M. Kunilovskaya, A. Kutuzov

A QUANTITATIVE STUDY OF TRANSLATIONAL RUSSIAN (BASED ON A TRANSLATIONAL LEARNER CORPUS)

Abstract. This research sets out to investigate the linguistic features of student translations from English into Russian against comparable non-translated Russian texts. The study is based on the translation universals hypothesis, a long-standing approach in corpus-based translation studies. Comparing large-scale frequencies of basic textual parameters helps to pin down general tendencies in student translations overriding the level of individual deviations revealed by error analysis. Our results show that translational Russian contained in RusLTC corpus differs from native speakers texts mostly in sentence length and lemma frequencies, while there is little variation between translations of different quality.

Keywords. Translational learner corpora, translation universals, Russian Learner Translator Corpus, translationese.

1. Introduction and setting the task

Translational learner corpora are mostly used in translation quality assessment research and in didactics of translation. There are but few attempts to approach learner translator production as a manifestation of some interlanguage, which is distinct in its linguistic qualities from non-translated *target language* (TL). To the best of our knowledge there is no corpora-based research either for learner translator Russian or for translational Russian in general.

It seems reasonable to assume that a translational learner corpus should contain TL which reflects the influence of several factors.

On the one hand all translations are expected to display textual features associated with the so-called translation universals or translationese, i.e. bear observable linguistic characteristics that are typical for translated texts and are explained by the nature of the

translation process itself. On the other hand it represents the output by people with limited knowledge and experience in translation.

Taking into account the nature of our data we put forward two research questions: 1) what are the quantitative features of learner translator language contained in the corpus and 2) is there a correlation between the quality of translations and their linguistic features measurable by corpus-linguistics methods?

2. Translation universals and their linguistic indicators

The hypothesis of translation universals is a fruitful corpus approach in descriptive translation studies, aimed at discovering and explaining the long-known linguistic difference between translations and TL original texts.

The early approaches to represent and explain regular linguistic features of translated texts as opposed to non-translations are centred on the idea of interference of the source language (SL), viz. «translationese» [Gellerstam 1986], and Baker's idea of translation universals as «features which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems» [Baker 1993: 243]. The list of universals usually includes about a dozen of linguistic features and a number of indicators used to validate them. The most established universals, repeatedly referred to in many overviews are 1) simplification, 2) normalisation, 3) explicitation, 4) interference, 5) convergence.

Arguably, SL interference is one of the most salient features of the learner translator language. Literal renditions, being individual, cumulatively, have a bearing on the general lexical profile of translational language. These features can be indicated by the differences in lexical frequencies of unigrams and word clusters, lists of collocations for most common vocabulary units.

On the grammatical level interference is manifested, for instance, in the overuse of predicative modality and passives and non-typical word order patterns in English ST Russian translations.

We hope that some associated irregularities can be represented by comparing the distributions of part-of-speech (PoS) tags and PoS n-grams.

Simplification is another generalisation about translationese. This assumption is usually tested through a range of quantitative indicators, initially suggested in [Laviosa 1998], that include lexical variety, calculated as a standardized type to token ratio and lexical density, represented as the ratio of content words to total text length. Simplification is also related to frequencies of basic units and can be studied by comparing the range and the proportion of most frequent words from the list-head, which account for at least 0,1% of the corpus. This measure is expected to show that the nucleus of the words most frequently used in the corpus is less varied, while the proportion of frequent versus less frequent words is higher. The syntactic manifestation of simplification can allegedly be measured by mean sentence length.

3 Research design and corpora used

The research is based on a subcorpus of the Russian Learner Translator Corpus (RusLTC[1]) [Kutuzov and Kunilovskaya 2014], which includes 1 257 translations of 187 English source texts that come from electronic mass media and represent informational and analytical genres of newspaper register. The corpus is multiple, i.e. in most cases it contains more than one translation for the same ST. To make sure that our results are not affected by the multiple character of the data we experiment with three different selections containing just one random translation of each source. The overall size of the translational subcorpus is 559 134 tokens.

The comparable native corpus consists of 8 492 texts extracted from Russian national corpus[2] by “publicist” and “article” metadata tags for the functional style and genre respectively. To

1 http://dev.rus-ltc.org/search

2 http://ruscorpora.ru

make this corpus even more comparable to RusLTC, we pruned the texts to the first 350 words each. The resulting native subcorpus counts 2 971 159 running words.

As one of the motivations for this research is to investigate the relations between quantitative parameters of translationese and the perceived quality of translations with regard to applicability of this approach in translation quality assessment we have classified student translations into four quality bands, based on the grade awarded by a human assessor.

4. Comparing the basics

In this section we present the results following the suggestions put forward above. All the differences are found to be statistically significant, with t-test *p* value close to 0, and all calculations involve lemmatised data, unless specified otherwise.

One of the frequently used measures to detect translationese is the variation in word clusters and frequency lists. To reliably demonstrate the difference between translationese and native language we computed Spearman's rank correlation coefficient (Spearman's rho) for the **word bigrams frequency ranks** produced on the arbitrary bisection of the native corpus (0,574) and compared it to the similar statistic for the translational corpus against the native corpus (0,417). The results show that there is a difference between native and translational languages based on this parameter.

We have compared word bigrams frequency for subcorpora, representing different quality bands, to the native data to find no connection between the perceived quality and frequency lists. The same is true for **lemma frequencies**: for any of the bands against the native corpus Spearman's rho is around 0,574. At the same time, there is a difference between translations and original texts in this respect. Spearman's rho for the whole translational corpus against the native one is 0,642, while calculations on the bisection of the native corpus give Spearman's rho value of 0,767. Thus,

translated corpus is more distant from native texts than they are from each other.

There is a statistically significant difference in the average counts of **type to token ratio** (TTR) between our translational and native data (0,61 and 0,65 respectively), which supports the simplification assumption that translations are lexically simpler and rely on a smaller vocabulary. The distance between the native and the translated texts is 4 times more than for the native corpus bisections and should be caused by translation itself. We did not find any significant difference between the quality cross-sections against each other, particularly between the polar 'good' and 'poor' translations ($p=0,597$), but we noticed significant difference in TTR between the collections of unique translations against each other and against the sections with multiple target texts ($p<0,05$). It means that the multiple nature of the corpus can be affecting these counts.

The **lexical density** parameter (calculated here as a ratio of content words to total text length) proved to be irrelevant for our data. We found that though there is a slight, but significant, difference between translated and non-translated corpora (0,776 and 0,781 correspondingly) which points towards less density in translations, but almost the same difference exists between halves of our original Russian corpus (0,782 and 0,778). This variation can be attributed to the domain inconsistency of our data.

Another measure that is used to characterise the lexical set up of the corpus deals with the **proportion and range of most frequent words**. We adopted the approach introduced in [Laviosa 1998], i.e. we looked into how many words form the list-head in translations and originals by covering individually more than 0,1% of the text. In contrast to expectations translations include 115 items into this list, while originals do with only 93. These findings do not support the simplification hypothesis which suggests that translations tend to rely on a smaller and simpler vocabulary. An important finding is that it is notably the more so, the higher the

perceived quality of the translation (the best translations include 135 words into the list-head, the average-quality ones – 124, the worst rely on 121 words).

For our material average **sentence length** across native texts was found to be 15,67 words, and 17,09 words across translations. To prove that this difference is in fact caused by translation, we again compared sentence lengths in the bisection of our native corpus. Average sentence lengths were 15,1 and 16,2, both considerably less than in the translated texts. Thus, translated sentences in our corpus are generally longer, which rejects the simplification assumption, but, probably, is in favour of explicitation: translators convey meaning in more words, thus making it more explicit. A more plausible explanation, however, is interference, which would suggest that translators do not resort to splitting, ignoring cross-linguistic differences. Validating this assumption requires cross-linguistic research on parsed data that we plan for the future.

Though our data is not syntactically parsed, we extracted information about differences in sentence structures through analysis of part-of-speech (PoS) and PoS n-grams distributions.

General **PoS distribution** across native texts and translations is identical, with one exception: in the translated texts, adjectival pronouns (*который*, *твой*) are more frequent than adverbs, while it is the other way round in the native texts. We believe the cause of this is the excessive usage of *который*, perhaps supporting the explicitation hypothesis. This is a characteristic feature of translated Russian texts, not an occasional fluctuation, as comparison of two randomly split parts of our native corpus has shown no differences. Within **PoS bigrams distribution**, not much variation can be found. Top six PoS bigrams, together accounting for 40–42% of all running bigrams in the texts, are identical in their ranking.

5. Difference in discourse markers

Moving further, we compared the frequency of discourse markers in the native corpus and in the translated one, using a list

of 50 discourse markers, most frequently mentioned in grammars as playing a role in organising the logical structure of discourse.

We collected counts of all discourse markers for every text in the corpora and normalized these counts to text length. The list below enumerates the items most heavily overused in translations, i.e. their average normalised frequency is more than 2,5 times higher in translational corpus than in native Russian:

1. *так как*
2. *так же*
3. *таким образом*
4. *в свою очередь*
5. *вследствие*
6. *итак*
7. *другими словами*

On the other hand, we discovered a set of discourse markers that are underused in translations, given that their frequency in the native texts is at least 3 times higher than in translation:

1. *во-вторых*
2. *кстати*
3. *наконец*
4. *впрочем*
5. *прежде всего*
6. *тем более*
7. *в общем*
8. *во-первых*
9. *стало быть*
10. *между прочим*
11. *в сущности*
12. *вдобавок*

6. Conclusions

In this research we have tested a number of linguistic indicators of translationese on learner translations from English into Russian. The investigation of the unigrams and bigrams frequency lists built on lemmatised corpora proved to be predictive of translationese, but irrelevant as to the quality of translations. Lexical variety calculations on our data supports the simplification hypothesis, while lexical density does not. We have pointed out that the former is a poor indicator due to the multiple character of the data, too. None of the two, naturally, correlates with the translation quality. While looking at the range of the so-called list-head, i.e. the very top of the word frequency list, which includes items with the individual text coverage of 0,1%, we have found results that run coun-

ter the expected: translations demonstrate more varied vocabulary, the more so, the higher the quality.

Furthermore, we have calculated mean sentence length and PoS and PoS bigrams distributions to find out, among other things, that sentences in translations tend to be longer, not the least, perhaps, due to the overuse of the adjectival pronouns such as *который*, which introduce attributive clauses. Both of the last two results require further research and can only tentatively support the explicitation hypothesis, as well as the general counts on the discourse markers studied.

References

1. *Baker M.* (1993), Corpus Linguistics and Translation Studies: Implications and Applications. [M. Baker, G. Francis & E. Tognini-Bonelli (eds.), Text and technology. In honour of John Sinclair]. Benjamins, pp. 233–250.

2. *Gellerstam M.* (1986), Translationese in Swedish novels translated from English. [Lars Wollin & Hans Lindquist, Translation Studies in Scandinavia]. CWK Gleerup, Lund, pp. 88–95.

3. *Kutuzov A., Kunilovskaya M.* (2014), Russian Learner Translator Corpus. [P. Sojka et al. (Eds.): TSD 2014, LNAI 8655]. Springer International Publishing Switzerland, pp. 315–323.

4. *Laviosa S.* (1998), Core patterns of lexical use in a comparable corpus of English narrative prose. [Meta], 43(4), pp. 557–570.

Maria Kunilovskaya
Tyumen State University (Russia)
E-mail: m.a.kunilovskaya@utmn.ru
Andrey Kutuzov
National Research University Higher School of Economics (Russia)
E-mail: akutuzov@hse.ru