This is the accepted manuscript of the article, which has been published in

Mikhailov Mikhail. (2017). Are Classical Principles of Corpus Compiling Applicable to Parallel Corpora of Literary Texts?. Teoksessa Zybatow Lew N, Stauder Andy, Ustaszewski Michael (toim.) Translation Studies and Translation Practice: Proceedings of the 2nd International TRANSLATA Conference, 2014 Part 1. Frankfurt am Main, Bern, Bruxelles, New York, Oxford, Warszawa, Wien: Peter Lang, 151-157. (Forum Translationswissenschaft 19).

http://dx.doi.org/10.3726/b10842

ARE CLASSICAL PRINCIPLES OF CORPUS COMPILING APPLICABLE TO PARALLEL CORPORA OF LITERARY TEXTS?

Mikhail Mikhailov / University of Tampere

Electronic text corpora of all kinds – collections of whole texts, text fragments, transcripts of recorded speech, etc. – are becoming so common that research that does not use corpus data arouses suspicion. Availability of online text archives (documents archives, newspapers, fiction books, etc.) makes it possible to automate collecting the data. Although the problem of corpus availability is still far from being resolved, monolingual corpus linguistics is progressing rapidly. 'National' corpora (BNC, ANC, the Czech National Corpus, the Russian National Corpus, etc.) include hundreds of millions running words, and Sketch Engine corpora are even bigger.

Research using multilingual corpora is less encouraging. Multilingual language resources are much more limited and more modest in size. The reason is obvious: it is far easier to obtain a large number of texts in one language than to find texts with corresponding versions in several languages. Besides, automation of collecting data for multilingual corpora is more difficult to handle. However, multilingual text corpora are being compiled, their sizes increase, the language repertoire is being improved, as well as their availability.

One thing is however important to note: the classical principles of compiling corpora which were outlined for monolingual text collections are so far presumed to be valid for all kinds of corpora.

They are as follows:

- The corpus should be representative, i.e. “represent the range of texts in the population” (Biber 1993).
- The corpus should be composed of samples of equal size (e.g. 2000 running words like in the Brown corpus of American English, see e.g. Francis 1992).
- The corpus should be by default synchronic, i.e. should represent a relatively short period in the functioning of the language.

Moreover, the research methods and software tools would also admittedly be the same, even for speech corpora.

However, even while compiling monolingual corpora researchers cannot consistently use all the above-mentioned principles. E.g. the Russian National Corpus is compiled of whole texts and it represents different historical periods (http://ruscorpora.ru/corpora-

structure.html). The only basic principle of corpus compiling which is being followed, remains aiming at all genres, topics, and authors relevant for the population in question.

These principles is much more difficult to comply with when compiling parallel corpora, i.e. collections of source texts and their translations into other languages (Teubert 1996). The scope of texts to be potentially included into a parallel corpus is limited to the texts which are translated from language A to language B. And it is evident that texts of some genres are translated regularly, some occasionally, some only sometimes, and some are never (or almost never) translated. Translated texts make only a relatively small part of all the texts of target language. Besides, public availability of texts of some genres is limited (e.g. official letter exchange between private companies) others are available but the text pairs are difficult to locate (e.g. newspaper articles).

It is easy to notice, that among parallel corpora are many corpora of literary texts. The reasons of their popularity are as follows:

- published literary texts are generally considered a reliable source of good language, which is not always true, but they are certainly better then brochures and manuals;
- literary texts and their translations is the kind of data which is easier to obtain in large quantities (of course depending on the language pair: for some language pairs only pseudoparallel texts are available, e.g. the only Kyrgyz-German literary texts available would be works by Chinghis Aitmatov most of which are likely to be translated into German via Russian);
- other kinds of publicly available parallel texts are either in specialist sublanguage or potentially of low quality.

Strangely, the main group of compilers and users of parallel corpora of literary texts are not translators, as it might be expected, but rather researchers in contrastive linguistics. They are interested mainly in linguistic issues: correspondences of grammar forms, syntax constructions, lexemes, etc. Therefore, the issues of translation not connected with comparative linguistics (e.g. the process of translating, translation strategies, skopos, etc.) are seldom studied using corpora.

Many of the classical principles of compiling corpora, when applied to parallel corpora, make them less useful in translation research and in practical translation work.

Parallel corpora of literary texts can be composed of samples, but full texts are preferable. The reason is that a translation of a certain literary work might easily become an object of a case study, and in such case the whole text would be needed.

Parallel corpora are very difficult to collect as synchronic corpora of present-day texts. In such case a parallel corpus of literary texts should be compiled only of recently published translations of recently published works, besides both the author and the translator should not belong to an older generation and should be native speakers of the language they write / translate in. This might be a difficult task even for pairs of major languages. Besides, even if we get a valid set of data for linguistic research of present-day language, we'll miss many cases that may be interesting from the point of view of translation studies, e.g. evolution of translation of Dostojevski's works from Russian into Finnish.

It might be useful therefore to include into parallel corpora of literary texts the following kinds of data:

- present-day texts and their translations,
- old texts and their present-day translations,
- old texts and their old translations,
- different texts by the same author and their translations by different translators,
- different translations of the same texts, so called retranslations.

The translations of classical works make an especially difficult case. It is not typical when a classical work had never been translated and is translated now for the first time into this language. Even so (let's imagine translating Petronius into the Komi language) there would be other translations into other languages the translator might most probably use. The majority of ‘new’ translations are in fact retranslations, which are influenced both by the original text and by previous translations. Therefore, one cannot study the language of retranslations in the same way as the first translations and even when studying the translations of certain contexts one should be aware of the possible influence of previous translations.

It is also important to include more texts by well-known authors and translators. The classical texts in the corpus are of practical use for translators of fiction: they might find translations of important quotations, get equivalents to the culture-bound lexical items, etc. Another reason is that classical texts are usually more often translated and thus yield more data for research than single translations of less known belletrists.

Compiling corpora along these lines makes the data “skewed”, and the corpus becomes less usable as a whole, i.e. general statistics only show tendencies. On the other hand, a user can and should work with subcorpora, which would provide sufficient amount of data of

different kinds. A parallel corpus also becomes a kind of searchable archive of translations, which is useful both for research and for practical translation work.

The above-mentioned principles are used in compiling parallel corpora of literary texts at the University of Tampere: *ParRus* (Russian-Finnish) and *ParFin* (Finnish-Russian).

For compiling *ParRus*, the following source texts are being collected:

- Classical authors of the XIX century: Pushkin, Lermontov, Gogol, Tolstoi, Dostojevski, Turgenev, Chekhov, etc.
- Classical authors of the XX century: Bulgakov, Sholokhov, Pasternak, etc.
- Fiction of the second half of XX century: Pristavkin, Belov, Dudintsev, etc.
- Modern popular fiction: Tolstaja, Ulitskaja, Marinina, etc.

When choosing the source texts an attention was paid to how often they were translated and who was the translator. The most well-known Finnish translators of the XX century were Juhani Konkka, Esa Adrian and Ulla-Liisa Heino. When working on historical dimension it is also important not to forget the works of famous Finnish translators of the past: Samuli Suomalainen, Arvid Järnefelt, and Martti Wuori. Special attention is paid to the works which were frequently retranslated, e.g. works by Gogol, Dostojevski, and Tolstoi.

References

Biber Douglas 1993. Representativeness of Corpus Design. *Literary and Linguistic Computing* vol. 8, no. 4.

Francis W. 1992: Language Corpora B.C. In: Jan Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991.* — Berlin – New York: Mouton de Gruyter, 17–35.

Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner (eds.) 2012: *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German.* De Gruyter, Mouton.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora*. Amsterdam: John Benjamins.

Teubert, Wolfgang 1996: Comparable or Parallel Corpora. *International Journal of Lexicography.* Oxford University Press. 9(3), 238–264.

Zanettin, Federico 2012. *Translation-driven corpora corpus resources for descriptive and applied translation studies*.