Towards Functionally Similar Corpus Resources for Translation

Maria Kunilovskaya
University of Tyumen
Tyumen, Russia
m.a.kunilovskaya@utmn.ru

Serge Sharoff
University of Leeds
Leeds, UK
s.sharoff@leeds.ac.uk

Abstract

The paper describes a computational approach to produce functionally comparable monolingual corpus resources for translation studies and contrastive analysis. We exploit a text-external approach, based on a set of Functional Text Dimensions to model text functions, so that each text can be represented as a vector in a multidimensional space of text functions. These vectors can be used to find reasonably homogeneous subsets of functionally similar texts across different corpora. Our models for predicting text functions are based on recurrent neural networks and traditional feature-based machine learning approaches. In addition to using the categories of the British National Corpus as our test case, we investigated the functional comparability of the English parts from the two parallel corpora: CroCo (English-German) and RusLTC (English-Russian) and applied our models to define functionally similar clusters in them. Our results show that the Functional Text Dimensions provide a useful description for text categories, while allowing a more flexible representation for texts with hybrid functions.

1 Introduction

Comparable corpora are an important prerequisite for translation studies (TS) and contrastive analysis. One wants to make sure that the corpus resources used to explore differences between languages or aspects of translational specificity in several target languages (TL) are comparable in the first place.

One of the common approaches to corpus comparability is to define it as the domain similarity and to rely on the vocabulary overlap as the measure of comparability. A brief summary of possible interpretations of the concept and comparability measures can be found in Li et al. (2018). The authors give a domain-based definition to cross-linguistically comparable corpora: "*document sets in different languages that cover similar topics*". While lexical similarity is an important factor in linguistic variation, we would argue that it does not capture all the translationally relevant features of texts. Neumann (2013), Kruger and Van Rooy (2010) and Delaere (2015) have also highlighted the importance of register and genre in studying translations by showing that different registers produce different types of translationese. Moreover, functional theories within translation studies (TS) insist that what matters in translation is functional adequacy. The target text (TT) is expected to fulfill the same communicative functions as the source text (ST) and meet the TL conventions expected in the situation of TL communication of the message (Nord, 2006). Both Reiss and Vermeer (1984) and Neubert (1985) build their theory of translation around genres or text types, while Shveitzer (1973) underlines the impact of the text functions hierarchy on the translator's linguistic choices.

The above suggests that translational comparability of corpus resources should take into account social and situational constrains of the communication and the the speaker's purpose along with the text topic. The functional and communicative variation of texts is usually interpreted through the concepts of register and genre. For the purposes of this research, we will accept the distinction between the two suggested by Lee (2001). Register, as a text-**internal** view with respect to text categorization, refers to the lexicogrammatic choices made by the author. This notion reflects the differences in the linguistic make-up of texts and it relies on frequencies of lexicogrammatic features such as passive voice, relative clauses or personal pronouns. It is assumed that the observed linguistic variation captures the possible combinations of field, tenor and mode, the most prominent factors of communication (suggested by Halliday (1985)).

On the other hand, genres are understood as conventionally recognizable text categories that can be established on a number of **external** criteria, referring to the function of the text and its situational constrains. According to Lee (2001), most existing corpora rely on the text-external approach to text categorization and the choice of parameters behind it is guided by practical considerations in each case. It has been shown how little comparability there is between the genre classifications used to annotate different corpora (Sharoff, 2018). TS researchers interested in register variation find that existing corpora provide "limited background information on the genres ... and how they were defined" and choose to set up annotation tasks to reorganize the corpora (Delaere, 2015).

One translationally relevant common footing to compare texts from corpora with divergent or absent genre annotation is to rely on their function. On the one hand, text function is an important factor in translation, as texts aimed at informing the reader are translated differently from texts aimed at promoting goods and services (Lapshinova-Koltunski and Vela, 2015). On the other hand, text functions can be used to produce continuous rather than discrete text descriptions and account for hybrid texts. In this research we explore the potential of Functional Text Dimensions (FTD), hand-annotated for English and Russian (Sharoff, 2018) to produce text representations and to build functionally comparable corpora for TS research.

The aim of the present study is solve a practical task of creating research corpora for the study of translational tendencies in English-German and English-Russian translation. To this end, we develop a method to build a reasonably big and functionally homogeneous intersection of the three text collections: CroCo, an English-German parallel corpus (Hansen-Schirra et al., 2006), and the students and professional collections from RusLTC, a English-Russian parallel corpus (Kutuzov and Kunilovskaya, 2014). Our major motivation for this research is to find a way to reconcile the diverging genre annotations that exist in these corpora (see Table 4). We want to reduce the probability that the differences observed in the translations are down to the differences between the sources and are not genuine translational or cultural effects.

The rest of the paper is structured as follows. Section 2 has a brief overview of the topological approach to the text characterization and the research related to corpus similarity and genre classification. In Section 3 we describe our approaches to FTDs modelling and report the results of the intrinsic evaluation of the models. A selection of BNC genres is used to evaluate the models against an independent judgment and to test the clustering approaches to be used in the real-life task (Section 4). Section 5 presents a study that showcases the application of the functional vectors to computing the most similar parts of the two corpora. In Section 6 we aggregate the analytic results and highlight important findings.

2 Related Research

The practical needs to describe and compare corpora have made 'corpusometry' a prominent area of research in corpus linguistics. Below we outline the two major approaches to measuring similarity and describing the corpora contents. The first one is based on lexical features and yields a thematic description of corpus texts. It is one of the most prominent methods of measuring similarity between texts and/or building comparable corpora. For example, Kilgarriff and Salkie (1996) put forward a corpus homogeneity/similarity measure based on calculating χ^2 statistics from frequency lists or N keywords. A lexical approach to estimate the corpus composition is taken by Sharoff (2013). This research compared the results of clustering and topic modelling as ways to represent a corpus content using keywords statistics. In Li et al. (2018), the authors compared the performance of several bilingual vocabulary overlap measures on a specifically designed corpus with known comparability levels and found that frequencies of words with a simple Presence/Absence weighting scheme outperformed other approaches.

Another approach to measuring corpora has to do with calculating frequencies of a range of lexicogrammatic features (tense forms, modals) that allegedly reflect linguistically relevant parameters of the communicative situations. This text-internal approach to the text categorization can be best exemplified by Biber's work (Biber, 1988). He used several dozens of hand-picked text-internal features to place a text along each of the six dimensions (e.g. involved vs informational production or abstract vs non-abstract information). Biber's multidimensional approach to describing text variation has been criticized for lack of interpretability and, more importantly, for being loosely related to any external and situational factors, which cab be a socially more important reality for text categorization than linguistic features. The latter can throw together texts that are perceived as belonging to different genres (Lee, 2001). The attempts to classify genres, particularly, as annotated in the BNC and limited to a selection of major 'tried and tested' three or four top-level categories, have shown that the 67 Biber's features can be an overkill for a task like that. Lijffijt and Nevalainen (2017) report over 90% classification accuracy for the BNC four major genres on just pairs of surface features (such as frequencies of nouns and pronouns, values of type-to-token ratio and sentence length). The results from Kilgarriff and Salkie (1996); Xiao and McEnery (2005) indicate that the most frequent words can cope with the four major BNC categories as well. More specifically, Xiao and McEnery (2005) show that keyword analysis can be used to replicate Biber's results. In effect they analyze differences in the frequencies of mostly functional words that are key to genre identification. In a setting similar to Biber's, Diwersy et al. (2014) use 29 lexicogrammatic features and mildly-supervised machine learning methods to tease apart genres annotated in CroCo. The visualizations they provide indicate that they have managed to clearly separate only fiction and instruction of the eight genres in their experiment.

This demonstrates that describing genres needs a sophisticated approach that takes into account a multitude of criteria such as topic and situated linguistic properties. This research continues the investigation of the functional aspect of genre shaped in Sharoff's Functional Text Dimensions (Sharoff, 2018). Sharoff's work establishes a text-external framework to capture human perception of the texts functionality (as distance to a functional prototype) and to link it to any text-internal representations, with the aim of predicting the functional setup of unknown texts. This work is particularly relevant to our task for three reasons: (1) it provides a solid theoretically grounded approach for comparing texts coming from different or unknown sources and for producing comparative descriptions for the corpora at hand, (2) it is focused on functional and communicative parameters of texts that are particularly important in TS, (3) this framework, like Biber's, provides a flexible way to represent texts functionality along a few dimensions instead of squeezing texts into the atomic genre labels. In effect, FTD framework is a way to produce functional text vectors that position each individual text in a multidimensional functional space and help to account for variation within and across text categories.

3 Modelling: Setup and Results

The annotated data from the FTD project was used to learn models that predicted 10-dimensional vectors for the English texts in our research corpora. Further on, we used these vectors to compare texts and to get functionally similar subcorpora for a subsequent TS research (Section 5).

The annotated data for English consists of 1624 chunks of texts that count about 2 mln tokens from two different sources: 5gthe Pentaglossal corpus (Forsyth and Sharoff, 2014) and ukWac (Baroni et al., 2009). We used the annotations for the 10 most prominent FTD described in Sharoff (2018). Each dimension received a score on a 4-point Likert scale that reflects the proximity of a text to the suggested functional prototype. The inter-annotator agreement is reported at Krippendorff's α >0.76. We refer the reader to Sharoff (2018) for more details on the FTD framework.

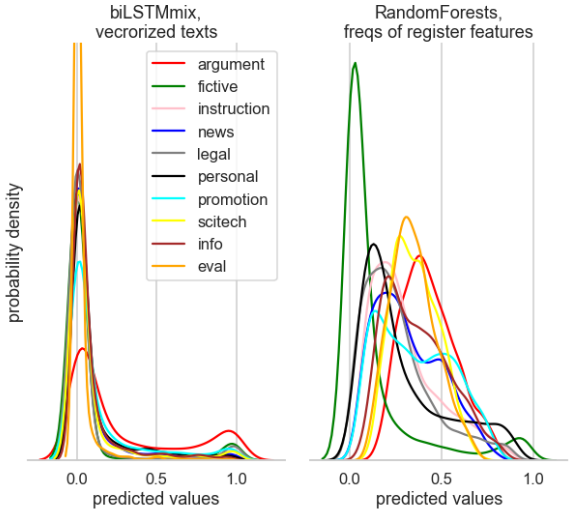
We used two modelling approaches to learn functional vectors from the annotated dataset: a multi-label task in a deep neural network architecture and a set of binary classifiers in a traditional machine learning setting. The respective models produced two types of functional vectors, which demonstrated comparable performance in several evaluation settings. This paper investigates the differences between, and adequacy of, these two types of functional vectors.

In the neural networks scenario, we used a bidirectional LSTM with an attention layer and two types of text input. Firstly, texts received mixed token-PoS representations suggested by Baroni and Bernardini (2006), biLSTMmix throughout this paper and in Table 1). The 1500 most frequent words were kept in their surface form, while the rest of the tokens were converted into their PoS. For example, a sentence "It was published in 1931 by one of New York's major publishers." was transformed into "It was VERB in [#] by one of PROPN PROPN major NOUN." The embeddings for PoS were initialized as random vectors and trained in the Embedding layer. Secondly, we used lemmatised texts, with stop words filtered out (biLSTMlex in Table 1). For both scenarios we used pre-trained word embeddings of size 300, trained on the English Wikipedia and CommonCrawl data, using the skip-gram model, from the WebVectores database (Kutuzov et al., 2017). The preliminary experiments showed that cross entropy as the loss function with the Adam optimizer performed best (Kingma and Ba, 2014). We trained the models for 10 epochs. In the ML case, we reformulated the task as the binary classification task and learnt a classifier for each FTD. To this end, we binarized the existing human annotations by converting '0.5' score to 0 and '2' to 1. To get the real-valued functional vectors we used the probabilities returned for the positive class for each FTD on the assumption that the model would return higher probabilities for texts with a clearer functional makeup. We experimented with features (TF-IDF and Biber's 67 text-internal register features) and with different algorithms (Support vector machines (SVM), RandomForest (RF), Logistic Regression (LogReg)). SVM and RF results below pertain to the experiments with the grid search optimized for macro F1 score. TF-IDF representation proved to be inferior to the Biber's features and was excluded from the results below. We added a dummy classifier which randomly predicts a class with respect to the class distribution as a baseline. For register feature extraction (the Biber's features) we used MAT for English (Nini, 2015).

To use a comparable performance metrics for the two learning approaches, the annotations and the models predictions were transformed into multi-hot vectors.

In Table 1 we report the standard measures averaged over 10 FTDs on the 10-fold cross validation for the six experiments. We accounted for the severe class imbalances in all training settings by using 'class_weight=balanced' option, stratified (multi-label) split with cross-validation and, at the evaluation stage, by choosing macro-averaging which penalizes model errors equally regardless of class distributions.

Figure 1. Distribution of predictions for the modeling approaches



From the statistics in Table 1, it follows that the deep learning approach is more accurate in determining the text functionality than the classical algorithms, and the mixed representations work best.

The difference between the models performance, however, is quite slim: it is in the second decimal digit only. A brief glance at the values of the functional vectors components (i.e. values predicted for each FTD) returned by the models reveals the differences in how the models arrive at the same overall result. Figure 1 shows the probability density for the values produced by the best performing models in each learning setup.

Figure 1 demonstrates that biLSTM, unlike the traditional ML algorithms, tends to predict near-zero values, with up to 7-11% of the training texts receiving values smaller than 0.2 on the strongest 'dominant' dimension.

In the next section we will show how the predictions of these two models correlate with the experiment-independent judgment we can suggest.

4 Evaluation on BNC Categories

4.1 Genre Classification

To test the functional vectors on the data outside the annotated dataset, we constructed a corpus with the 'known' genre composition. To this end, we followed Lee's scheme for the BNC text

	FTD perspective			FTD minority class	Samples perspective	
	P	R	F1	F1	F1	neg_humming_loss
biLSTMmix	.804	.767	**.776**	.609	.640	.902
biLSTMlex	.787	.747	.757	.576	.596	.895
RF	.732	.756	.723	.532	.517	.846
SVM	.667	.531	.510	.095	.059	.844
LogReg	.664	.734	.659	.480	.527	.753
dummy	.504	.504	.504	.169	.134	.737

Table 1. Results of FTD modelling experiments

categories (Lee, 2001) to select the genres that are, in our opinion, most functionally distinct. The genres that use domain as the major underlying principle were deliberately excluded (religious texts, subcategories of academic texts). Table 2 has the basis parameters of the resulting BNC subcorpus. To find out how well the functional vectors reflect the structure of this functionally-motivated BNC subcorpus, we classified the six genres on the functional vectors produced by our best-performing models and compared their performance to several alternative representations: the raw statistics for Biber's features and log likelihood values for the 446 most common keywords. For brevity, in Table 3 we report the macro-averaged 10-fold cross-validated results only for RandomForest. Several other algorithms (SVM, LR) return approximately the same results.

genre	texts	words
academic(sci)	43	1.3M
editorial	12	115K
fiction(prose)	431	19M
instruct	15	492K
non-acad(sci)	62	2.8M
reportage	87	3.6M
Total	650	30M

Table 2. Basis statistics on the six functionally distinct categories from BNC

From Table 3 it follows that the models learnt on Biber's features coped with the selected BNC genres better than any other representations. However, the best performing pair of surface features from Lijffijt and Nevalainen (2017) — frequencies of nouns and pronouns, which return the reproducible result of over 90% accuracy on the four 'tried and tested' registers of English, — fail in the face of the more fine-grained categories.

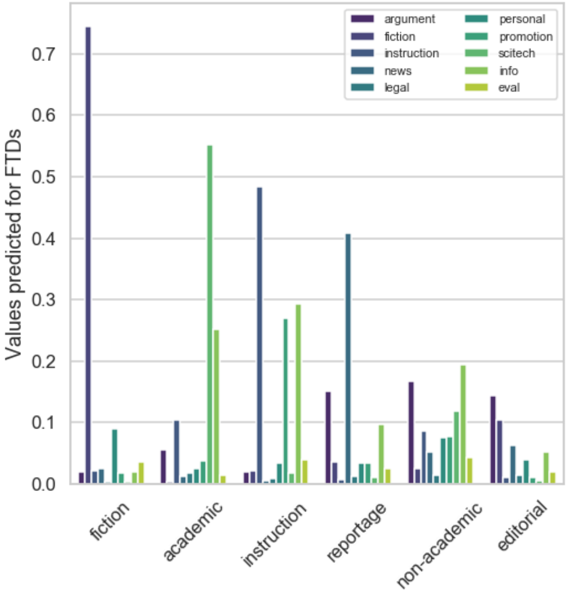
The analysis of the interrelations between the BNC genre labels and the predicted dominant functions suggest that only two categories can be easily mapped to the list of FTDs by all models: fiction and academic texts. The most problematic genres for all models are non-academic writings and editorials. For these texts the models either return no score above the 0.2 threshold on any of the dimensions or similar (relatively low) scores on several dimensions, especially on argumentative, evaluative, informational and personal. Editorials and non-academic texts stand out as functionally hybrid: in Figure 2, which shows the distribution of the FTD values predicted by biLSTM across the six genres, they do not have a functional focus, but integrate several text functions. Their hybrid status is evident from the more uniform distribution of average values for FTDs and from the diversity of text dominants predicted for these genres as well as from the higher percentage of the strong second function in the vectors.

	P	R	F1
biLSTMmix	.84	.75	.79
biLSTMlex	.88	.66	.69
RF	.88	.90	**.89**
Baselines			
67 Biber's features	.93	.88	**.90**
Nouns+Pronouns	.76	.74	.74
keywords	.91	.79	.84

Table 3. BNC classification results

The analysis of the predicted dominant functions against the actual genre labels shows that the best overall fit for BNC is produced by the RandomForest-based model, not the least because it does not produce vectors consisting of very low values only, which results in failure to define texts at all, given the accepted threshold of 0.2 necessary to signal a function. For all genre categories (except editorials) 60-95% of texts can be referred

Figure 2. Average values on the 10 FTDs by BNC genres predicted by biLSTMmix model



to the true genre category following the strongest prediction, if we map genres to FTDs as follows: fiction : fictive, academic : scitech, reportage : news, instruction : instruction, non-academic : argumentative. However, reducing a functional vector to just the strongest component would be unfair to the functionally hybrid texts that fall under the genre labels of non-academic and editorial in our BNC slice.

4.2 Testing the Clustering Method on BNC

The ultimate goal of this work is to produce a functional intersection of two corpora, i.e. to find functionally comparable texts in several sets. In this section we apply two clustering techniques to the BNC selection to determine which text representation and clustering approach is better at matching the annotated genres as class labels.

In the first clustering scenario we ran Affinity Propagation on a square matrix of pair-wise correlations pre-computed as euclidean similarities for the 650 BNC texts. This approach is attractive because it does not require the value of k, which is difficult to deduce in the real application context. We searched through the combinations of parameters to get the highest score for the Adjusted Rand Index (ARI), a clustering metric, which returns the proportion of text pairs that were assigned to the correct cluster, given the gold standard classes. The best clustering solution, with ARI=0.92 and 4 clusters on our BNC selection, was returned for both biLSTMmix and RandomForest vectors at damping=0.9 and preference=-12. The clusters, quite predictably, were built around (1) fiction, (2) reportage (3) non-academic + editorial (4) academic + instructions. Vectors learnt on lemmatized embeddings (biLSTMlex) were not able to converge to this solution.

An alternative clustering technique used in this research was KMeans algorithm with the Elbow Criterion method to determine k. The latter is based on measuring the sum of the squared distances between each member of the cluster and its centroid. k is defined as the number of clusters after which this value drops abruptly. However, this method needs to be applied with regard to the task at hand and some understanding of the data. For BNC k was automatically put at 2, because of the imbalance in our collection towards fiction: more than half of the texts were fiction. The best KMeans result (ARI=0.92) was registered on the RandomForest model vectors for k=5. It was superior to the best biLSTM result in that it separated the instruction cluster. Clustering on Biber's features and keywords did not achieve ARI of more than 0.2 for either Affinity Propagation or KMeans.

5 Case Study: CroCo and RusLTC

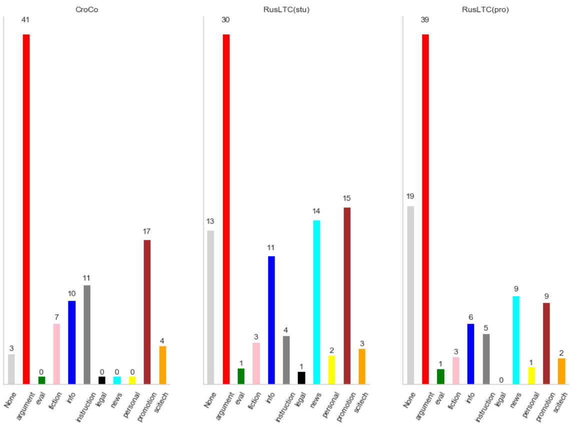
In this section we report the results of a case study where we used the functional vectors to get comparable functional clusters from several text collection. Our data comes from the English parts of the three parallel corpora: RusLTC, including student and professional translations subcorpora that have different English sources, and the CroCo corpus. As can be seen from Table 4, the three text collections vary in size, have diverging genre setup, and there is no way to tell whether the same categories include the same texts. In this work CroCo was chosen as the normative corpus, i.e. the starting point for the comparison and clustering operations.

The first step in solving our practical task with KMeans was to determine k. K-value was identified as n+1, where n is the number of the most populous groups formed by the texts with a specific FTD as the dominant function (see Figure 3, which show the ratio of texts with a specific dominant function). For the tree corpora in our experiment it seems possible to set k to 5 or 6. Our ex-

	CroCo	**RusLTC(stu)**	**RusLTC(pro)**
	EN>GE	EN>RU	EN>RU
Tokens	240 K	213 K	1.2 M
Texts	110	360	517
Genres		Acad(12)	Media(417)
	Essay(29)	Adverts(12)	Popsci (100)
	Fiction(10)	Educat(58)	
	Instr(10)	Essay(131)	
	Business(13)	Fiction(12)	
	Popsci(11)	Info(143)	
	Speech(14)	Interview(3)	
	Tourist(11)	Letters(3)	
	Web(12)	Speech(12)	
		Tech(15)	

Table 4. Parameters of the parallel corpora

Figure 3. Ratio of texts by the dominant FTD in the research corpora as predicted by biLSTM



periments showed that the difference was not critical for both types of functional vectors: it did not affect the makeup of the most populous cluster in CroCo. The clusters we received for the normative corpus (CroCo) were not at odds with the existing genre annotation (see Appendix). Both models succeed in grouping together instructions and fiction. The difference between the clusterings is in how the models interpret hybrid texts such as popular scientific texts and what aspects of texts they focus as secondary functions. biLSTM, which was learnt on the vectorized patterns of 1500 most frequent words and PoS for other tokens, produces vectors that highlight the fictional, narrative nature of pop-sci, throwing these texts together with fiction (Cluster3), while the classifiers learnt on frequencies of lexicogrammatic features (including inter alia lists of amplifiers and downtoners, private, public and suasive verbs (Quirk et al., 1985)) prioritize the informational and scientific component of pop-sci and group it with tourist informational leaflets (Cluster2).

Taking into account the size and homogeneity of the CroCo clusters, it makes sense to target Cluster 1 in finding functionally similar subsets from RusLTC(stu) and RusLTC(pro). These two collections were clustered with KMeans, the centriods for each cluster were calculated and compared to the CroCo centroids using Euclidean similarity measure. The most similar subsets of the two corpora are the clusters with most similar centroids. In determining k for RusLTC, we looked for a reasonable balance between the similarity and homogeneity scores. For RusLTC(stu) k = 8 and for RusLTC(pro) k = 10 return the best combination of the two.

To triangulate the results from KMeans, we compared them to the results for Affinity Propagation with the parameters tested on the BNC selection. The algorithm returned clusters which shared 85-98% of the files with the most successful KMeans result for all the experiments.

6 Discussion of Results

The primary goal of this project was to test the applicability of the text vectors learnt from the annotated text functions to the task of producing functionally similar subsets of two arbitrary corpora. This involved decisions on the input text representation, learning approach, clustering method and similarity metric. We have found that, first, the functional vectors learned on sequences, patterns and lexicogrammatic properties of texts were more effective in genre/function detection than those learnt on lexical features. Our results from neural networks modelling demonstrated that the functional properties of texts were better captured by the mixed sequences of the most frequent words and PoS than by lemmatized embeddings with stop words filtered out. The purely lexical features (TF-IDF) and keywords statistics proved inferior to lexicogrammar in the alternative ML setting, too.

However, the patterns of the most frequent words and PoS can be more successful with identifying some functions, but not other. In particular, it seems that the functional representations based on the vectorized texts did not quite capture the evaluative, personal and informational FTDs. This can be explained by two factors: first, these functions can rely on lexical features for their expression and, second, they are often annotated as a second strong dimension, unlike the mutually exclusive (genre-pivotal and relatively easy to predict) FTD such as fiction, instruction, news and scitech. To support this argument: in the classification task on the six hand-picked BNC genre categories, the raw Biber's features performed a bit better than the functional vectors learnt on them, while the biLSTMmix vectors demonstrated even less skill in recognizing our select BNC genres, where the majority of texts are of the easy-to-recognize type. On RusLTC(pro) corpus which consists of mass media texts and popular scientific texts, biLSTMmix returns no reliable predictions for the staggering 19% of texts. This analysis shows that FTD detection can benefit from combining vectorized and statistical register features, which we leave for future work.

Second, though our modelling approaches per se are not directly comparable, because they had different objectives and operated on different text representations, we can evaluate their usefulness for the practical task of predicting functional properties of texts. The inspection of the real-valued vectors indicates that the vectors learnt within the classification task setting overestimated the texts functionality (i.e. produced noisy predictions) and were less adequate in determining the functions hierarchy as manifested in the human scores. The two approaches had very similar overall performance in the intrinsic evaluation and in the BNC genre classification task, though in the real application they produce only partially overlapping clusters. This is probably because the models are focusing different properties of texts that are equally relevant for fulfilling text functions, but are more or less pronounced in individual real texts. It seems reasonable to use the union of the two sets for practical purposes. Besides, the models are different in terms of processing effort required, with the model on Biber's features less easily applicable to big corpora.

Third, the effectiveness of the functional representation was ultimately tested in the BNC clustering task. While for Affinity Propagation on pair-wise similarities the type of functional vectors did not matter, the better-performing KMeans proved to be sensitive to the difference in the functional vectors and managed to find a good fit to the BNC genres (5 clusters, ARI=0.92) only for the RandomForest vectors. The functional vectors learnt on embedded mixed representations achieved ARI=0.58 for any k in the range from 4 to 8. Note, however, that any functional vectors were by far better in this task than the baselines: we failed to produce any good clustering results for our BNC selection on the lexical and on the raw register features.

7 Conclusions

This paper presents an approach to deal with a practical issue of constructing functionally comparable corpus resources. We proposed a method to measure functional comparability of the resources at hand and to produce their functionally homogeneous intersection. The method offers a way to verify the researcher's judgments about the corpora comparability which are usually based on pre-existing corpus annotation schemes and researcher's intuition. We show that texts can be described externally via a reference to a number of communicative functions and that the functions are reflected via text-internal linguistic features. We found that functional text representations offer a better clustering result for a corpus with 'known' functions in comparison to keywords and linguistic register features. They can be effectively used to identify functionally homogeneous subsets of texts in a given text collection and to match them to functionally comparable sets from another corpus. The cross-linguistic extension of this research (left for future work) is supposed to equip a researcher with a corpus of non-translations in the TL functionally comparable to the ST. Such a reference corpus would effectively represent the expected TL textual fit (Chesterman, 2004) that is necessary to estimate specificity of translations.

Acknowledgments

The first author has been supported by the Russian Foundation for Basic Research within Project No. 17-06-00107.

References

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3):259–274. https://doi.org/10.1093/llc/fqi039.

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.

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

Andrew Chesterman. 2004. Hypotheses about translation universals. *Claims, Changes and Challenges in Translation Studies* pages 1–14. https://doi.org/10.1075/btl.50.02che.

Isabelle Delaere. 2015. *Do translations walk the line? Visually exploring translated and non-translated texts in search of norm conformity*. Phd, Ghent University.

Sascha Diwersy, Stefan Evert, and Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In *Aggregating dialectology, typology, and register analysis. linguistic variation in text and speech*, Walter de Gruyter, Berlin, pages 174–204.

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

Michael A. K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

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*. Trento, pages 35–42.

Adam Kilgarriff and Raphael Salkie. 1996. Corpus similarity and homogeneity via word frequency. In *Proceedings of Euralex*. volume 96.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Haidee Kruger and Bertus Van Rooy. 2010. The features of non-literary translated language: a pilot study. *Proceedings of Using Corpora in Contrastive and Translation Studies, England, July 2010* .

Andrei Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling*. Linköping University Electronic Press, pages 271–276.

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 and Mihaela Vela. 2015. Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. *Proceedings of the Second Workshop on Discourse in Machine Translation* (September):122–131.

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.

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.

Albrecht Neubert. 1985. *Text and Translation*. Enzyklopdie.

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).

Christiane Nord. 2006. Translating as a purposeful activity: a prospective approach. *TEFLIN Journal* 17(2):131–143.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Katharina Reiss and Hans J Vermeer. 1984. Groundwork for a general theory of translation. *Tubingen: Niemeyer* 101.

Serge Sharoff. 2013. Measuring the distance between comparable corpora between languages. In *Building and Using Comparable Corpora*, Springer, pages 113–130. http://www.tausdata.org/.

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

Alexandr Shveitzer. 1973. *Translation and Linguistics: Informational newspaper and military publicist texts in translation [Perevod i lingvistika: O gazetno-informacionom i voienno-publicisticheskom perevode]*. Voenizdat.

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.

Appendix: Supplementary material

CroCo clusters as determined for the two alternative functional representations against the existing annotation

	biLSTMmix				**RandomForest**			
	texts	description	homo	genres	texts	description	homo	genres
Cluster 0	12	instr:0.91, promo:0.19, info:0.09	.685	INSTR 9, WEB 3	15	instr:0.71, promo:0.57, info:0.43	.687	INSTR 10, WEB 5
Cluster 1	**43**	argum:0.83, new:0.11, personal:0.08	**.706**	ESSAY 27, SPEECH 12, SHARE 2, POPSCI 1, WEB 1	**47**	argum:0.7, personal:0.62, new:0.54	.626	ESSAY 22, SPEECH 13, SHARE 7, POPSCI 3, FICTION 1, TOU 1
Cluster 2	12	info:0.59, promo:0.19, eval:0.14	.588	TOU 7, WEB 2, POPSCI 1, SPEECH 1, SHARE 1	39	scitech:0.5, new:0.5, info:0.48	.487	TOU 10, POPSCI 8, WEB 7, ESSAY 7, SHARE 6, SPEECH 1
Cluster 3	24	fiction:0.27, scitech:0.13, argum:0.1	.406	FICTION 10, POPSCI 8, WEB 3, ESSAY 2, SPEECH 1	9	fictio:0.85, eval:0.48, personal:0.35	.672	FICTION 9
Cluster 4	19	promo:0.7, info:0.15, argum:0.1	.527	SHARE 10, TOU 4, WEB 3, POPSCI 1, INSTR 1				