**Corpus of Learner Translations and its role in communication, translation and research**

*dr Julia Ostanina-Olszewska, Uniwersytet Warszawski* *jostanina@uw.edu.pl*

*Janusz Parfieniuk, Glosbe Parfieniuk i Stawiński spółka jawna* *janusz@glosbe.com*

It has already been over 20 years since the moment when corpus linguistics and translation studies started collaborating. The beginning of that cooperation goes to the times when Mona Baker published her seminal work 'Corpus Linguistics and Translation Studies: Implications and Applications’. Today we know that the language researchers often use quantitative and corpus methods/methodology in their research.

It should be pointed out that translation studies keep up with the progress in corpus linguistics. Translators and other linguists are working with collections of electronic texts, databases and other corpora. The use of corpus significantly facilitates the translator’s work, however it also makes them responsible for the quality of the translation. It concerns the professional translators as well as translation educators. The relatively big reference corpora serve as unique and extensive source of information about the language use, it’s register, connotations and collocations, phraseology, grammar constructions, etc. (Pędzik 2000). Among these corpora, national corpora plays a very important role. Today, practically every language in Europe has its own national corpus (Narodowy Korpus Języka Polskiego – NKJP[[1]](#footnote-1), Rosyjski Korpus Narodowy – RKN[[2]](#footnote-2) , British National Corpus – BNC, Corpus of Contemporary American English– COCA to name but a few). It’s often the case that parallel corpora is created within national corpora (Russian- Polish, English and other languages within ruscorpora, Polish-Russian corpus created by Marek Laziński form the University of Warsaw and his collaborators, other corpora of Slavic languages (ParaSol, Intercorp, etc.).

Other corpora used by translators are specialized corpora (based on a particular author, for instance A corpus of Gribojedov’s or Chechov’s language). For the translator, it’s vital to have an access to specialized corpus concerning different areas, since it reflects the actual state of professional jargon there. Such corpora are often used for frequency lists and collocation lists which in turn is later used as language teaching material in the language and translation courses.

However, one of the most important type of corpus used by translators is a parallel corpus. An example of such corpus is the system EURAMIS , used in European institutions. Regardless of which one of the 24 languages the original text was written in, according to language diversity and multilingual communication in EU, this text should be translated to all the other languages. Since the translators use CAT tools, all the translations are aligned and available for search in any language pair. It’s used mostly for terminology search as well as whole text fragments search. The analysis of such corpora allows for the terminological unification as well as lexical enrichment. On the basis of parallel corpora we can:

-create specialized and multilanguage databases as well as work on terminology unification,

-create translation memory (TM), analyze the translation, considering the equivalence, reduce the translation time in case of the identical text fragments, search the existing TM, creating terminological database.

- educate: teach courses on translation, teach special language courses, etc.

Parallel corpora are important source for analysis: the aligned texts are crucial for comparison and contrastive analysis.

Another type of special type of corpus is a learner corpus. It is a collection of texts written by the language learners and therefore prone to non-standard language and mistakes.

Such type of corpora was developed at the CECL (Centre for English Corpus Linguistics) led by Sylviane Granger and her team. She’s been working in this area since 80-90s. The initial objective for creating the learner corpus was the need of vast and diverse material for error analysis. Later on it became clear, that this data can provide the opportunity for systematic description of the most typical and frequent mistakes made by the foreign language learners (in this case learners of English).

This idea was an inspiration for creating the Corpus of Learner translations (CoLT) of the Polish students learning Russian language at the Institute of Applied Linguistics UW.

CoLT is meant to be a corpus and a tool (system) which will support the teaching/learning process and improve teacher/student communication.

The system allows the teacher to transfer the texts for translation and for the students to receive those texts, avoiding numerous email messages. Students’ translations may be accompanied by parallel and comparable texts, which are essential in translator’s work. A list of dictionaries (eg. Glosbe) and corpora (ruscorpora, NKJP, COCA, etc.) as well as the tools for creating wordlists are also available at hand.

The teacher’s comments and corrections are saved automatically, which in case of repeated errors saves teacher’s time during reviewing and correction.

Description, classification and analysis of errors will help the teacher to choose the right text for the translation in the future as well serve as needed material in his/her research work.

Here are the steps that the teacher and the students will go through in the process of translation and creating the learner corpus:

1. The teacher creates a translation project, gives a name to define the languages ​​to be translated, and assigns it to the students.

2. The system performs automatic segmentation of text (dividing it into translation units such as paragraphs and sentences).

3. The teacher validates segmentation and if it is done in the right way, starts a project and sends out the relevant notifications to the students.

4. A student performs translation of the document in the system (sentence by sentence, as in systems CAT). those translations can be annotated (adding footnotes, links, explanations). After translating the entire document, marks it as ready which triggers the next step - checking.

5. The teacher gets a notification about the document being ready to be checked. The system automatically searches the document to check whether there are any errors that have already been identified in the database errors. Potential errors are flagged, and the teacher decides whether this is actually an error in the given context. In addition to reviewing the errors detected by the system, the teacher himself marks the errors (which adds errors to the database).



6. Student reviews the errors and fixes them (or leaves them, justifying the choices made).

7. Steps 5-6 are repeated until the error-free translations.

8. If the translation contains no errors can be downloaded as a PDF / TXT / DOC.

Architecture client - server, using a web browser, could also be available from mobile devices such as smartphones, tablets, etc.

Any translation made by the student is recorded in a database. Improved versions are stored separately, so we know how translation is evolving - building a corpus of translations.

Each note submitted by the teacher is also stored in a database – thus the translation errors database is built.

In effect CoLT will improve the organization of the translation process as well as error correction since the full history of translations and corrections will be accessible for analysis.

The study of language interference could be facilitated due to the abundant materials that will be gathered over years.

In the future, the prototype of the corpus of students translations could also be applied in the work of large translation units, for instance in the European institutions, ministries, etc. Even professional translators are not free from making mistakes. This type of corpus could serve as a useful tool for the translation correction, and common mistakes collection.

1. http:// nkjp.pl [↑](#footnote-ref-1)
2. http://www.ruscorpora.ru [↑](#footnote-ref-2)