Available Corpora and Error-Annotated Student Translations in Translator Education

Maria Kunilovskaya[1], Natalia Morgoun[2]

Abstract

The aim of this proposal is to describe the practical applications of the online error-tagging environment and freely available corpus resources in translator education. Our teaching methodology includes 1) the use of comparable reference corpora for source text analysis and selection of the most natural sounding renditions and 2) the use of error-tagged translations (both current and stored in the database) at the editing and revision stage of target text production. Thus, we rely on ready-made corpora to teach searching skills and use corpus technology to provide students feedback and access to peer translations stored online.

While our applications of available ready-made reference corpora in translation generally follow the suggestions made in the didactic-oriented corpus-based translation studies, we see our contribution in pulling together a set of corpus resources for the English-Russian language pair and sharing our experience of introducing them at the undergraduate level. The truly original part of this paper is the design and potential of the error-tagging component of the methodology. It fits in the broad context of learner corpus research within its less popular strand of learner translator corpora. To the best of our knowledge, there are about five projects of the kind available online today, including Russian Learner Translator Corpus (RusLTC), and their practical applications still remain unclear.

In this paper on the basis of our experience we will focus on the most effective ways to employ the select reference corpora in the English-Russian translation classroom and discuss benefits and drawbacks of translation error annotation as a method of student-teacher interaction.

1. Aims and motivation

The aim of this paper is two-fold: we set out to describe our experience of introducing open corpus resources to undergraduate translation students, and report the use of our on-line error-annotation tool for providing teacher's feedback and facilitating self-editing.

Despite recent advances in corpus linguistics, increased availability of large corpora and growing awareness of their benefits for translators, schools of translation are slow to include corpus skills and technologies into their curricula, at least in Russia. In designing our trial module we proceed from the assumption that translation is a problem-solving activity. On the one hand it requires a clear understanding of the expected outcome and its communicative potential, and on the other hand it implies multiple solutions to the same problem and various methods to arrive at them.

The second part of the paper is based on our experience of using corpus annotation for translation error mark-up. Although error analysis is a typical part of instructors' teaching responsibilities, there is a lot of scepticism as to whether it is a useful teaching method. At the same time computerised implementation of error analysis makes it an invaluable source of information on prevalent error types and learner translation choices. This data can be used to reconstruct translation process and generalise about weaker components of translation competence of the current student population as well as to make inferences about major language difficulties that students face.

2. Related work

The literature on the subject (notably multiple publications by Bernardini, Zanettin and Frankenberg-Garcia, among many others) proves beyond doubt that corpora query skills are an important part of translator's technical competence today and would especially benefit for novice translators working with specialised texts into their L2. Corpora enable quick access to a wider range of natural solutions, facilitate creativity and variety and boost confidence to resort to less literal renditions. These benefits are not so much due to question-answering, but to thought-provoking potential of corpus use [1]. Therefore, many authors call for special training in search techniques for translators, which helps introduce corpora into translators' daily practice and overcome the initial chill of using a resource that is less intuitive than a dictionary [2,3], but acknowledge lack of systematic information on how to use

[1] Tyumen State University, Russian Federation
[2] Lomonosov Moscow State University, Russian Federation

corpora in translation practice [3]. An attempt to bridge this gap is made in Pastor and Alcina (2009) who offer a translator-oriented approach to search techniques [2]. We find descriptions of teaching actual modules on corpora in translation practice useful [4, 5], and that is where we hope to contribute.
Manual translation error analysis and annotation has acquired new dimensions with the arrival of relatively user-friendly text annotation tools. In the area of translation didactics computer-assisted error annotation is used 1) for assessment and ranking purposes [6]; 2) for research into the translation process and learners' problem-solving strategies with implications for curriculum and material design; 3) and for quality description, when students get structured and personalised feedback on their submissions. In the context of the KORTE project the authors suggest a grading procedure, which is based on the difference of positive points awarded for "good solutions" and negative points for errors. Espunya (2013), who uses data from UPF learner translation corpus, shows how frequency of items associated with specific error tags can be informative for both translation studies and cross-linguistic contrastive analysis [7]. Longitudinal studies, objective feedback on the students' progress and access to multiple translations of the same text to identify common problems and explore variation have been in the heart of Translation Tracking System (2003) [8]. The latter is one of the first initiatives, which sets out to adopt the advances in learner corpora research to translator education. However, this technological transfer is far from being direct. Unlike language errors translation mistakes are rarely binary and their mark-up is more subjective, which limits reliability of corpus analysis. According to results reported in Kunilovskaya (2015) raters tend to disagree on the rigor of analysis, seriousness of errors, notably on the notion of "good solution". They spot a mistake in the same place in text in only about 50% of cases, and out of that they agree on the type of mistake in 80% of cases at Krippendorf's α=0.605. The same coefficient for overall translation quality evaluation (based on a 20-point scale) is a bit higher - 0.734 [9]. These results suggest that reliability of translation error annotation is only in the vicinity of acceptable, even provided that the raters have prior experience with the annotation scheme.

3. Hands-on introduction to using ready-made corpora in translation practice

In this part of the paper we share our experience of teaching a module on corpora use in translation practice for undergraduate students (English<>Russian) as a prerequisite for the subsequent practical translation course. It has 18 hours of contact time and aims at introducing junior students to basic corpus skills and ready-made online corpora that are most feasible for general domain translation.
The course is structured around corpus resources starting from simple ones like Just The Word and SkELL to more demanding ones. Our instruction is limited to three corpus interfaces (SketchEngine, byu.com and that of Russian National Corpus (RNC)) and two types of corpora monolingual national corpora (COCA, BNC and RNC) and comparable web-corpora (Aranea [10], a multilingual collection, which hosted by Comenius University in Bratislava).
This module suggests a hands-on approach to acquiring corpus skills. The unavoidable theoretical concepts and corpora descriptions are offered on the fly as required with minimum formality, while all activities involve real-life problems set in the translation context borrowed from RusLTC [11]. Methodologically the tasks are arranged in progression from *a model example which demonstrates an interface function to evaluating and editing somebody else's solutions to generating one's own*. Our major teaching objective is to lure students into corpora, to develop a happy habit to check their choices with appropriate corpora and crave more heuristic, yet confident, ways of expression. One of the important effects of corpus practice is that it raises translators' awareness of typical translation problems and boosts creativity, even if does not offer quick solutions.
The second tier of the course structure is *interface-specific functionality*. Simpler resources at the beginning of the session give more space for theoretical intervention, so towards the end of the course students are supposed to be familiar with the basic concepts and can absorb more technical skills like regular expressions and CQL syntax. And finally, the practical assignments for independent work are built around typical *translator needs*, associated with 1) source text analysis, 2) generating possible translation variants, 3) editing target texts and evaluating solutions.
Independent corpora use at the source text analysis stage is particularly challenging. In our experience students stumble because they have no hunches about possible implications and intended pragmatic effects. They often take texts at par value, and therefore, lack incentives to refer to corpora.
The more straightforward and popular uses of corpora include finding appropriate collocations and building synonyms lists. Students readily grasp the idea and feel that it does contribute to their productivity and quality. Frequency lists, collocational profiles of synonyms, generic analysis and study of concordance lines come into play when students have to choose between several variants or evaluate and edit translations.

4. How error annotation can be useful in translator education

We motivate students to apply these corpora skills while working on translation assignments during senior year. The course of general translation (76 contact hours over two semesters) offers fragments from a range of text genres for translation. Students are supplied with full source text versions and a translation brief. All classes are scheduled to computer labs with internet access. In pre-translation activities students are asked inter alia about linguistic aspects of the text and are motivated to support their intuitions with evidence from any relevant source. Then there is a discussion stage when the group compares their translations and decides on their acceptability. Most of students' submissions are manually error annotated by the instructor, and there can be a mandatory revision or editing task.

The error mark-up is technically performed in the customised version of the text annotation program called brat [12]; marked-up translations are available on-line and are part of RusLTC. Its site [11] has a description of the predefined error typology used for the annotation.

As of now (April 2016) the collection of error-tagged translations consists of 456 texts with 9619 error-tags in English-Russian subcorpus only. Around 2/3 of tags are supplied with the instructor's notes, which contain explanations, leading questions, pointers to references. The system is in place for two academic years, and we are using the data acquired in three ways.

Firstly, it is used directly as an additional teacher-student communication system, which offers individualised feedback and increases students' autonomy at the editing stage of translation production. It is important that students have access to anonymised peer translations, including from previous years. The comparative analysis shows that revised versions are significantly better, and this improvement is down to students' self-study guided by the received feedback.

Secondly, we have devised an assessment methodology based on error statistics, which is used for ranking translations [9]. Initially we hoped to arrive at error-types-and-weights-into-points scheme, but after a series of inter-rater reliability tests and looking at the differences in text size, difficulty and translation conditions decided that a universal approach was hardly possible. The ranking procedure helps to avoid imposing absolute values on errors and takes into account the performance of the group as a whole.

And finally, we use quantitative and qualitative error analysis to modify syllabi of translation-related modules and even the curriculum as a whole to address the current students' problems. For example, we have introduced a course in translation-oriented discourse analysis. The overview of standard problems in English<>Russian translation as part of Translation Studies course is no longer based on theory, but is data-driven.

5. Conclusions

This paper illustrates the applications of corpus resources and technologies in translation education. Our approach instantiates the use of available corpora and custom-made error annotation on-line environment in practical translation.

We have described the educational context and structure of a practical course for undergraduate translation students designed to provide them with basic corpus-searching skills and, more importantly, a habit to analyze linguistic choices made by the ST author and TT producer against the backdrop of existing language practice. Teaching about corpora can be seen as an effective practical approach to exposing students to a variety of entertaining linguistic issues and discovery procedures that they can use as translators and language learners.

Another component of our teaching methodology is providing students with personalised feedback, which comes in the form of error-annotated translations. We have shown three possible applications of this technology. But given the subjective nature of translation error annotation, we doubt that error statistics can be reliable data for quantitative quality measurements, so cost-effectiveness of this approach is open to question.

References

[1] Zanettin, F., Bernardini, S., & Steward, D. (2014). Corpora in translator education. Routledge.

[2] Pastor, V., & Alcina, A. (2009). Search Techniques in Corpora for the Training of Translators. In Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning, 13–20.

[3] Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: Challenges and reactions by a group of thirteen students at a UK university. In Corpora, 10(3), 351–380.

[4] Boulton, A. (2012). Beyond concordancing: Multiple affordances of corpora in university language degrees. Procedia - Social and Behavioral Sciences, 34(0), 33–38.

[5] Frankenberg-Garcia, A. (2012). Raising teachers' awareness of corpora. In Language Teaching, 45(December 2010), 475–489.
[6] Wurm, A. (2013). Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE), 381–419. Retrieved from http://www.trans-kom.eu/bd06nr02/trans-kom_06_02_06_Wurm_Eigennamen.20131212.pdf
[7] Espunya, A. (2013, June). Investigating lexical difficulties of learners in the error-annotated UPF learner translation corpus. In Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011) (Vol. 1, p. 129). Presses universitaires de Louvain.
[8] Bowker, L., & Bennison, P. (2003). Translation Tracking System: A tool for managing translation archives, 503–507.
[9] Kunilovskaya, M. (2015). How far do we agree on the quality of translation? In English Studies, 1(1), 18-31.
[10] Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 257-264.
[11] Russian Learner Translator Corpus (RusLTC) http://www.rus-ltc.org/
[12] Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S. & Tsujii J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 102–107.