Linguistic Corpora as International Cultural Heritage: The Corpus of Bulgarian and Ukrainian Parallel Texts

Olena Siruk[1] and Ivan Derzhanski[2]

[1]Taras Shevchenko National University of Kyiv, Ukraine
[2]Institute of Mathematics and Informatics — Bulgarian Academy of Sciences, Sofia, Bulgaria
olebosi@gmail.com,iad58g@gmail.com

Abstract. The paper relates about our ongoing work on the creation of a corpus of Bulgarian and Ukrainian parallel texts. We discuss some differences in the approaches and the interpretation of some concepts, as well as various problems associated with the construction of our corpus, in particular the occasional 'nonparallelism' of original and translated texts. We give examples of the application of the parallel corpus for the study of lexical semantics and note the outstanding role of the corpus in the lexicographic description of Ukrainian and Bulgarian translation equivalents. We draw attention to the importance of creating parallel corpora as objects of national as well as global cultural heritage.

Keywords: Text Corpus, Corpus Linguistics, Parallel Texts, Translation Equivalents, Cultural Heritage.

1 Parallel Corpora as an Instrument for Linguistic Research

In discussing the changes that the world of lexicography underwent at the end of the last century as a result of the increase in computer power, researchers note the importance of the use of massive corpora, that is, vast electronic collections of authentic text, for the study and analysis of linguistic phenomena by new methods that were impossible earlier, and call the development of such corpora 'the most tangible and farthest-reaching consequence of these changes' and the impact of corpora on lexicographers' work 'revolutionary' [1]. Neither has the influence of corpora bypassed multilingual lexicography: at the current stage of development of linguistics, corpora of parallel (original and translated) texts are used with increasing frequency for the contrastive analysis of lexical semantics and for interlanguage research in general as well as for the creation of bilingual and multilingual dictionaries of various types [2]. Such corpora are developed for many languages, especially Slavic (e.g., bilingual corpora with Ukrainian texts within [3], the Corpus of Parallel Russian and Bulgarian Texts [4], etc.). Ukrainian and Bulgarian, however, have not as yet been subjected to comparative corpus analysis, nor have texts in these languages been brought together in a parallel corpus. This paper relates about our experience of designing such a corpus and some aspects of its application[1]. The methodological principles of constructing parallel corpora, and the corpus itself as a systematically ordered and annotated collection of texts in a particular language or languages, are a treasure trove not only for the linguist but also for any specialist who needs linguistic or extralinguistic data contained in the texts of a corpus. Parallel corpora are of significant value as objects of the national and the global cultural heritage.

2 Text Corpora as National Cultural Heritage

The general understanding of a concept and its interpretation by legislative bodies do not always coincide. Such is the case with the understanding of the notion ‘cultural heritage’. The Ukrainian Law ‘On the Protection of Cultural Heritage’ defines this term as the set of objects of cultural value inherited by humanity from previous generations [6], especially material ones, as is evident from the enumeration. Among the kinds of cultural heritage objects are objects of science and technology—unique industrial, manufacturing, engineering, transport, mining facilities, which reflect the level of science and technology of a certain age, scientific domains or sectors of industry. The same law states that the protection of cultural heritage is one of the priorities of the central and local governments.

In a broader sense objects of cultural heritage are understood as both immovable property bearing works of art, science and technology, and intangible cultural heritage objects. This includes customs, ways, ideas and expressions, knowledge and skills as well as associated instruments, objects, artefacts and cultural areas recognised by communities, groups and sometimes even individuals as part of their cultural inheritance. Under the auspices of UNESCO a new nomination of ‘intangible cultural monument’ was introduced, which covers various manifestations of traditional folk culture—folklore, folk arts and crafts, household traditions, etc. Intangible cultural heritage objects are of undeniable value as a source of information on history, archaeology, architecture, urban planning, science and technology, art, aesthetics, ethnology and anthropology, social culture, etc. In recent years, the international community has been attaching great importance to the protection of intangible cultural monuments.

Linguistic corpora as vast collections of various types of texts (especially fiction, folklore, dialect) contain information about popular traditions, customs, lifestyles, knowledge concerning nature and the universe, skills in traditional crafts. Several synchronous corpora, which represent the language of specific time intervals, may constitute a diachronic corpus that reflects the language of a longer historical period. At the same time, corpora are information retrieval computer systems, whose texts are combined according to certain criteria and undergo a series of pre-processing stages, in the course of which they receive additional linguistic and extralinguistic annotation. On the one hand, such linguistic information retrieval systems are monuments of material culture as objects of science and technology; on the other hand, they are also

[1] The paper is partly based on our article [5].

intangible cultural monuments, as they contain text that reflect the life of a community, its linguistic and cultural traits.

3 The Composition of the Corpus of Bulgarian and Ukrainian Parallel Texts (CUB)

The bilingual corpus consists of Bulgarian and Ukrainian parallel texts available in electronic libraries or scanned and recognised by us from paper editions. This motivates the prevalence of fiction in the corpus, in particular novels, which dominate in such sources.

Because original and translated parallel texts for Ukrainian and Bulgarian languages are hard to come by, especially in computer-readable form and online accessibility, we decided to also use Bulgarian and Ukrainian literary translations from other languages as corpus material. That is, we do not restrict the notion of parallel texts to pairs 'original : translation' as do, *inter alia*, the designers of parallel corpora within the National Corpus of the Russian Language [7], the Russian–Bulgarian and Russian–Slovak corpora [8], or lexicographers such as Lendau [1]. Note that other researchers equate the term 'parallel texts' with the term 'bitext', which means simply two versions of a text, usually in different languages [9]. Our concept of a parallel corpus covers both 'the union of a subset of original texts and a subset of their translations into (an) other language(s)' [10] and what researchers call a mutual parallel corpus, which 'contain originals as well as translations into the languages constituting the corpus' [10]. Thus our corpus has several sectors, each of which covers parallel Bulgarian and Ukrainian texts translated from the same language. All sectors are roughly equal in size. At present the Ukrainian part contains approximately 700 thousand tokens in each sector. The Bulgarian one is larger by approximately 15 percent. This is due in part to the differences in the grammatical makeup of the two languages (analytic and synthetic, respectively), in part to the stylistic preferences of many translators (the following example is from G. Boccaccio's *Decameron*; note that the original Italian sentence and its very faithful Bulgarian translation by N. Ivanov and D. Petrov are twice longer than the rather more vivid Ukrainian rendering by M. Lukash: *Calandrino, essendogli il vino uscito dal capo, si levò la mattina; e come scese giù guardò e non vide il porco suo e vide l'uscio aperto* || *На следната утрин, когато главата му се избистрила от виното, Каландрино станал, слязъл долу, огледал се и видял, че прасето е изчезнало, а вратата – отворена* || *Прочумався рано-вранці Каландріно після випивки, встав, дивиться – кабана нема, а двері одчинені).* We envisage adding the originals of all translated texts to the corpus. In addition to allowing the construction of a corpus of useful size, this approach provides a basis for research in a wider range of topics of comparative linguistics (through the emergence of indirect counterparts).

We started the extension of the set of original languages by including translations from closely related languages. It turned out that Polish and Russian were the languages with the greatest quantity of translations into both Ukrainian and Bulgarian available online, so at this stage we limited ourselves to these four Slavic languages.

Other European languages were considered, and parallel translations from English, German, French and Italian were located and added to the corpus. It would be interesting to expand the corpus by adding translations from other Western European languages, and possibly non-European ones as well, although it is predictable that the availability of parallel texts will be problematic.

Obviously, a greater distance between the languages entails more substantial mismatches between the parallel Bulgarian and Ukrainian translations. On the other hand, in the process of translation from a closely related language the translator may consciously or unconsciously choose too literal a translation of some expression. If the source and the target languages are only distantly or not at all related, this risk seems considerably lower. Identifying lexical correspondences is more difficult, but at the same time more interesting, as they are more ‘direct’ because the translator is less affected by the original language, and even where he is, the nature of this effect is of interest in itself.

It should be noted that texts found on the Internet often contain many errors of the OCR, and editing them (with consultation of the paper source as needed) is a very time-consuming process. Currently, we continue to augment the corpus with scanned books from paper libraries, although the quality of the print is frequently too poor for OCR, which reduces the likelihood that the text can be used at all.

4 Aligning Texts in CUB

The texts are segmented into sentences through *ad hoc* software tools. In the case of some texts with particularly long sentences it seems appropriate to also treat semicolons, and occasionally colons, as end-of-sentence punctuation. Otherwise the matching portions of text can turn out to be impractically large, especially if the sentence boundaries do not coincide, which happens quite often (such texts include *Doctor Faustus* by T. Mann, *Decameron* by G. Boccaccio, *One Hundred Years of Solitude* by G. G. Marquez).

The texts segmented into sentences are aligned using the program Hunalign [11]. The partial automation of this process helped identifying the problem of ‘nonparallelism’ of original and translated texts. The numerous differences between the juxtaposed Bulgarian and Ukrainian texts can be due to a reduction of the original text in translation (in three basic varieties: deletion of individual sentences, reformulation of sentences with the effect of shortening a paragraph with preservation of the general meaning, and deletion of large portions of text: for example, P. Kâneva’s translation of P. Zahrebelny’s novel *Let's Come to Love* is missing the entire inbuilt play). Or it can amount to rearranging entire paragraphs and even chapters (for example, chapters 2, 3 and 10 of A. Gulyashki’s novel *Midnight Adventure* correspond to chapters 9, 1 and 8 in O. D. Ketkov’s translation) or to complete content divergence of the texts. We can only guess at what stage the text was transformed in each case and whether it was the translator’s or the editor’s decision, perhaps reflecting their notions of the audience’s interests or expectations. It may also be that in some cases there were different original editions, and we found one but the translation we have was made from another. But this is a matter for a separate exploration.

Such 'not-quite-parallel' texts evoke contradictory feelings. On the one hand, they are material for research of transitions in translations, translation history etc., but on the other they complicate the corpus processing of texts, because semantic changes are often accompanied by formal rearrangements. Where sentences or paragraphs are omitted or moved, the text has to be aligned by hand, which takes extra time.

Items that are of interest for translation theory but not subject to automatic processing include comparable but nonparallel passages of text. For example, U. Eco's novel *The Name of the Rose* contains a sentence in early (10th century) vernacular Italian (a text known as the *Placito Cassinese*), and in N. Ivanov's Bulgarian translation its meaning is rendered literally in the contemporary language, whereas in M. Prokopovych's Ukrainian translation it is replaced by a passage from a Ukrainian text of a similar age (in accord with to the author's intent, which is that the words should be understandable only in part and their meaning of no significance to the scene: interestingly, William Weaver's English translation available on Google Books is missing this sentence altogether).

Poetic insertions in prose texts are also problematic. Since the regularity of the poetic form is usually achieved at the expense of the accuracy of the translation, we decided to excise them and only work with the prose. Reduction may also be in order in the case of certain characters' deliberately distorted language (for example, Salvatore in the same novel by U. Eco 'spoke all languages, and no language' in the original and in the Ukrainian translation, that is, expresses himself in a hodge-podge of words and structures of different languages; in the Bulgarian translation his speech is virtually standard).

Albeit seldom, unconscious divergences happen too, when two translators differ in their understanding of some genuinely ambiguous expression. Thus the Polish *Wybiła godzina* (S. Lem, *Fiasco*) was understood and translated as *Час настав* 'The time came' by the Ukrainian translator D. Andrukhiv and as *Удари един часът* 'The clock struck one' by the Bulgarian translator L. Vasileva. A related, more frequent (and more interesting) phenomenon is the inevitable divergence in the translation of ambiguous lexical items of the original language, such as English *you* (singular or plural) or *cherry (Prunus avium* or *Prunus cerasus*, considered different plants and named by different words in the Slavic languages).

Finally, imprecise and incorrect translations occur. An example of imprecision is the rendition of French *cochon de lait* 'suckling pig' (J. Verne, *The Mysterious Island*) as simply *прасенце* 'piglet' in Y. Petrov's Bulgarian translation. A translation error can be seen in the comparison of the parallel *sentences Czas biegu sygnałów nie może być dłuższy od czasu reakcji składników komputera* 'The travel time of the signals could not be longer than the reaction time of the components' || *Времето за преминаване на сигнала не трябваше да бъде по-голямо от времето, за което реагират съответните съставни части от компютъра* ditto || *Швидкість руху сигналів не може бути більшою від швидкості реакції складових елементів комп'ютера* lit. 'The speed of the signals could not be greater than the reaction speed of the components' (the greater the speed, the less time it takes to travel, so it should have been 'The speed […] could not be less […]').

5 The Problem of Balance of the Corpus of Parallel Texts

In the process of accumulation of texts we observed a correlation between languages and genres, which does not improve the balance of the corpus. Even if we found many translations from a certain language, they may all belong to a single much translated author. And this is undesirable for the purposes of statistical lexical studies. In some extreme cases the most frequent words include proper names invented by authors (e.g., from I. Efremov's fantasy novels). If a language is represented by two or three authors, it is very likely that they worked in the same genre. As a result, at present the Bulgarian sector is dominated by political detective (A. Gulyashki, B. Raynov), the Polish one by historical novels (B. Prus, H. Sienkiewicz), the Russian one by science fiction (A. Belyaev, I. Efremov). The balance of the corpus is also affected by the small number of translators (their individual tastes, especially in vocabulary and phraseology, influence the style of the translation). Clearly it will be difficult to go beyond fiction and find parallel Ukrainian–Bulgarian texts of other genres (journalism, science, art criticism, memoirs, biographies, etc.) in representative quantities, but we should strive to reduce the imbalance at least within the fictional genre.

6 Corpus-based Research of Lexical Translation Equivalents

The parallel corpus provides rich opportunities for statistical research of interlingual lexical correspondences in order to clarify the meanings of words or correlations between words in certain meanings and conditions of use.

An interesting issue is to analyse the correlations between the frequency of pairs of translation equivalents and in the original language, because 'even when the lexical unit of the target language can be used as a translation equivalent of the lemma of the source language (which is not always the case), the translation of this lexical item is not always the lemma of the original language' [1].

A corpus analysis of translation equivalents can be done for any lexical semantic units. We performed it using a working version of the corpus of texts with Slavic-language originals (about 2½ million tokens on the Ukrainian side) for a group of Bulgarian and Ukrainian time nouns [12, 13].

7 Corpus-based Identification of Lexical Translation Counterparts

Another interesting application of the parallel corpus is the automatic identification of lexical translation equivalents, culminating in the automated construction of a bilingual dictionary. This procedure is based on finding pairs of words that occur most frequently in corresponding sentences in the parallel corpus.

The quality of such a dictionary rises with the size of the corpus, although this increases the demands on computing resources. The factors which worsen the quality of the dictionary include:

- inaccuracy of translation, which is characteristic of fictional text (especially for the sectors that contain translations from third languages);
- the great length of sentences characteristic of some authors or created automatically due to different placement of sentence boundaries in the aligned texts;
- the significant difference in the grammatical makeup of the languages involved.

Since both Ukrainian and Bulgarian are inflecting languages with large paradigms, it will be highly expedient to use grammatical vocabularies (lemmatisers) at the initial stage of automatic detection of lexical correspondences in order to conduct the search in the space of pairs of lexical items and not in the much larger space of word forms.

8 Perspectives for the Project

The work on the corpus of Bulgarian and Ukrainian parallel texts continues in three main directions.

The first is the development of the quantitative and qualitative composition of the corpus. This means adding new sectors (original languages), new texts and, where possible, new genres in order to increase the variety of genres and reduce the correlation between genres and languages.

The second is betterment of the quality of the text and its presentation. This covers a wide range of tasks from locating and correcting errors to morphological and syntactic markup of the texts performed in accordance with the accepted standards.

The third is the development of supporting auxiliary software, namely lemmatisers (based on grammatical dictionaries) and a search engine, as well as an online version of the corpus.

References

1. Lendau, S.I.: Dictionaries: the Art and Craft of Lexicography (in Ukrainian). Kyiv (2012)
2. Cysouw, M., Wälchli, B. (eds.) Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie and Universalienforschung STUF 60.2. (2007)
3. National Corpus of the Russian Language, http://ruscorpora.ru
4. Corpus of Parallel Russian and Bulgarian Texts, http://rbcorpus.com
5. Siruk, O., Derzhanski, I.: Lexical Translation Equivalents in Bulgarian and Ukrainian Parallel Texts (in Ukrainian). Ukrajins'ke movoznavstvo: Mizhvidomchyj naukovyj zbirnyk. V. 43, pp. 75–86 (2013)
6. Ukrainian Law, http://zakon4.rada.gov.ua/laws/show/1805-14
7. Dobrovolskij, D.O., Kretov, A.A., Sharov, S.A.: A Corpus of Parallel Texts: Architecture and Possibilities of Use (in Russian). National Corpus of the Russian Language: 2003–2005, pp. 263–296. Moscow (2005)

8. Garabík, R., Zakharov, V.P.: A Parallel Russian‒Slovak Corpus (in Russian). Proceedings of the International Conference "Corpus Linguistics 2006", pp. 81–87. Saint Petersburg (2006)
9. Vitas, D., Krstev, C., Laporte, E.: Preparation and Exploitation of Bilingual Texts. Lux Coreana №1. pp. 110‒132. Han-Seine (2006)
10. Demska, O.: Text Corpus: an Idea of a Different Form (in Ukrainian). Kyiv (2011)
11. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. Parallel corpora for medium density languages. Proceedings of the RANLP, pp. 590‒596 (2005). The program Hunalign: http://mokk.bme.hu/resources/hunalign/
12. Derzhanski, I.: Time Words in Bulgarian and Ukrainian (Using Evidence from Parallel Texts) (in Bulgarian). In: A. Burova, D. Ivanova, E. Hristova, S. Dimitrova, Ts. Avramova (eds.), Time and History in Slavic Languages, Literatures and Cultures. Proceedings of the Eleventh National Slavic Studies Conference, 19–22 April 2012. Volume One: Linguistics, Sofia: St Kliment Ohridski University Publishing House, pp. 229–237 (2013)
13. Derzhanski, I., Siruk, O. Brief Time Words in Bulgarian and Ukrainian (Using evidence from parallel texts). The Eight International Conference "Formal Approaches to South Slavic and Balkan Languages": Book of Abstracts. Zagreb (2012)