Functional Text Representations for Building Cross-Linguistic Comparable Corpora in English and Russian

Maria Kunilovskaya
University of Wolverhampton
maria.kunilovskaya@wlv.ac.uk

Shiva Taslimipoor
University of Wolverhampton
shiva.taslimi@wlv.ac.uk

Tatyana Ilyushchenya
University of Tyumen
t.a.ilyushhenya@utmn.ru

Abstract

In this contribution we report the results on cross-linguistic building of functionally comparable corpora. Functional similarity of corpus resources is an important prerequisite for translationese studies, which traditionally reveal translations as texts deviating from the conventions of the intended genre in the target language. Therefore, measuring translationese is directly contingent on the corpus of non-translated target language selected to represent the expected norm for a given genre. Functional similarity of the corpora is also key for contrastive analysis. We propose a solution based on representing texts with functional vectors and comparing texts on these representations. The vectors are produced by a recurrent neural network model trained on the hand-annotated data in English and Russian from the Functional Text Dimensions project. Our results are verified by an independent annotation experiment, and the tests run on an evaluation corpus. The latter experiments are set to investigate whether the vectors capture traditionally recognised genres and the expected cross-linguistic degree of text similarity. We apply this approach to describe the functional similarity of the 1.5 million token English and Russian subsets of the respective hundred-million word Aranea corpora, the comparable web-corpora project.

1 Introduction

Corpus-based translation studies and contrastive analysis typically require intra- and inter-linguistically comparable corpora. The comparability of the resources is usually ensured by collecting texts from similar sources (e.g. the same institutions, websites, or corpora), and by using the same chronological and sociolinguistic sampling frame. Alternatively, researchers can rely on the pre-existing register/genre annotation. Sometimes, the description of the resources comparability is limited to a phrase such as 'the BNC sample was chosen so as to mirror the makeup of the TEC' or 'reference corpus made comparable to the parallel data in terms of register'. The assumed comparability of monolingual and cross linguistic resources is typically a point of criticism. For example, in his overview of research on explicitation, Becher (2011) questions the comparability of materials used in numerous cases. The importance of building an adequate reference corpus is also reflected in the fact that some corpora (like CroCo) that are designed for translationese or contrastive research, include the untranslated reference texts as their integral part (Hansen-Schirra et al., 2012). It is a well known fact that different registers/genres trigger different type of translationese: Lapshinova-Koltunski (2017) shows that register is one of the major factors explaining variation in translation along with translation method and expertise. Neumann (2013) revealed the specificity of German-English translations observed in some registers but not others.

The above demonstrates that the concept of corpus comparability in translation studies or contrastive analysis is not based on the domain or 'aboutness' of the texts, but has to do more with the 'context of situation'. It is the interplay of various parameters of the communication event that are important for defining genres. Therefore, despite most research in corpus comparability defining comparable corpora as texts in the same topic domain — e.g. they are harvested on a set of seed terms (Kilgarriff et al., 2011); comparability is calculated based on the lexical features, such as vocabulary overlap or bag-of-words representations (Li et al., 2018). This research interprets comparability as a functional or genre-related property, similarly to how bilingual comparable corpus is described in Kutuzov et al. (2016), or how it is traditionally defined in corpus-based translation studies (Zanettin, 2012).

This paper aims to test whether abstract and language-independent functional properties of texts can be used as a text-external approach to cross-lingual text categorisation. Namely, we explore the usability of the Functional Text Dimensions, a set of text functions hand-annotated for English and Russian web texts (Sharoff, 2018), as a training data to produce vectorised representations of texts functionality. Text functions, which reflect the speaker's communicative goal, are among of the major descriptors of a communicative event and are invariably present in the genre definition. Besides, it is one translationally relevant aspect of texts that can be used to build cross-lingual comparable resources for translationese studies.

In addition to the intrinsic evaluation of the models' performance, we provide results of the external evaluation in two aspects. First, we evaluate the effectiveness of the functional vectors for genre classification against alternative text representations. For these purposes we use a selection of 'known' genres extracted from the national corpora in the two languages. Second, the cross-linguistic comparability of the models' output is tested by measuring the average functional similarity of text pairs coming from subcorpora with varying degrees of similarity.

The rest of the paper is structured as follows. Section 2 outlines the research on genre identification and text functionality that we draw upon. In Section 3 we describe our training data, the settings of the modelling experiment, including the architecture of the recurrent neural network model, as well as experimental results. Then, we predict functional vectors for our evaluation corpora and estimate these vectors against expected standard in Section 4. Section 5 has a brief description of the application of the functional vectors to the description of the English and Russian samples of the Aranea web-corpora. The final section (Section 6) summarises the results.

2 Related Research: Register Studies

Apart from the domain-based text categorisation typical for NLP tasks, there are two major approaches to describe text variation in register/genre studies. The text-*internal* approach to text categorisation is based on calculating frequencies of lexicogrammatic features ('register' features, such as conjunctions, passives, modals, pronouns, tenses), that allegedly reflect linguistically relevant parameters of the communicative situations. One of the best known implementations of this approach is Biber's work (Biber, 1988).

The text-*external* approach draws on the audience's perception of the author's communicative aims and known circumstances of the text production (the author's social role, mode of speech, degree of the participants' interaction), and uses genres as a loose set of culture-specific categories to explain text variation.

There is no arguing that these views are complementary. Calculating frequencies of tokens (lexis-based catagorisation typical for domain-oriented approach), can be as effective in genre classification as the more elaborate register features. Xiao and McEnery (2005) show that keywords analysis can be as effective in detecting both similar (everyday conversation vs official speech) and distant genres (spoken genres vs. academic prose) as Biber's features.

There have been also numerous attempts to establish a link between genres and their linguistic features, while ignoring domain differences inside genre categories (including Lee and Myaeng (2002) and Braslavski (2010)). However, the researchers have to use a pre-existing genre typology, which pigeonholes texts in accordance with the accepted convention in the given language community, and does not allow for a more flexible and realistic reflection of the evolving text-type variety or for reliable cross-linguistic comparisons. Moreover, simple solutions, which work for the major text categories, fail in the presence of more subtle distinctions. For example, we have found that the impressive and reproducible results from Lijffijt and Nevalainen (2017), where they achieved F1 = 90% in the classification of the four 'tried and tested' top-level categories from BNC using pairs of the simple register features like frequencies of nouns and pronouns, gets reduced to only F1 = 71% on a less balanced six categories subcorpus described in Kunilovskaya and Sharoff

(2019).

One approach to avoid the atomic genre labels and to work around the culture-specific nature of genre categorisation is to represent texts as vectors in a multi-functional space, where a text can get relative scores on several dimensions. An attempt to define such a space using the speakers' perception of the candidate text proximity to the recognizable functional prototypes was made in the Functional Text Dimensions (FTD) annotation project presented in detail in Sharoff (2018). This framework is particularly appealing for our purposes because it is based on translationally relevant functional properties of texts and offers a theoretically reasonable tertium comparationis required for cross-linguistic corpus building.

3 Modelling Functional Text Representations

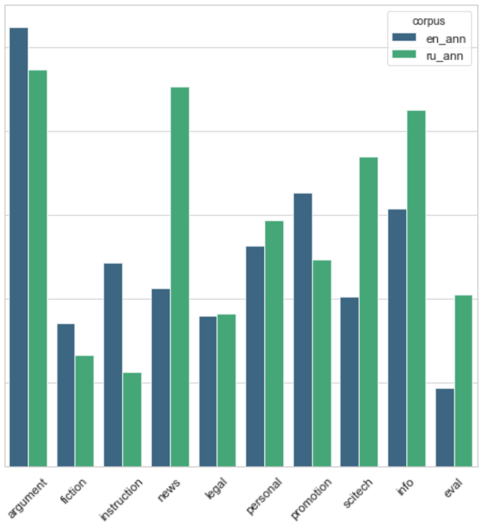
In this section we describe the annotated data in English and Russian from the FTD project and illustrate the neural network approach that we used for modelling text functionality.

3.1 Data

The annotated data consists of 1,624 chunks of texts that count about 2 million tokens for English; the Russian part of the project includes 1,930 texts (2.4 million tokens). For both languages the texts come from two sources: 5gthe Pentaglossal corpus (Forsyth and Sharoff, 2014) and ukWac (Baroni et al., 2009). We used the annotations for the ten most prominent FTD described in Sharoff (2018). Sharoff (2018) also has a detailed description of the original annotation experiment and reports the inter-annotator agreement at Krippendorff's alpha >.76. The annotators were asked to score each text on the 4-point Likert scale (0, 0.5, 1, 2) depending on how much the text resembles the suggested functional prototype for each dimension. While referring the reader to the original paper for more details, a few examples of the labels and prototype texts used in the annotation project are: A1 (argument) blogs, editorials, opinions; A7 (instruction) tutorials or FAQ; A8 (hardnews) report of events, inc. future events.

The original dataset was augmented by splitting longer texts into additional instances. The text length used for training was set to 1000 words. This re-sampling helped the distribution of the FTD labels to be more even than in the original dataset and normalized the text length. Figure 1 depicts the distribution of texts assigned to different functions by annotators for both English and Russian. The corpora of the two languages follow the same building frame: the data come from the same multilingual resources, which makes it possible to assume their comparability.

Figure 1. How often each function receives a positive score in the annotated data (proportion to all texts)



3.2 Model

Our model is a bidirectional LSTM with an attention layer on top. The input to the model are the embeddings of the words in the text, pretrained on the Common Crawl data from the fastText project (Grave et al., 2018). The output is a 10-dimensional functional vector for each document, namely, the functional representation of the document. In this experiment the model was set up to recognise the functions manifested in the text, rather than learn the scores assigned by the annotators. To this end, the annotations for each FTD were binarised (0.5 is set to 0, and 2 is set to 1). Consequently, we had a multi-hot vector for each document as our target.

As a simple baseline, we used a classifier which attempted to learn the binarised values for each FTD separately. In addition we set up a multitask learning scenario in which the model learns all 10 binary values simultaneously. In this case, our learning model back-propagated based on the accumulation of the loss functions for all 10 labels. In another experiment we enriched the embedding features with the Biber's register features of the documents. For extraction of these features we relied on MAT for English (Nini, 2015), and the framework provided in Katinskaya and Sharoff (2015) for Russian.

3.3 Results

For all our models, we used CuDNN LSTM and trained the model for 20 epochs. We used Adam optimizer and 0.2 dropout after embedding and 0.5 dropout after the LSTM layer. The loss function was binary cross entropy since we predict binarised valued of FTD columns.

In Table 1 we report the performance measures on the 10-fold cross validation for the main models (biLSTMa), and the models which use the Biber's features together with embeddings (biLSTMa-bib) – both in multi-task settings, compared to the baseline (biLSTMa-10b) for both languages.

We evaluate the performance of the models in predicting FTD values in three ways: first, we established how well they predict individual functions on average. Second, given the sharp imbalances between the positive and negative classes for each function, we report the F1 measure for the minority class. Finally, the last two columns in Table 1 have the F1 and accuracy statistics for individual samples. Since our target is a multi-hot 10-component vector, accuracy (which counts an observation as correctly predicted only if all the 10 classes are correctly classified) is very strict. Instead, we opt for the negative hamming loss, i.e. the ratio of all correctly predicted labels for an instance to all labels. To deal with severe class imbalances, we use stratified (multi-label) split with cross-validation and at the evaluation stage by we choose macro-averaging which penalises model errors regardless of class distributions. The results in Table 1 show that our multi-task architectures (rows indicated with multi) outperform the baseline in which the 10 values for each text are learned independent of each other. The better performance of adding register features to our model can be seen only in the case of Russian.

These results show the effectiveness of our models in estimating the probability that a text fulfills the corresponding functions. and leads us to further use these vectors as functional representations for text. In the next two sections we first demonstrate how the predictable functional representations correlate with the text's general functional and genre properties which are external to the initial annotation experiment. We also demonstrate the application of the functional vectors for corpus comparison in Section 5.

4 External Evaluation

4.1 Functional vectors for BNC/RNC genre categories

To determine whether our functional representations are useful to distinguish the text categories outside the original annotated corpora, we designed a 'known' genre composition corpus. To compile this genre evaluation corpus, we used the metadata in the British National Corpus (BNC) and the Russian National Corpus (RNC), described in Lee (2001) and Savchuk (2006) respectively. We focused on the genre categories which approximate some of the prototype texts described in the annotation guidelines and are annotated in both national corpora. We extracted the written texts that were longer that 400 words by the tags (in the order shown in Table 2).

For English we extracted all texts tagged as follows: *ac:nat science, fict prose, nonAc: nat science, newsp brdsht nat: report, newsp other: report, biography, advert*. For Russian the texts for each category were selected by the tags combinations: academic (sphere=science and education, type=article, topic=science and technology, audience != big, level=professional/high), fiction (sphere=fiction, type=short story, story, novel), reportage (sphere=publicist, type=info message), personal (sphere=publicist, type=memoirs/biography), promotion (sphere=promotion). [1] Unlike the BNC, the RNC has no separate text type for non-academic texts. To remedy this incompatibility, we used chapters from 14 popular Russian scientific books in academic domains such as linguistics, biology and anthropology published between 2010 and 2017. The books were split into 1000-word chunks; a random selection of 100 of those chunks was used as part of the Russian evaluation corpus. The resulting collections were balanced in terms of the number of texts per each category: we retained 100 random documents for

[1] For Russian we additionally limited the sampling frame to include only the texts published after 2004, neutral of style and marked as intended for a large audience, with no restriction by age or education level (with the exception for academic texts).

		FTD overall			FTD minority	Samples	
		P	R	F1	F1	F1	hamming loss
baseline	biLSTMa-10b (EN)	.810	.853	.824	.655	.683	.927
	biLSTMa-10b (RU)	.799	.878	.825	.644	.709	.924
multi	biLSTMa (EN)	.824	.862	.841	.722	.483	.930
	biLSTMa (RU)	.818	.871	.841	.724	.504	.922
multi+bib	biLSTMa-bib (EN)	.814	.861	.835	.711	.472	.927
	biLSTMa-bib (RU)	.829	.875	.849	.742	.522	.926

Table 1. Results for FTD modelling experiments

each category that counted more than 100 texts. We further truncated the texts to the first 1000 words, if the selected texts were longer. Table 2 shows the basic parameters of the BNC and RNC subcorpora used for evaluation in this study. We also include the FTD which is expected to be dominant in the texts of each category.

To investigate the reliability of the predicted functional vectors for genre analysis, we classified the texts into six categories that are listed in Table 2. We report the results received in the same settings for the alternative text representations, namely, the raw Biber's features and the keywords statistics.

The Biber's features were extracted with MAT and MDRus analyzer (Nini, 2015; Katinskaya and Sharoff, 2015). The keywords features for each text were calculated using the log likelihood (LL) measure against all of the data from the respective national corpus used in this experiment. Prior to keyword extraction the data was lemmatized, and functional words were filtered out, leaving us with a vocabulary of 24k and 40k content lemmas for English and Russian respectively. To reduce the sparsity of the data, we limited the list of keywords to the top 100 with LL >6.63 (the standard 1% significance level) for each text and, further, to only those which occurred in 3% of the texts. The number of the resulting keywords was 408 for the BNC selection and 489 for the RNC.

In Table 3 we report the macro-average stratified 10-fold cross-validation results for a RandomForest classifier with the default scikit-learn parameters (n_estimators=10, criterion='gini', bootstrap=True) (Breiman, 2001). The results show that the classification using functional representations (vectors) outperforms the classification using alternative ways of representation, given our selection of genres and the classification settings. Interestingly, the combination of functional and the Biber's features yield a 2% increase in the performance of the classifier.

It can be seen that the results for English in this experiment were consistently better than those of Russian. In-depth analysis of the classifier performance per category (omitted here for brevity) showed that the algorithm struggled with different genres for different languages. In Russian *reportage* proved to be the most challenging genre (the F1 score for predictions of items in this category was 0.56), while *fiction* returned the highest results (F1 = 0.79); In English non-academic texts were comparatively difficult to solve (F1 = 0.63), while reportage and promotion were comparatively easily recognised (F1 = 0.87 and F1 = 0.83).

4.2 Cross-linguistic comparability of English and Russian functional vectors

In this section we test whether the texts in English and Russian that are expected to be functionally similar in the real world receive similar functional representations in our experiments. We have seen that our predicted vectors are able to detect the generic properties of texts reflected in the handcrafted text category metadata of the two national corpora. However, it is not clear whether the vectors produced by the models learnt on the English and Russian data are effective in measuring their cross-lingual comparability.

To explore this aspect of the functional representations, we measured and compared similarity between the four sets of text pairings which are expected to display decreasing degrees of functional similarity: (1) aligned parallel texts of the four genres; (2) texts from the same genres in the parallel corpus that are not translations of each other; (3) random text pairs from the comparable categories of the national corpora (described in Table 2) and (4) random text pairs for texts from

		academic	fiction	nonac	pers	promo	rep	total
BNC	texts	43	100	62	100	59	88	452
	words	38k	86k	56k	88k	47k	79k	394k
RNC	texts	100	100	100	100	100	82	582
	words	92k	94k	104k	94k	75k	52k	k 511k
FTDs		A14	A4	A1	A8	A11	A12	

Table 2. The composition of the genre-balanced comparable evaluation corpus

	EN	RU
vectors	.77	.68
Biber's	.73	.64
keywords	.66	.60
vectors+Biber's	.79	.70

Table 3. Classification results for the six categories in each national corpus selection

	texts	words
fiction	170	9.6m
media	132	133k
ted	100	259k
popsci	100	826k

Table 4. Composition of the parallel component of the evaluation corpus

different genres in the national corpora (the negative similarity material). We used the Euclidean measure of similarity to compare the functional vectors between the text pairs in each set. This measure takes into account the magnitudes of the vectors components, which are meaningful in our representation.

The highest degree of the expected cross-lingual functional similarity is represented by the professional translations and their sources. Within the functional theories of translation, which underlie the current professional norm, good translations are expected to reproduce the functional hierarchy of the source. The texts in this section of our evaluation corpus are extracted from the parallel English-Russian component of the RNC[2] (fiction, mass-media texts) (Dobrovolskij et al., 2005) and from the professional translations segment of the RusLTC project[3] (including TED talks and popular scientific texts) (Kutuzov and Kunilovskaya, 2014). The descriptive statistics for the parallel component of the evaluation corpus are given in Table 4 (the word count is based on the English sources).

In Table 5 we report average pair-wise similarity for the documents of different categories in the four aforementioned sets. As can be seen in the table, as the level of comparability decreases from the top set to the bottom set, the calculated similarity also decreases.

The fluctuation in the similarity values for the different genres (see Table 5) indicates that the translations of popular-scientific books are the most functionally faithful (see the results for aligned translations). In the comparable part of the evaluation corpus the academic texts demonstrate the highest similarity of 0.396, while the least functionally similar genres are non-academic (popular-scientific) texts (0.127), personal writings, such as biographies and memoirs (0.139), and promotional texts (0.145).

4.3 Results from an independent annotation effort

To test the generalisation power of the models on external data, we ran an independent annotation experiment, following the guidelines described in Sharoff (2018) and summarised in Section 3.1. We had three trained linguists assign 10 functional scores to each of 70 English texts selected randomly from two parallel collections: CroCo (Hansen-Schirra et al., 2006) and the RusLTC corpus mentioned in Section 4.2. After discussing the results on the first 10 texts, which revealed some differences in the task interpretation, the three raters reached the overall agreement of Krippendorff's $\alpha = 0.537$. In cases of triple disagreements (48 items out of 700), the three different values assigned by annotators were averaged and rounded to the closest score.

We used the same evaluation strategy as in Section 3.3 and compared the binarised human scores

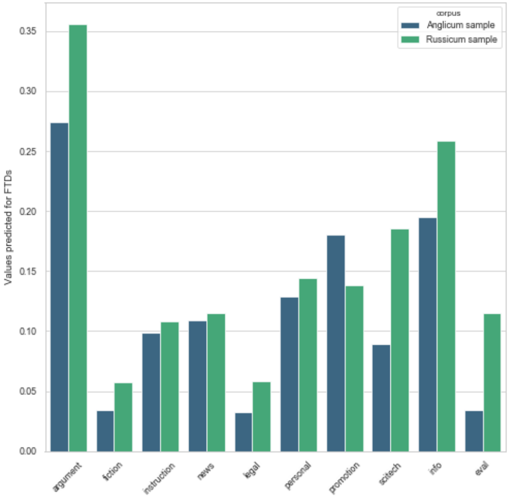
[2] http://www.ruscorpora.ru/new/corpora-structure.html

[3] https://www.rus-ltc.org/static/html/about.html

set	category	similarity	mean
aligned	fiction	.432	.470
	media	.476	
	ted	.456	
	pop-sci	.514	
unrelated	fiction	.315	.305
	media	.263	
	ted	.323	
	pop-sci	.317	
same genre	academic	.396	.214
	fiction	.259	
	non-academic	.127	
	personal	.139	
	promotion	.145	
	reportage	.216	
unrelated	academic::fict	-.190	.004
	non-ac::promo	.116	
	pers::report	.085	

Table 5. Average functional similarity measures for the four components of the evaluation corpus

Figure 2. The overview of the texts functionality based on the average values for FTDs: Anglicum Minus vs Russicum Minus



and the English model predictions. The average result over the 10 FTDs reached macro F1 score of 0.732. The lower performance of the model on this data can be explained by a different distribution of the text types in the annotated 70 texts as compared to the training set.

5 Application to Aranea

In this section we demonstrate the application of the vectors to analyse the genre composition of comparable web-corpora from the Aranea project. We randomly selected 1% of texts from the 120-million token Araneum Minus Anglicum and Araneum Minus Russicum (Benko, 2014)[4] and represented them with functional vectors. The samples include around 4.5k texts that are over 450 tokens long, and count 1 and 1.5 million tokens, respectively. Since there is no prior information on the internal generic structure of the corpora we can not measure their overall similarity directly (as we did with the national corpora). The purpose of this exercise is only to provide a comparative description of the corpora genre composition.

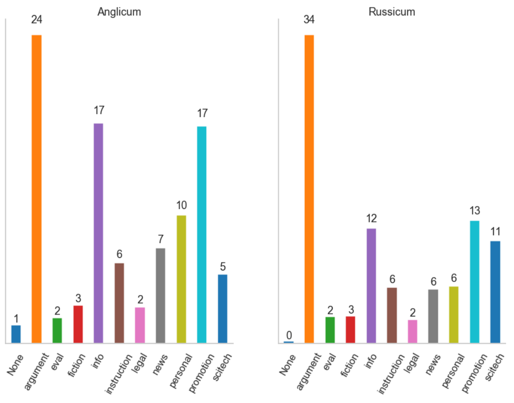
We tried two ways of capturing the contents of the corpora: 1) the average value for each function and 2) the ratio of the texts with a given function as the predicted dominant function. The dominant function is defined as the FTD with the highest probability value returned by the model for a given text. In either case the general picture comes at the price of losing the functional hierarchy and the possible hybrid nature of texts. From this perspective the Russian texts in the analysed slices of the Aranea web-corpora have higher scores for evaluative, informative and argumentative functions (see Figure 2).

In the second approach, for each text we used only the highest functional value (i.e. their dominant function) and characterised the corpora by the ratios of texts with these dominant functions. Figure 3 shows that the Russian corpus (compared to the English one) has fewer texts with the informational and promotional as dominant functions, but it has more texts that come across as primarily scientific.

The ratios seem to be more directly comparable than the averaged probabilities, but they neglect the polyfunctional nature of many texts in Aranea: 40% of the English texts and 70% of the Russian texts have the second strong prediction (we set the threshold for the ratio between the highest value in a functional vector and the second high value at 0.7). These numbers reflect the proportion of hybrid texts in the training corpus: The human subjects assigned high scores to two (or more) functions in 40% of texts in the English part of the experiment and in 53% of texts in Russian.

[4] http://unesco.uniba.sk/aranea/index.html

Figure 3. Ratio of texts per predicted dominant function for Anglicum Minus and Russicum Minus



6 Conclusions

This paper reports the experimental results on learning functional text representations for English and Russian and describes extensive tests on their cross-linguistic comparability. We used the hand-annotated data released within the Functional Text Dimensions project to train a multi-label binary classifier based on recurrent neural networks. The average performance of the classifier is estimated at F1 >0.84 for both languages.

We evaluated the quality of the functional vectors by using them to represent texts from the six comparable text categories of the British and Russian national corpora and running a simple RandomForest classifier on the resulting data. The six-class classification returned the F1-score of 0.77 and 0.68 for English and Russian respectively. This outperformed the classification results in the same settings with the alternative representations (the Biber's and the keywords features). We saw a steady increase in the quality of the genre classification when we combined our functional vectors with Biber's features.

To evaluate the cross-linguistic comparability of the models output, we measured the Euclidean similarity between text pairs with the expected various degrees of similarity. The functional vectors learnt independently by the English and Russian models for the translationally related text pairs returned the highest similarity score of over 0.45. It is a relative score which can be interpreted in the context of the scores for the text pairs that were expected to be less similar. For example, the most dissimilar text pairs – English and Russian texts from categories with different genre labels returned 0.04. These experiments show that the functional vectors are an adequate representation of the texts functionality, a major criteria for genre identification, and can be used for measuring similarity between texts in the two languages as well as for building bilingual comparable corpora.

Acknowledgments

This work has been partly supported by the Russian Foundation for Basic Research within Project No. 17-06-00107/19.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3):209–226.

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. http://ediss.sub.uni-hamburg.de/volltexte/2011/5321/pdf/Dissertation.pdf.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) web corpora. In *International Conference on Text, Speech, and Dialogue*. Springer, pages 247–256.

Douglas Biber. 1988. *Variations Across Speech and Writing*. Cambridge University Press.

Pavel Braslavski. 2010. Marrying relevance and genre rankings: an exploratory study. In *Genres on the Web*, Springer, pages 191–208.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

Dmitrii Dobrovolskij, Aleksei Kretov, and Sergei Sharov. 2005. Parallel corpus: the architecture and usage potentiona [Korpus parallel'nyh tekstov: arhitektura i vozmozhnosti ispol'zovanija]. In *Russian National Corpus: 20032005 [Nacional'nyj korpus russkogo jazyka: 20032005]*, Indrik, pages 45–49. http://ruscorpora.ru/sbornik2005/17dobrovolsky.pdf.

Richard Forsyth and Serge Sharoff. 2014. Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing* 29:6–22.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proc 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing at EACL*. Association for Computational Linguistics, Trento, pages 35–42.

Silvia Hansen-Schirra, Erich Steiner, Sandra Hansen, Marlene Kast, Kerstin Kunz, Karin Maksymski, and Mihaela Vela. 2012. *Cross-Linguistic Corpora for the Study of Translations*.

Anisya Katinskaya and Serge Sharoff. 2015. Applying Multi-Dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres. *The 5th Workshop on Balto-Slavic Natural Language Processing* (September):65–74. http://www.aclweb.org/anthology/W15-5311.

Adam Kilgarriff, P V S Avinesh, and Jan Pomikálek. 2011. Comparable Corpora BootCaT. *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011, 10-12 November 2011, Bled, Slovenia* .

Maria Kunilovskaya and Serge Sharoff. 2019. Towards functionally similar corpus resources for translation. In *Proceedings of RANLP 2019*. in print.

Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko, and Lyubov Ivanova. 2016. Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*. pages 3–10. http://arxiv.org/abs/1604.05372.

Andrey Kutuzov and Maria Kunilovskaya. 2014. Russian Learner Translator Corpus: Design, Research Potential and Applications. In *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014, Proceedings*. Springer, volume 8655, page 315.

Ekaterina Lapshinova-Koltunski. 2017. Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. *Empirical Translation Studies. New Theoretical and Methodological Traditions* pages 207–234. https://doi.org/10.1515/9783110459586-008.

David Lee. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3):37–72.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 145–150.

Bo Li, Eric Gaussier, and Dan Yang. 2018. Measuring bilingual corpus comparability. *Natural Language Engineering* 24(4):523–549. https://doi.org/10.1017/S1351324917000481.

Jefrey Lijffijt and Terttu Nevalainen. 2017. A simple model for recognizing core genres in the bnc. In *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, University of Helsinki, VARIENG eSeries, volume 19.

Stella Neumann. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German*. Mouton de Gruyter, Berlin, Boston.

Andrea Nini. 2015. Multidimensional Analysis Tagger (v. 1.3).

Svetlana Savchuk. 2006. The corpus of texts from the first half of the 20th century: status quo and perspectives [korpus tekstov pervoj poloviny xx veka: tekushhee sostojanie i perspektivy]. *Russian National Corpus[Nacional'nyj korpus russkogo jazyka]* 2008:27–45.

Serge Sharoff. 2018. Functional Text Dimensions for annotation of Web corpora. *Corpora* 13(1):65–95.

Zhonghua Xiao and Anthony McEnery. 2005. Two Approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics* 33:62–82. https://doi.org/10.1177/0075424204273957.

Federico Zanettin. 2012. Bilingual Comparable Corpora and the Training of Translators. *Meta: Journal des traducteurs* 43(4):616.