Universal Dependencies-based syntactic features in detecting human translation varieties

Maria Kunilovskaya
Institute for the Humanities and Social Sciences
University of Tyumen
Tyumen, Russia
m.a.kunilovskaya@utmn.ru

Andrey Kutuzov
Department of Informatics
University of Oslo
Oslo, Norway
andreku@ifi.uio.no

Abstract

In this paper, syntactic annotation is used to reveal linguistic properties of translations. We employ the Universal Dependencies framework to represent learner and professional translations of English mass-media texts into Russian (along with non-translated Russian texts of the same genre) with the aim to discover and describe syntactic specificity of translations produced at different levels of competence. The search for differences between varieties of translation and the native texts is augmented with the results obtained from a series of machine learning classifications experiments. We show that syntactic structures have considerable predictive power in translationese detection, on par with the known low-level lexical features.

1 Introduction

This research aims to detect distinctive syntactic properties of learner and professional translations from English into Russian when compared to the originally authored texts in Russian. The contrasts between them can provide insights into translation quality and be informative in translator education as well as machine translation design.

It is known from previous studies that translations differ from non-translations at all levels of language hierarchy. These linguistic differences are usually referred to as *translationese* (Gellerstam, 1986), and text production processes behind them are explained within the theory of translation universals (Baker, 1993). Quantitative specificity of translated texts is used in translationese detection and classification. It has been shown that learning systems can achieve high performance on shallow data representations (Baroni and Bernardini, 2006), and character n-grams work best (Popescu, 2011).

However, features useful for machine learning algorithms are often difficult to interpret linguistically. At the same time, it is important to know what gives translations their peculiar foreign sound. This knowledge will promote our ability to counteract it, if we want to produce more natural texts in the target language, as well as our awareness of typical linguistic behavior in the situations of language contact.

The concept of translation quality is inherently connected to the idea of translationese. In the most common case of informational texts, we expect translations to blend well with the rest of genre-comparable texts in that language. *Fluency*, the property of translation to read as natural as a non-translation, is one of the three major criteria of translation evaluation, along with *adequacy* and *fidelity* (Secară, 2005). It means that we can use proximity to the reference non-translations as a measure for this component of translation quality. The question remains whether all machine detected differences between translations and non-translations reflect a reduction in fluency and readability. Therefore, it is useful to test the findings on the basis of some external labels or markers of quality. In our research setup we represent translation quality classes by professional and learner translations assuming that translations produced at different levels of competence differ in terms of quality.

In this work we explore the use of syntactic features as possible indicators of translationese for modeling a classifier able to distinguish between novice and professional translators (or between translations and non-translations). Machine classification is used here as an exploratory technique: after we establish the appropriateness of syntactic representations for the purposes of text classification, we identify the most

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 27–36,
Prague, Czech Republic, January 23–24, 2018. Distributed under a CC-BY 4.0 licence. 2017.

Keywords: human translation varieties, translation classification, syntactic features, UD

informative features and test their validity in contrastive and comparative linguistic analysis of our data. This is one of the reasons why we don't use modern neural methods (like LSTMs, for example) working directly on sequences of words or characters: we need interpretable features in this setup. Thus, we rely on more old-fashioned classifiers like SVM.

We use Universal Dependencies (UD) framework (Nivre et al., 2016) in our syntactic analysis. It is a linguistically-motivated initiative aimed to facilitate multilingual research by offering a universal approach to represent and compare sentence structures. Besides, UD provides a better account for free word order languages such as Russian (Jurafsky and Martin, 2014), and gives direct access to annotated treebanks. Unlike previous work in the field of translationese detection and classification, we make use of truly syntactic properties of sentences as defined in Dependency Grammar, not their PoS n-gram emulations or similar quasi-syntactic approaches.

2 Research questions

We are designing a learning system as a heuristic approach to establish syntactic specificity of translations with the view of using their most distinctive syntactic properties as tools in translation quality assessment. Our research questions can be put as follows:

1. Can translated texts be distinguished from non-translations based on syntactic features, given their UD-based representation described below?

2. Are there machine-learnable syntactic differences between translations produced by learner translators and by professionals?

3. If yes to any of the above questions, which features are most correlated with the text class?

4. How can these features be explained by contrastive analysis and translation universals theory?

We resort to comparative and contrastive analyses to offer linguistic explanation for the experimental findings. To this end, we analyze the distributions of features in the sentences, compare them with the respective source segments and typify the results. In this part we are guided by findings within corpus-based translation studies and contrastive knowledge for the given language pair.

3 Related work

Previous work on translation quality assessment (TQA), translational expertise, translationese detection, translation universals and parsing is abundant. There is research that establishes links between the areas of study above. For example, Aharoni (2015) demonstrates that accuracy of translationese detection depends on the quality of machine translation.

One particularly relevant study on machine classification of translations produced at different levels of expertise is Rubino et al. (2016). To solve the task of distinguishing student and professional translations from each other and from originally authored texts, the authors use four distinct feature sets: traditional surface characteristics of sentences (words with mixed-case characters, sentence length, number of punctuation marks) and three sets inspired by information density theory and machine quality estimation. The research is focused on evaluating feature importance and returns mixed results as to what can be used to predict translation experience. In the binary classification (learners vs professionals) their approach achieves the average F1 score of 58.5%.

The assumption that levels of competence (defined extra-linguistically) and practices used in the process influence the quality of the product are corroborated in Carl and Buch-Kromann (2010), who also show that the differences between learners and professionals lie mostly in text fluency. Lapshinova-Koltunski (2017) finds that differences between translational varieties (represented in the author's research as human and machine translations) with regard to the degree and types of cohesion are smaller than between translations and originally authored texts.

Research in translationese detection increasingly relies on utilizing linguistically reasonable (interpretable) features of text as opposed to 'unreasonably effective' character n-grams (Volansky et al.,

2015). Research of this kind uses delexicalized syntactic features to solve the tasks related to translationese detection (Laippala et al., 2015) and classification (Rubino et al., 2016; Rabinovich et al., 2017). One of the feature sets in Laippala et al. (2015) consists of PoS bigrams and trigrams enriched not only with morphological features, but also with syntactic relations extracted from dependency grammar based syntactic trees. This feature set, however, performs slightly worse than PoS with morphological features.

4 Data, features and experimental setup

4.1 Corpus resources and parsing

Our experiments are based on two aligned parallel corpora that contain learner and professional English-to-Russian translations of mass-media texts in a variety of topical domains and a genre-comparable collection of non-translated data.

1. Learner component was sourced from the *Russian Learner Translator Corpus*[1] (Kutuzov and Kunilovskaya, 2014) via filtering by genre.

2. Professional translations were collected from a range of well-established digital mass media such as *Nezavisimaya Gazeta* and *InoSMI.RU* or Russian editions of global mass media such as *Forbes*. All professional translations either have the translator's name or are endorsed by the editing board. Originals for both translational collections come from roughly the same pool of English and American editions (*The Guardian, the New York Times, the Economist, Popular Mechanics*, etc) and were published between 2001 and 2016.

3. The reference corpus consists of the texts from the *Russian National Corpus*[2] (further RNC) belonging to the '*article intended for large adult non-specialist readership*' type; all texts are written after 2003 and are marked as style-neutral.

The Russian texts were tagged and parsed with the *UDPipe 1.2* model (Straka and Straková, 2017) which we trained on the *SynTagRus* treebank from the Universal Dependencies 2.1 release (Dyachenko P.V., 2015; Droganova and Zeman, 2016). The model achieves UAS 89.96 and LAS 87.42 on the corresponding UD2.1 test set. Sentences shorter than 3 words, with disconnected dependency trees, or containing '*root*' relations only, were filtered out, as well as punctuation and null nodes (in case of ellipsis).

Table 1 presents the statistics of the corpora used. With regard to the average sentence length, the translational corpora are significantly different from the RNC at 0.05 level of confidence, while there is no such difference between learner and professional translations.

	Learners translators		**Professional translators**		**RNC**
	sources	targets	sources	targets	
Size (tokens)	222 911	204 787	345 843	320 198	3 215 242
No. of sentences	10 345	9 899	14 595	14 427	153 691
Sentence length (averaged over texts)	23.56	22.41	24.15	22.67	21.29
No. of texts	200		200		1 562

Table 1: Basic corpora statistics (after preprocessing and parsing)

4.2 Methodology

We represent texts as feature vectors, produced by averaging the feature vectors of individual sentences in the text. The majority of our features are the UD syntactic relations. Values for syntactic relations are represented as their sentence-level probabilities, i.e. the ratio of the number of occurrences of a given

[1] https://www.rus-ltc.org/
[2] http://www.ruscorpora.ru/en

relation in the sentence to the number of occurrences of all other relations in the same sentence, averaged over all sentences in each text in a corpus. Given this approach to normalizing the data, the *root* relation actually contains only the information on the sentence length: there is only one *root* in each sentence and its probability is contingent on the number of other relations in the sentence, which in its turn equals the number of words in this sentence. As our aim is to detect purely syntactic relations useful for translation classification, we excluded *root* from the feature set. Additional features included basic graph statistics for dependency trees.

Here we present the full list of our 45 features:

- 34 UD dependencies
  - normalized to represent sentence-level probabilities of each particular relation;
- 7 features characterizing abstract structural properties of the dependency graph:
  - average out-degree, maximal out-degree, number of communities in the graph (by the Newman's leading eigenvector method), average community size (in nodes), average path length, density and diameter of the graph;
- 4 other tree complexity measures, calculated from the parsed data:
  - mean hierarchical distance (MHD), suggested in Jing and Liu (2015);
  - mean dependency distance (MDD), defined as 'distance between words and their parents, measured in terms of intervening words' (Hudson, 1995);
  - probability of non-projective arcs;
  - average number of non-projective sentences.

Machine learning classifiers were trained to separate non-translations from translated texts as a single class and to distinguish different translation varieties from each other and from non-translations. We attempt classification into learner and professional translations to see whether we can find a way to predict professional expertise based on the features suggested.

After a series of development experiments we chose the SVM multinomial classification algorithm with balanced class weights. It was shown to score high in various NLP tasks, including translationese detection, in a number of publications, starting with the ground-breaking Baroni and Bernardini (2006). Before training, the feature values were standardized to have zero mean and unit variance of 1.

For comparison, we also report results of a simple baseline system similar to syntactic component in Pastor et al. (2008). It uses bags of part-of-speech trigrams ('*SCONJ PROPN VERB*', '*NOUN NOUN ADJ*', etc) as feature vectors for each document, with the values of features being the frequencies of particular trigrams in a given document. Pastor et al. (2008) refer to Nerbonne and Wiersma (2006) motivating their choice of n-gram size. These values were standardized in the same way as the syntactic ones and then fed to the same SVM classifier. Note that this approach produces thousands of features, and thus is considerably more computationally expensive than the one with the syntax features.

5 Results

We calculated macro-F1 score for each classification task using stratified 10-fold cross-validation. The results are presented in Table 2.

The classifiers based on syntactic features perform better (and are trained about 70 times faster) than the PoS-trigrams baseline in all scenarios except when discriminating learner translations from professional ones. In the case of 3-class classification with the full set of features, the two approaches are on par, with the syntactic feature set still outperforming the baseline when only 10 best features are used. Thus, English-to-Russian translations are indeed different from non-translated Russian in their syntactic structures. However, translations produced at different skill levels in addition demonstrate differences in the tier of surface word type sequences. Note also that all our results are considerably higher than those

	Binary classification				3-class
	translations/RNC	*learners/RNC*	*prof/RNC*	*learners/prof*	
			10 best features		
PoS trigrams baseline	0.735	0.738	0.658	**0.791**	0.603
Syntactic features	**0.818**	**0.796**	**0.740**	0.721	**0.635**
			all features		
PoS trigrams baseline	0.820	0.820	0.797	**0.806**	0.707
Syntactic features	**0.866**	**0.841**	**0.871**	0.703	0.707

Table 2: Macro-F1 scores for the classifiers on different feature sets

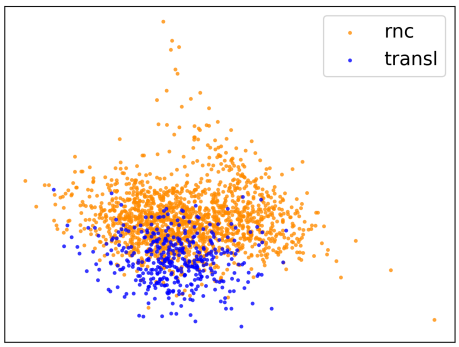

Figure 1: Non-translations (RNC) and translations, syntactic feature space

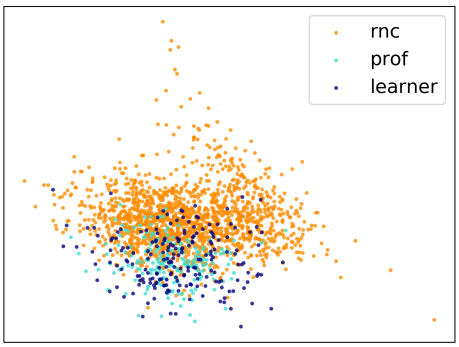

Figure 2: Non-translations (RNC), professional and learner translations, syntactic feature space

reported in Rubino et al. (2016) with a rich set of diverse features including complexity and perplexity in the language models (but with no 'deep' syntactic features)[3].

Figures 1 and 2 visualize the documents in our training data projected from the initial 45-dimensional syntactic feature space into 2 dimensions with PCA (Tipping and Bishop, 1999). For comparison, figures 3 and 4 present similar projections from the PoS-trigrams 2704-dimensional feature space. It can be seen that the texts represented with PoS are much less discernible with regard to our classes: all documents are densely grouped together, with little difference between instances of different types. At the same time, with the syntactic features the documents are distinct from each other and the instances are distributed across the feature space much more uniformly. One can observe a clear tendency for translations to be 'shifted' to a region where non-translations are very rare, and vice versa.

5.1 Best features

The three classifiers that compare translations with the Russian reference corpus rely on the same set of features. The most useful features (in terms of their ANOVA F-value against the class of the text) are listed in Table 3 along with the ratio of their probabilities for each pair of corpora, which indicate the direction and size of discrepancies (all of them are statistically significant).

The set of features that were identified as most useful reflects various aspects of more complex syntax typical for translated sentences (for example, higher probability of clauses). Only three of the features highly correlated with 'translation/non-translation' classes appeared useful in the more difficult task of classifying translational varieties ('*nsubj:pass*', '*xcomp*' and '*acl:relcl*').

[3]Of course, their results are not directly comparable to ours, as they worked with English-to-German translations.

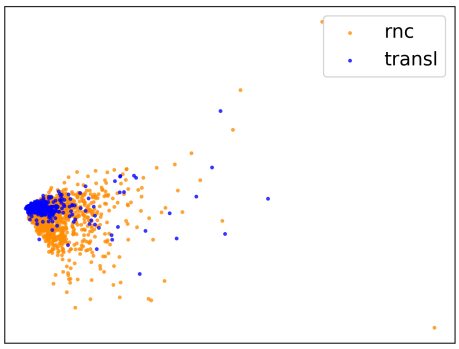

Figure 3: Non-translations (RNC) and translations, PoS feature space

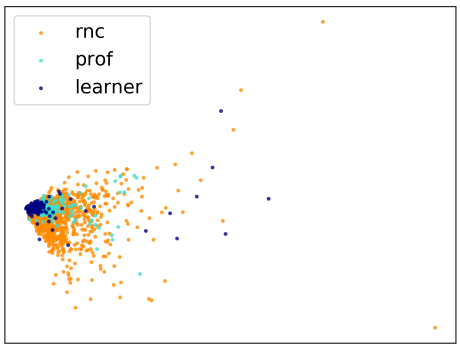

Figure 4: Non-translations (RNC), professional and learner translations, PoS feature space

feature	learners/RNC	prof/RNC
mark	1.86	1.85
ccomp	2.09	2.28
acl:relcl	1.92	1.68
advcl	1.73	1.68
nsubj	1.29	1.34
parataxis	0.66	0.74
aux	2.08	2.29
xcomp	1.44	1.60
obj	1.29	1.33
nsubj:pass	0.62	0.46

feature	learners/prof
nmod	1.11
aux:pass	1.63
nsubj:pass	1.35
iobj	0.82
flat:foreign	0.60
parataxis	0.89
fixed	1.16
acl:relcl	1.14
cc	0.93
xcomp	0.90

Table 3: Most useful features and the ratio of their probabilities in the data

6 Case studies

6.1 Syntactic complication: more fully expressed subordinate clauses

The most useful features in translations vs originally authored texts classifications (including 3-way classification) have to do with the higher probability of dependent clauses (*mark, ccomp, advcl, acl:relcl*). The strong correlation between *mark*, *nsubj* and relative and adverbial clauses suggests that translators tend to produce complex sentences more often than in naturally occurring Russian texts. They reproduce explicit pronominal subjects in the subordinate clauses, though in Russian they can often be left out. Five lexical items that head the frequency list of *nsubj* dependents (который ('which/that)', это ('this/it'), он ('he/it'), они ('they'), вы ('you')) are 2 to 3 times more frequent in translations that in non-translations. Example 1 gives a typical student translation that transfers the English structural pattern.

(1) ...человек на улице не думает о ЕС когда он входит в торговый центр
...the man in street not think about EU *when* *he* enters in shopping center
Source: '... the man on the street is not thinking about EU *as he* enters a shopping centre.'

Besides, the clauses are more often joined with explicit subordinating conjunctions to the detriment to other options such as punctuation. This finding corroborates the explicitation hypothesis in translational behavior (Blum-Kulka, 1986) and aligns well with extensive research on cohesive explicitation in translation (Kamenická, 2007; Cartoni and Zufferey, 2011; Becher, 2011).

6.2 *xcomp*: transfer of compound verbal predicates, particularly modal ones

The probability of open clausal complement (*xcomp*) in translations with regard to all other relations in the corpus is on average 1.5 times higher than in the reference corpus. This dependency describes relations between a verb and its adjectival or non-verbal complement. For English it captures complex object constructions and strings with catenative verbs as heads (including verbs with modal semantics such as *need to, have to, be going to, be able to*, but excluding modal auxiliaries) (Huddleston and Pullum, 2002). In Russian, it includes relations with the modal verb мочь ('can') and combinations with aspect and causative catenatives among others (начать учить петь ('to start to teach to sing'). The parser routinely assigns this relation to a verb and a deverbal noun (хоронить погибших ('to bury the deceased'). Despite these discrepancies in the parsing strategies, the cross-linguistic comparison shows that English uses this dependency 1.4 times as often as Russian (1.5% and 1.1% respectively, with the average probability of this relation for the translational corpora being 1.7%).

To find out which constructions drive up the probability of *xcomp* in translations, we looked at the semantic types of the top 25 head verbs in this relation. These cases account for 83% of all occurrences of this dependency in the learner corpus, for 77% in professional translations and for 73% in the RNC. English head nodes in this dependency are much more varied and lexically unrestricted than in Russian. The same 25 head nodes make up 59% of all occurrences of this relation. The structure of the frequency lists for translational corpora is a clear indication of the translational simplification in the form of higher lexical repetitiveness. Translations manage to cover more text with a smaller and less varied set of items.

We found that in translations, notably in learner translations, modal auxiliaries make up 55% of this top of the list in the learners corpus, with мочь ('can') alone covering 45% of all *xcomp*, while in professional translations it is 37% and in non-translated Russian text it is 32%. Another explanation for the increase of *xcomp* relations is the tendency to reproduce English non-finite constructions, especially with causative and aspect verbs as in example 2 from student translations:

(2) Многие десятилетия терроризм продолжал ассоциироваться...
Many decades terrorism *continued* *to-be-associated*...
Source: 'Terrorism *continued* for many decades *to be associated* primarily with the assassination of political leaders and heads of state.'

6.3 Passives: more analytical structures

In both translational corpora there are fewer dependencies marked *nsubj:pass* than in the non-translated reference corpus. Learner translations are 1.6 times short of this relation, while professional translations have 2.2 times less of it. This feature is among the 10 most well-correlated with the predicted class in two binary classifications (professionals/RNC and learner/professionals), as well as in the 3-way classification.

It makes sense to consider the values for *aux:pass* together with the above feature. This relation is more probable in student translations than in the output of professionals. The translational varieties appear to be at different sides from the reference corpus, with learners slightly overusing passive auxiliaries (1.4 times more of this dependency) and professionals underusing them (1.2 times less). This discrepancy between two translation varieties makes it one of the most useful features for predicting expertise.

In Russian, the choice of passive constructions is dependent on morphological properties of verbs, particularly on their aspect. The relations between semantic subject and object are mostly realized either by verb forms with the special formant -ся/-сь (imperfective verbs) or by passive participle in the short form with or without the analytical verb быть (*to be*) (perfective verbs). This gives a translator a variety of choices to render the single English grammatical meaning of passive, if she decides that this meaning needs to be rendered. For example, '*The house was built*' has options 'Дом построен' ('*the house is built*'), 'Дом был построен ('*the house was built*'), 'Дом строился' ('*the house was being built*').

To untangle the reasons behind the discrepancies in the distribution of passive auxiliaries and subjects, we looked at the proportions of analytical and the two morphological passives in translated and originally authored Russian. Contrastive analysis showed that English mass-media texts have less passive verbs than comparable Russian discourse. In our data, passive occurred in 15.9% of English sentences, while

original Russian texts had 18.6% of passive sentences on average. Analytical passives were used only in every forth passive construction.

In translational data, however, the proportion of passives dropped to 15% and 11% to the number of sentences in the corpus for learner and professional translations respectively. Both groups of translators use more analytical passives, driving up their ratio to all passives from 25% in the RNC to 38% and 35% in learners' and professional data. With that, professionals, when choosing between passive forms, tend to use more forms with -ся/-сь than short past participles. In this corpus, their ratio to other passives is 3% higher than in the learners' output.

7 Discussion

We showed that syntactic representation of translational data is a useful way to approach automatic classification, and the features useful for the classifiers lend themselves to linguistic interpretation. One major finding yielded by this research is the tendency to increase the number of clauses (particularly relative clauses) typical for translated Russian. With that, these clauses tend to express all structural components, particularly conjunctions and subjects, explicitly.

Cross-linguistic comparisons confirm that strings of non-finite verbs joined with two consecutive *xcomp* arcs in the sentence tree ('*Krugman added cartoons to try to make opponents look silly*') are more common in English than in Russian. In English-to-Russian translations this type of syntactic relation tends to be overrepresented (the average sentence-level probabilities of this relation are 1.7% and 1.1% for translations and Russian non-translations respectively), indicating a possible translationese-prone area. It is particularly true for sentences with the compound modal or aspect predicate.

The specificity of verbal elements in translations included a distinctive distribution of passive forms in translated Russian. We revealed higher proportion of analytical passives in professional, and especially in learner translations. Both translation varieties had less passives than the comparable Russian non-translations, with professional translations being further away from them and having 1.6 times less passive constructions. This trend may reflect the translational norm to avoid passives whenever possible or to rely more on syntactic rather than analytical forms. Educational and normative guidelines on English-to-Russian translation often warn against the overuse of passives (Moiseenko, 2012).

8 Conclusion

This research used syntactic annotation in the task of translation classification with the view to reveal syntactic specificity of translation varieties represented by learner and professional translations. We have compared our results with the PoS-trigrams baseline and have shown that syntactic representations are a fruitful way forward. We focused our attention on predicting translation expertise which is a fairly new area of research, exemplified by (Rubino et al., 2016) only.

The few cases tackled in this study just scratched the top of possibilities offered by the approach. We plan to continue research on syntactic properties of translations in several ways. First, it seems reasonable to use more refined morphosyntactic features as suggested in (Lapshinova-Koltunski, 2017) to provide algorithms with better learning material. Second, the UD framework makes it possible to take into account the linear order of heads and dependents in a relation and the order of relations in the sentence, which looks promising. Another possible extension is studying the role of disconnected parse trees in telling translations from non-translations. Finally, we would like to employ parallel nature of our corpora in a more meaningful way and describe translationese-prone areas in English-Russian translation based on cross-linguistic analysis of the aligned data.

Acknowledgements

This work was supported in part by a grant (Reference No. 17-06-00107) from the *Russian Foundation for Basic Research* (RFBR).

References

Roee Aharoni. 2015. *Automatic Detection of Machine Translated Text and Translation Quality Estimation*. Ph.D. thesis.

Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and Technology: In honour of John Sinclair*, J. Benjamins, Amsterdam, pages 232–250.

Marco Baroni and Silvia Bernardini. 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.

Viktor Becher. 2011. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* pages 17–35.

Michael Carl and Matthias Buch-Kromann. 2010. Correlating translation product and translation process data of professional and student translators. *14 Annual Conference of the European Association for Machine Translation, Saint-Raphaël, France* (May).

Bruno Cartoni and S Zufferey. 2011. How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web* (June):78–86.

Kira Droganova and Daniel Zeman. 2016. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies. Technical report, Institute of Formal and Applied Linguistics (ÚFAL MFF UK) Faculty of Mathematics and Physics, Charles University.

Iomdin L.L. Lazursky A.V. Dyachenko P.V. 2015. A deeply annotated corpus of Russian texts (SynTagRus): contemporary state of affairs. *Trudy Instituta Russkogo Yazyka im. V. V. Vinogradova* pages 272–299.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia* .

Rodney Huddleston and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Richard Hudson. 1995. Measuring syntactic difficulty. *Manuscript, University College, London* .

Yingqi Jing and Haitao Liu. 2015. Mean Hierarchical Distance Augmenting Mean Dependency Distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. pages 161–170.

Dan Jurafsky and H. James Martin. 2014. *Speech and Language Processing*. Pearson London.

Renata Kamenická. 2007. Defining explicitation in translation. *Sborník prací Filozofické fakulty Brněnské univerzity, Řada anglistická: Brno Studies in English* 33:45–57.

Andrey Kutuzov and Maria Kunilovskaya. 2014. Russian learner translator corpus: Design, research potential and applications. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Springer, volume 8655, pages 315–323.

Veronika Laippala, Jenna Kanerva, Anna Missilä, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. 2015. Towards the classification of the Finnish Internet Parsebank: Detecting translations and informality. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Linköping University Electronic Press, Sweden, pages 107–116.

Ekaterina Lapshinova-Koltunski. 2017. Cohesion and translation variation: Corpus-based analysis of translation varieties. *New perspectives on cohesion and coherence* .

Georgiy Moiseenko. 2012. *Translator and Editor Guide*.

John Nerbonne and Wybo Wiersma. 2006. A measure of aggregate syntactic distance. *Proceedings of the Workshop on Linguistic Distances* pages 82–90.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC-2016*.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA'08)*. October, pages 21–25.

Marius Popescu. 2011. Studying Translationese at the Character Level. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (September):634–639.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 530–540.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef Van Genabith. 2016. Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification. In *HLT-NAACL*. pages 960–970.

Alina Secară. 2005. Translation evaluation – a state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE Workshop, Leeds*. pages 39–44.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. pages 88–99.

Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(3):611–622.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities* 30(1):98–118.