КОРПУС НЕСОВЕРШЕННЫХ ПЕРЕВОДОВ КАК ИНСТРУМЕНТ ДЛЯ ПЕРЕВОДОВЕДЧЕСКИХ ИССЛЕДОВАНИЙ

Кутузов А. Б. (akutuzov72@gmail.com),
Lionbridge Technologies, Inc., Тампере, Финляндия

Куниловская М. А. (mkunilovskaya@gmail.com),
Ощепков А. Ю. (aoschepkov@gmail.com),
Чепуркова А. Ю. (AnnaChepurkova@yandex.ru)
Тюменский государственный университет,
Тюмень, Российская Федерация

Ключевые слова: параллельные корпусы, учебные корпусы, составление корпуса, переводоведение, эрратология

RUSSIAN LEARNER PARALLEL CORPUS AS A TOOL FOR TRANSLATION STUDIES

Kutuzov A. B. (akutuzov72@gmail.com)
Lionbridge Technologies, Inc, Tampere, Finland

Kunilovskaya M. A. (mkunilovskaya@gmail.com),
Oschepkov A. Y. (aoschepkov@gmail.com),
Chepurkova A. Y. (AnnaChepurkova@yandex.ru)
Tyumen State University, Tyumen, Russia

The paper presents a project aimed at the development of a Russian Learner Parallel Corpus, discusses the existing analogues, describes the current status and the tasks in which it could be used. The existing parallel corpora contain (comparatively) "correct" translations; whereas the aim of the present project is to create a sufficiently large corpus of imperfectly translated Russian and English texts together with their sources and use it as a tool for translation studies, especially those related to translation mistakes. The new corpus will be a valuable resource for computational linguistics as it provides another way of getting data for evaluation which could be used to improve machine translation systems. As of now, the corpus is available online, it already contains nearly half a million word tokens and is growing. The main source of material is translations made by student translators in Russian universities.

Keywords: parallel corpora, corpus building, translation studies, translation mistakes, learner corpora

1. Introduction

The present paper is about the project of **Russian Learner Parallel Corpus (RLPC)**, which is currently under development by a group of Tyumen-based linguists. We will discuss the feasibility of such a corpus, similar projects that have been carried out elsewhere and describe the progress we have made in corpus building as well as the tasks which lie ahead and possible applications of this corpus.

2. Project outline and preceding research

Corpus linguistics has been enjoying the justified attention of translation studies for quite a long time, which is supported at least by the fact that there exist functioning translation and parallel corpora. The cooperation of corpus linguistics and translation studies gave birth to a wide range of translational corpora types. At the moment there are monolingual translational corpora (e. g., Translational English Corpus by Mona Baker) and multilingual aligned corpora with texts created by professional translators (e. g., parallel sub-corpus of Russian National Corpus). The aims of such corpora to provide verified data on translational conformities account for the fact that they do not include amateur or non-professional translations a priori containing translational errors. Such translations are considered to be nuisance, a kind of noise undesirable in parallel corpora.

However, translational mistakes are of interest for researchers in the fields of translation studies and psycholinguists, as well as for methodologists and teachers directly involved with vocational translation training, as will be shown below. That's why the modern range of corpora certainly needs more multilingual translational corpora compiled from non-professional translations.

This is especially true for translations to and from Russian. It seems that there exist(ed) only one such corpus. Since 2004 the Department of Applied Linguistics of the Ulyanovsk State Technical University (Russia) carried out a similar project [Sosnina 2005]. Unfortunately, this corpus is not available on-line, which makes it impossible to make use of it, moreover, 1 million word tokens is not a sufficient number for universal deductions.

There is also the infamous MeLLANGE Learner Translator Corpus (http://corpus.leeds.ac.uk/mellange/ltc.html), the project completed in 2007 by a pull of European scientists and promoted by Department of Intercultural Studies and Applied Languages of the University Paris Diderot and the Société Française des Traducteurs (cf. Kübler 2008). This corpus is compiled of 440 student and professional translations, 350 words long on average, in at least 3 languages (English, German, French) covering 4 text types that were inter alia annotated for errors according to a customised error typology (http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg). The project has a developed query interface which is compatible with larger parallel corpora developed at University of Leeds Centre for Translation Studies. Though this project is a powerful tool for creating training materials and

educating translators, it does not include texts in Russian and is quite limited in terms of text types and overall scope.

Thus, there is a certain vacuum, which our project is supposed to fill.

3. Aims of the project

Within the current project we aim to design and compile **Russian Learner Parallel Corpus** which will include non-professional (student) translations, aligned with their source texts. It would be large enough to give reliable data on translation mistakes and easily-available via the Internet to be used in translation studies research, to create training materials and to inform translation pedagogy. The target corpus size is 10 million word tokens. It will include source and translated texts in two directions: "En-Ru" and "Ru-En". Certainly, there can (and do) exist multiple translation variants of the same source.

The Corpus is compiled of translations made by Translation Studies university students with various levels of translation experience from several Russian universities. Their translations do regularly contain mistakes, they already exist in digital form and they are comparatively easy to obtain.

It should be stressed that our Corpus is after translation mistakes. Though it is intuitively clear what sort of imperfections can a learner translation carry we will offer a few examples that are in no way meant to be exhaustive or limitative or even representative. They are included here merely for the sake of illustration.

Source 1 (from an essay on car industry):

(1) "*The problem with the car, a century on, is that we have allowed Henry Ford and not Henry Royce (of Rolls-Royce fame) to be the winner. For the most part, the car is now an ugly, commonplace form of transport rather than a beautiful toy or exquisite work of engineering. Mass-market cars may be democratic, but with a few shining exceptions (the Mini; the Fiat 500; the Volkswagen 'Beetle') — the vast majority are about as inspiring as a dishwasher*".

Translation 1.1.

(2) "*Сейчас, век спустя, проблема автомобилей в том, что мы позволили Генри Форду, а не Генри Ройсу (основатель компании Ролс-Ройс) стать победителем. Автомобиль сейчас, по большей части, — просто средство передвижения, а не красивая игрушка или привлекающее внимание произведение инженерного искусства. Автомобили массового производства, за редким исключением, весьма неплохи. Но некоторые представители этого класса (например, Мини, Фиат 500, Фольтсваген «Жук») вызывают не больше эмоций, чем посудомоечная машина*".

It seems that the message in the Russian variant of the first sentence is unclear, the implication of "winning" require some explanation such as the one offered in the version below. Besides the overall wording of the sentence in 1.1 is inelegant and clumsy which is especially obvious in the collocation "проблема автомобилей". A much more grievous mistake can be found in sentence 3. The message in translation is opposite to that in the source (cf. translation 1.2 below) and contradicts the logic of the sentence immediately above which contributes to the general target text incongruity. The situation is further aggravated by the extension this mistake has in the closing sentence. This type of translation error is usually referred to as content mistake and is opposed to language (or form) mistakes such as inappropriate collocation, wrong lexical choice (ex. «игрушка» in this context) or grammar mistake.

Translation 1.2.

(3) "*Но дело в том, что сегодня у нас были бы совершенно другие машины, если бы сто лет назад в конкурентной борьбе победил не Генри Форд, а Генри Ройс (основатель марки Rolls-Royce). По большей части, современный автомобиль — это заурядное средство передвижения, а не изысканное инженерное творение. Конечно, автомобили массового производства доступны, но, за исключением ярких примеров, таких как Mini, Fiat, Volkswagen (Жук), в большинстве своем, они вызывают не больше эмоций, чем посудомоечная машина*".

Source 2 (from a company profile description):

(4) "*Развитие малых предприятий тормозит недостаток оборотных средств*".

Translation 2.1. "*Development of small businesses prevents lack of circulating assets*".

A typical mistake for trainee translators working into a foreign language is copying the syntax of the mother-tongue source as is the case in the example above. Here a form mistake interferes with the content. There is also a terminology mistake in rendering "оборотные средства" which are usually referred to in English as "current assets". Below there are two more example of glaring literalisms in translation: a syntactic and a lexical ones respectively:

Source 3:

(5) "*Предпринимателю необходимо в вежливой форме выяснить, на каком основании производится проверка*".

Translation 3.1.

(6) "*To the businessman it is necessary in polite form to find out, on what basis check is made*".

Source 4:

(7) "*Юридические лица в пределах своей компетенции*"

Translation 4.1.

(8) "*Legal persons*"; "*inside its competence*".

4. What one can do with this corpus?

The study of translated texts versus sources yields valuable results for computer-powered monolingual and contrastive research. A new scope to such studies can be added by the analysis of *'non-professional'* *'imperfect'* translations based on the adequate and representational corpus which our project is aimed at.

Below we will deliberately focus on the most workable applications of the Russian Learner Parallel Corpus among many more other possible uses. They seem to fall into three major categories:

1) Translation Studies Research in translation mistakes

The analysis of a variety of imperfect translations against originals can be used to see which of the mistakes are translation-wise relevant and which can be dismissed as formal linguistic faults having no detrimental effect on the translation in general. With any luck this analysis can be formalized on the basis of frequencies in translation memory-like systems.

For example, mistakes related to verb aspect choice (or noun case, or subject and predicate agreement) in Russian translations usually do not corrupt semantic equivalence to the source. At the same time, lexical mistakes in translating connotations or metaphors lead to text tonality change and distortion of text pragmatics. Similarly modelling syntax after foreign patterns often damages semantics of the text.

Further on, generalizations on the translational methods applied in erroneous units can show which of the strategies can be considered inapplicable in particular text-types due to the distortions in pragmatic and semantic plains of the source text and what sort of mistakes more often cause distortions of the source textual and interpersonal meta-functions.

The analysis of imperfect translations can be used to discover translational universals and conformities or translational units that pose no problems in a certain language combination in a certain text-type.

2) Translator Training Methodology

Even wider and immediately practicable applications of the Corpus (which so far lacks translation mistakes mark-up and is basically a collection of student translations) can be seen within translator training sphere. This collection can be used as material in post-editing classes and in the translation studies courses where students practice to spot mistakes, explain their nature and offer other variants as well as to compare different ways of expressing the same idea. It also provides ample material to develop learners insight into the much-discussed issue of the quality in translation. It helps to teach the difference between standard and individual translational problems and generalize about ways of overcoming them. On the basis of the Corpus a methodologist can also make didactic conclusions and target mistakes typical for Russian translation students in their training, including target language mistakes induced by source texts and revealed through the comparative analysis of the translation subcorpus and original target language texts.

3) Computational Linguistics

The Corpus provides the material for comparative analysis of mistakes typical for machine and human translations and draw conclusions on the differences in the processes involved.

The machine-translation end of this Corpus applications also involves the possible use of the "negative" material (generalizations about wrong translations of certain language units for the given language combination) in developing software. Comparative frequencies of corresponding words in two languages and in translational texts can also give useful insights about evaluating and improving machine translation systems.

The Corpus can be used to test or develop algorithms for evaluating the quality of text, such as BLEU (Bilingual Evaluation Understudy) [Papineni 2002].

And finally, the accessibility of the Corpus via the Internet implies the possibility of verification of the findings based on it and also the possibility of offering other explanations by appealing to parameters that may have been downplayed or ignored in previous studies [Baker, 2004].

5. Quantitative and qualitative data about the Corpus

As of now the collected corpus (available on-line in raw form — http://tc.utmn.ru/files/trc.zip) consists of 1146 texts, English and Russian, sources and corresponding translations. The accurate numbers are given in the table below.

Texts

	Source	Translation	Total
English	100	513	613
Russian	25	508	533
Total	125	1021	1146

These texts contain **467513 word tokens, 51% of them being English and 49 percent Russian**. We continue to add texts and will do so until we reach 1 million tokens. This corpus will be made available on-line with query and feedback tools. After that we plan to analyse all the feedback we will have to this moment from our users: researchers of translation errors and conformities and teachers using the "negative" data from the Corpus to create translator training materials. It is especially important for use to receive comments about corpus structure, meta-data, GUI options and actual cases of corpus usage.

After making adjustments to the Corpus structure if necessary we plan to continue expanding the Corpus to reach 10 million word tokens. We are going to set up an on-line contribution option for volunteers to submit source and translated texts, thus employing the power of crowd-sourcing. However, we will still strive to preserve the present ratio of English and Russian translations.

At present we collected texts from Tyumen State University (about 900 texts), Moscow State University (about 70 texts) and Udmurt State University (about 100 texts). Texts from Chelyabinsk State University and Nizhny Novgorod State Linguistic University are still being processed.

Technically, the Corpus is organized in plain text files and header files, according to standards of Translational English Corpus (http://www.llc.manchester.ac.uk/ctis/research/english-corpus). Header files contain student and text meta data for corresponding text files which is arranged in the following fields which present information that can actually be found in the Corpus:

1. Translator' sex
2. University level (Year of study)
3. Assessed translation (Mark received for this translation)
4. Draft or final version
5. Genre of the text (Academic/News(informational)/Essay/Interview/Tech/Fictional/Educational/Encyclopaedia/Speech/Business Letters
6. Type of translation (exam or routine work)
7. Situation of translation (home or classroom)
8. Translation date (by year)
9. Corpus subset (whether the text is a source or a translation)
10. University which has submitted texts

This annotation does not include translation mistakes mark-up which we only plan to develop and introduce at some later stage. Such mark-up, however, will be available for a limited subset within the Corpus because it will have to be made by hand (see below for details).

We plan to use TEC Browser tool (http://modnlp.berlios.de) in order to allow users to query particular sub-corpora. For example, one could look for draft translations made by male students of the Tyumen state university in 2009 and compare them with translations made by female students.

This web application (based on Modnlp libraries) will allow to query the corpus without downloading and installing any software. Queries in bilingual corpus are organized in 3 stages:

1. Looking up text string as per user query.

2. Strings found during the first stage are checked for their translations with the help of bilingual TMX file.
3. All the translations found during the second stage are checked for their equivalents in the other language.

The result of a query is the text string mentioned in the query and all variants of its translations.

Texts in the corpus will be POS-tagged and syntactically marked-up, possibly with the help of Mystem and Synan respectively.

6. Further research

At present the Corpus is still under development and within the next few months we will have to pass the following milestones to reach our objective of providing a comprehensive tool to translator trainers and translation studies academia:

1. Increase the Corpus scope while keeping its structure intact.
2. Develop graphical query interface available on-line. It would be based on TEC Browser client, but we need to take considerable steps to expand its functionality (presently it doesn't work with aligned texts).
3. Provide morphological and syntactic mark-up to make it possible to use the corpus in grammatical contrastive and translation research.
4. It is ambitious but not altogether impossible to supply a smaller sub-corpus with descriptive linguistic mark-up employing XML. Such mark-up can be based on the most basic and universally recognized types of mistakes such as form mistakes and content mistakes discriminated on the degree they interfere with rendering the source message.
5. One can go on and create a GUI which will allow a reviser (such as a teacher or a critic) to mark mistakes electronically and therefore make a revised copy of a student's translation easily and automatically assessable and a ready-made material for further extension of the Russian Learner Parallel Corpus into the bargain.

References

1. *Baker M.* (2004), A corpus-based view of similarity and difference in translation, International Journal of Corpus Linguistics, Volume 9, Number 2, 2004, pp. 167–193
2. *Kübler N.* A Comparable Learner Translator Corpus: creation and use. Proceedings of the Comparable Corpora Workshop of the LREC Conference, May 31 2008, Marrakech, Maroc, pp 73–78
3. *Papineni K., Roukos S., Ward T., Zhu W.* BLEU: a method for automatic evaluation of machine translation. ACL 02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Association for Computational Linguistics. Stroudsburg, PA, USA, 2002, pp 311–318
4. *Sosnina E.* Russian Translation Learner Corpus: The First Insights. The proceedings of the 6 international scientific conference «Interactive systems: problems of human-computer interaction»,Ulyanovsk: UlSTU, 2005