Chapter 5

Testing target text fluency: A machine learning approach to detecting syntactic translationese in English-Russian translation

Maria Kunilovskaya
University of Tyumen

Andrey Kutuzov
University of Oslo

This research is aimed at the semi-automatic detection of divergences in sentence structures between Russian translated texts and non-translations. We focus our attention on atypical syntactic features of translations, because they have a greater negative impact on the overall textual quality than lexical translationese. Inadequate syntactic structures bring about various issues with target text fluency, which reduces readability and the reader's chances to get to the text message. From a procedural viewpoint, faulty syntax implies more post-editing effort.

In the framework of this research, we reveal cases of syntactic translationese as dissimilarities between patterns of selected morphosyntactic and syntactic features (such as part of speech and sentence length) in the context of sentence boundaries observed in comparable monolingual corpora of learner translated and non-translated texts in Russian.

To establish these syntactic differences we resort to a machine learning approach as opposed to the usual statistical significance analyses. To this end we employ models that predict unnatural sentence boundaries in translations and highlight factors that are responsible for their 'foreignness'.

Maria Kunilovskaya & Andrey Kutuzov. 2017. Testing target text fluency: A machine learning approach to detecting syntactic translationese in English-Russian translation. In Katrin Menzel, Ekaterina Lapshinova-Koltunski & Kerstin Kunz (eds.), *New perspectives on cohesion and coherence*, 75–104. Berlin: Language Science Press. DOI:10.5281/zenodo.814466

For the first stage of the experiment, we train a decision tree model to describe the contextual features of sentence boundaries in the reference corpus of Russian texts. At the second stage, we use the results of the first multifactorial analysis as indicators of learner translators' choices that run counter to the regularities of the standard language variety. The predictors and their combinations are evaluated as to their efficiency for this task. As a result we are able to extract translated sentences whose structure is atypical against Russian texts produced without the constraints of the translation process and which, therefore, can be tentatively considered less fluent. These sentences represent cases of translationese.

1 Introduction

This research is an attempt to use machine learning algorithms to identify cases of less-than-typical syntactic structures in learner translations (syntactic translationese). We aim at developing a robust methodology, which could be used to look into differences between standard Russian and its translated variety and to select the linguistic features that are best in signalling these contrasts. It can be used to test researchers' intuitions as to the tendencies in translational behaviour and provide data for contrastive analysis. Solutions to both tasks (establishing typical deviations from the reference corpus and describing them in terms of predictive linguistic features) are applicable in translator training (the purpose we are immediately after) and in designing machine translation systems to improve fluency.

Linguistic peculiarities of translations distinguishing them from original texts in the same language are described within corpus-based translation studies. Typical research in this domain is usually designed to test linguistic indicators that reveal some tendencies in translations and to disentangle various factors that can be associated with certain translational behaviour, including extralinguistic ones. The aim is to arrive at a clearer understanding of the motivations behind translators' linguistic choices. While this is a possible and tempting extension for current research we refrain from making explicit claims as to why specific patterns are observed in our data. We proceed without a specific "universal" hypothesis in mind, beyond the assumption that the two corpora are significantly different (the argument that has been supported in our previous research on the same data in Kutuzov & Kunilovskaya (2015). That said, we do rely on previous work in this strand of corpus-based translation studies in selecting linguistic indicators of syntactic translationese, making use of suggested ways to implement their detection computationally and provide tentative descriptions of detected tendencies in line with some of the well-known concepts within this theory.

An important aspect of our task is its focus on syntactic properties of translations. On the one hand, it is due to the role of sentence structure in the overall textual efficiency, in how easily a text is processed by the reader, how effectively it gets its message across. It has been shown that both structural integrity, interpreted as cohesion, and conceptual and pragmatic connectivity of corresponding discourse units (coherence) can be affected if target language specific (i.e. natural and conventional) sentence patterns are compromised in translation (e.g. with regard to failure to split sentences in translation, see Ramm 2006; Solfjeld 2008; Fabricius-Hansen 1999; Gile 2008; with regard to cohesion means, see Kachroo 1984; Hatim & Mason 1990). On the other hand, syntactic features of texts are less obvious to the naked eye, but are particularly informative in comparing corpora. There is ample evidence from corpus linguistics that functional and grammatical properties of words and surface characteristics of sentences (number and types of discourse markers, number of conjunctions and finite verbs, PoS, sentence length), which are typically used to operationalise syntactic or stylistic features of texts, are useful for the whole range of similar comparative tasks (for detecting translationese, see Baroni & Bernardini 2005; Pastor et al. 2008; in learner language studies Hinkel 2001; in authorship attribution and stylometry van Halteren 2007 and in text classification Koppel, Argamon & Shimoni 2002).

To achieve our goal, we use a traditional monolingual comparable corpora set-up: we exploit genre-comparable sub-corpora of the Russian National Corpus (**RNC**) and the Russian Learner Translator Corpus (**RusLTC**). The former is a reference corpus, which contains arguably representative sample of Russian language used to model dependencies that are then tested on translated data, the latter contains student translations that are viewed as particularly suitable for this task. They provide a strong case of human-produced translationese, because novice translators are notorious for generating disfluent texts that stand out for carrying foreign-sounding unnatural wording and structures. The corpora are described in detail in Section 3.

Methodologically, we follow the ideas of **multifactorial comparative analysis** of corpus data implemented within a supervised machine-learning approach suggested by Gries & Deshors (2014). One of the important improvements on previously used methods discussed in this work consists in ensuring contextual comparability of the phenomena under study. We tried to identify syntactic differences between the same corpora in previous experiments (Kutuzov & Kunilovskaya 2015) using de-contextualised PoS n-grams, but, against intuitive expectations and extensive theoretical evidence, failed to come up with meaningful results. Therefore, we introduce sentence boundaries (**SB**) as a structural

'anchor' to avoid over-generalisations of de-contextualised lexical and PoS frequencies and to ensure comparability of these features. Sentence boundaries are also an important linear syntactic event, which is traditionally used to gauge a number of textual properties such as sentence length and structural complexity.

We treat sentence boundaries as a surface feature of text structure and define it as an orthographically marked position, at which a sentence ends. It is typically marked with one of the four punctuation marks (full stop, dots, exclamation, question mark) or their combinations, and followed by a space and a capital letter. Effectively, sentence boundaries mark-off more or less independent chunks of information to be processed successively, thus encoding procedural information that guides pragmatic inference as to whether two informational constituents should be interpreted as a whole or individually, and how each of them relates to the topic and the intentional structure of the discourse (Guzmán & Klin 2000; van Dijk 1976; Carston & Behrens 2007; Unger 2011).

A meaningful study of semantic and pragmatic processes involved with speakers' motivations to start a new sentence (i.e. the analysis of regularities behind text/discourse segmentation into sentences *per se*) requires consideration of high-level linguistic phenomena (such as discourse and information structure), which are well beyond the scope of the present study. Instead we offer an account of typical and unnatural combinations of surface linguistic features at sentence boundaries as indicative of syntactic translationese.

At the same time, revealing unnatural sequences at sentence boundaries and sentences with atypical properties in Russian translations (in the present study limited to translations out of English) is potentially predictive of problematic text cohesion. Unlike English, non-emphatic Russian relies on word order as a primary means of structuring information. It has a strong tendency to place rhematic, new or focused elements in the sentence-final position (Grenoble 1998). This typological difference between the two language systems gives rise to the well-known structural deficiency of learner translations attributed to interference: they often contain prepositional phrases, which lack logical stress, at the end of the sentence (such as *никогда не слышал о нем* 'never heard of him'; *покарает его за это* 'will punish him for it'; *не успел избавиться от них* 'didn't have time to get rid of them' and adverbials (*купить по дешевке в России* 'to buy on the cheap in Russia').

The importance of maintaining cohesion in translation in ways licensed by the target language can hardly be overestimated. It was repeatedly stressed in translation studies (Blum-Kulka 1986; Hatim & Mason 2005; Baker 2011) on the grounds that faulty information structure and cohesion inadequacies can give

rise to extra processing efforts for the reader entailed by the necessity to handle inconsistencies in co-reference, they and also lead to inappropriate topicalisations and induce misleading interpretations of either content or the speaker's intentions. This claim is corroborated by psycholinguistic research, which finds that during text processing 'due to limited attentional resources, precedence may be given to processes involved in building a locally coherent representation [...] there may not be sufficient resources left for more global processes, such as integrating the current sentence with information from earlier in the passage' (Guzmán & Klin 2000: 728). The recent trend in statistical machine translation and natural language generation research seeks to enrich existing architectures with text-level linguistic data in attempt to overcome their cohesion and coherence limitations (Meyer & Popescu-Belis 2012). So, current research can yield useful comparative information to be applied in translation quality assessment and machine translation as well as provide insights on cross-linguistic contrasts and translator behaviour. Teaching translator trainees about typical translational choices that deviate from standard language can be a useful consciousness-raising exercise, while linguistic indicators of possible translationese can be used to develop tools to range translations by the degree of their 'nativeness'.

The rest of the paper is structured as follows. Section 2 offers a brief overview of research on translation universals (it seems that this term is well-established in the field despite its limitations and will be used as such further on), especially at the level of syntax and in the area of methodology, while Section 3 introduces multidimensional analysis as our primary approach, describes our corpus data and comments on the principles and process of feature selection. It is followed by the empirical results in Section 4, where we report, compare and interpret the performance scores of the first-step model on both corpora. This part of the paper also describes how these results are used to train the second model, which effectively predicts errors of the first model, i.e. strong cases of syntactic dissonance with the reference corpus as well as informs of the linguistic features associated with them. In Section 5, we interpret our findings trying to isolate patterns that can be explained from contrastive and translational perspectives and present some considerations on model-fitting for future work. Section 6 concludes the work with some general considerations of its applicability and scalability in terms of accommodating more sophisticated features and their combinations to target higher-level linguistic phenomena.

2 Related work

As stated above, our research is set in the framework of the so called **translation universals theory**, which posits that translations differ from non-translations in the same language in a number of statistically measurable ways, while bearing some common features regardless of the source language. It focuses on empirically assessable properties of translated language known as *translationese* or *third code*, which are allegedly manifestations of *translation universals* or laws of translation. Without going into terminological details and the history of this paradigm of contemporary translation studies, now well-established, we merely outline main concepts of this approach and survey some studies that deal with the syntactic indicators of translationese and ways of their computational implementations.

Over the last 20 years or so research in this area has thrown up about a dozen of hypotheses about translational behaviour and a number of linguistic indicators to validate them. The most widely discussed tendencies include explicitation, interference and transfer, standardization (or levelling-out), simplification, normalisation, atypical patterning and over- and under-use of items. Most of these features can be revealed both at lexical and syntactic levels (Zanettin 2013).

In terms of methodology the study of universals is closely related to the Contrastive Interlanguage Analysis described in seminal works by Sylviane Granger (Granger 2010; Štěpánek & Pajas 2010). It can be built around either of three types of comparisons, surveyed in several papers, including Chesterman (2010) and Xiao, He & Yue (2010), or a combination thereof (i.e. on data from complex multi-corpora architectures, which enables the researcher to account for several factors simultaneously like in Pastor et al. (2008); Hansen-Schirra (2011); Dai & Xiao (2011); Bernardini (2007));

1. It can be based on monolingual comparable corpora and compare translations to non-translations in the same language (e.g. Laviosa (1998); Olohan (2001); Xiao, He & Yue (2010));

2. a less common approach is taken in Rayson et al. (2008), where lexical translationese is revealed as difference between texts translated by Chinese translators into English and versions of the same texts hand-corrected by English native speakers;

3. research into universals can require a parallel corpus component to reveal similarities and differences between sources and their translations (see an

almost unique research based on multiple parallel corpus in Castagnoli 2011);

4. finally, translations can be compared to translations into other languages or genres or by different translators (Baker 2004, among others).

Our research draws upon the results obtained in the pioneering work by Baroni & Bernardini (2005), who apply machine learning based on text classification to detect translationese. Their results are inspiring: they find that one can computationally learn the difference between high quality translations and very comparable non-translations by relying on distributions of some classes of function words. They also found out that humans are outperformed by machines in their ability to tell translations from non-translated language (Baroni & Bernardini 2005).

These findings, on the one hand, stress the objective nature of translationese and at the same time underline the unreliability of human assessment. 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 and is not immediately obvious to the naked eye. The authors present convincing evidence that 'machine learning is reaching a stage in which it is no longer to be considered simply as a cheaper, faster alternative to human labour, but also as a heuristic tool that can help to discover patterns that may not be captured by humans alone' (Baroni & Bernardini 2005: 38). So, it makes sense to work towards employing computer technology in revealing and describing translationese as well as in evaluating target text fluency.

In corpus-based linguistics it is common practice to model language in studied corpora as PoS n-grams. This approach is implemented as part of an experiment to attest specific indicators of simplification and convergence in (Pastor et al. 2008), where shallow-parsed multiple corpora are represented as frequency vectors of PoS 3-grams. Other indicators of similarity in the same research include sentence length in tokens and the type of sentence identified as the number of finite verbs (and their corresponding verbal constructions) in it.

Our previous inquiry into translationese on the same data in (Kutuzov & Kunilovskaya 2015), which was set on lexical level and within a more conventional framework of statistical significance analysis, revealed opposing trends in the frequency of discourse markers - almost the same number of items were significantly overused or underused in translations. These findings can be interpreted in line with the third code hypothesis supported in (Hansen-Schirra 2011). Hansen-Schirra used carefully designed and annotated corpus resources and proved

hybrid character of translationese, which manifested opposite tendencies of normalisation and interference for individual register features.

Finally, to the best of our knowledge translated Russian is yet to be investigated in the corpus-based framework, though there has been extensive previous work in the pedagogical and prescriptive area. There is not much research on comparative analysis of Russian corpora either (however, see Mikhailov 2003 where a Russian-Finnish parallel corpus is described, and Kutuzov & Kuzmenko 2015, where machine learning methods are used as well, together with distributional semantics). But we can rule out frequency distribution of PoS n-grams, mentioned in many English-based studies, as a useful indicator of differences between copora due to the fact that word order in Russian is relatively more flexible. It can hardly be used as a crude substitute for syntactic information either, because it does not signal syntactic relations. At the same time it is crucial for structuring information, i.e. for arranging theme and rheme progressions and providing text cohesion (Alekseyenko 2013).

Taking the previous work on corpus-based studies of translated text into consideration, in the next Section we describe our experimental set-up and define the set of linguistic indicators chosen to represent our corpora in the machine learning task.

3 Applying multidimensional analysis to translations

As shown above, our main research question can be formulated as follows: are there any regular differences between translated and non-translated corpora in the typical linguistic environment of sentence boundaries, and which linguistic features (from the set we employ) will the machine learning algorithm mostly draw upon to calculate this difference? In other words, we aim at achieving a twofold objective. First, we want to detect whether a machine is able to learn contrasts between translations and naturally produced texts on the basis of representations of the two corpora built around the lexical and grammatical properties of tokens to the right and to the left of sentence boundary. Second, we want to reveal the indicators that are most informative for this task.

To tackle this, we roughly follow the multidimensional analysis approach established by Gries & Deshors (2014). They explore differences between native speakers and learners or non-native speakers through studying statistical interactions in corpus data. They establish a two-step procedure: a model trained on native data is applied to non-standard texts in order to find cases where their authors made decisions, distinct from what a native speaker would probably do in

the same linguistic situation. This approach was successfully applied to a comparison of differences in the usage of *may* and *can* between native English speakers and French and Chinese learners of English.

In the present research, texts translated from English into Russian are considered a kind of a specific Russian language variety that can be compared to a standard or native language. We hypothesize that while translating, native Russians construct sentences differently, and their deviating choices can be revealed through the statistical evaluation of the set of at-the-sentence-boundary-factors offered below. We argue that these features can be used to predict sentence boundaries as a formal structural event indicative of sentence structure. We use data from two corpora:

1. the well-known monolingual Russian National Corpus (further **RNC**) containing non-translated texts by native Russian speakers and extensively described in the literature[1];

2. the parallel Russian Learner Translator Corpus (further **RusLTC**) described in Kutuzov & Kunilovskaya (2014), containing translations from English into Russian and backwards done by Russian translation students from 8 different universities[2]. There are no reference translations in the corpus, but one source can be accompanied by multiple translations.

The RNC represents 'native' Russian language, while the second corpus is arguably a strong case of a non-standard variety ('translationese' in the current research context). From each corpus, we extracted a sub-corpus containing texts belonging to mass-media expository genres, so that the material is as comparable as possible. Overall, our 'standard' corpus consists of 7 679 documents and 8 289 884 word tokens, while translations corpus consists of 1 332 documents and 586 935 word tokens.

In order to evaluate differences between non-translated and translated texts, we employ a number of contextual features in sentence boundaries environments. They were used to train a machine learning model to predict these boundaries. We will now briefly describe the essential details of the process. Our training set (a mass-media sub-corpus of the RNC) lacks manual sentence mark-up. Thus, we first trained a *Punkt* model on the whole RNC (about 150 million tokens). *Punkt* (Kiss & Strunk 2006) is a well-known unsupervised algorithm to learn abbreviations, collocations and typical sentence-starters. After initial training, it

[1] See http://ruscorpora.ru/corpora-biblio.html
[2] Available at http://rus-ltc.org

can then be used on raw text to detect sentence boundaries with high accuracy. We applied the trained model to our sub-corpus to split it into sentences. This segmentation is accepted as ground-truth and used further.

We are interested in how various linguistic features correlate with the event of a sentence boundary. Thus, in our approach, word tokens in the text are observed as instances with various linguistic features (attributes). Each instance belongs to exactly one of two classes: either it is the last in the current sentence or not. If it is, it means that its class is 'boundary', otherwise it is a regular token.

Then, the problem is to build a binary classifier which predicts boundary class depending on token features. It is important to note that tokens in our case include punctuation, but not end-of-sentence punctuation marks: those were ignored during training and testing. This is because we are after linguistic features, not trivial orthographic predictors like a full stop or a capitalized word (all tokens were lower-cased). Because of punctuation, the total number of instances in our data sets is slightly higher than stated above: 9,422,955 instances for the RNC corpus and 631,361 instances for the translation corpus.

Initially, we extracted a total of 82 features:

1. current token (instance itself);
2. lemma of the current token[3];
3. part of speech of the current token (one of 19 categories, including punctuation);
4. token length in characters;
5. lemma length in characters (because of rich inflectional system in Russian, it is often quite different from the token length; also, functional words are usually shorter than content ones);
6. accumulated sentence length in tokens (up to the current token);
7. accumulated sentence length in characters;
8. accumulated number of finite verbs in the current sentence;
9. accumulated number of Nominative nouns and pronouns;

[3] Lemmatisation and PoS-tagging was performed with the help of state-of-the-art *Mystem* morphological analyser for Russian, described in Segalovich (2003)

10. accumulated number of coordinate conjunctions (including multi-word entities, 26 conjunctions in the list);
11. accumulated number of subordinate conjunctions (including multi-word entities, 56 conjunctions in the list)
12. lemmas of five tokens to the left and five tokens to the right of the current token (further 'neighbours');
13. lengths of lemmas and tokens of the neighbours;
14. binary feature 'is a coordinate conjunction' for all the neighbours;
15. binary feature 'is a subordinate conjunction' for all the neighbours;
16. binary feature 'is a discourse marker' for all the neighbours (discourse markers list comprises 86 elements and includes words like *итак* 'thus', and multi-word entities);
17. part of speech for all the neighbours;
18. binary class attribute (sentence boundary or not), with about 6% of all tokens being boundary.

Not all features possess equal predictive power. First of all, we had to filter out string features (lemmas and tokens themselves). Using text strings as predictors is principally possible, but only with corpora much larger than ours, to overcome the sparsity problem (the majority of words are rare). Also, most classifiers do not work with string attributes: we managed to train Bayes multinomial and stochastic gradient descent models (essentially vectorizing text attributes and then treating them as numerical ones), but performance was much worse than with other features (numerical and categorical/nominal). Thus, we leave this possibility for a future work.

After removing problematic string features, we performed basic feature selection by measuring *information gain* (mutual information, MI) with respect to sentence boundary class for each feature independently in the RNC. Below is a list of the most promising features in descending order, with respective information gain values and identifiers:

1. 0.031049 PoS of the current token (**pos**);
2. 0.022271 PoS of the first token to the right (**pos1R**);

3. 0.010838 length of the current token in characters (**token_length**);
4. 0.010205 length of the current lemma in characters (**lemma_length**);
5. 0.009188 PoS of the first token to the left (**pos1L**);
6. 0.008043 accumulated sentence length in characters (**sent_char_length**);
7. 0.007313 accumulated sentence length in tokens (**sent_length**);
8. 0.005357 accumulated number of finite verbs in the current sentence (**finite_verbs**);
9. 0.005047 PoS of the second token to the right (**pos2R**);
10. 0.004097 is the first token to the right a discourse marker? (**dm1R**);
11. 0.003592 length of the first token to the right (**token_length1R**);
12. 0.002896 is the first token to the right a coordinate conjunction? (**conj1R**);
13. 0.002832 length of the first lemma to the right (**lemma_length1R**);
14. 0.002556 accumulated number of coordinate conjunctions in the current sentence (**conjunctions**);
15. 0.001879 PoS of the third token to the right (**pos3R**).

Additionally, *CfsSubsetEval* the (Correlation-based Feature Subset Selection) algorithm, described in Hall (1998), was used to discover the best subset of features. This information is important, because features may be (and certainly are) interdependent and improve or degrade performance of each other. Bidirectional evaluation of 621 subsets (only globally predictive features[4]) returned the following set of 4 features as the best one:

1. PoS of the current token (**pos**);
2. is the first token to the right a discourse marker? (**dm1R**);
3. is the first token to the right a coordinate conjunction? (**conj1R**);
4. PoS of the first token to the right (**pos1R**).

[4] It means that we measured their performance over the whole dataset. This effectively eliminates features which are very predictive at some particular parts of the data (for example, in texts by one author), but useless in the majority of other parts.

Based on this data, we conclude that the best-predicting features are parts of speech for both the current token and its immediate right and left neighbours, length of the current token, the accumulated sentence length in characters and the number of finite verbs. It turns out to be important to look at the functional status of the neighbours: the property of being a discourse marker or a conjunction for the first token to the right ranks high as a predicting feature in our experiments. On the other hand, the features manifesting the length of neighbour tokens do not contribute much to the prediction, but slow down the training. Therefore, these features as well as accumulated number of Nominative nouns and pronouns were filtered out.

The last feature seemed promising initially, but did not provide enough predictive power. We believe the reason is grammatical homonymy: in Russian, Nominative and Accusative forms often coincide for inanimate nouns, and this ambiguity is not resolved by *Mystem*, not without syntactic parsing anyway. We considered a noun to be Nominative only when it was the only possible morphological interpretation, and this is only the case for animate nouns. Thus, in fact this feature reflected the accumulated number of Nominative *animate* nouns. Note that most information potentially delivered by the number of Nominatives is probably already contained in the number of finite verbs (and this feature is closely correlated with boundary class), so, the loss was not big.

The remaining 48 features were used to train a REPTree model (Reduced Error-Pruning Tree, introduced by Quinlan 1987) to predict sentence boundaries in native non-translated mass media texts. Unlike regression used by Gries & Deshors (2014), this algorithm belongs to the family of decision tree learners; we use its implementation in the open-source Weka software package Hall et al. (2009). A decision tree approach was chosen because it allows training on various types of features (predictors): numeric, binary or nominal/categorical. Additionally, decision trees are more human-readable than the output of other machine learning classifiers, though, of course, with large amount of data the model becomes more complex, with tens of thousands branches or more, which makes it not feasible to try to 'read' it directly.

In order to avoid over-fitting and improve accuracy, we used REPTree with the *Bagging* meta-algorithm suggested in Breiman (1996). It essentially multiplies training data through bootstrapping and then trains models on each of resulting sets ('bags'). The predictions from each model are averaged before final output. In our task, it substantially improved performance of the classifier. Thus, we have a model that classifies tokens into boundary (final) and non-boundary ones based on the above mentioned set of features. For each classification (prediction) the

model additionally outputs the degree of its confidence in the range {0…1}. We will comment on the performance of this model in Section 4.

Example 1 below illustrates the model's predictions on a piece of Russian text:

(1) *…но & и & алмазодобывающим. & Сейчас…*
[non-boundary & non-boundary & boundary & non-boundary]
…but also diamond-producing region. Today…

The next step is to use this model to 'predict' sentence boundaries in our translation corpus. We expect the model to perform slightly worse, because translations (let alone learners' translations!) are well-known to be linguistically different from non-translations in the same language. The results of testing the previously trained model on translated texts may be used for two purposes: first, to manually inspect cases of the model failing to predict sentence boundaries and possibly gain insights on the reasons, and second, to train another model which predicts not sentence boundaries, but inconsistencies between the first model decisions and what a translator did in a particular context.

In other words, we try to find out which of the above mentioned linguistic features or their combinations are associated with 'non-typical' (or outright erroneous) sentence boundaries in translations. This answers one of the important questions in translation studies (and in cross-linguistic research in general): what patterns of linguistic elements and their characteristics make translations or learner speech in L2 sound non-fluent, foreign and unnatural? Experimental results are described in Section 4.

4 Experimental results

Table 1 shows performance of the first trained model tested on the native corpus (RNC) and on the translation corpus (RusLTC). Overall F_1 (harmonic mean of precision and recall) is a weighted value over both predicted classes, boundary and non-boundary; boundary F_1, precision and recall are the respective values for boundary class only. Performance on detection of the non-boundary tokens is much higher than on the boundary ones, because the first class is much more frequent: it is easier to detect an in-sentence token than a final one. This is the reason behind the difference between overall and boundary performance.

We report precision and recall results, not only purely statistical values like coefficient of determination (R^2) or likelihood ratio. We believe it is more important to evaluate real predictions of the model on the data rather than abstract

Table 1: Performance of sentence boundary detection model

	Overall F_1	Boundary F_1	Boundary precision	Boundary recall
RNC	0.955	**0.584**	**0.873**	**0.439**
RusLTC	0.956	0.522	0.708	0.413

goodness or the regression fit: one is interested in how much noise is present in the model's predictions for each class (precision), and what fraction of instances belonging to this or that class was correctly classified as such. Simply reporting the overall accuracy (percentage of correctly classified instances) is not enough.

Quite often we deal with binary classification tasks, where instances of class **A** are much rarer than instances of class **B**. For example, in our data, sentence boundary tokens occur 15 times rarer than the non-boundary ones. The same is true for usage of *can* and *may* in Gries & Deshors (2014): *can* is 2 or 3 times more frequent. In this situation, a classifier can be very reliable for the majority class, but though showing poor quality for the minority class. However, because of larger number of majority class instances, the overall number of correct predictions will be high and accuracy would seem to be quite satisfactory, notwithstanding the fact that the model actually almost never correctly predicts the minority class (and this 'marked' class is often the aim of the whole research). Thus, it is very important to report precision and recall for each class separately, especially for the minority one.

Getting back to our results, we see that despite high overall F_1, the model is not quite perfect in detecting sentence boundaries even in the native corpus it was trained on: more than half of the boundary tokens are not detected as such. However, precision is very high: there is almost no noise in the detected boundary events (Baroni & Bernardini (2005) faced the same situation). It means that not all sentence boundaries correlate well with the features we chose. This is expected and quite natural: Russian sentence structures are highly variable due to relatively flexible word order. Also, sentence boundaries are often influenced by other higher-level linguistic phenomena, such as syntactic dependencies, or semantic and pragmatic structure of the discourse.

However, quite a lot of boundaries are predicted by the formal and morphological characteristics of the elements we employed. As stated earlier, boundary tokens comprise no more than 6% of all instances in the data set (both in native and translated corpora). Consequently, F_1 of the boundary class detection in our model is more than 4 times better than expected $F_1 = 0.1$ of random baseline

(choosing one of two classes with equal probability). Thus, our features do provide some signals which are meaningful for predicting sentence boundaries. It means that in non-translated Russian texts there are relatively stable patterns marking such boundaries, which makes it feasible to compare these patterns to ones found in the translation corpus.

It is also encouraging that performance does not drop significantly when the model is applied to the translated corpus: the same regularities generally hold in translated texts as in native ones (they are still in the Russian language, after all). However, both precision and recall are slightly lower, which means that the model makes wrong predictions more often than on the native texts, and thus, the aforementioned patterns of features behave slightly differently in the translated corpus. This also seems quite logical: as stated earlier, translated texts represent a special non-standard variety of Russian, and sequences of items in these texts deviate from the standard ones the model was trained on.

In order to learn which linguistic features from the general list above are associated with these deviations, once again we follow Gries & Deshors (2014)'s approach and compile a dataset with all instances from our translated corpus, their respective features and a new class attribute. This time, instances are divided into two classes, depending on whether the model made a correct or incorrect prediction.

Then, we remove all instances where confidence of the model prediction was below 0.9 to filter out 'weak' decisions[5]. This step leaves us with 548 231 instances, out of 631 thousand total.

For this dataset we perform feature selection as well: from the linguistic point of view, we look for combinations of features that typically accompany non-native behaviour of the text producer. The following features are found to correlate best with the probability of error (the correlation is again calculated as *information gain*):

1. 0.0069041 **pos**;
2. 0.002582 **pos1R**;
3. 0.0025574 **sent_char_length**;
4. 0.0023149 **sent_length**;
5. 0.0022355 **token_length**;

[5] Studying weak predictions and correlating them with real translators' decisions also seems promising, but we leave it to future research.

6. 0.0018903 **lemma_length**;
7. 0.0017348 **pos1L**;
8. 0.0014444 **finite_verbs**;
9. 0.0011813 **conjunctions**.

Additionally, the best set of features selected using *CfsSubsetEval* includes **pos**, **token_length**, **sent_char_length**, **finite_verbs**, **conjunctions**, **subconj1L**, **pos1R**, and **subconj2R**. **dm4R** was selected as a locally predictive feature: it predicts an error only in some contexts, while other features do this globally.

Thus, it is part of speech of the token itself and its immediate neighbour to the right that mostly mark non-native behaviour of learner translators in our RusLTC corpus; accumulated sentence length (it seems that one can safely use either token length or character length) is also among the best predictive features, as well as the length of the current token and, to some extent, the number of conjunctions and finite verbs in the sentence.

Note that if we look at the predictions that the model made in the native texts (RNC corpus) at test time and try to find features correlated with correctness of decisions made, the set of most effective predictors would be different and much weaker. Only one feature (**pos**) achieves the information gain value of 0.002[6], while other features' correlations are an order of magnitude lower and can be considered non-existent. Thus, in native texts, correctness of our model's decisions is not directly dependent on particular features, and its errors are caused by external factors (preprocessing or lemmatising issues, higher linguistic constraints on sentence boundaries, etc). At the same time, in the translation corpus the models' mistakes are often determined by the feature patterns found in the data, rather than by noise or factors outside our reach.

The reference corpus is 15 times larger than the translational one, so it is very unlikely that the model has not seen some patterns of the selected features. We suppose that the model's failure to predict sentence boundaries in translations can be safely attributed to sentence boundary pattern deviations from the standard, found in translations.

Thus, applying the model trained on the comparable reference corpus to the translated texts reveals that they possess intrinsic characteristics different from those of non-translations. Lexical and grammatical features of tokens in the immediate context of sentence boundaries are found to be stably different in cor-

[6] Still 3.5 times lower than in the translations.

pora of non-translations and translations. In Section 5 we discuss examples and implications of these findings.

5 Discussion and future work

The analysis of the algorithm's performance on the translation corpus and error modelling led to several interesting insights and observations, described below.

Manual inspection of correlation between instances' parts of speech and the first model errors on the translation corpus indicates that some of PoS yield more errors on the same amount of instances than the others. It means that they are more often included in non-standard sentence boundary patterns in translations. As shown in Figure 1, the parts of speech of the current token that are apt to defy standard Russian regularities include nouns and pronouns in non-nominative cases (**S** and **SPRO**) and tokens for which *Mystem* was not sure about their PoS (**UNKN**). Other parts of speech are more conforming and cause less mistakes, signalling that translators make more natural choices.

Linguistically speaking, it means that there are **contextually identical** situations, in which standard Russian texts usually feature sentence boundary, while translated texts do not (or vice versa). This difference in sentence patterns is most frequently associated with non-nominative nouns and pronouns.

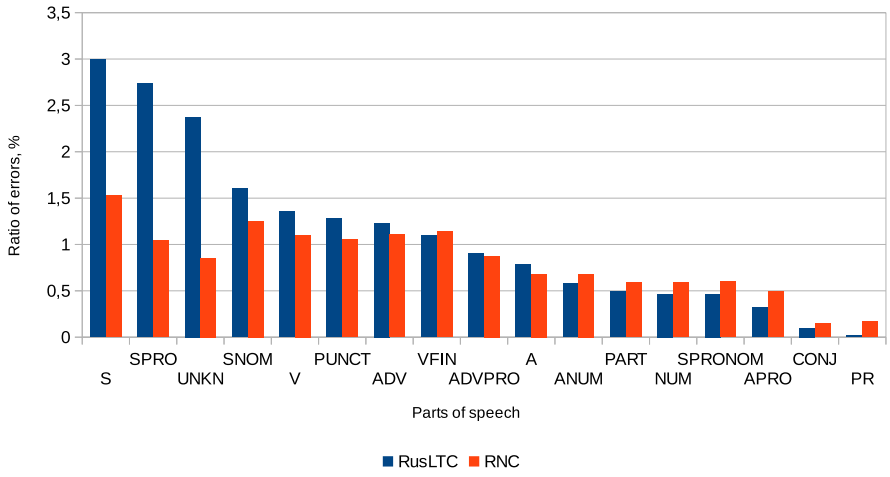

Figure 1: Error rates for PoS values of current token

It is quite logical that the model makes mistakes on 'strange' tokens with unknown PoS (mostly they are foreign words in Latin alphabet, digits or rare abbreviations).

Additionally, such atypical patterns are often caused by interference from English word order. In Example 2 translators routinely reproduce the structure with the final non-nominative pronoun, which is less frequent, but not unacceptable, in non-translated Russian texts (see more detailed explanation below, in the description of **PR_SPRO** pattern).

(2) *Trees rustled above him.*
Деревья шумели над ним.

Note that the mistakes are rarer on the native texts (see **RNC** bars in Figure 1) for almost all parts of speech where error ratio exceeds 1%, and are on par with the translations in the other cases. Also, non-nominative nouns (S) and pronouns (SPRO) seem to be not so variable as to their positions within a sentence in the reference corpus as in the translation corpus: in the RNC corpus the error ratio for them is almost equal to their nominative counterparts.

As it is clear from the precision/recall metrics and confusion matrix, most model errors occur when the model does not predict an actual sentence boundary in the translated texts (false negatives). Sentence boundaries predicted in the middle of running sentences (false positives) are far less frequent errors: they account for only 5% of all model failures. It means that the model does cover some real contextual patterns where sentence boundaries are typical for RNC, but it does not observe these patterns in translated data, given our feature set. For the purposes of this exploratory work we decided to prefer precision to recall and did not try to cover other (numerous) cases, when sentence boundaries are not described by our features.

Figure 2 illustrates this with the **pos1R** feature (PoS of the first token to the right of the current one). Bottom parts of the chart bars represent cases where the actual SB was missed by the model, because the observed sequence of linguistic features is problematic for the model trained on the standard language variety (false negatives), while the top ones represent cases where SB was predicted after tokens that actually were not final in translations (unlikely non-boundary tokens, false positives).

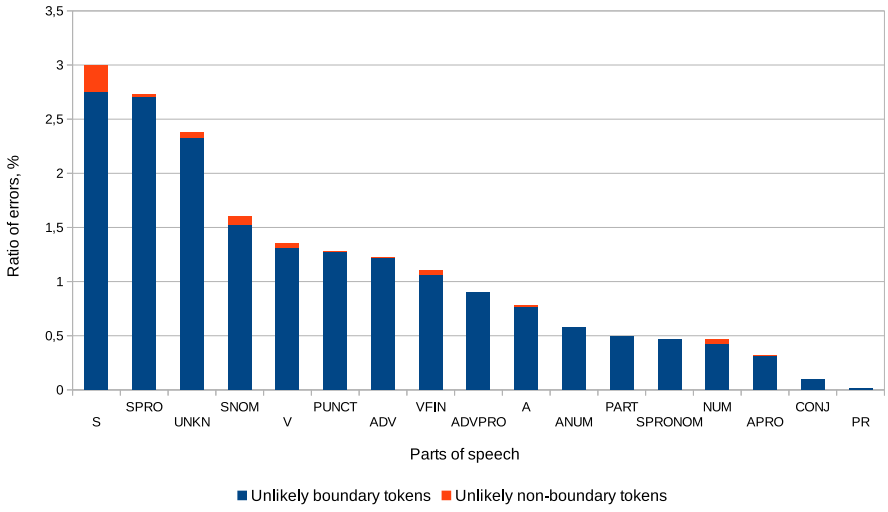

Figure 2: Error rates for PoS values of the first token to the right; evaluation on translational data

Interestingly, the ratio of false positives for some PoS of the nearest neighbour to the right is unusually high (higher than 5% of all errors, which is the mean value over the whole corpus): precisely, for **S** and **NUM**, and to some extent for **SNOM**. Thus, translators comparatively more often continue sentences with numeral words (including lexical units like *оба* 'both' or *полтора* 'one and a half'), while in the same situation in the native texts we would expect the sentence to end, and a new sentence to start with this numeral.

Similar observation can be made concerning particular binary features, which also seem predictive of non-standard translators' behaviour. For example, the probability of an error is almost two times higher (2% probability) when the next token to the right belongs to the set of discourse markers (like *в сущности* 'in fact', *наверное* 'perhaps'), manifested in the feature **dm1R**. These errors are distributed almost evenly between false negatives (69%) and false positives (31%), leading to a false positives ratio that is 6 times higher than the average over the corpus. This is because under the same circumstances in standard Russian the sentence would end, and the new sentence would be started with a discourse marker, but translators decide to continue the sentence, joining it with the next. Thus, the model yields a false positive in detecting a sentence boundary token

immediately to the left of the marker. Note that when the **dm1R** feature takes the 'False' value (the first-to-the-right token is not a discourse marker) the distribution of false negatives and false positives is quite standard: 95% vs 5%.

Despite the fact that RusLTC contains more sentences starting with one of the discourse markers from our list (7,28% of all sentences) than RNC (5,66%, the difference is statistically significant), it also contains sentences with atypical in-sentence position of typical sentence-initials. Thus, our strategy of revealing translationese overcomes limitations of the traditional statistical significance analysis.

Consider the translation in Example 3 to the English source text:

(3) *The findings have broken down some of the illusions commonly associated with burglaries; with four out of five revealing burglary was not opportunistic,* ***instead*** *returning to a property a number of times before breaking in (Daily Mail, Nov. 1, 2011).*

Результаты исследования разрушили некоторые мифы, касающиеся краж со взломом, ***так*** *например, четыре из пяти раскрытых преступлений не были незапланированными,* ***напротив****, грабители несколько раз возвращались на место потенциального взлома прежде, чем вторгнуться в чужой дом.*

The information units after the English semi-colon and after 'instead' are both rendered as well-formed separate discourse units, each with their own discourse markers, but these potential sentences are unreasonably jammed into one formal structure.

The difference is even more striking with the feature **subconj1R** (whether the next token is a subordinate conjunction or its equivalent). When this feature takes the 'True' value, the ratio of false positives is close to 50%. It means that the model expects to observe more sentences that start with a subordinate conjunction (e.g., *затем* 'then' or *если* 'if') than is the case with the learner translations. It seems to speak in favour of the normalisation hypothesis in translation. Traditional stylistics frowns upon starting a sentence with a subordinate conjunction and translators are opposed to using these less standard opportunities of the language system, which leads to a flatter, less varied expression typical for translations and to lower frequencies of more peripheral elements in them.

Note that our specific interest to false positives is also rooted in the expectations from our previous research Kutuzov & Kunilovskaya (2015), which showed that sentence length in translations is significantly higher than in non-translated texts (from the same sub-corpora). Our belief was that an algorithm like the one

reported here should return more false positives for longer sentences, especially as sentence length is among the best predictors in both models. The experiment indeed shows that there is a strong (0.72) exponential correlation between sentence length in characters and the number of false positives; for false negatives this correlation is even higher and reaches the value of 0.8. Thus, statistical modelling approach seems to support the observation that (learner) translations tend to over-use long sentences and this leads to a 'foreign' flavour of the produced texts. In the future, we plan to conduct a more thorough investigation into how and why error rate increases in correlation with sentence length.

Such analysis can be easily made more granular and multi-factorial: we can test for correlation between *sets* of features and non-standard language usage. For example, after ranking patterns **pos+pos1R** by the probability of false negatives, the sequence **SPRO+CONJ** (non-nominative pronoun followed by conjunction) is found on top of the list, with the model failing to predict sentence boundary in almost 10% of its occurrences. Examples of such contexts include sequences like '*которые попадаются у него на пути или похожи на* ***них****. И такие поступки бросают...*'[7] (boundary token is given in bold). It seems that when preceded by a non-nominative pronoun, such a sentence start is rather unnatural: if the first sentence instead ends in a nominative pronoun, the model makes mistakes in less than 2% of such cases. As expected, there are no false positives for both of these patterns.

Another interesting pattern is **pos1L+pos**. The top of the list is dominated by patterns like **V_SPRONOM**, **VFIN_SPRO**, **V_SPRO** (pronouns preceded by verbs) and **PR_SPRO** (pronouns preceded by prepositions). 5-6 % of all their instances produce false negative results. This can be explained by English-based interference: typical English sentences ending in non-rhematic (prepositional) phrases get diligently copied into Russian translations. See the following examples 4 - 7 of sentence ends:

(4) *...until you can clearly define and understand what is being conveyed you cannot hope* ***to translate it****.*
...пока вы не можете ясно определить и понять то, что имеется в виду, не надейтесь ***перевести это****.* (**V_SPRONOM**)

(5) *...with which he* ***identified himself****.*
..с которыми он ***ассоциировал себя****.* (**VFIN_SPRO**)

[7] ...which are on his way or similar to **them**. And such actions make...

(6) *...even sometimes obliging a Great Power to tail along* ***after him****.*
...иногда даже заставляющим великие державы ***подчиняться ему****.*
(V_SPRO)

(7) *It was the end of books* ***for me****.*
Книги перестали существовать ***для меня****.* **(PR_SPRO)**

In all these cases putting the rhematic verb in the end of the sentence, after the pronoun, would sound much more natural and close to a native text. Such cases of translationese are detected by our approach: the model trained on the native corpus 'stumbles' at these sequences and rejects to acknowledge that this is the end of the sentence. Thus, this is another example of morphosyntactic feature sets that are perceived by a native speaker as somewhat unnatural, and that are computationally detectable in our approach.

There is one **pos+pos1R** pattern in which the ratio of false positives exceeds the average over the corpus, comprising more than 6% of all errors. It is **S+ADVPRO** (non-nominative noun followed by an adverbial pronoun). False positives in this pattern are often due to translators' punctuation errors. For example, in the fragment '*морского побережья, открытых земель, мест обитания и* ***мест*** *куда художники и обычные люди могли бы*'[8] it would be correct in Russian to insert a comma after '*мест*'. Without it the model supposes a sentence boundary (perhaps, the lack of finite verbs in this sentence is another reason for the wrong prediction).

We have also detected the tendency for learner translators to overuse pronouns, such as *это* 'this, it' and *здесь* 'here', *так* 'so' at the end of the sentence, which can be the English source text 'shining through'.

Given above are only some examples of 'translationese' discovered by our approach; in fact, this list can be continued and expanded. It is, however, already clear that a researcher can draw numerous insights analysing the output of an algorithm modelling 'native speaker' (in our case, an author of a non-translated text) applied to translations. For example, one can find translations which are most different from native text by simply calculating the density of model mistakes in the given documents. Interestingly, in our material, such procedure revealed several student translations which, upon manual inspection, were obviously produced by machine translation (students cheated).

We emphasize that these differences in the structure of native and translated texts are not always the sign of 'lower quality' of the latter. Differences can be caused by one of translation universals (see example with normalization above)

[8] ...of seaside, open lands, habitats and **places** where artists and common people could...

and do not necessarily negatively impact the language of translation. However, detecting 'syntactic translationese' can still be helpful in many settings.

At the same time, manual error analysis brought to our attention several issues with the model design to be addressed in future work. First of all, the model does not distinguish between different punctuation signs, and fails to recognize sentence boundaries before inverted commas opening a sentence; a lot of mistakes come from inverted commas used to set off trademarks, titles and some proper names.

Much noise comes from the binary features that involved multiword discourse markers, which were considered as one lexical unit. The latter proved to be sometimes homonymous to nominal phrases with preposition, and this led to unreasonable predictions. To be a truly reliable feature, these elements need to be disambiguated. Also, some normalization for numbers is needed: as of now, all numbers written in figures are referred to unknown category, which makes a good deal of instances less usable.

We believe that the model would benefit from adding at least several lexical features as strings. As stated above, for now we excluded all string features because of computational complexity and their high dependency on semantics of the utterance. However, a number of words typically accompanying sentence boundaries can be selected and employed.

Thus, our future work in this area should include attempts to decrease the noise in the output through more thoughtful formatting and add new and better-motivated features to the corpora representation, including syntactic ones.

6 Conclusion

The work described above is an attempt to apply multi-factorial statistical analysis to study a variant of the Russian language instantiated in learner translations. We trained machine learning models that detect cases of dissonance between translated and non-translated texts based on a set of formal and morphological features and sentence properties. The approach is tested on traditional for this task monolingual corpora (the reference corpus of non-translated Russian texts and a corpus of comparable learner translations from English into Russian).

Differences between translated and non-translated texts are detected with reference to sentence boundaries, an important structural event, which serves here as a comparability factor. We hypothesize that sentence boundaries in the two corpora are dissimilar in terms of their morphosyntactic environments, and support this claim with empirical evidence.

We analysed variation in sentence patterns between learner-translated and non-translated Russian mass-media texts on the basis of surface and morpho-syntactic parameters of sentence boundaries context. We employed a sliding window of 10 tokens (5 to the left and 5 to the right of a possible sentence boundary) and their associated features to train a classifier which tries to predict whether the current token is the end of the sentence or not. The trained model was then applied to translated texts to find out differences in typical sentence boundaries patterns.

In our experiments, the model trained on the native texts served as a 'mechanical intelligence' representing an average native speaker of Russian making decisions about whether the sentence is going to end in this particular position or not. Comparing this models' decisions with real sentence boundaries in the translated texts allowed to automatically reveal several repeating patterns of features, frequently pointing at cases of 'translationese' typical for learner translators. Thus, this two-step methodology proved fruitful for our aims.

In the future we plan to enrich it with higher-level indicators, such as syntactic dependencies, anaphoric and co-referential chains, semantic data or, maybe, discourse relations, to build up knowledge about sentence boundary as a discourse structural event. Meanwhile, our approach makes it possible to detect sentence boundaries atypical for native texts. This is another step towards an automatic translationese spotter, widely sought in the field of computational translation studies.

7 Acknowledgements

This work has been partly supported by the Russian Foundation for Basic Research within Project No. 17-06-00107. The authors thank the anonymous reviewers for their helpful comments, which were crucial in guiding our work into the right direction. However, all mistakes and inconsistencies remain the responsibility of the authors alone.

References

Alekseyenko, Nataliya V. 2013. *A corpus-based study of theme and thematic progression in English and Russian non-translated texts and in Russian translated texts*. Kent State University PhD thesis.

Baker, Mona. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* 9(2). 167–193.

Baker, Mona. 2011. *In other words: A coursebook on translation*. Amsterdam/Philadelphia: Routledge.

Baroni, Marco & Silvia Bernardini. 2005. 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.

Bernardini, Silvia. 2007. Collocations in translated language: Combining parallel, comparable and reference corpora. In *Fourth Corpus Linguistics Conference held at the University of Birmingham*, 27–30.

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

Breiman, Leo. 1996. Bagging predictors. *Machine Learning* 24(2). 123–140.

Carston, Robyn & Bergljot Behrens. 2007. Making connections–linguistic or pragmatic? In Randi Alice Nilsen, Nana Aba Appiah Amfo & Kaja Borthen (eds.), *Interpreting utterances: Pragmatics and its interfaces*, 51–78. Oslo: Novus Press.

Castagnoli, Sara. 2011. Exploring variation and regularities in translation with multiple translation corpora. *Rassegna Italiana di Linguistica Applicata* 43(1). 311–332.

Chesterman, Andrew. 2010. Why study translation universals? In R. Hartama-Heinonen Kiasm & P. Kukkonen (eds.), *Acta translatologica helsingiensia*, 38–48. Helsinki: Helsingin yliopisto, Suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos.

Dai, Guangrong & Richard Xiao. 2011. ''SL shining through" in translational language: A corpus-based study of Chinese translation of English passives. *Translation Quarterly* 62. 85–108.

Fabricius-Hansen, Cathrine. 1999. Information packaging and translation: Aspects of translational sentence splitting (German–English/Norwegian). *Studia Grammatica* 47. 175–214.

Gile, Daniel. 2008. Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum* 6(2). 59–77.

Granger, Sylviane. 2010. Comparable and translation corpora in cross-linguistic research: Design, analysis and applications. *Journal of Shanghai Jiaotong University* 2. 14–21.

Grenoble, Lenore A. 1998. *Deixis and information packaging in Russian discourse* (Pragmatics & Beyond New Series 50). Amsterdam & Philadelphia: John Benjamins Publishing.

Gries, Stefan Th. & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard target: Two suggestions. *Corpora* 9(1). 109–136.

Guzmán, Alexandria E. & Celia M. Klin. 2000. Maintaining global coherence in reading: The role of sentence boundaries. *Memory & Cognition* 28(5). 722–730.

Hall, Mark. 1998. *Correlation-based feature subset selection for machine learning*. Hamilton, New Zealand: University of Waikato PhD thesis.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorer Newsletter* 11(1). 10–18. DOI:10.1145/1656274.1656278

Hansen-Schirra, Silvia. 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* 12. 133–162.

Hatim, Basil & Ian Mason. 1990. *Discourse and the translator*. London & New York: Longman.

Hatim, Basil & Ian Mason. 2005. *The translator as communicator*. London/New York: Routledge.

Hinkel, Eli. 2001. Matters of cohesion in L2 academic texts. *Applied Language Learning* 12(2). 111–132.

Kachroo, Balkrishan. 1984. Textual cohesion and translation. *Méta: Journal des traducteurs* 29(2). 128–134.

Kiss, Tibor & Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32(4). 485–525.

Koppel, Moshe, Shlomo Argamon & Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4). 401–412.

Kutuzov, Andrey & Maria Kunilovskaya. 2014. Russian learner translator corpus. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.), *Text, speech and dialogue* (Lecture Notes in Computer Science 8655 8655), 315–323. Springer International Publishing. DOI:10.1007/978-3-319-10816-2_39

Kutuzov, Andrey & Maria Kunilovskaya. 2015. A quantitative study of translational Russian (based on a translational learner corpus). In *Proceedings of Corpus Linguistics 2015 Conference*, 33–40. Saint Petersburg State University.

Kutuzov, Andrey & Elizaveta Kuzmenko. 2015. Comparing neural lexical models of a classic national corpus and a web corpus: The case for Russian. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text process-*

ing (Lecture Notes in Computer Science 9041), 47–58. Springer International Publishing. DOI:10.1007/978-3-319-18111-0_4

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta: Journal des traducteursMeta:/Translators' Journal* 43(4). 557–570.

Meyer, Thomas & Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, 129–138. Association for Computational Linguistics.

Mikhailov, Mikhail. 2003. *Parallel'nye korpusa xudo estvennyx tekstov: Principy sostavlenija i vozmožnosti primenenija v lingvističeskix i perevodovedcheskix issledovanijax*. University of Tampere PhD thesis.

Olohan, Maeve. 2001. Spelling out the optionals in translation: A corpus study. *UCREL Technical Papers* 13. 423–432.

Pastor, G. Corpas, Ruslan Mitkov, Naveed Afzal & Viktor Pekar. 2008. Translation universals: Do they exist? A corpus-based NLP study of convergence and simplification. In *8th AMTA conference*, 75–81.

Quinlan, J. Ross. 1987. Simplifying decision trees. *International journal of man-machine studies* 27(3). 221–234.

Ramm, Wiebke. 2006. Dispensing with subordination in Translation–Consequences on discourse structure. In Torgrim Solstad, Atle Grønn & Dag Haug (eds.), *A festschrift for Kjell Johan Sæbø*, 121–136. Oslo: University of Oslo.

Rayson, Paul, Xiaolan Xu, Jian Xiao, Anthony Wong & Qi Yuan. 2008. Quantitative analysis of translation revision: Contrastive corpus research on native English and Chinese translationese. In *XVIII fit world congress*.

Segalovich, Ilya. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, 273–280.

Solfjeld, Kåre. 2008. Sentence splitting and discourse structure in translations. *Languages in Contrast* 8(1). 21–46.

Štěpánek, Jan & Petr Pajas. 2010. Querying diverse treebanks in a uniform way. *International Journal of Learner Corpus Research* 1(1). 1828–1835.

Unger, Christoph. 2011. Exploring the borderline between procedural encoding and pragmatic inference. In, vol. 25, 103. Leiden: Brill.

van Dijk, Teun A. 1976. Philosophy of action and theory of narrative. *Poetics* 5(4). 287–338.

van Halteren, Hans. 2007. Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP)* 4(1). 1.

Xiao, Richard, Lianzhen He & Ming Yue. 2010. In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In Richard Xiao (ed.), *Using corpora in contrastive and translation studies*, 182–214. Cambridge: Cambridge Scholars Publishing.

Zanettin, Federico. 2013. Corpus methods for descriptive translation studies. *Procedia-Social and Behavioral Sciences* 95. 20–32.