to researchers that could have facilitated qualitative as well as quantitative research, satisfying the demands of modern translation studies research in Hungary.

Keywords: corpus linguistics, translation corpus, parallel, comparable, corpus research

1. Translation studies and corpus research

All knowledge in the world, not formulated in one's native language, may only be accessed through language mediation, i.e. translation. Therefore, it is not tenable to treat translations as the defective, inferior products of secondary communication, unworthy of scientific research in their own right (Baker 1993). Empirically based, descriptive translation studies disregards all such views; however, it does not deny the differences between translated and non-translated texts. One concept capturing this difference is the so-called *third code* (Frawley 1984). The idea of the third code states that the language of translations differs from both the code of source text and that of the target text, despite being created under the influence of both (Frawley 1984: 168). Therefore, the task of translation researchers is to explore the nature of these differences and to examine the universal characteristics of the translational text (Károly 2007).

For descriptive translation studies, which aims to explore the specific general features of translated as well as interpreted language, it is essential to study the translated texts for their own sake. Furthermore, it is also vital to compare translated with authentic, i.e. non-translated texts, as espoused by Baker (1993), in order to fully account for the universal tendencies of translated texts, which emerge when compared to authentic texts. Baker in her seminal paper (1993) named three types of corpora that can be useful for both translation studies and translator training: 1) *parallel corpora*, suitable for studying and teaching translational behaviour, translation strategies; 2) monolingual *comparable corpora*, which accommodate the comparison of translated and non-translated texts; and, finally, 3) *multilingual corpora*, which facilitate investigations of lexicography with a view to equivalence.

Responding to Baker's (1996) call for the use of corpora in translation studies, research projects were set up in many countries around the world to compile parallel and comparable corpora in order to provide statistically significant empirical findings to test the hypotheses formulated about the universal features of the translational text. The spread of computer-readable electronic corpora, facilitating automatic queries, allowed for corpus-based methods to be applied to the examination of translated texts. This means that through these analyses it is possible to uncover the universal linguistic patterns hypothesised to be specific to translations, thus establishing the research area of *corpus-based translation studies* (CTS). To date, many corpora have been compiled, even exceeding the three basic corpus types set up by Baker (1996). These new types can contain bilingual components, creating *bidirectional parallel corpora*, also suitable for comparable text analyses, or translated and interpreted components in *interpretational and intermodal corpora*, the latter containing texts from both written and spoken discourse, as well *as audiovisual* texts, catering for the needs of one of the latest trends in translation research.

Despite the wide use of corpus-based methods in translation research, no corpus, comprising millions of words, has been compiled in Hungary that would allow for the corpus-based study of a wide range of translational activities. To date, all translational Hungarian corpora have been self-assembled and relatively small, designed for the specific aims of the given research (Pápai 2001, Seidl-Péch 2011, Robin 2015). Klaudy (2012) notes how unfortunate it is that, despite numerous previous calls for deploying corpora in translator training (e.g. Kohn 1999), a large Hungarian translational corpus has yet to be compiled and made available to a wide community of translation scholars. Ideally, in order to be representative of the Hungarian translation industry, such a corpus would contain both literary and technical texts. Bringing corpus-based approaches to Hungarian translation studies would benefit both the practice and theory of translation. Significant results derived from a representative corpus could offer more valid information that is rooted in empirical evidence on translation strategies to translators. Similarly, by identifying tendencies, rules, and regularities of Hungarian language use, translation studies could contribute to the development of the Hungarian language (Klaudy 2001).

2. Corpus research in Hungarian translation studies

Pápai (2004) was the first to perform automated analyses on a Hungarian–English parallel and a Hungarian comparable corpus (Arrabona Corpus), examining explicitation in Hungarian translated texts. She compared translations of fiction and sociological texts with the source language originals and comparable authentic texts, examining their type–token ratio and lexical variability. The results of the statistical analysis supported Laviosa's previous results (1998, 2000), as Pápai found a lower type–token ratio in translations than in original texts, meaning that translated texts show less lexical variability. Pápai (2004: 160) concluded that there is a strong relationship between simplification and explicitation: explicitating shifts inevitably lead to an increase in the number of words and lexical repetition – for example, addition of connectives, pronouns, and cataphoric references –, giving rise to less varied vocabulary in translated texts.

Seidl-Péch (2011) similarly examined a self-compiled and annotated translational corpus composed of four sub-corpora, including public, fictional, religious, and scientific texts. She demonstrated cohesive shifts in translated texts through lexico-grammatical analyses. The analyses explore the typical lexical features of authentic and translated Hungarian texts. The corpus only contains texts which are in the public domain, thereby avoiding any copyright problems. The research shows that original and translated texts differ in terms of the use of cohesive devices, which means that the cohesive patterns of translated Hungarian can be traced back to the effects of translation. Furthermore, the research brought a significant result by proving that the examination of cohesive shifts can be automated with tools of corpus linguistics (WordNet).

While examining translation universals in revised texts, Robin (2015) performed general statistical machine analyses on a revisional corpus consisting of the translated and revised versions of ten English language novels. Later on, she compared the results with the statistical data of ten novels originally written in Hungarian (2016). The average length of sentences, differences between type–token ratios, lexical frequency profiles, lexical density, and the standard deviation of data were examined. From the results, it may be assumed that revisers – whose task is to revise translations and fine-tune them in accordance with the target language norms – perform a significant amount of operations, thereby creating more explicit and less redundant texts with a richer vocabulary. In the majority of cases, due to revisional operations, the features of translated texts seem to approximate those of authentic texts, i.e. the norms of the target language. At the same time, some universal editing strategies may be observed, typical of revision.

Table 1. *Translation corpora in Hungarian translation studies*

	Pápai (2001)	**Heltai (2007)**	**Szabó (2011)**	**Seidl-Péch (2011)**	**Robin (2015)**
Name of the corpus:	Arrabona Corpus		HunOr	Hungarian Lexical Cohesion Project	
Type:	parallel and comparable	parallel	parallel, bidirectional	comparable	parallel, revisional
Size (number of words):	2,400 sentences, 45,000 words	1.1 million words	130,000 words	4 million words	2.8 million words
Languages:	English–Hungarian; Hungarian	English–Hungarian	Russian–Hungarian, Hungarian–Russian	Hungarian–Hungarian	English–Hungarian
Annotated/ Metadata/ Type:	no/ yes	yes/ yes/ POS-tagging, headers	yes/ yes	yes/ partly/ WordNet	no/ yes

	Pápai (2001)	**Heltai (2007)**	**Szabó (2011)**	**Seidl-Péch (2011)**	**Robin (2015)**
Text types/ **Sub-corpora:**	fiction and scientific prose	technical texts/ economy, agriculture, environmental protection, EU texts, science, biology, human sciences	fiction, scientific, official	4 sub-corpora public (EU), fiction, scientific, religious	popular literature 3 sub-corpora original, translated, revised
Date of publication:		1970–2008 (sub-corpus 2002–2008)			after 2000
Other:	The first one hundred sentences of each work. Each of the sub-corpora contains 8 works (original–translated–comparable).	Complete texts. The complementary corpus contains 105 texts of 4,000–5,000 words (these are translations of students of translation), this sub-corpus contains 630,000 words.	Complete texts	Only publicly available texts have been collected (in order to avoid copyright problems).	10 pairs of translator–proofreader. Quantitative and qualitative methods. Categorization of grammatical and lexical transfer operations based on exp. and imp.

Only corpora compiled individually and with a predefined research goal served as the basis of the aforementioned examinations. The characteristics of these corpora are summarized in *Table 1.* In Hungary, there have not been any corpora similar to the English TEC or the Finnish CTF, which could be utilized for a wide range of purposes, nor any corpora containing translations which could give a representative overview of translation activities. The Language Institute of Szent István University started to build a parallel corpus of technical texts in 2001, which was the first project of its kind in Hungary (Heltai 2007). The project aimed at using the results of corpus research in translator training. Prior to compiling the corpus, the research group had defined the fields where texts should be collected from in order to cover a range of translation activities as wide as possible. Also, the texts were categorized according to their level of translation quality. It was regarded as a novelty that the corpus contained not only translations from professional translators but translations of university students as well, providing an opportunity to examine translation quality and competence. Unfortunately, the project was advancing very slowly with building the corpus; then the process got halted partly because of technical reasons, partly due to the difficulties of collecting translated texts; the research group did not achieve their goal as the corpus remained unfinished and inaccessible for researchers. Therefore, Hungarian translation studies still remains without a translation corpus which could facilitate a wide range of research goals.

3. Critical views of corpus-based translation studies

One of the basic methodological problems pointed out by critics concerns how texts are chosen for a particular corpus (Tymoczko 1998). It is not entirely clear on what criteria one chooses texts to be included in the corpus. What should be considered a translation at all? In what type of texts can phenomena assumed to be universals or can be measured at all? Is it legitimate to ignore differences in quality? Can we assume that the potentially universal characteristics resulting from the research are present in all types of translations? Chesterman (1993) also discussed these questions, and he concluded that general descriptive laws can be set up in connection with any kind of translation, on one condition: the behaviour and its result can be described as translation if a connection can be identified between the source and target texts (cf. Toury 1995, Károly 2007). Chesterman (2010) also emphasized that it is worth paying attention to connections between universals and text quality and also to incorporate a quality variable when compiling the corpora.

Bernardini and Zanettin (2004) questioned the way corpora were compiled. They criticized the usage of monolingual comparable corpora. Such corpora became very popular since examining exclusively the target texts excludes bias originating from the source texts. However, they raised the questions of comparability and opposed the idea of ignoring the source texts. They argued that if one intends to compare the characteristics of a translation corpus with that of a corpus originally written in the target language, then it is also necessary to examine the status of the source language text, using a corpus compiled from texts which were originally written in the source language.

Pym (2008) also laments the exclusion of the source language texts, mainly in connection with Baker's (1995) corpus research, arguing that monolingual, comparable corpora are not sufficient when it comes to accounting for interference affecting translation; therefore, conclusions drawn from research using such corpora cannot be deemed as valid and/or universal. Becher (2010) holds similar views in connection with Olohan and Baker (2000), criticizing the "dogma" of the so-called translation-inherent explicitation. His criticism can be generally applied to corpus-based research, similarly to that of several other researchers (Jantunen 2004, Bernardini & Zanettin 2004). Becher (2010) maintains that monolingual translational corpora only suffice for setting up hypotheses and not for providing evidence in themselves.

The debate around corpus data leads back to the conflict between approaches preferring either competence or performance, the fundamental difference of opinion between applied linguistics and generative grammar, based on the fact that the empirical data sourced from corpora might be corrupted as performance unlike competence could be ungrammatical. Corpus research is also criticized because statistical measurements only examine superficial phenomena and do not explore the reasons behind these (Károly 2003: 20). The solution seems to be that quantitative research needs be complemented with qualitative methods (Robin 2015) in order to account for the textual transfer operations causing the patterns identified by quantitative analyses. Furthermore, critics point out how important it is to have comparable data because they provide a point of reference for research results (Saldanha & O'Brien 2013: 67). For example, frequency can only be meaningfully explored if other benchmarks are known for the frequency of the given item or phenomenon, i.e. comparable data are required to put the frequency measured in a given corpus into perspective.

4. The Pannonia Corpus Project

The project was initiated by the researchers of the Translation Studies Doctoral Programme at Eötvös Loránd University with the aim of compiling a so-called mega-corpus of translated Hungarian. Beyond the compilation of this corpus, the project also intends to describe the properties of translation behaviour in general. Such a corpus must be able to accommodate quantitative and qualitative research as well. The compilation of the corpus started within the framework of a doctoral seminar course on translation universals in the spring of 2016. The work has since continued and expanded with the support of the Department of Translation and Interpreting at Eötvös Loránd University, as MA students have been taking part in developing the interpretational and audiovisual sub-corpora.

The research project and the compilation of the Pannonia Corpus has aroused the interest of the Hungarian research community. We have reported on the progress made in the compilation process in various articles and conference papers (Robin et al. 2016; Götz 2016a, 2016b; Robin 2017; Szegh 2016; Robin & Szegh 2017). Beyond the compilation of the corpus, empirical research is continuously conducted on its texts with regard to the properties of translated and interpreted texts; furthermore, dissertations are under way, based on corpus-based analyses of the collected texts.

4.1 The components of the Pannonia Corpus

The Pannonia Corpus lives up to the standards set for modern-day electronic corpora supporting valid research in translation studies: it is multimodal, meaning that it contains *translated*, *interpreted*, *and audiovisual* texts as well in parallel and comparable components, which allows for studying the varied translation activities of the Hungarian translation industry. The texts of the corpora were chosen to reflect modern Hungarian language use as all translated texts were created after 2000. The aim is to build a translational corpus of tens of millions of words from various text types to ensure that the corpus remains useful for future Hungarian translation research. During the compilation of the texts, we kept in mind all the critical views discussed above concerning the methodology of corpus research (Károly 2003, Bernardini & Zanettin 2004, Pym 2008), choosing texts (Tymoczko 1998) and the variety of genres (Heltai 2007).

The Pannonia Corpus is made up of a parallel and a comparable component, as shown in *Figure 1*. The comparable component contains texts written originally in Hungarian, which can be broken down into *translational*, *interpreted*, *and audiovisual* sub-corpora, mirroring the make-up of the parallel corpus, so it may be considered translation dependent (Zanettin 2000). The parallel corpus comprises texts translated into Hungarian, mainly from English, and texts translated from Hungarian. The Pannonia Corpus is a *bidirectional* corpus as Hungarian texts translated into other languages are also included in the comparable component. It is planned that when the corpus reaches its final size, these texts will comprise half of the main comparable corpus.

The parallel corpus consists of three sub-corpora: *translational*, *interpretational*, and *audiovisual*. The translational corpus contains written, published texts, whereas the interpreting corpus, similar to EPTIC, consists of EP speeches and their transcribed and normalized versions as well as the simultaneously interpreted and translated versions. In this sense, this is rather a *pseudo-parallel* corpus, like EPTIC, since the written version and the speech of the interpreter cannot be always deemed as strictly parallel, although they are very closely connected. Currently, the interpretational corpus contains only simultaneously interpreted texts though the addition of consecutive interpretation is planned. Similarly to the interpretational, the audiovisual corpus includes the subtitles, the spoken text, and the dubbed versions of movies and television series as well as the original and translated subtitles and the voice-over versions of documentaries.

An important novelty of the parallel corpus is that it contains a number of complementary elements: 1) draft translations of certain translated texts incorporated in the parallel corpus, both from fiction and technical texts; thus, it is possible to build a *revisional* corpus, enabling the researcher to explore differences of quality between revised versions and draft translations and to

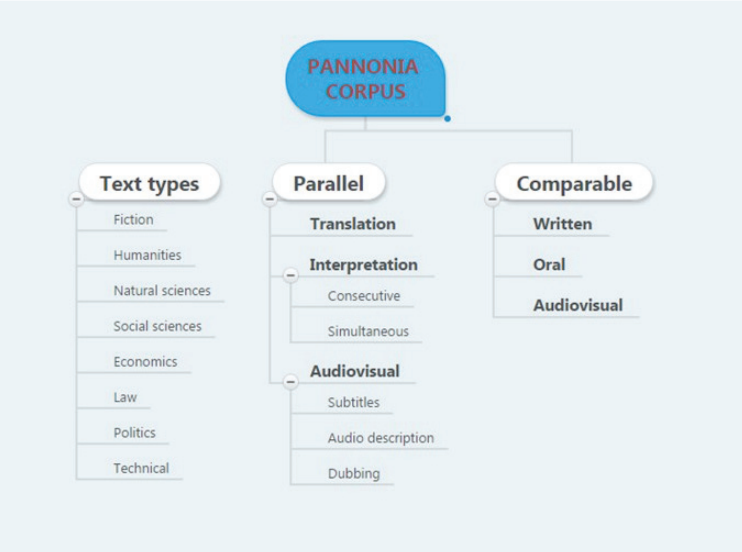


Figure 1. *The make-up of the Pannonia Corpus*

examine revision as such; 2) the qualifying translations of university students, serving as a complementary didactic corpus, can also be of help when making comparisons of quality or examining translators' competence; hopefully, later on supplemented by interpreted texts as well; 3) previous translations of high literary pieces, created before 2000, are also included, constituting the basis for a *retranslational* corpus; although the main aim of the project is to represent modern language in translation, the inclusion of re-translated texts opens up the possibility for diachronic research as well. Amongst the audiovisual texts, the researcher may find the work of fan translators, providing even more opportunities for the examination of translation quality.

4.2 Collection of texts and representativeness

The technical and complementary, didactic corpus of the Language Institute of Szent István University and Robin's (2015, 2016) revisional and comparable corpora served as an example for our corpus. We have collected texts from the vast amount of texts available on the Internet, and we have contacted different publishers and organizations in order to ask for translations for the Pannonia Corpus to use them – with their consent – for scientific purposes. Although in some cases publishers rejected our request, many publishers and organizations supported the project and provided us with original and translated texts in a digital format. We are grateful to every translator and reviser and the following publishing companies for supporting the project with texts: Könyvmolyképző Kiadó, Szak Kiadó, HVG Könyvek, Tempus Közalapítvány, Gondola Kiadó, and Corvin Kiadó.

Table 2. *The texts of the Pannonia Corpus according to sub-corpora and text types*

Comparable	**humanities-related**	81,971
	business	27,151
	engineering	39,084
	popular fiction	924,994
	social sciences	166,386
Comparable – Total		1,239,586
Parallel	**humanities-related**	199,102
	business	603,251
	legal	559,715
	engineering	816,231
	political	149,723
	literature	3,630,079
	popular fiction	3,819,721
	social sciences	1,069,804
	science	481,298
Parallel – Total		11,425,130
Parallel, Comparable	**humanities-related**	50,748
	business	624
	legal	148,393
	engineering	954
	political	14,929
	literature	224,659
	popular fiction	193,394
	science	30,999
Parallel, Comparable – Total		664,700
Total		13,329,416

We have also collected original and translated texts publicly available on the Internet, in each case from webpages of organizations that permit the free use of their content if bibliographical data and references are indicated properly, which we have done, too. Among our most important sources are ELTE Reader, Amnesty International, Greenpeace International, the homepage of TED Talks, and the database of the European Parliament containing translated and interpreted texts. We have processed the texts of the audiovisual corpus by transcribing the oral texts. In each of the cases, we collected complete texts, books, studies, films, or speeches so that later researchers can decide if they wish to analyse complete texts or only parts of texts. The corpus reflects the work of numerous translators, interpreters and revisers; it consists of altogether 800 text files but does not contain more than 200,000 words from any of the authors.

The aim is to collect texts from as many genres as possible in order to ensure that the corpus appropriately represents the Hungarian translation activities, thereby ensuring representativeness. *Table 2* shows the current distribution of the different text types of the corpora, which still needs to be balanced out. Now, the Pannonia Corpus contains approximately 14 million words: almost half of the corpus is made up of technical texts, following the methodological concept according to which research in translation studies must not be limited to fiction (Heltai 2007). The final size is expected to be around 30 million words.

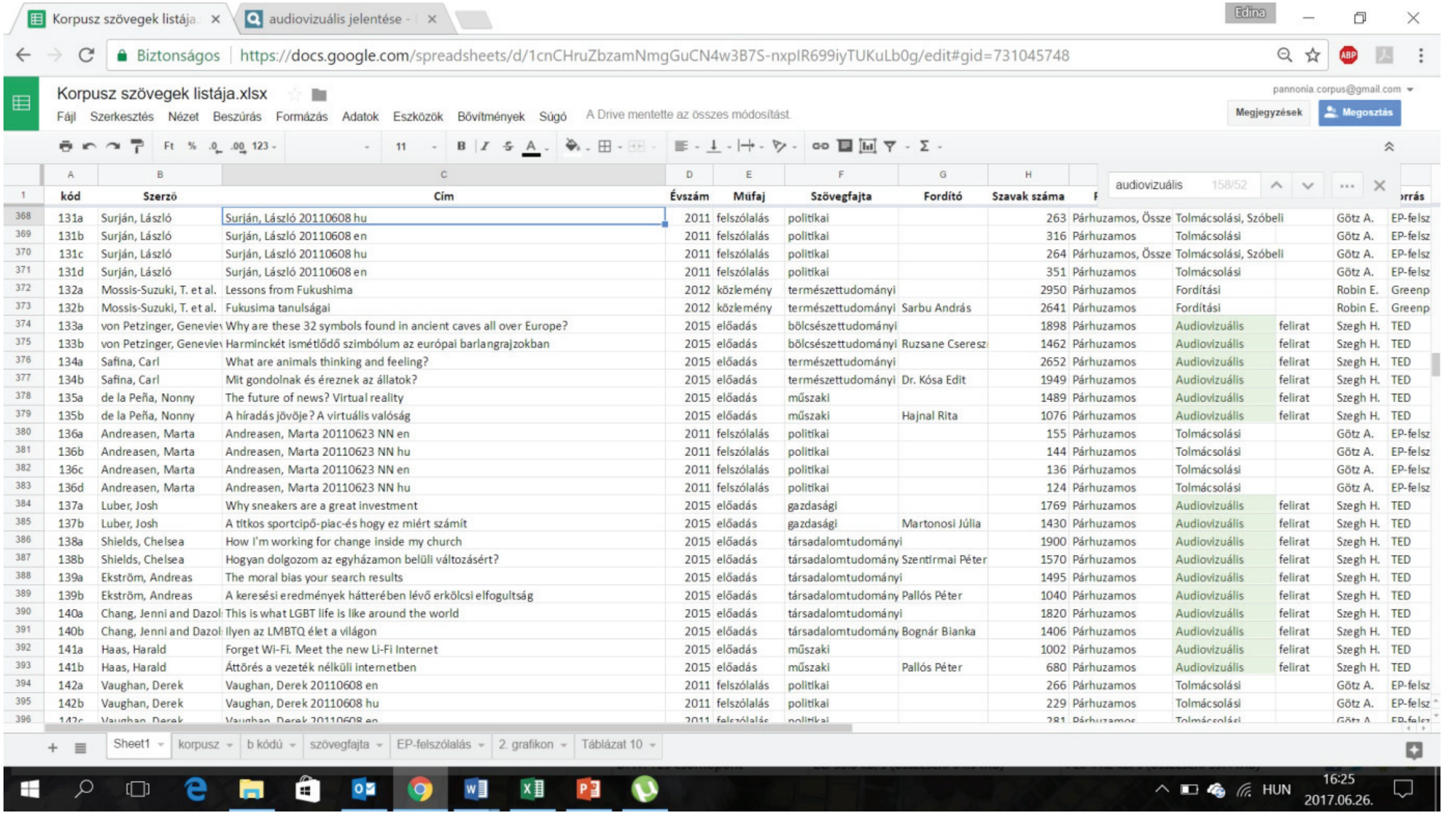
4.3 Technical background of the corpus

The corpus is completely digitized. Currently, it is stored in a cloud storage service. The texts can be searched semi-automatically with the help of a spreadsheet, where the researcher can choose from the texts according to their author, title, year of publication, genre, text type, and translator. This helps if the researchers do not want to search the whole corpus but would like to compile their own sub-corpus instead, based on their own criteria. The search result points to a link with an individual code showing the original text as well as its translated or interpreted version.

The documents are accessible in .txt format, and their metainformation is available in files containing separate headers. *Table 3* shows what kind of information the headers contain on each text, e.g. the name of the translator, the title of the translation, the type of the translation process, the author's name, and the source text's title.

Furthermore, another document containing the bibliographic data is also part of the corpus. This document ensures the searchability of the texts and the protection of copyrights.

In its current state, the Pannonia Corpus can be analysed manually, semi-automatically, and automatically. The translated and interpreted texts are saved in a .txt format, which can be examined with the help of Wordsmith Tools 6.0, Lex Tutor, and AntConc – all of them are computer-based analysing programs. This way, based on the texts in Pannonia Corpus, it is possible to query lists of frequency, and it is also possible to establish frequency profiles (Xiao et al. 2010) and the type–token ratio, the average length of sentences, and numerous other statistical data can be identified – also for each genre or text type separately.


Korpusz szövegek listája
audiovizuális jelentése -
Edina
Biztonságos
https://docs.google.com/spreadsheets/d/1cnCHruZbzamNmgGuCN4w3B7S-nxplR699iyTUKuLb0g/edit#gid=731045748
Korpusz szövegek listája.xlsx
Fájl Szerkesztés Nézet Beszúrás Formázás Adatok Eszközök Bővítmények Súgó
A Drive mentette az összes módosítást.
pannonia.corpus@gmail.com
Megjegyzések
Megosztás
audiovizuális 158/52
Sheet1
korpusz
b kódú
szövegfajta
EP-felszólalás
2. grafikon
Táblázat 10
HUN
16:25
2017.06.26.

1	kód	Szerző	Cím	Évszám	Műfaj	Szövegfajta	Fordító	Szavak száma	F				rrás
368	131a	Surján, László	Surján, László 20110608 hu	2011	felszólalás	politikai		263	Párhuzamos, Össze	Tolmácsolási, Szóbeli		Götz A.	EP-felsz
369	131b	Surján, László	Surján, László 20110608 en	2011	felszólalás	politikai		316	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
370	131c	Surján, László	Surján, László 20110608 hu	2011	felszólalás	politikai		264	Párhuzamos, Össze	Tolmácsolási, Szóbeli		Götz A.	EP-felsz
371	131d	Surján, László	Surján, László 20110608 en	2011	felszólalás	politikai		351	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
372	132a	Mossis-Suzuki, T. et al.	Lessons from Fukushima	2012	közlemény	természettudományi		2950	Párhuzamos	Fordítási		Robin E.	Greenp
373	132b	Mossis-Suzuki, T. et al.	Fukusima tanulságai	2012	közlemény	természettudományi	Sarbu András	2641	Párhuzamos	Fordítási		Robin E.	Greenp
374	133a	von Petzinger, Genevie	Why are these 32 symbols found in ancient caves all over Europe?	2015	előadás	bölcsészettudományi		1898	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
375	133b	von Petzinger, Genevie	Harminckét ismétlődő szimbólum az európai barlangrajzokban	2015	előadás	bölcsészettudományi	Ruzsane Cseresz	1462	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
376	134a	Safina, Carl	What are animals thinking and feeling?	2015	előadás	természettudományi		2652	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
377	134b	Safina, Carl	Mit gondolnak és éreznek az állatok?	2015	előadás	természettudományi	Dr. Kósa Edit	1949	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
378	135a	de la Peña, Nonny	The future of news? Virtual reality	2015	előadás	műszaki		1489	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
379	135b	de la Peña, Nonny	A híradás jövője? A virtuális valóság	2015	előadás	műszaki	Hajnal Rita	1076	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
380	136a	Andreasen, Marta	Andreasen, Marta 20110623 NN en	2011	felszólalás	politikai		155	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
381	136b	Andreasen, Marta	Andreasen, Marta 20110623 NN hu	2011	felszólalás	politikai		144	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
382	136c	Andreasen, Marta	Andreasen, Marta 20110623 NN en	2011	felszólalás	politikai		136	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
383	136d	Andreasen, Marta	Andreasen, Marta 20110623 NN hu	2011	felszólalás	politikai		124	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
384	137a	Luber, Josh	Why sneakers are a great investment	2015	előadás	gazdasági		1769	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
385	137b	Luber, Josh	A titkos sportcipő-piac-és hogy ez miért számít	2015	előadás	gazdasági	Martonosi Júlia	1430	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
386	138a	Shields, Chelsea	How I'm working for change inside my church	2015	előadás	társadalomtudományi		1900	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
387	138b	Shields, Chelsea	Hogyan dolgozom az egyházamon belüli változásért?	2015	előadás	társadalomtudomány	Szentirmai Péter	1570	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
388	139a	Ekström, Andreas	The moral bias your search results	2015	előadás	társadalomtudományi		1495	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
389	139b	Ekström, Andreas	A keresési eredmények hátterében lévő erkölcsi elfogultság	2015	előadás	társadalomtudomány	Pallós Péter	1040	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
390	140a	Chang, Jenni and Dazol	This is what LGBT life is like around the world	2015	előadás	társadalomtudományi		1820	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
391	140b	Chang, Jenni and Dazol	Ilyen az LMBTQ élet a világon	2015	előadás	társadalomtudomány	Bognár Bianka	1406	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
392	141a	Haas, Harald	Forget Wi-Fi. Meet the new Li-Fi Internet	2015	előadás	műszaki		1002	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
393	141b	Haas, Harald	Áttörés a vezeték nélküli internetben	2015	előadás	műszaki	Pallós Péter	680	Párhuzamos	Audiovizuális	felirat	Szegh H.	TED
394	142a	Vaughan, Derek	Vaughan, Derek 20110608 en	2011	felszólalás	politikai		266	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
395	142b	Vaughan, Derek	Vaughan, Derek 20110608 hu	2011	felszólalás	politikai		229	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz
396	142c	Vaughan, Derek	Vaughan, Derek 20110608 en	2011	felszólalás	politikai		281	Párhuzamos	Tolmácsolási		Götz A.	EP-felsz

Figure 2. *The Excel spreadsheet containing the basic data of the texts for quick search*

The corpus needs its own website and online storage space, where, beyond storing the details of the texts, an interface would allow for automated searches carried out on the corpus. This could allow researchers to carry out keyword searches on the corpus and its selected components. In the future, the corpus will be automatically annotated, which requires the purchase of a software (POS-tagging, HUMor, WordNet program) and the development of a search interface, which will allow for further linguistic analyses, concordance queries, accommodating analyses of lexicogrammar and cohesion to explore the properties of translated texts – without compromising the availability of qualitative research.

Table 3. *Header of the Pannonia Corpus for recording the metainformation of the texts*

TEXT	
File name	
Main corpus	*parallel, comparable, revisional*
Sub-corpus	*translation, interpreting, audiovisual, written, spoken*
Text type(s)	*spoken, interpreted, normalized, translated, original, revised, translator's version, retranslated, original subtitle, translated subtitle, dubbing*
Genre of the text	*fiction, entertaining literature, human sciences, natural sciences, social sciences, economic, legal, political, technical*
TRANSLATOR	
Name	
Sex	
Nationality	
Competence	*professional, student, volunteer*
TRANSLATION	
Translation's title	
Target language	
Qualification	
Publisher	
Place of publication	
Year of publication	
THE TRANSLATION PROCESS	
Direction	*into native or foreign language*
Type	*consecutive, simultaneous, subtitles, dubbing*

Revision	*revised, translator's version*
CAT-tool	*memoQ, Trados, Google*
Project	*group or individual work*
AUTHOR	
Name	
Sex	
Nationality	
Command of the language	*native, non-native*
SOURCE TEXT	
Original title	
Source language	
Genre	*novel, study, press article, monograph, declaration, informative text, contract, presentation, speech, TV series, movie, an act of law, documentary, decree, guideline*
Publisher	
Place of publication	
Year of publication	
NOTE	
	EP, TED, Amnesty, EU, etc.

5. Conclusions and research possibilities

The work in the present research project has two goals: *corpus compilation* and *corpus-based research*. Work on the Pannonia Corpus has just started; nevertheless, its size with nearly 14 million words is already substantial. Its final size is planned to reach 30 million words. As shown in *Table 1*, the number of texts in certain text types needs to be balanced out. Primarily, additional legal, political, humanities-related, and science texts are needed. The comparable component of the corpus requires further work as all text types require additional texts. The wider research community can only be granted access to the corpus after it has been balanced out. In the future, individual researchers will be granted access to the corpus after having signed the terms and conditions regarding the copyright and appropriate use of the texts. Access will be granted by the lead researcher of the project or the Head of the Translation and Interpreting Doctoral Programme at Eötvös Loránd University.[1]

The Pannonia Corpus is a multimodal, parallel, comparable corpus, specifically established for the purposes of translation research. As set out among the

[1] For access and further inquiries: pannonia.corpus@gmail.com.

objectives of the project, the corpus will soon be accessible for all researchers of translation studies to examine translated texts. The corpus can be combined with other corpora for individual purposes (e.g. Götz 2016b) in order to further enrich our knowledge on translation, and it can be used for compiling education material in translator training. Although the development of Pannonia Corpus is not completed, it supports a plethora of examination in its current state. For example, analyses can already be carried out on translated, interpreted, and audiovisual texts, as well as for intermodal comparisons. In addition, textual operations of translation and interpretation can be investigated, and operations of literary as opposed to technical translation can also be contrasted. Furthermore, the effect of editing can be investigated in terms of the effect of editorial operations on translated texts – not only in literary but also in technical translations as well, in multiple text types. Universals of translation and interpreting can be further explored in relation to the Hungarian language as well as other concepts of translation research such as the re-translation hypothesis.

Acknowledgements

We are grateful to Professor Kinga Klaudy, Head of the Translation Studies Doctoral Programme for providing us with the opportunity to start the work on the corpus within the programme's framework. Furthermore, we would like to thank Dr Ildikó Horváth for giving us support by making it possible for us to involve the MA students at the Department of Translator and Interpreter Training.

References

Baker, Mona. 1993. Corpus linguistics and translation studies: implications and applications. In: Baker, Mona–Francis, Gill–Tognini-Bonelli, Elena (eds), *Text and technology: in honour of John Sinclair*. Amsterdam/Philadelphia: Benjamins. 233–250.

1995. Corpora in translation studies. An overview and suggestion for future research. *Target* 7(2): 223–245.

Becher, Viktor. 2010. Abandoning the notion of 'translation-inherent' explicitation. Against a dogma of translation studies. *Across Languages and Cultures* 11(1): 1–28.

Bernardini, Silvia–Ferraresi, Adriano–Miličević, Maja. 2016. From EPIC to EPTIC. Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28(1): 58–83.

Bernardini, Silvia–Zanettin, Federico. 2004. When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In: Mauranen, Anna–Kujamaki, Pekka (eds), *Translation universals: do they exist?* Amsterdam: Benjamins. 51–62.

Chesterman, Andrew. 1993. From "is" to "ought": Laws, norms and strategies in translation studies. *Target* 5(1): 1–20.

2004. Beyond the particular. In: Mauranen, Anna–Kujamaki, Pekka (eds), *Translation universals: do they exist?* Amsterdam: Benjamins. 33–49.

2010. Why study translation universals? In: Hartama-Heinonen, Ritva–Kukkonen, Pirjo (eds), *Kiasm. Acta Translatologica Helsingiensia* 1: 38–48. Helsinki: Helda.

Dam, Helle Vronning. 2010. Consecutive interpreting. In: Gambier, Yves–van Doorslaer, Luc (eds), *Handbook of translation studies 1*. Amsterdam: John Benjamins Publishing Company. 75–79.

Eskola, Sari. 2004. Untypical frequencies in translated language: a corpus-based study on a literary corpus of translated and non-translated Finnish. In: Mauranen, Anna–Kujamaki, Pekka (eds), *Translation universals: do they exist?* Amsterdam: Benjamins. 83–97.

Frawley, William. 1984. *Translation: literary, linguistic and philosophical perspectives.* Delaware: University of Delaware Press.

Götz, Andrea. 2016a. Az intermodális tolmácsolási korpusz felépítése és kutatási lehetőségei. Elhangzott: A Papp Ferenc Baráti Kör Találkozója. Budapest, ELTE BTK FTT. (19 december 2016).

2016b. *Vajon* in translated Hungarian. Diverging patterns in two fiction genres. *Acta Universitatis Sapientiae, Philologica* 8(3): 31–41.

Heltai, Pál. 2007. Párhuzamos szaknyelvi korpusz munkálatai a Szent István Egyetemen. In: Feketéné Silye, Magdolna (ed.), *Porta Lingua*. Debrecen: DE ATC. 285–293.

Károly, Krisztina. 2003. Korpusznyelvészet és fordításkutatás. *Fordítástudomány* 5(2).

2007. *Szövegtan és fordítás*. Budapest: Akadémiai Kiadó.

Kenny, Dorothy. 2000. Lexical hide-and-seek: looking for creativity in a parallel corpus. In: Olohan, Maeve (ed.), *Intercultural faultlines. Research models in translation studies I: Textual and cognitive aspects*. Amsterdam: St. Jerome. 93–103.

Kohn, János. 1999. Párhuzamos szövegek számítógéppel segített elemzése a fordításoktatásban (1. rész). *Fordítástudomány* 1(1): 67–78.

2000. Párhuzamos szövegek számítógéppel segített elemzése a fordításoktatásban (2. rész). *Fordítástudomány* 2(1): 5–16.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4): 557–570.

Mauranen, Anna. 2000. Strange strings in translated language. A study on corpora. In: Olohan, Maeve (ed.), *Intercultural faultlines. Research models in translation studies I: Textual and cognitive aspects*. Manchester: St. Jerome Pub. 119–141.

Mauranen, Anna–Kujamaki, Pekka (eds). 2004. *Translation universals: do they exist?* Amsterdam: Benjamins.

Olohan, Maeve–Baker, Mona. 2000. Reporting *that* in translated English: evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1(2): 141–158.

Øverås, Linn. 1998. In search of the third code: an investigation of norms in literary translation. *Meta* 43(4): 571–588.

Pápai, Vilma. 2001. *Az explicitációs hipotézis vizsgálata angol–magyar és magyar–magyar párhuzamos korpuszok egybevetésével*. (Unpublished doctoral dissertation). PTE–SZIE, Pécs–Győr.

2004. Explicitation: a universal of translated text? In: Mauranen, Anna–Kujamaki, Pekka (eds), *Translation universals: do they exist?*. Amsterdam: Benjamins. 143–164.

Pym, Anthony. 2008. On Toury's laws of how translators translate. In: Pym, Anthony–Shlesinger, Miriam–Simeoni, Daniel (eds), *Beyond descriptive translation studies: investigations in homage to Gideon*. Amsterdam: John Benjamins. 311–328.

Rabadán, Rosa–Belén, Labrador–Ramón, Noelia. 2009. Corpus-based contrastive analysis and translation universals: a tool for translation quality assessment English → Spanish. *Babel* 55(4): 303–328.

Robin, Edina. 2015. *Fordítási univerzálék a lektorált szövegekben.* (Unpublished doctoral dissertation). ELTE, Budapest.

2016. Lektorált fordítások és eredeti magyar szövegek gépi összehasonlítása. *Fordítástudomány* 18(1): 19–30.

2017. Korpusznyelvészet és fordításkutatás: a Fordítástudományi Doktori Program korpusza. Elhangzott: *XIV. Fordítástudományi PhD-konferencia*. Budapest, ELTE BTK. (30 March 2017).

Robin, Edina–Szegh, Henriett. 2017. A Pannónia Korpusz audiovizuális korpusza. Elhangzott: *Fordításoktatás – szakmai nap*. Budapest, KRE BTK. (10 February 2017).

Saldanha, Gabriela–O'Brien, Sharon. 2012. *Research methodologies in translation studies*. Manchester: Routledge.

Scarpa, Federica. 2006. Corpus-based quality-assessment of specialist translation: a study using parallel and comparable corpora in English and Italian. In: Gotti, Maurizio–Sarcevic, Susan (eds), *Insights into specialized translation–linguistics insights*. Bern: Peter Lang. 155–172.

Seidl-Péch, Olívia. 2011. *Fordított szövegek számítógépes összevetése. Autentikus magyar szövegek és fordítás eredményeként létrejött célnyelvi magyar szövegek* *lexikai kohéziós mintázatának összehasonlító elemzése*. (Unpublished doctoral dissertation). ELTE, Budapest.

Shlesinger, Miriam–Ordan, Noam. 2012. More spoken or more translated? Exploring a known unknown of simultaneous interpreting. *Target* 24(1): 43–60.

Szabó Martina, Katalin–Schmalcz, András–Nagy T., István–Vincze, Veronika. 2011. A HunOr magyar–orosz párhuzamos korpusz. In: Tanács, Attila–Vincze, Veronika (eds), *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: JATEpress. 341–347.

Szegh, Henriett. 2016. Az audiovizuális alkorpusz létrehozásának módszertani kérdései. Elhangzott: *A Papp Ferenc Baráti Kör Találkozója*. Budapest, ELTE BTK FTT. (19 December 2016).

Szirmai, Mónika. 2005. *Bevezetés a korpusznyelvészetbe. A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában.* Segédkönyvek a nyelvészet tanulmányozásához XLVI. Budapest: Tinta.

Tirkkonen-Condit, Sonja. 2002. Translationese – a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target* 14(2): 207–220.

2004. Unique items – over- or under-represented in translated language? In: Mauranen, Anna–Kujamaki, Pekka (eds), *Translation universals: do they exist?* Amsterdam: Benjamins. 177–186.

Toury, Gideon. 1995. *Descriptive translation studies and beyond.* Amsterdam: Benjamins.

Tymoczko, Maria. 1998. Computerized corpora and the future of translation studies. *Meta* 43(4): 653–659.

Xiao, Richard–He, Lianzhen–Yue, Ming. 2010. In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In: Xiao, Richard (ed.), *Using corpora in contrastive and translation studies*. Newcastle: Cambridge Scholars Publishing. 182–214.

Zanettin, Federico. 2000. Parallel corpora in translation studies: Issues in corpus design and analysis. In: Olohan, Maeve (ed.), *Intercultural faultlines. Research models in translation studies I: Textual and cognitive aspects*. Manchester: St. Jerome Pub. 105–118.

2011. Translation and corpus design. *SYNAPS – A Journal of Professional Communication* 26: 14–23.