Russian Learner Translator Corpus:
Design, Research Potential and Applications

Andrey Kutuzov[1] and Maria Kunilovskaya[2]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] Tyumen State University, Tyumen, Russia
{akutuzov72,mkunilovskaya}@gmail.com

Abstract. The project we present – Russian Learner Translator Corpus (RusLTC) is a multiple learner translator corpus which stores Russian students' translations out of English and into it. The project is being developed by a cross-functional team of translator trainers and computational linguists in Russia. Translations are collected from several Russian universities; all translations are made as part of routine and exam assignments or as submissions for translation contests by students majoring in translation. As of March 2014 RusLTC contains the total of nearly 1.2 million word tokens, 258 source texts, and 1,795 translations. The paper gives a brief overview of the related research, describes the corpus structure and corpus-building technologies used; it also covers the query tool features and our error annotation solutions. In the final part we make a summary of the RusLTC-based research, its current practical applications and suggest research prospects and possibilities.

Keywords: corpus building, learner corpora, multiple translation corpora, query tool, linguistic mark-up, mistakes annotation, corpus-driven translator education.

1 Introduction

This paper aims to provide a detailed description of an English-Russian learner translator corpus, which brings together traditional parallel corpus-building methods and a specific type of annotation used in translator training, a combination never attempted before.

Russian Learner Translator Corpus (RusLTC) is a parallel corpus of translation trainees' target texts aligned with their sources at sentence level. It includes translations into English and out of English. The learners' mother tongue is Russian. The project sets out to create a representative and reliable resource to be used in translation studies research and to inform translation pedagogy. The corpus is available under Creative Commons license at http://rus-ltc.org website. The corpus is enriched with meta data searchable via the query interface on different external and linguistic factors that can affect students' performance.

RusLTC is a large corpus which is not designed with a specific research purpose in mind, but rather for a broad research agenda in translation studies, including

1. exploring variation and choice in translation, when different translations of the same source are compared;

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 315–323, 2014.

2. comparing learner translator output to native data which can bring to conclusions about non-nativeness and 'translationese'; the translator inter-language and consequences of the constraints of translation as communicative activity as opposed to free speech;
3. exploring interdependence between the translation characteristics and various meta data (direction and conditions of translation, source text genre);
4. analysis of concordances of multiple translations (comparing several translations to sources) that can help to develop and test hypotheses about error-prone linguistic items ("problem areas");
5. computer-aided error analysis of the most common translation errors to draw conclusions about the weaker components in the current translator population competences; this strand of research can be extended to include the study into the didactics of translation quality assessment (TQA);
6. apart from learner corpora and translation studies research, the results of which can be applied in the curriculum and materials design, there are numerous ways how the corpus can be directly used as a teaching and learning aid.

The present paper opens with a brief overview of the related research in Section 2; Section 3 describes the corpus design, its content and types of annotations as well as query interface; the next part of the paper (Section 4) outlines the RusLTC-based research and its classroom use. In Section 5 we provide the outlook for further research and developments.

2 Current Learner Translator Corpora Projects and Research

The use of learner translator corpora in research and translator education seems to have been first reported by Robert Spence [13], who compiled a Corpus of Student L1–L2 Translations. It was followed by a learner translator corpus within the PELCRA project by Raf Uzar and Jacek Waliński [15] and the Student Translation Archive by Lynne Bowker and Peter Bennison [2]. These corpora vary in the number of languages they involve, directionality of translation and design technologies used, but they are similar in being unavailable and relatively small in size.

A new excitement to the field was added by two learner corpora projects which were online. The first one – ENTRAD – was introduced by Celia Florén in 2006 [6]. The corpus is a bank of about 45 English sources translated into Spanish by trainees whose mother tongue is marked as mostly Spanish or French. The corpus is text-level aligned and provides no multiple concordances. The query can be narrowed down by the meta data, including the translator's age, gender and mother tongue. The error annotation is based on a colour code and graphical marks and is not machine-readable.

Another notable achievement in multiple LTC development is multilingual MeLLANGE LTC, which is well-documented and easily available online. It provides extensive searchable meta data and proper error-tagging based on prior linguistic annotation. It comes with an elaborate but user-friendly query interface, which allows to retrieve contexts containing specific error category. MeLLANGE seems to be the only LTC which provides one-to-many concordances and reference translation [4]. Sara Castagnoli is also the author of Multiple Italian Student Translation Corpus (MISTiC) which

stands out for being compiled for specific research purposes. Her PhD thesis reported in [3] sets a benchmark for multiple learner corpora research in translation studies and demonstrates its potential.

In connection with the present English-Russian project we should mention a similar project undertaken at the Department of Applied Linguistics, Ulyanovsk State Technical University (Russia) [12]. Sosnina's RuTLC is reported to be 1 mln tokens in size. It stores translations of technical texts by part-time translator trainees, who carried out translations as their final paper of 10000 words each. Alignment is not reported. The corpus bears HTML-based error-annotation which allows automatic analysis.

The recent developments in the field include the project LTC-UPF presented by Anna Espunya, Andrea Wurm's KOPTE[1] and NEST developed by Anne-Line Graedler. The first two corpora are enriched with state-of-the-art linguistic annotation; both corpora provide proper error annotation. The related publications describe the error typologies and discuss technical problems which are very familiar [5]. While LTC-UPF is relatively small in size (10 sources, 194 targets), KOPTE is considerably larger (77+ sources, 971+ targets) and has an informative site. The third corpus mentioned above is still fledging; the interface is not yet in place, no linguistic or error annotation is reported. The corpus size is limited to 18 sources in Norwegian, which have been rendered into the translator's foreign language (English) [7].

The overview above makes it possible to define RusLTC as **the third learner translator corpus available** online after ENTRAD and MeLLANGE LTC; it is a large corpus, which provides **sentence-level aligned multiple concordances** and searchable meta data. It is **versatile** in terms of source text genres and is balanced as to **directionality** of translation in English-Russian pair. Besides, it can be used both for a "traditional" corpus research and for error-analysis.

3 Corpus Design

3.1 Data Source

We collect translations from 10 Russian universities offering professional translator training. All translations are made as part of routine and exam assignments or as submissions for translation contests. There are, however, translations from trainees who study translation as a supplementary course or study translation part-time. We also include translations made as internship tasks by students majoring in translation and as graduation translation projects by part-time students. All these trainee populations are described in the searchable corpus meta data.

As of March 14, 2014 RusLTC contains the total of nearly 1.2 mln word tokens, 258 source texts (41 Russian sources and their 589 translation; 217 English sources and their 1,206 translation). The number of translations to a single source varies from 1 to more than 60. All translations and meta data are anonymised.

[1] http://fr46.uni-saarland.de/index.php?id=3702

3.2 Corpus Alignment

The parallel nature of the corpus poses some difficulties and does not allow to re-use many popular corpus engines. Thus, we had to design a lot ourselves.

The corpus source texts are aligned with their translations at sentence level. The alignment is done with *hunalign* library [16] and then manually corrected with a TMX-editor *Okapi Olifant*[2]. Sentences that were added by the translator are included in the preceding unit of alignment; untranslated segments of the original are left out, which means that the corpus has no segments with blank translation. When there are several translations of one and the same text, we build several pairs of source and target texts with one and the same source and align them pair-wise. Then redundancy is removed (see below).

Aligned bitexts are stored in **TMX** (Translation Memory eXchange) format, with source and translation segments identified by corresponding XML attributes. The principal unit here is the sentence, however, links to original files are also preserved which allows the interface to retrieve the whole text at need.

Originally TMX format is intended to store translations of one and the same sentence to several languages ('translation unit variants'). We use it differently, at the same time retaining full compatibility. In our TMX base translation unit variants are translations of one and the same sentence made by different students, with different corresponding meta data. Technically this is achieved by our home-grown script which takes the output of *hunalign* and searches for pairs with identical source sentences. Then it joins them into one translation unit, thus avoiding redundancy.

3.3 Meta Data

All translated texts in the corpus are equipped with meta data falling into three major groups: personal, on the source text and translation situation. Personal data include:

1. Trainee's gender
2. Trainee's experience (year of study/education programme or contest)
3. Trainee's affiliation

Translation situation data include:

1. Grade for the translation
2. Conditions of production (routine/exam; home/classroom)
3. Year of production

It should be noted that translations were made under very different conditions (with different amount of stress, under tight or no time limit, as huge projects that had taken weeks to complete and as mini-tasks). No restrictions on use of reference materials was ever reported. The researcher might want to be very specific about the sample he or she is querying. For example, home routine translation are usually done after the source and draft translations have been discussed in class and the translator is relatively at ease as

[2] https://code.google.com/p/okapi-olifant/

far as time factor is concerned, while class exam translations are made under significant time constraint (usually 400–450 words in two hours).

Source text data are limited to its genre. We approach genre as a category assigned on the basis of external criteria such as intended audience, purpose, and activity type. RusLTC includes translations of sources classified into the following eleven genres: academic, informational, essay, interview, tech, fiction, educational, speech, letters, advertisement, review.

Meta data are stored as separate plain text files (headers). Up-to-date automatically updated statistics over the whole corpus can be accessed online at http://dev.rus-ltc.org/statistics.

3.4 Linguistic Mark-Up

All the texts in the corpus are tokenized and POS-tagged with the help of Freeling library [11]. We are now in the process of designing a query tool which will make proper use of this mark-up. Experimental version which allows to search for particular part-of-speech is available at http://dev.rus-ltc.org/search.

There is also a small, but rapidly growing, subcorpus equipped with error annotation. The process of error annotation is powered by customized *brat* text annotation framework [14] and is based on the pre-defined error classification which is a 3-level hierarchy of 31 types. This scheme is based on the fundamental distinction between content-related and language-related errors. Each of the mentioned main categories includes sub-categories such as (wrong) referent, cohesion and pragmatics and lexis, morphology, syntax, spelling and punctuation with further subdivisions respectively. We tried to make our types mutually exclusive, but have found out that in many cases double annotation is possible. In this case the annotators were instructed to evaluate the damage to content transfer and, if present, to tag the mistake as content-related rather than language-related. Apart from this major type of mistakes description, the annotators can use two extra mistakes characteristics, such as weight and technology. The former refers to the seriousness of mistakes and includes a scheme of three points (critical, major and minor), and the latter offers a non-exhaustive variety of possible reasons behind a mistake such as “too literal”, “lack of SL knowledge”, “lack of background information”, etc. It is also important that the error-annotation interface offers an option of a Note by annotator in which he or she can explain the mistake. The note is revealed by hovering with the mouse over error-tags. Because of length limits we do not elaborate much on mistakes classification here, but it is available in full on the corpus site[3].

As of March 2014, we have annotated 241 translation (198 translations into Russian and 43 translations into English), but as the annotation tool is in everyday classroom use, this data is rapidly growing. So far this part of the Corpus can be queried only with standard *brat* means and is available at http://dev.rus-ltc.org/brat/#/rusltc/.

[3] http://rus-ltc.org/classification.html

3.5 User Interface and Query Tool

Java-based query tool at http://rus-ltc.org that we developed supports lexical search for both sources and targets and returns all occurrences of the query item in respective texts along with their targets/sources aligned at sentence level. The search for the query actually happens in the above-mentioned TMX file which allows to return aligned pairs. There is also the possibility to export search results in CSV format and to view full texts and corresponding meta data. Unfortunately, this tool does not support neither search for all morphological variants of the query nor search for words belonging to a particular part of speech. Also, one can search for particular mistakes only in *brat* interface (see subsection 3.4).

As mentioned above, we are also developing a new morphology-enriched query tool in Python which will eventually replace the current one.

Beta status of query tools is mitigated by the fact that the whole Corpus is available under Creative Commons Attribution-ShareAlike license. It can be downloaded both in TMX format or as a compressed set of plain text files and their headers with meta data.

4 Use in Translation Studies and Translator Training

4.1 RusLTC-Based Translation Studies Research

Though our corpus is relatively young (it is under development since 2011), its data have been used in two translation studies research projects. The first one deals with gender asymmetry in translated texts. This work, based on the statistical analysis of the translations in the corpus, has shown that translations made by males and females reproduce the same gender asymmetry (in terms of lexical variety calculated as type to token ratio) that is known for Russian originals. The author has found, though, that texts translated by females tend to contain longer sentences than those translated by males, which doesn't reflect similar statistics for originals [10].

Another RusLTC-based research was to study one of the techniques in translation which is usually referred to as sentence-splitting [9]. It offers detailed analysis of typologically justified types of syntactic splitting in English-Russian translation. More importantly, it describes translation mistakes associated with splitting. Most of these mistakes result in damage to target text coherence and erratic discourse structure. Careless splitting most often leads to loss or misinterpretation of semantic relations between propositions, issues with anaphora resolution and greater communicative value acquired by upgraded sentences. These findings are used to draw students' attention to the textual quality of their production and are implemented in editing exercises aimed at detecting incoherence at semantic and pragmatic levels and editing them.

As part of the corpus involves translation error-tagging, which is claimed to be essentially subjective, there was an attempt to measure inter-rater reliability of our error-annotation system and define most subjective and relatively objective areas in marking translations [8]. The measure was based on annotations provided by three different evaluators who have been trained to use the error typology. It was found out that the experts agree about location of the target text span containing a mistake on average in 60% of cases and demonstrate different target language rigour. The percent

agreement on the type of mistake (in terms of main categories in our hierarchy) is 80,5%, with slightly more agreement on content errors than on language errors. But disappointingly little agreement (34,8%) was seen on the weight of these mistakes and on good decisions offered by trainees. Besides, the statistical data showed that while awarding grades for translations the evaluators tend to agree more on poor translations, rather than on good ones. We have taken steps to improve our typology and reduce disagreement, although it is possible for different teachers to use customized and compatible versions of the classification.

4.2 Classroom Use

Apart from these research applications RusLTC is used as a source of data for awareness-raising and revision/editing exercises aimed at prevention of most frequent translation mistakes and introducing trainees to other corpora usefulness in translation practice. The usefulness of a learner corpus-driven approach in foreign language and translator training has been convincingly shown in a number of works ([1] among others). We have developed Python scripts to generate mistakes statistics visualised in a bar chart for any sample of texts, which helps to detect and address most common errors and underlying weaker translator competences for each group or a translator and to register their progress. Language mistakes most often signal lack of transfer skills when a student produces a literal translation, or they can be blamed on the lack in TL competence when a student is unfamiliar with the language use in a particular sphere. In both cases checking the translator's variants with national corpora helps. Editing these mistakes is aimed, therefore, not only at raising trainees' awareness, but also at the study and use of query syntax of the national and special corpora of the working languages. Content mistakes are usually provoked by lack of background information. It can be remedied by extending the scope of world knowledge through information search and raising awareness of cultural differences. Another common reason behind content errors is text analysis and text comprehension problems. We put special stress on these categories of mistakes which, despite having lower frequency ranks, are most dramatic in terms of harm done to the overall quality of translation. Editing these mistakes involves the study and practice in information-mining technology. We feel particularly bad about logic errors which result in incoherent unreadable target texts with vague or non-existent message, and, consequently, defeat the very purpose of translation. Editing these mistakes often requires the analysis of significant chunks of the source and target texts and builds up translators' textual competence.

5 Conclusion and Future Work

With multiple learner translator corpora remaining a fairly new and scarce resource in translation studies, RusLTC seems a valuable contribution to the scene. In the very least it can be utilised as a significant well-organised and documented source of raw data on translational behaviour, including variation and problem-solving techniques. This corpus is a large bidirectional corpus that provides one-to-many concordances of source/target sentences, containing the query item, aligned with their respective

translation unit variants in the other language. There are a lot of challenging tasks lying ahead, both technological and linguistic, as well as methodological. We are committed to improve the interface to integrate the functions that are now isolated. Translator training demands error-tag query and easy-to-use statistics. It would be practicable to provide the user with an opportunity to create subcorpora using meta data filters. We are looking for technical solutions to produce parallel concordances of error-annotated subcorpus.

As for the corpus application we are working on a more efficient technology to create tailor-made exercises to address individual translator learning issues that could be used as self-study materials. The corpus will also be used for applied research into translation quality assessment and to measure changes in translation quality and strategies over time. RusLTC-driven research can also help to describe "problem areas" in English-Russian and Russian-English translation, which can be accounted for in the curriculum design.

Acknowledgements. The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2014.

References

1. Bernardini, S., Castagnoli, S.: Corpora for translator education and translation practice. In: Rodrigo, E. (ed.) Topics in Language Resources for Translation and Localisation. Benjamins translation library: EST subseries, vol. 79, pp. 39–57. John Benjamins Publishing Company (2008)
2. Bowker, L., Bennison, P.: Student translation archive: design, development and application. In: Zanettin, F., Bernardini, S., Stewart, D. (eds.) Corpora in Translator Education, pp. 103–117. Saint Jerome Publishing (2003)
3. Castagnoli, S.: Variation and regularities in translation: insights from multiple translation corpora. In: UCCTS 2010 - Using Corpora in Contrastive and Translation Studies (2010)
4. Castagnoli, S., Kunz, K., Kübler, N., Volanschi, A.: Designing a learner translator corpus for training purposes (2006)
5. Espunya, A.: Investigating lexical difficulties of learners in the error-annotated upf learner translation corpus. In: Granger, S., Gilquin, G., Meunier, F. (eds.) Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings. Presses Universitaires de Louvain (2013)
6. Florén, C., Sanz, R.: The application of a parallel corpus (english-spanish) to the teaching of translation (entrad project). In: Muñoz Calvo, M., Buesa-Gómez, C., Ruiz-Moneva, M.A. (eds.) New Trends in Translation and Cultural Identity, pp. 433–443. Cambridge Scholars Publishing (2008)
7. Graedler, A.L.: Nest – a corpus in the brooding box. In: Huber, M., Mukherjee, J. (eds.) Corpus Linguistics and Variation in English: Focus on Non-Native Englishes. Studies in Variation, Contacts and Change in English, University of Giessen (2013)
8. Ilyushchenya, T., Kunilovskaya, M.: Inter-rater reliability in student translation evaluation). In: Proceedings of International Conference on Translation Studies Ecology of Translation: Interdisciplinary Research and Perspectives, Tyumen, Russia, pp. 105–115 (2013) (in Russian)

9. Kunilovskaya, M., Morgoun, N.: Gains and pitfalls of sentence-splitting in translation. In: Perm National Research Polytechnic University Herald, pp. 152–166. Linguistic and Pedagogy, Perm National Research Polytechnic University (2013)
10. Kutuzov, A.: Is there a difference between male and female translations (based on the rusltc data). In: Proceedings of International Conference on Translatology, Problems of Translation and Methods of Teaching Translation, vol. 1, pp. 97–104. Nizhny Novgorod, Russia (2012) (in Russian)
11. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul (2012)
12. Sosnina, E.: Russian translation learner corpus: The first insights. In: The Proceedings of the 6 International Scientific Conference Interactive Systems: Problems of Human-computer Interaction (2005)
13. Spence, R.: A corpus of student l1-l2 translations. In: Granger, S., Hung, J. (eds.) Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 110–112. The Chinese University of Hong Kong (1998)
14. Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: EACL, pp. 102–107 (2012)
15. Uzar, R., Waliski, J.: Analysing the fluency of translators. International Journal of Corpus Linguistics 6(1), 155–166 (2001-12-01T00:00:00)
16. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Recent Advances in Natural Language Processing (RANLP 2005), pp. 590–596 (2005)