LEARNER VS. PROFESSIONAL TRANSLATIONS INTO RUSSIAN: LEXICAL PROFILES

Alexey Pariy
alexpariy@gmail.com
Maria Kunilovskaya
m.a.kunilovskaya@utmn.ru
Tyumen State University, Tyumen, Russian Federation

Abstract

One of the relatively recent trends in learner corpora research is building and exploiting learner translator corpora. Within corpus-based translation studies translations in general are approached as a special variety of the target language, while learner translations represent an even more specific dialect of the said variety. They are believed to demonstrate heavier translationese features due to the assumed lack of professional skill and comparatively poor source language competence. However, this claim remains widely unsupported within corpus-based translation studies, while typical linguistic features of learner translations as opposed to professional ones are only tentatively described. The aim of this research is to contrast learner and professional corpora of Russian translations of English mass media texts to the reference target language corpus to reveal the lexical differences between the three. We find that learner translations consistently show more distance from non-translations than their professional counterparts, while the two undoubtedly form a special type of discourse which is linguistically different from naturally occurring language. These findings might help define (un)professionalism in translation in linguistic rather than social terms and shed light on its relevance to the so called translation universals, as well as be informative in translator education.

Аннотация

Давно замечено, что переводы по своим лингвистическим характеристикам отличаются от текстов на том же языке, созданных вне ситуации межъязыкового посредничества. Выявление и описание этих различий – одна из задач корпусного переводоведения. В рамках этого направления исследований интерес представляет не традиционное соотношение оригинала и перевода, а лингвистические свойства переводов как совокупности текстов, в целом составляющих особую разновидность (диалект) языка перевода – переводной дискурс. При этом по аналогии с обширными исследованиями в области учебных текстов, учебные переводы можно считать особой формой переводного дискурса, обладающей собственными особенностями. Теоретически учебные переводы должны демонстрировать гораздо более явные черты так называемого языка переводов (или переводного дискурса) чем профессиональные ввиду недостаточной сформированности профессионального мышления, ограниченности текстового и жизненного опыта студентов. Вместе с тем эти интуитивно ощущаемые особенности студенческих переводов объективно не описаны и не получили пока эмпирического подтверждения. Непонятно, в чем конкретно они состоят, действительно ли студенческие переводы представляют собой более явную форму переводного дискурса, статистически противопоставленную профессиональным переводам. Целью данной работы является выявление и сравнение лексических особенностей учебного и профессионального переводных дискурсов, ограниченных переводами материалов СМИ, на основе их сопоставления со «стандартным» непереводным русским языком – корпусом текстов из российских СМИ, изначально написанных на русском языке. Результаты нашего исследования подтверждают объективность существования переводного дискурса как такового – обе разновидности переводного дискурса по ряду показателей противопоставлены непереводным текстам, при этом студенческие переводы демонстрируют более резкие отличия от не-переводов, чем их профессиональные аналоги. Эти данные могут быть использованы для выработки объективно-статистических метрик, позволяющих автоматически выявлять переводные тексты, классифицировать переводы по степени их

соответствия характеристикам непереводного русского текста определенного жанра, а также для повышения осведомленности студентов о закономерностях, свойственных их переводам в отличие от профессиональных и непереводных текстов.

Keywords: learner translator corpus, professional translation, translation universals, translationese, corpus-based translation studies

Aims, motivation and key concepts

There is hardly a textbook in translation that does not discuss translation quality. Whatever the evaluation criteria, it seems obvious that novice translators perform worse than seasoned professionals and their texts should reflect the difference. The prescriptive and evaluative bias in translator education is to some extent offset by the descriptive approach which posits that all translations, regardless of translator's professionalism, appear to differ from non-translations in their linguistic characteristics, even if they are not easily detected by the naked eye. In this paper we employ corpus linguistics methods to find out whether and how much students' output differs from published and socially accepted translations in linguistic terms and whether these two sociolinguistic varieties can be grouped together as representing different degrees of translationese in comparison with non-translations.

Throughout this research 'translationese' is used as a general non-evaluative term to refer to the quantitative linguistic features of translations which set them apart from non-translations in the same language. This interpretation is in line with the tradition in European corpus-based translation studies and metonymically, with the notion of 'translated discourse' offered by Garbovskyi (Garbovskyi 2012) at home. Yet, Russian translated discourse lacks attention of corpus linguists (but see Krasnopeeva, 2015) and remains largely undescribed, and therefore, attractive.

'Professional translation' is another key concept here. For the purposes of this research professional translations are those that have been published under translator's own name by an official trustworthy information agency or endorsed by the editorial board. We assume that the translators hold down a paid job with the respective agencies. Learner translators are also defined socially rather than in terms of linguistic or professional competence – they are full time students enrolled in the Masters in Linguistics programs with the translation and translation studies focus. Professional output of translational Russian has been an object of corpus-based scrutiny in the recent dissertation by Krasnopeeva (2015), who found statistically significant lexical dissimilarities between original Russian belles-lettres prose and its comparable translated counterpart.

Our comparative and contrastive analysis is limited to some lexical parameters of the corpus data, i.e. we mostly use frequency statistics of lexical features to accomplish our task. In that we rely on the corpus methodology suggested to reveal translationese or learner language characteristics, which distinguish these varieties from standard language produced by native speakers in monolingual communication.

To this end we use a complex corpus architecture, which includes subcorpora of learner and professional translations in Russian with respective source texts in English and a reference corpus of non-translated Russian texts from Russian National Corpus (RNC). All texts included in our data are newspaper articles in a variety of topical domains published in public media over the last decade (naturally, except learner translations).

In what follows we define our general theoretical viewpoint as regards translationese and review the various linguistic indicators introduced in previous research to profile translations along with their computational implementation. The next section has the description of our corpus data and research set-up. In the final part of the paper we report and discuss our results.

Related work: translationese and learner translator corpora

This research is set within the framework of descriptive translation studies and is inspired by the theory of translation universals, which has been central to corpus-based investigations into the linguistic properties of translations. This strand of research is rooted in the idea of “translationese”

introduced by Gellerstam (1986) to pin down the dissimilarity of translations to native discourse in terms of statistical large-scale frequencies. In the tradition set by him and maintained within the descriptive translation studies this term lost its negative connotations, but is used in a neutral, non-pejorative way to characterize translated language as deviating from naturally occurring one. The quantitative nature of lexical differences has been confirmed in the ground-breaking research by Baroni and Bernardini (2006). They have shown that humans are outperformed by machines in their ability to tell translations from non-translated language (Baroni, Bernardini 2006). These findings, on the one hand, stress the objective nature of translationese and, at the same time, underline the unreliability of human assessment. Thus, translationese is not a traditional error insofar as it is not located in a specific part of the text but is manifested cumulatively; it is distributed in the text making it different from non-translations in quantitatively measurable features. The translation universals theory attempts to see specific tendencies in translator linguistic behavior which explain the quantitative differences and which are known as 'universals'. The researchers bring on board qualitative analysis and complex combinations of correlated comparable and parallel corpora to show that raw frequencies can mask opposite trends in translation (Hansen-Schirra 2011) and that some trends might be less universal, i.e. more pronounced in a specific language pair (Mauranen 2004).

Basically, research into universals applies and develops the idea of lexical profiling from Crystal (1991), who defined it as "the identification of the most salient features" of a personal or register-specific discourse. This approach seems useful for our two-fold purposes: 1) it helps to provide a general linguistic profile of translations against non-translations and 2) define learner translations as opposed to professional ones. It stands to reason that all translations are expected to display the textual features associated with translationese (or translation universals) attributed to the nature of the translation process itself. At the same time, learner translations hypothetically represent a special flavor of translationese. They are produced by people of another social standing - with limited knowledge and experience in translation and in text production in the target language. We hypothesize that this is a confounding factor that should tell on the measurable textual properties of our data. Probably we can expect a gradient in the features contrasting translations and non-translations, which would make learner output "more pronounced translationese" than professional product. This idea was tried out in Federica Scarpa (2006), but our own previous study aimed to describe the quality of learner translations on the basis of lexical parameters attempted in Kunilovskaya, Kutuzov (2015) yielded only modest results due to the scarcity and unreliability of translational data arranged in the expert-defined quality bands. Lexical profiling of translated discourse against non-translations is done in Garside, Rayson (2000) and Hansen-Schirra et al (2013) among others.

Generally, translational learner corpora research is a fairly new method in corpus-based translation studies. The first learner translator corpora stem back to the early 2000s (e.g. Spence 1998; Bowker 2003 and Wurm 2013, Graedler 2013 for most recent additions to the field). They are multiple parallel corpora based on many translations produced by a group of translator trainees for one source text under various conditions documented in metadata and usually featuring some translation error annotation. These corpora are mostly used in translation quality assessment research to define more feasible error typologies and scoring methods, including in MT (e.g. see Vela et al. 2014), and in didactics of translation to inform teachers on problem areas and progress of a particular student population and provide a helpful source of data for teaching material design. There are but few attempts to approach learner translator production as a manifestation of some dialect of translational interlanguage or "third code", which is distinct in its linguistic qualities from both professional translations and non-translated language.

Datasets and methodology

For this study we have compiled a genre-specific research corpus which consists of three major components:

1) The **subcorpus of multiple learner translations** (612839 tokens, 1441 texts) consists of

translations produced by senior students majoring in translation studies at several Russian universities either as part of their routine training exercises or independent test translations as well as in the setting of several student translation contests. The texts for this subcorpus were extracted from Russian Learner Translator Corpus (RusLTC)[1] to fit the selected genre parameter: all of them are Russian translations of newspaper or magazine articles published in well-established American or English mass media. This subcorpus is also truly parallel - all translations are aligned at sentence level with their sources. While the multiple nature of the corpus can be useful for some experiments (see e.g. Castagnoli 2009), it is certainly a factor to be considered when producing keyword lists and assessing frequency distributions of lexical items as well as when calculating type-to-token ratio (TTR). Therefore, from this data we have isolated **a unique student translation collection** (147523 tokens, 194 text fragments), which consists of one random translation to each original.

2) The comparable **component of professional translations** (144618 tokens, 141 texts) includes translations published by 10 central Russian news portals such as *ng.ru, globalaffairs.ru, inosmi.ru, polit.ru, www.forbes.ru,* most of which carry links to originals, and we used them to compile a corpus of sources to most of these translations (144618 tokens, 100 texts).
3) To build our **reference subcorpus** we filtered Russian National Corpus (RNC) by the metadata to include only texts described as 'publicistic style', 'article', published after 2003, aimed at general adult audience, each counting over 400 words. The filter gave us a corpus of 1598 texts, which in total amount to 2 691 142 tokens.

The corpora are represented in one-sentence-per-line format: sentence segmentation was done with *Punkt* (Kiss & Strunk 2006) from NLTK (which has a pre-trained English model) and a model trained for Russian on all of RNC (ca. 150 mln tokens). All corpora are stripped of sentences which contain less than 5 words (mostly headlines, publication dates and names of authors taking a separate line). We also made sure that there are only few sentence splitting mistakes by screening the corpora for odd sentences that start with a lowercase letter.

All our experiments are based on lemmatized data for both languages. For linguistic annotation we used *TreeTagger* with the supplied English model (Schmid 1995) and the Russian model trained by Sharoff and based on MULTEXT-East tagset (Sharoff et al. 2008); lemmatization in the Russian components was optimized with the lemma-prediction tool CSTlemma by Bart Jongejan.

Methodologically, this research is based on three types of comparison used to characterize translation discourse and mentioned in Chesterman (2010): 1) translations against non-translated naturally occurring discourse; 2) translations against sources; 3) translations against translations. Where direct comparisons are not enlightening we use the reference corpus as third comparationis, the standard by which we measure the two types of translationese.

Experiments and results

Sentence length

As it is shown above we have filtered out pseudo-sentences such as titles and dates, which could upset the counts, given the considerable difference in the average text size and number of texts in the subcopora. Unlike text sizes, the values for sentence length in all three sets fit the normal probability distribution, which means that we can use *t-test* to statistically assess the difference between them. Our calculations demonstrate that learners tend to produce longer sentences than non-translators: the difference in sentence length between learner translations and non-translations is statistically significant ($p = 0$). The same is true for professional translations, though the probability of mistake according to t-test is much higher: $p = 0.01248$. At the same time there is no evidence that learner translations differ from professional ones in this respect ($p = 0.3775$). Thus, **learner and professional translations do group together as linguistically different from non-**

[1] http://www.rus-ltc.org/

translations, but we are unable to make reliable conclusions about the relationships between the two according to this parameter (except the level of confidence for Learner-RNC difference is much higher ($p = 0.000000298$) than for Professional-RNC difference ($p = 0.01248$)).

The figures for sentence-length in cross-linguistic perspective suggest that longer sentences in both social dialects of translationese can be due to explicitation (the usual suspect) and interference. Sentence length in our two source text collections is significantly different - and judging by the average higher - than in translations. This inference is confirmed by the statistically significant contrast between English sources and non-translated Russian (on average 21.5384 and 17.9445 words per sentence respectively, $p = 0$), which is easily explained typologically.

Given that we cannot compare sentence lengths pairwise (our data is not aligned), we investigated the number of sentences per text from the cross-linguistic perspective. To determine the difference between the two dependent variables (number of sentences per text in translations and in sources) we employed *Wilcoxon's matched pairs test*, which is recommended for non-parametric data. The results indicate that **professional translations in our data are usually 3 sentences longer than respective sources**, i.e. there is a statistically significant median increase in the number of sentences in professional translations as compared to their sources ($W = 755.5$, $z = -4.3733$, $p = 0$), **while learner translations do not feature this difference** ($W = 4589$, $z = -0.7197$, $p = 0.4715$). On this basis we can tentatively suggest that professionals are more open to structural changes, while students tend to avoid deviations from sentence-to-sentence correspondence.

Basic linguistic indicators of lexical translationese

Corpus sampling

Most lexical profile descriptors used in this study are known to be vulnerable to the corpus or/and text size, therefore we sampled our data to produce comparable chunks, which represent each subcorpus. These chunks include random 100 texts from respective component each, and the texts included are pruned to only the first 400 words (this is about the median for learner texts). The sampling procedure gave us five comparable collections, ca. 40 000 tokens each. For languages with developed morphology (such as Russian) the counts are only reliable if based on lemmatized data. Besides, all counts for learner translations are based on the non-multiple section of the subcorpus, unless specified otherwise.

Lexical variety

One of the standard ways to profile a corpus is to measure lexical variety. In translation studies this indicator has been consistently used (since Laviosa 1998) to demonstrate that translations are lexically simpler than comparable texts in the same target language. This parameter can be gauged as type-to-token ratio (TTR), which is basically the number of different words (*types*) over the total number of words (*tokens*) in a text. According to Hansen-Schirra et al (2013) TTR "can be taken as an indicator of semantic precision and information load density, with indirect consequences for explicitness". Statistical implementation is based on the mean value for texts in each chunk. The relevant results are reported in Table 1.

Table 1. Average TTR counts in the sampled Russian components and p-value for t-test

	Learner	**RNC**	**Professional**
Sample size (tokens, python-style)	40000	40000	40000
TTR	0.54215	0.563425	0.560225
Standard deviation	0.04196	0.04703	0.05428
p-value	0.002325		
p-value		0.65805	
p-value	Learner vs professional: 0.004788		

As can be seen from Table 1, **learner translations have the lowest lexical diversity** and, if our calculations are to be trusted, they are significantly different from professional texts and non-

translated ones. As regards professional translations, we do not have enough evidence to say whether they are any different from RNC.

Another method to implement lexical variety computationally is to measure the proportion of most frequent words in each corpus and their range. Within this approach lexical variety can be operationalized, firstly, as the ratio of the so-called *list head* to full list frequency. Secondly, corpora can be compared on the number of words that make up the list head. According to Laviosa (1998), who introduced this operator, 'list head' is the top of the frequency list which includes items that individually cover at least 0,1% of the corpus. It means that the first item not included in this list is the first item down the full frequency list with the frequency less than 'corpus size/1000'. Lexical variety in this case is represented as ratio of the sum of list head items frequencies to the corpus size. If this ratio is comparatively high (e.g. over 50%) the language of the corpus is characterized as 'repetitive'; the speaker makes excessive use of the same vocabulary or in case with translations of standard typical ways of expression to the detriment of lexical scope.

The values for both measures related to list head are given in Table 2. To compare the cumulative frequencies we used *Pearson's chi-square test*, which determines whether the distribution of the frequencies observed in the translational corpora deviate significantly from that in the reference corpus (or the expected value), following the statistical approach suggested in Gries (2010).

Table 2. Size of list heads, proportion of high frequency items and the chi-square contingency table

	Learner	RNC	Professional
Corpus size (AntConc-style, inc. punctuation, numbers and other non-alphabetic symbols)	39897 6665	39901 8244	39901 7654
Number of lemmas in the list head	109	89	99
Cumulative frequency and proportion of list head	19090 (47.85%)	16765 (41.91%)	18373 (46.05%)
Cumulative frequency of low frequency words	20807	23136	21528
chi-square, p-value	274.2014, $p < 0.05$		
		131.4774, $p < 0.05$	
	learners vs professionals: 26.0016, $p < 0.05$		

These calculations do not support the findings based on translational English (Laviosa 1998) and translational Chinese (Xiao 2010) and do not fully confirm that the translation component has "a higher proportion of high frequency words and its list head covers a greater percentage of text with fewer lemmas than the non-translational component." If anything, they suggest the opposite, at least for the first part of the statement. It is still true that translations demonstrate significantly higher repetitiveness, and judging by the counts it is even more so for student translations. Note also that the distance between translated discourse and non-translated one is greater than between the two dialects of translationese. For now we have to refrain from any interpretation of this matter, except typological, but will definitely look into it in the future.

Lexical density

Another well-known lexical parameter is lexical density. It measures the ratio of content words to grammatical words and is sometimes termed the measure of information load, i.e. it shows how content-rich a text is, based on the assumption that it is lexical words that convey information. We implemented the counts using NLTK lists of functional words (stoplists) for both languages.

As can be seen from our results both translational subcorpora are significantly different at $p<0.01$ from non-translational one and feature lower density: content words account for 74.71% and 74.75% of the overall size of learner and professional subcorpora respectively, while non-translations have 76.65% of them. Overall cross-linguistic comparisons based on means do not

make sense due to the typological differences: English will show significantly lower density, except one can do correlation analysis to establish whether the feature of translations reflect the sources in this respect. Surprisingly, sources to learner translations differ significantly from sources to professional translations (0.6085 vs 0.6241, $p<0.01$), which is an interesting result in terms of educational strategies and curriculum.

Frequency distribution of word classes

Granger and Rayson (1998) suggest that "one way of characterizing a language variety is by drawing up a word category" and show that non-native speech differs from standard English in distributional characteristics of some categories such as determiners, pronouns and adverbs.

Following the methodology consistently developed by Rayson (see e.g. Rayson et al 2008), we looked into keyness statistics for individual grammar forms and their sets, measured by log likelihood. We proceeded from the top 50 items that were thrown by Antconc log likelihood calculator as key for translationese against the reference corpus. Table 3 lists items from top 50 according to AntConc and has alternative LL values that come from Rayson's LL calculator[2]. All frequencies are normalized per 10 000. Below we mostly quote the statistics for the highest scoring item from the group, which comes first.

Table 3. Statistics for forms most key to translations in comparison with non-translations

	Frequency, keyness and overuse		Frequencies in the reference corpus	Examples
	Learner stranslations	Professional translations		
Corpus size (in PoS tags)	657502	157740	2936436	
Adpositions (esp. locatives) **Sp-l** Sp-g, Sp-a, sp-d, sp-i	27084 107365.485 107.28 (-)	6644 43836.533 13.81 (-)	129559	на, в
Finite (personal) verbal forms, esp. present tense **Vmip3s-a-e** Vmip3p-a-e, vmis-sfa-p, vmps-p-psp	11355 45013.111 734.02 (+)	2649 17477.871 167.23 (+)	37637	действует, ведет, знает, обещает; смогла, заплатила,
Infinitives **Vmn----a-p** Vmn----a-e, vmn----m-e	8924 35376.222 1294.78 (+)	1956 12905.518 224.67 (+)	25134	задобрить, обеспечить ослабить
Passive Verbs: **Vmip3p-m-e**, inc. analytical forms: **vmps-p-psp**	2673 10596.217 156.49(+) 1163 4610.326 10.52 (-)	454 2995.452 1.72 (-) 185 1220.614 55.88 (-)	8996 5758	скупаются, добываются, снижаются; (быть) освобождены, объяснены
Adverbial and nominative pronouns (inc. neuter) **P-----r** **P--nsnn** P---pna	6226 24680.901 141.53 (+) 4242	1480 9764.911 35.42 (+)	23399 11942	Где, так, сейчас всегда, тогда; Это, то, что; Который, любой, самый

[2] http://ucrel.lancs.ac.uk/llwizard.html

P--msna (22 forms)	16815.994 616.57 (+)	1053 6947.602 207.39 (+)		
Particles q	18537 4499.877 891.56 (+)	4037 809.165 90.85 (+)	64150	не, ли, же, ведь
Discourse markers (dm)	3308 543.311 49.96 (+)	629 33.352 5.30 (-)	12847	на самом деле, по крайней мере, однако

The results indicate that students, unlike professionals, tend to overuse synthetic passives and, notably, discourse markers. In both translational varieties of Russian there is significant overuse of many types of pronouns, especially in the nominative case, infinitives and particles. These findings seem consistent with the general typological differences between Russian and English and known tendencies in translational behavior. However, within the scopes of this work we do not attempt explanations (cross-linguistic or translation studies-based) and save this for the future.

Conclusion

Generally, this research shows that while translated discourse in our data as a whole is indeed opposed to naturally occurring language. Statistically significant differences are observed in terms of basic lexical and surface structural features such as sentence length, lexical variety and lexical density

Students produce less natural longer sentences, being unable to come up with compact succinct reformulations, while professionals show more structural flexibility. Students use less varied vocabulary and rely on less content words than professionals. Lower lexical variety and density are well-known features of translationese established for many language pairs, and our findings are in line with it. Thus, we can conclude, that student and professional translations are indeed two different dialects of translationese. They bear common features of translated discourse without being indistinguishable. Learner texts and non-translated discourse form extremes of the continuum, while professional translations come in the middle.

References

1. Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274.
2. Bowker, L., & Bennison, P. (2002). Translation Tracking System: A tool for managing translation archives. *Proceedings of LREC*. 503–507.
3. Castagnoli, S. (2009). Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation. PhD Dissertation, University of Pisa.
4. Chesterman, A. (2010). Why Study Translation Universals? *1*. 38–48
5. Crystal, D. (1992). Profiling linguistic disability.
6. Dai, G., & Xiao, R. (2011). SL "shining through" in translational language: a corpus-based study of chinese translation of english passives. *Translation Quarterly* 62. 85–108.
7. Garbovskyi N.K. (2012).Russian translated discourse: myth or reality. *Proceedings of the III International research conference Language and culture in the mirror of translation*. 130–136.
8. Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation Studies in Scandinavia*.
9. Graedler, A., (2010). NEST – a corpus in the brooding box http://www.helsinki.fi/varieng/series/volumes/13/graedler/ Accessed March 10, 2016.
10. Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. *Learner English on computer*, 119-131.

11. Gries, S. T. (2010). Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: Selected approaches*, *66*, 269-291.
12. Hansen-Schirra S., Neumann S., Steiner E. (2013). Cross-linguistic corpora for the study of translations: insights from the language pair English-German. Walter de Gruyter. T. 11.
13. Hansen-Schirra, S. (2011). Between normalization and shining-through. specific properties of English-German translations and their influence on the target language. *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*. 133–162.
14. Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32(4). 485–525.
15. Krasnopeeva, E.S. (2015). Lexical features of Russian translated discourse (a corpus-based comparative study of contemporary narrative prose). *Dissertation,* Chelyabinsk.
16. Kunilovskaya, M., & Kutuzov, A. (2015). A quantitative study of translational Russian (based on a translational learner corpus). In *Proceedings of corpus linguistics 2015 conference*, Saint Petersburg State University. 33–40.
17. Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta: Journal des traducteursMeta:/Translators' Journal* 43(4). 557–570.
18. Maurenen, A. (2004). Corpora, universals and interference. *Translation universals: Do they exist?* / edited by Anna Mauranen, Pekka Kujamäki. 65-82
19. Rayson, P., & Garside, R. (2000, October). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora* (pp. 1-6). Association for Computational Linguistics.
20. Rayson, P., Xu, X., Xiao, J., Wong, A., & Yuan, Q. (2008). Quantitative analysis of translation revision: contrastive corpus research on native English and Chinese translationese. In *Xviii fit world congress*.
21. Scarpa, F. (2006). Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in English and Italian. *M. GOTTI. Insights into specialized translation–linguistics insights. Bern: Peter Lang*, 155–172.
22. Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, *43*, 28.
23. Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., & Divjak, D. (2008). Designing and Evaluating a Russian Tagset. In *LREC*. 279-285.
24. Spence, R. (1998). A Corpus of Student L1-L2 Translations. *Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 110–112.
25. Vela, M., Schumann, A. K., & Wurm, A. (2014). Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. *EACL 2014*.
26. Wurm, Andrea. (2013). Proper names and specific cultural items in a corpus of student translations (KOPTE) // trans-kom 6 [2]: 381–419 http://fr46.uni-saarland.de/index.php?id=3702 Accessed March 10, 2016.
27. Xiao, R., He, L., & Yue, M. (2010). In pursuit of the third code: using the Zju corpus of translational chinese in translation studies. *Using Corpora in Contrastive and Translation Studies*. 182–214.