Article:
Sharoff, S (2021) Genre Annotation for the Web: text-external and text-internal perspectives. Register Studies, 3 (1). pp. 1-32. ISSN 2542-9477

https://doi.org/10.1075/rs.19015.sha

Genre Annotation for the Web: text-external and text-internal perspectives

Serge Sharoff

Abstract

This paper describes a digital curation study aimed at comparing the composition of large Web corpora, such as enTenTen, ukWac or ruWac, by means of automatic text classification. First, the paper presents a Deep Learning model suitable for classifying texts from large Web corpora using a small number of communicative functions, such as Argumentation or Reporting. Second, it describes the results of applying the automatic classification model to these corpora and compares their composition. Finally, the paper introduces a framework for interpreting the results of automatic genre classification using linguistic features. The framework can help in comparing general reference corpora obtained from the Web and in comparing corpora across languages.

Keywords: Automatic genre identification, Deep learning, Interpreting neural networks

1 Introduction

John Sinclair once provided a fairly convincing justification of corpus studies: "language looks a lot different when you look at a lot of it at once" (Sinclair, 1991). However, this statement hides the presence of variations which are specific to individual text types or genres. Looking at language *at once* can be misleading, we also need to look at different varieties of language, since "language may vary across genres even more markedly than across languages" (Biber, 1995). Arguably, Web corpora measuring in billions of words provide a better window for "looking into a lot of language" simply because they offer much more data in comparison to traditional corpora, such as the British National Corpus (BNC). However, because of the method for their construction they lack curated categories, the only label available for each page is its web address. This does not provide information to study linguistic variation by comparing their subsets. For example, ukWac (Ferraresi et al., 2008), enTenTen (Jakubíček et al., 2013) and Aranea-En (Benko, 2016) are three large Web corpora for English, all consisting of millions of Web pages and all produced by crawling the Web, but without metadata suitable to study their linguistic variation. We can assume that there are different kinds of texts **within** each corpus. We can also assume that there is some difference **between** these corpora, because they have been produced using different techniques for crawling the Web and for getting texts out of the Web pages.

The aim of digital curation proposed in this paper is to add value to collections of texts in large Web corpora by adding metadata to each text via automatic text classification. Traditional corpora, such as the Brown Corpus or the BNC, have been curated by design, because in the process of their compilation text variety labels have been assigned to each text. In the end, with the Brown corpus one can investigate the difference in the use of passives in academic texts vs fiction (Evert, 2006). Similarly, patterns of grammatical markers can be compared against the range of genres recorded in the BNC (Szmrecsanyi, 2009). Studies of this kind are not possible with Web corpora like enTenTen or ukWac unless some kind of curation is applied

to describe the varieties of their texts. If the curation process uses the same set of labels, such corpora can be also compared to each other.

This naturally leads to three questions:

1. What is a way to curate a large Web corpus with respect to text varieties?
2. What are the differences between available Web corpora with respect to these varieties?
3. Which linguistic features are associated with each variety?

The first question assumes a typology suitable for curation on the Web scale. Genre typologies are often compared to the jungle, cf (Lee, 2001; Kilgarriff, 2001); this alludes to a multitude of categories and little compatibility between various typologies. The very term to describe the text varieties, *genre, register, style, text type*, is used in different ways even in closely related studies, thus necessitating the need to clarify the kind of curation addressed in this paper.

The setup of the current study focuses on the duality of $\langle form \leftrightarrow function \rangle$. John Sinclair has summarised this duality by considering the text-internal and text-external perspectives on corpus design (Sinclair and Ball, 1996). A **text-internal** perspective starts from the linguistic features present in a text, such as categories of words or syntactic constructions, e.g., the use of the first person pronouns or *that* deletion. A **text-external** perspective starts from the function a text serves in communication, e.g., for *informing*, *reporting* or *entertaining*. A similar distinction between the linguistic features of texts vs their situational profile has been made in (Biber, 1995). Any text exhibits this duality, while analysis can start from either of the two perspectives: a text, which is described through its text-external function in communication, can be linked to its text-internal linguistic features or vice versa.

This study follows the Automatic Genre Identification (AGI) tradition (Santini et al., 2010) by using the term 'genre' to refer to the label describing a text from the text-external perspective before going into analysis of any linguistic features. The notion of genres in the AGI tradition stems from a Machine Learning task to predict a label for a text, while the features for prediction are not necessarily interpretable. For example, useful features for making predictions for genre labels are character n-grams (Sharoff et al., 2010) and word embeddings (see Section 2), which do not have natural linguistic interpretations. When the link between text-external communicative functions and text-internal linguistic features is analysed, the term 'register' is more appropriate to describe the varieties. More specifically, in this paper we are interested in interpreting predictions in the AGI tradition using linguistic features, thus shifting from the genre perspective when the focus is on classification accuracy (Section 2) to the register perspective when the focus is on analysis of linguistic features (Section 4).

1.1 Text-external communicative functions

Studies in the AGI tradition usually focus on a small number of genre categories, for example three (Petrenz and Webber, 2010) or ten (Stamatatos et al., 2000), while fine-grained genre typologies often consist of hundreds or thousands of labels (Adamzik, 1995; Crowston et al., 2010; Görlach, 2004).

One reason for the existence of this jungle is the sheer variety of text types when corpora are collected from a large number of sources. There is also a variety of ways to curate a text collection with genre labels. For example, the BNC genre typology includes a single genre of fiction, while six genres of fiction are used in the Brown Corpus family. At the same time, the British Library catalogue lists more than 150 genres of fiction, such as *Picaresque* or *Robinsonades* (Lee, 2001). Similarly, different domains can be associated with different styles of academic writing. The linguistic features of research articles in philosophy, political

sciences, medicine or chemistry are substantially different from each other, for example, this concerns the rates of nouns, prepositions and type-token ratio (Nesi and Gardner, 2012). This can be an argument to annotate such varieties with different genre labels, as this has been done in the BNC genre typology (with its six academic genres, such as *Arts, Medicine* or *Tech.Eng*), but not in the Brown Corpus typology, which uses a single genre label *Learned* for all academic texts. It is natural that the British Library wanted to create a more precise account for the range of fiction texts in its collection. However, the differences in the design decisions for fiction or academic texs in the Brown Corpus or the BNC are fairly arbitrary, and this hinders the comparison of their composition.

Another consideration for the genre inventory is that traditional corpora came from a relatively small number of well controlled sources. Corpora collected from the Web come from a far larger number of sources: there are no genre categories in the written part of the BNC which cannot be found on the Web, while the Web has many other sources, such as Wikipedia articles or personal blogs, which cannot be fully described in terms of the BNC genre categories. This assumes that we need more categories to add metadata to a Web corpus. At the same time, statistical studies prefer smaller genre inventories because of the need to have a sufficient number of samples in each category. The number of texts in Web corpora is far larger than what is available in traditional corpora, more samples can be found for any appropriate category, hence leading to the possibility to deal with bigger genre typologies. However, curation of Web corpora requires manually constructed training sets, this constrains the number of labels to be used for annotation.

Yet another reason for the difficulty in curating a large corpus comes from genre hybridism (Santini et al., 2010). Even with strict editorial control, the authors may express different communicative purposes or to combine different styles of writing in a single text, such as mixing reportage and expressions of opinions in a newspaper article. This leads to possible proliferation of genre labels, e.g., *editorial, column, opinion, analytic, feature article*. On the Web there are far fewer explicit gate-keepers, such as editors or reviewers, and far more authors with varying levels of expertise or willingness to express themselves according to traditionally accepted ways which are recognised as genres by the gate-keepers. This blurs the more clearly defined genre boundaries of traditional corpora and leads to numerous novel hybrid genres, such as citizen journalism or research blogs. From the annotation perspective, if only one annotation per text is allowed (as it is the case with traditional corpora), different annotators can interpret a hybrid Web text in different ways, thus producing different annotations for stylistically similar texts. This annotation noise can confuse Machine Learning tools. Experiments with human annotation show that hybrid texts can account for 25% of annotated Web pages (Biber and Egbert, 2016).

The automatic classifier in this study uses a compact text-external annotation scheme (Sharoff, 2018), which is based on **Functional Text Dimensions** (FTDs). For each FTD, there is a test question for the degree of presence of communicative functions and a list of prototypes, which are commonly found on the Web. For example, the FTDs for the three major categories of newspaper texts, coded in the Brown corpus as `A`, `B` and `C`, are defined as:

news (A8) To what extent does the text provide an informative report of recent events? (For example, a newswire item).

argument (A1) To what extent does the text try to persuade the reader? (For example, an argumentative blog entry or a newspaper opinion column).

review (A17) To what extent does the text evaluate a specific entity? (For example, a review of a product, location or performance).

A text can score on the *News* FTD (the answer to the test question is **Strongly**) if it is sufficiently similar in its function of informative reporting to prototypical newswire articles (the prototype is given in brackets

after the test question). A text can receive more than one **Strongly** score if it is functionally similar to more than one prototype, thus establishing that this text is a hybrid. One-word labels like *News* provide a useful short-hand for listing the FTDs, which are fully defined by the test question and are judged by the similarity to one of the prototypes. Even shorter numerical codes can be used for reporting the FTDs in tables. The full list of the FTDs with their test questions is provided in Appendix 2.

1.2 Text-internal linguistic features

From the viewpoint of functional linguistics, text-external communicative functions are realised through text-internal lexicogrammatical features, e.g., temporal rhetorical relations, or functional roles, e.g., phenomenon identification, for more details and specific examples see (Matthiessen, 2015). However, for reliable automatic processing of texts on a Web scale, we need features, which can be extracted from millions of texts easily and with reasonable accuracy. Our study relies on the text-internal lexicogrammatical features which were introduced for describing register variation (Biber, 1988). The features include the following categories:

Lexical features such as:

- public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny...*
- time adverbials = *afterwards, again, earlier, early, eventually, formerly, immediately,...*
- amplifiers = *absolutely, altogether, completely, enormously, entirely,...*

Part-of-speech (POS) features such as:

- nominalisations (nouns ending in *-tion, -ness, -ment*)
- prepositions
- past tense verbs

Syntactic features such as:

- *be* as the main verb
- *that* deletions
- pied piping

Text-level features such as:

- average word length
- average sentence length
- type/token ratio (TTR)

This set was designed specifically for English. However, some of its features are nearly universal, for example, this concerns the text-level features, even though their exact values are language-dependent. Many lexical features are comparable across languages if they can be translated reliably, for example, public verbs. Many part-of-speech features can be used across a number of languages as well, for example, nominalisations, while many syntactic features are comparable only across a smaller set of closely related languages, for example, pied piping. Out of the total set of 67 features in the original study (Biber, 1988), we selected 45 features with compatible functions in English, French, Russian and Spanish. The full list of features is

provided in Appendix 1. Cross-lingual comparison is done only for English and Russian, since for these two languages we have corpora annotated with the **same** genre inventory.

The contribution of this study consists in presenting:

- a machine learning model for predicting functions of texts from large general purpose corpora, see Section 2;
- a comparison of the composition of large Web corpora using this model, Section 3;
- a model for interpreting the functions as registers via text-internal linguistic features, Section 4.

2 Automatic genre identification

2.1 Text classification model

The text classification model used in this study combines a Deep Learning architecture (Goodfellow et al., 2016) with a mixed representation which is based on keeping the most common word forms and replacing other words with their POS tags, see (Baroni and Bernardini, 2006). For example, a sentence from a hybrid text expressing the functions of review and promotion:

(1) *It won the SCBWI Golden Kite Award for best nonfiction book of 1999 and has sold about 50,000 copies.*

can be converted into a mixed representation as

(2) *It won the* PROPN ADJ NOUN NOUN *for best* NOUN NOUN *of* [#] *and has sold about* [#] NOUN.

Traditional neural models rely on vector representations (known as embeddings) produced for word forms. With respect to AGI studies, it has been shown that while word form models can be more accurate on the same corpus, their accuracy decreases under any domain shift (Petrenz and Webber, 2010). The model which mixes word form embeddings with vector embeddings for the POS tags is designed to capture genre-specific features in a stable way without relying too much on keywords specific to the training corpus.

The Machine Learning model in this study is based on a bi-directional Long Short-Term Memory (biLSTM) classifier (Yogatama et al., 2017) with the attention mechanism (Liu and Lane, 2016). The model parameters are as follows: word embeddings of 300 dimensions at the input, a bi-directional LSTM of 128 dimensions in each direction followed by an attention layer and a fully-connected layer of 10 output neurons for predicting the FTDs. For efficiency of training, the document length was capped by 1000 words. The training parameters were set for the 15 epochs, the learning rate of 0.001 with the Adam optimiser. These parameters were chosen via tuning their performance on a validation set.

This model provides a way of (1) detecting similarities between word forms via pretrained word embeddings, (2) linking long-distance dependencies via biLSTM neural networks and (3) selecting the most relevant words or constructions via attention. More specifically, word embeddings are vectors in a multi-dimensional space in which neighbouring words are likely to be similar in their meaning. The embedding spaces are themselves built using shallow neural networks aimed at predicting masked words from their contexts (Mikolov et al., 2013). For example, the masked word in a context like *I went to the ... and bought some milk* is likely to be one of *store, supermarket, shop, market*, which leads to producing similar embedding vectors for those words. When the words are represented by their embeddings, the LSTM model

Table 1: Training corpora

FTD Code	Short-hand label	Prototypes	En	words	Ru	words
A1	Argument	Argumentative blogs or opinion pieces	375	493921	345	507392
A4	Fiction	Novels, myths, songs, film plots	103	262856	97	199471
A7	Instruction	Tutorials or FAQs	221	149655	96	114776
A8	News	Reporting newswires	207	100567	538	173630
A9	Legal	Laws, contracts, terms&conditions	95	195509	105	285112
A11	Personal	Diary-like blog entries	161	158933	284	194191
A12	Promotion	Adverts	350	135805	331	147383
A14	Academic	Academic research papers	126	344426	223	577119
A16	Information	Encyclopedic articles or specifications	244	279838	313	635672
A17	Review	Reviews of products or experiences	102	73860	257	104877
		Total	1562	2195370	1930	2939623

is trained by building recursive connections of sequences of embeddings to provide a sentence representation, which captures some information about the syntactic structures (Gulordava et al., 2018). Finally, the attention mechanism is added on top of biLSTM to select words (or POS tags), which are most predictive for a given task.

2.2 Datasets for training

This study used an existing corpus with Web pages in English and Russian annotated with the same set of 10 basic communicative functions.[1] According to the corpus collection and annotation procedure (Sharoff, 2018), this set provides a balance between the coverage of all common Web texts and the presence of a reasonable number of annotated examples for each function. The texts in the training corpora originated from a range of Web pages, including a random subset of Web corpora, such as ukWac, as well as targeted genre collection to ensure the presence of less frequent text varieties (Forsyth and Sharoff, 2014).

It has been shown that the FTD annotation in terms of dimensions (*To what extent does the text do X?*) leads to higher inter-annotator agreement when measured in terms of Krippendorff's interval α (Krippendorff, 2004). Out of other inter-annotator agreement measures, this measure assigns more importance to larger disagreements on the FTD scale in comparison to using fixed genre labels. The annotators were asked to operate on the basis of test questions and prototypes (listed in Appendix 2 in this study) to avoid misunderstanding of the short-hand labels. For example, *News* can be interpreted in many different ways, while the answer to its test question refers to the specific FTD. According to the original study (Sharoff, 2018), the value of Krippendorff's α four our FTD set ranges from 0.78 for A12 ('commercial promotion') to 0.97 for A9 ('legal texts'), all above the threshold of 0.67 for statistically reliable annotations.

The training corpora were processed with UDPIPE (Straka and Straková, 2017) to extract the lexicogrammatical features and to produce the POS tags, which replaced the word forms for the neural classifier. The neural genre identification architecture is especially suited for the setup assumed in the FTD framework, as the neural model can be trained to predict a vector of probabilities for each FTD. Depending on the threshold, either the highest scoring FTD or several highest scoring FTDs can be used for predicting the functions of a text.

Table 2: Precision and recall scores for predicting the FTDs

Models:	En: Neural		En: Baseline		Ru: Neural		Ru: Baseline	
Accuracy:	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
A1.argument	0.65	0.85	0.65	0.55	0.71	0.75	0.63	0.51
A4.fiction	0.82	0.82	0.78	0.80	0.89	0.66	0.65	0.57
A7.instruct	0.79	0.76	0.61	0.56	0.89	0.59	0.64	0.57
A8.news	0.71	0.61	0.76	0.72	0.94	0.93	0.86	0.83
A9.legal	0.77	0.40	0.78	0.82	0.82	0.81	0.72	0.66
A11.personal	0.66	0.74	0.70	0.70	0.67	0.76	0.65	0.55
A12.promotion	0.89	0.90	0.64	0.67	0.93	0.90	0.73	0.76
A14.academic	0.67	0.82	0.72	0.76	0.86	0.87	0.63	0.48
A16.information	0.85	0.62	0.59	0.66	0.64	0.60	0.67	0.78
A17.review	0.88	0.72	0.63	0.62	0.71	0.86	0.66	0.52
Hamming loss	0.048				0.068			

2.3 Prediction accuracy

Because of our focus on hybridisation via multi-label classification, the accuracy of the neural classifier has been measured using the Average Precision score for predicting each FTD, as well as using the overall accuracy measured via the Hamming loss, which computes the proportion of *irrelevant* predictions (Sorower, 2010), thus the lower the better. Table 2 reports the precision and recall scores in a 10-fold cross-validation setup. The neural network model produces a reasonably confident classifier, predicting the FTDs with the average precision of around 0.77 for English and of 0.75 for Russian.

The baseline model is based on Logistic Regression (LR), which predicts the text functions using the linguistic features. In Section 4 we use the same LR model to interpret the decisions of the neural classifier. The LR model is slightly behind the neural one with the exception of *A9.legal* and *A14.academic* texts for English. Nevertheless the accuracy of the LR predictions is considerably better than the random baseline (of 10%), so it still provides a suitable basis for analysing the neural predictions. Some of its **recall** values are relatively low, for example, 0.55 for *A1.argument*. However, this does not affect our register study reported in Section 4 below, because the study focuses on **precision** in its ability to associate the Web pages (which have been reliably detected by the neural classifier) with any linguistic features extracted by the LR classifier for the same page, see more details on the interpretation procedure in Section 4.

Tables 3 and 4 offer a more fine-grained picture of the distribution of errors made by the neural classifier. The greatest confusion of the classifier is between argumentative texts (A1) and two kinds of reporting, news reporting (A8) and personal reporting (A11), as well as with reference information (A16). The errors are similar across the languages. The errors are also symmetric in the sense that the most common error in recognising texts annotated as A8 or A11 is in predicting them as A1 texts. This is partly because A1 is the majority class. However, other functions, such as fiction (A4), instruction (A7) or legal texts (A9), have less potential for being confused with argumentation.

In addition to predicting the accuracy on the training corpus via cross validation, it is important to check how the classifier predicts the genre of texts not included in the training set. One option is to apply the resulting classifier to corpora with known composition and to compare the labels. Table 5 lists the most

Table 3: Confusion matrix of predictions for English

Predicted→ Reference↓	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
A1.argument	310	2	2	16	2	12	10	6	3	2
A4.fiction	7	84	1	0	0	6	1	1	1	1
A7.instruct	12	4	167	1	2	10	7	12	3	2
A8.news	53	0	1	122	1	8	7	0	7	2
A9.legal	18	0	14	6	37	1	2	13	2	0
A11.personal	22	4	6	2	0	114	3	1	2	0
A12.promotion	7	0	5	8	1	6	314	2	3	2
A14.academic	9	1	2	2	0	5	0	102	4	0
A16.info	35	6	8	11	5	2	7	14	148	1
A17.review	6	2	5	3	0	10	0	1	1	71

Table 4: Confusion matrix of predictions for Russian

Predicted→ Reference↓	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
A1.argument	187	3	3	9	3	15	3	4	11	13
A4.fiction	6	50	0	0	0	12	0	0	2	6
A7.instruct	4	0	41	1	1	3	7	4	6	2
A8.news	13	0	1	446	2	3	2	1	8	4
A9.legal	4	0	0	0	59	2	1	2	5	0
A11.personal	13	3	0	2	0	134	3	0	4	18
A12.promotion	3	0	0	3	0	8	276	5	6	7
A14.academic	9	0	1	0	3	4	2	161	6	0
A16.info	19	0	0	11	4	10	1	10	90	6
A17.review	6	0	0	4	0	9	1	0	3	136

Table 5: Prediction of FTD categories against BNC genre classes

Texts%	Texts#	FTD, mixed	BNC genres	Texts%	Texts#	FTD, words	BNC genres
9.74%	395	Fiction	W.fict.prose	9.20%	373	Fiction	W.fict.prose
6.14%	249	News	W.newspaper	4.37%	177	Argument	W.ac
4.86%	197	Argument	W.ac	3.82%	155	News	W.newspaper
3.35%	136	Personal	S.conv	3.58%	145	Information	W.ac
3.08%	125	Argument	W.misc	3.50%	142	Information	W.misc
2.98%	121	Personal	S.consult	3.26%	132	Argument	W.newspaper
2.98%	121	Argument	W.newspaper	3.16%	128	Personal	S.conv
2.76%	112	Information	W.ac	2.64%	107	Argument	W.misc
2.64%	107	Information	W.misc	2.39%	97	Personal	S.interview
2.47%	100	Promotion	W.misc	2.34%	95	Promotion	W.misc
2.05%	83	Academic	W.ac	2.17%	88	Promotion	W.non.ac
1.95%	79	Personal	S.interview	2.07%	84	Personal	S.meeting
1.87%	76	News	W.misc	2.02%	82	Information	W.newspaper
1.60%	65	Argument	S.meeting	1.75%	71	Personal	S.consult
1.46%	59	Personal	S.meeting	1.46%	59	Instruction	W.misc

common combinations of the predicted FTDs and the corresponding BNC genre labels (some BNC labels such as `W.newsp.brdsht.arts` have been generalized to `W.newspaper`). The proportion of texts is given against the total number of texts in the BNC (4054 texts). When the genre labels of the BNC match the respective functions, for example, Fiction to `W.fict.prose` or Personal reporting to `S.conv`, the functions are detected reliably. On the other hand, argumentation as one of the most pervasive text functions can be found in a number of sources, since the BNC genre typology does not distinguish newswires from editorials unlike the use of the `A` and `B` categories in the Brown corpus.

The left half of Table 5 presents the prediction results for the neural classifier as described in Section 2.1, the right half presents the results of a neural classifier which is based on the same model with the only difference in using plain word embeddings instead of a mixed representation. The classifier with the simple word embeddings is worse in its ability to predict the functions in the BNC. For example, fewer newspaper texts from the BNC are predicted as either news reporting (249 texts with the mixed POS model vs 155 texts with the word-level model) or argumentation (197 vs 132), which supports the results from (Petrenz and Webber, 2010) concerning the topical bias of word-based genre classifiers.

The neural classifier is also able to assign more sensible interpretations to texts labelled as `W.misc`, which is the most common genre label in the BNC. This covers a range of genres not recognised in the BNC typology, such as political pamphlets (predicted as *Argument*), news reports not coming from newspapers (*News*), advertising materials (*Promotion*), fact sheets and reference materials (*Information*), see Table 6 for a sample of titles with their predicted functions.[2]

The neural classifier takes care of hybrid texts as well. While tables assigning an FTD label to a text are useful for presentational purposes, this is a simplification for representing a text in functional dimensions, where the hybrid texts are represented via several functional dimensions at the same time. The neural network predicts a vector of probabilities for each text. For example, newspaper articles from the BNC include clearly reporting texts (A8H in Table 7 with low probabilities for any other dimension), clearly

Table 6: Predictions for texts coded as W.misc in the BNC

FTD	ID	Bibliographic description
Argument	AM8	The best future for Britain. Conservative Central Office, 1992
Argument	AMA	It's time to get Britain working again. The Labour Party, 1992
News	A0E	The seventh Birmingham International Film&TV Festival, 1991
News	CFD	Action. World Assoc for Christian Comm, 1991
Promotion	A0C	Caterer & Hotelkeeper. Reed Pub. Group, 1991
Promotion	B2B	The Nottingham Graduate. Univ of Nottingham, 1992
Information	A0A	CAMRA fact sheets. Campaign for Real Ale, n.d.
Information	A11	British Rail in the eighties. St John Thomas, David, et al 1990

argumentative texts (A3B) or hybrids of these two functions (A3S, A53 or AKG). For example, Text A3S begins with:

(3) *KEITH GAUNT is no Pericles. But when the citizens of the Association of Futures Brokers and Dealers cast their stones into the pot last week and elected the managing director of Amalgamated Metal Trading on to their council with the biggest number of votes, they elevated the nearest that divided body has seen to a populist. The AFBD is divided for two reasons. First, despite a brave attempt to encourage candidates to stand at last week's annual general meeting, the council remains grossly under-represented. Neither the very big and important financial futures business, which accounts for about 155 of the AFBD 's 400 members, nor the equally important oil market...*

The prediction model was able to capture the hybrid nature of this text, which combines reporting about recent events with argumentation and expression of opinions about these events. Out of the 4054 BNC texts, 1068 are detected as hybrids, when the output value of the second most likely text function is at least half of the value of the most likely one, see the predictions in bold in Table 7. The most common hybrid functions in the BNC combine argumentation with news reporting (127 texts), personal reporting (124 texts) or reference information (108 texts).

Table 7: Distribution of FTD predictions for news texts in the BNC

ID	BNC genre	Bibliographic description	Top two predictions
A3B	W.newsp.brdsht.nat.misc	Independent, 19891007, Feature material	**A1** 0.950 A8 0.090
A3S	W.newsp.brdsht.nat.commerce	Independent, 19891009, Business material	**A1** 0.949 **A8** 0.781
A53	W.newsp.brdsht.nat.misc	Independent, 19891012, Title material	**A8** 0.930 **A1** 0.489
A8H	W.newsp.brdsht.nat.commerce	Guardian, 19891123, City material	**A8** 0.998 A1 0.010
A9S	W.newsp.brdsht.nat.misc	Guardian, 19891211, World affairs material	**A8** 0.983 A1 0.340
AKG	W.newsp.brdsht.nat.social	Daily Telegraph, 19920413, Social material	**A8** 0.360 **A1** 0.309

3 Comparing large Web corpora

The previous section has demonstrated that the neural model is reasonably reliable when tested through cross-validation and when applied to corpora with known composition. In the next step this model has been

Table 8: Composition of large Web corpora

FTD	ukWac		Aranea-en		en10[10]		ruWac	
A1.Argument	18.27%	366412	20.85%	1521994	22.44%	2867605	18.20%	222741
A4.Fiction	0.81%	16176	0.22%	16419	0.20%	25457	1.55%	18919
A7.Instruction	8.62%	172896	8.98%	655931	5.06%	646136	1.02%	12446
A8.News	11.78%	236233	12.21%	891047	14.61%	1866694	5.77%	70689
A9.Legal	0.66%	13232	0.60%	43706	0.30%	38408	1.23%	15088
A11.Personal	4.45%	89199	9.28%	677121	6.60%	843070	44.29%	542111
A12.Promotion	30.54%	612482	26.67%	1946777	21.92%	2800618	5.59%	68432
A14.Academic	2.62%	52516	2.74%	199803	5.17%	661213	4.77%	58410
A16.Information	18.41%	369152	16.11%	1176313	20.57%	2628022	11.71%	143324
A17.Review	3.84%	76905	2.35%	171293	3.14%	401185	5.87%	71905

applied to several large Web corpora:

- ukWac (Ferraresi et al., 2008), a corpus of 2 billion words, 2.5 million Web pages, produced by crawling the `.uk` Internet domain;
- ruWac (Sharoff et al., 2017), a corpus of 2.5 billion words, 1.5 million Web pages, produced by crawling Russian language websites;
- enTenTen (Jakubíček et al., 2013), a corpus of 9 billion words, 16 million Web pages, produced by crawling English language websites;
- Aranea-En (Benko, 2016), a corpus of about 8 billion words, 8 million Web pages, produced by crawling English language websites using a publicly released set of tools;[3]

These Web corpora have been produced using similar methods. However, they have started from different seed URLs for crawling, used different tools and slightly different parameters for corpus processing. They have been also produced within the span of about 10 years: starting from ukWac (2007) through ruWac (2010) and enTenTen (2013) to Aranea (2016).

The composition of large Web corpora is compared in Table 8, which shows the total number of their documents with the respective dominant FTD as detected by the neural classifier. The communicative functions of *Argument, News, Promotion* and *Information* are the most common FTDs for almost all corpora, and taken together they cover about 75% of the Web (when measured by the number of Web pages).

A manual check for the typical sources of pages for each predicted function (using a random sample of 25 Web pages per function and listing the sources contributing more than 50% of instances) shows that the most common kinds of pages predicted as *Argument* are newspaper columns, argumentative blogs and discussion forums. The neural model predicts the function of *Promotion* predominantly in the case of e-shops and spam pages. The most common kinds of pages tagged as providing reference *Information* are encyclopedic articles or specifications. *Argument, News* and *Information* are also common in traditional corpora, such as the BNC, see Table 5. The most significant differences between the Web corpora and the BNC concern *Fiction*, which provides a considerable share of traditional corpora, while one can expect to see less texts of this kind on the Web, and *Promotion*, which is to some extent available in the BNC (see examples in Table 6), but it is much more common on the Web.

The most significant differences between the Web corpora (according to the χ^2 test) are as follows:

- While texts aimed at *Promotion* are ubiquitous on the Web, they are considerably more common in ukWac in comparison to enTenTen, primarily this concerns spam. One of the reason for the larger amount of spam in ukWac is because the pipeline for collecting enTenTen includes a specific spam detection component (Kilgarriff and Suchomel, 2013), which was not prominent in collecting ukWac. The corpus collection of ukWac also suffered from a Search Engine Optimisation competition run in 2004, making a non-sensical phrase *Nigritude ultramarine* into its common keyword. *Fiction* is also slightly more common in ukWac because a substantial portion of its texts came from a single website for out-of-copyright fiction.[4]

- Web pages which report *Personal* experience are more common in enTenTen, since there are fewer blog hosting options in the `.uk` domain (the majority of personal reporting in ukWac is from the `blogspot.co.uk` domain). There are also fewer instances of reporting *News* in ukWac, since out of the major UK newspapers only The Guardian allowed easy crawling.

- The composition of ruWac is different from ukWac and other Web corpora in the list by the amount of *Personal* reporting, which can be explained by the proportion of texts coming from LiveJournal, a popular social media website. ruWac also has a markedly lower amount of *Promotion*, which is not a dominant function in single-authored social media messages. Nevertheless, ruWac still contains thousands of examples of the communicative functions in our list to investigate the linguistic features associated with each function.

4 Communicative functions vs linguistic features

4.1 Detection of linguistic features

The text-external output as predicted by the neural model answers the question on which communicative functions are more common in a given corpus, i.e., presenting the AGI perspective on variation. However, this output does not answer the question about linguistic features associated with those functions, i.e., it does not present the register perspective. This is the purpose of this section.

While the neural network model is reasonably efficient in text classification, it functions as a black box: the parameters it learns from the training corpus are not transparent for linguistic analysis. In contrast, Logistic Regression (LR) is a fast and transparent Machine Learning method, which is defined as:

$$ln\frac{p}{1-p} = w_0 + w_1x_1 + ... + w_nx_n$$

It fits a linear model to predict the log-odds ratio, where p is the probability of a text having a particular communicative function, e.g., being argumentative, x_i are interpretable variables, e.g., the proportion of the first person pronouns. Since the model is linear, the relative contribution of each feature can be determined through its weight w_i for detecting this function. To assist in comparing the weights, the variables have been standardised prior to fitting the logistic regression with respect to their values and dispersion, so that for each feature its mean is zero and its standard deviation is one.

Table 9: Text-external features in ukWac

A1.Argument		A8.News		A11.Personal		A12.Promotion		A16.Information	
D13.whQuestions	0.298	K55.publicVerbs	0.351	C06.1persProns	0.668	J43.TTR	0.522	I40.attrAdj	-0.082
E14.nominalizations	0.275	L54.predicModals	0.251	I42.ADV	0.395	E16.Nouns	0.401	G19.beAsMain	-0.108
J44.wordLength	0.212	A01.pastVerbs	0.229	C09.impersProns	0.102	C07.2persProns	0.255	C09.impersProns	-0.122
I39.preposn	0.185	B05.timeAdverbials	0.182	C11.indefProns	0.093	J44.wordLength	0.148	C11.indefProns	-0.155
K45.conjuncts	0.184	H34.sncRelatives	-0.173	C12.doAsProVerb	0.093	I40.attrAdj	0.102	N59.contractions	-0.160
I40.attrAdj	0.149	P67.analNegn	-0.177	A01.pastVerbs	0.090	I42.ADV	0.069	C12.doAsProVerb	-0.167
C10.demonstrProns	0.147	I40.attrAdj	-0.198	K58.seemappear	0.072	B05.timeAdverbials	-0.075	L54.predicModals	-0.195
K56.privateVerbs	0.125	E14.nominalizations	-0.251	C08.3persProns	0.067	N59.contractions	-0.115	D13.whQuestions	-0.241
K57.suasiveVerbs	0.110	E16.Nouns	-0.272	H36.concessives	0.057	C08.3persProns	-0.117	E16.Nouns	-0.244
E16.Nouns	0.100	D13.whQuestions	-0.337	E16.Nouns	-0.140	I39.preposn	-0.129	C08.3persProns	-0.274
C07.2persProns	-0.144	I42.ADV	-0.378	K55.publicVerbs	-0.161	C06.1persProns	-0.133	K55.publicVerbs	-0.288
J43.TTR	-0.261	C06.1persProns	-0.428	E14.nominalizations	-0.571	C12.doAsProVerb	-0.176	C06.1persProns	-0.619
A01.pastVerbs	-0.294	C07.2persProns	-0.971	J44.wordLength	-1.056	K55.publicVerbs	-0.313	C07.2persProns	-0.667
A4.Fiction		**A7.Instruction**		**A9.Legal**		**A14.Academic**		**A17.Review**	
A01.pastVerbs	0.394	C07.2persProns	0.672	E14.nominalizations	0.542	J44.wordLength	0.235	A03.presVerbs	0.165
C08.3persProns	0.300	E16.Nouns	0.374	L54.predicModals	0.379	J43.TTR	0.167	G19.beAsMain	0.124
C07.2persProns	0.195	I42.ADV	0.314	P67.analNegn	0.305	K45.conjuncts	0.093	N59.contractions	0.117
K55.publicVerbs	0.191	H37.conditional	0.120	I39.preposn	0.300	I39.preposn	0.093	I42.ADV	0.108
J43.TTR	0.154	L52.possibModals	0.115	A01.pastVerbs	0.274	E14.nominalizations	0.067	C08.3persProns	0.047
L53.necessModals	0.145	K45.conjuncts	0.105	H23.WHclauses	0.241	L54.predicModals	-0.116	L53.necessModals	-0.204
K48.amplifiers	0.132	C10.demonstrProns	0.099	C10.demonstrProns	0.160	C12.doAsProVerb	-0.139	K55.publicVerbs	-0.212
C09.impersProns	0.097	L53.necessModals	0.058	L53.necessModals	0.145	D13.whQuestions	-0.164	L52.possibModals	-0.214
C11.indefProns	0.075	H34.sncRelatives	0.053	K56.privateVerbs	-0.093	C08.3persProns	-0.176	L54.predicModals	-0.247
C06.1persProns	-0.196	A01.pastVerbs	-0.140	B04.placeAdverbials	-0.153	K55.publicVerbs	-0.190	I39.preposn	-0.304
I39.preposn	-0.334	I39.preposn	-0.169	B05.timeAdverbials	-0.191	C11.indefProns	-0.266	K45.conjuncts	-0.331
K45.conjuncts	-0.386	C08.3persProns	-0.205	K49.generalEmphatics	-0.217	N59.contractions	-0.367	E14.nominalizations	-0.441
E16.Nouns	-0.470	J44.wordLength	-0.225	J43.TTR	-0.516	C06.1persProns	-0.551	C06.1persProns	-0.522
J44.wordLength	-2.285	K55.publicVerbs	-0.316	J44.wordLength	-0.661	C07.2persProns	-1.005	J44.wordLength	-0.705

Another advantage of logistic regression over other machine learning methods is that it has been well investigated from the statistical viewpoint, thus allowing a number of tests to determine the significance of each feature. One of the approaches is by using the likelihood ratio test, which compares the likelihood of the data under the full model against the likelihood of the data under a model with one of the features removed (Hosmer Jr et al., 2013). If the behaviour of the logistic regression model predicting a function changes significantly when a feature is removed, the feature can be considered as more significant for this function.

The framework proposed in this study investigates the distribution of text-internal features for each FTD predicted by the neural model by fitting a logistic regression model over the output of the neural model applied to a corpus. In the end, the linguistic features of each communicative function can be examined via their weight and significance with respect to the likelihood ratio test.

4.2 Mapping linguistic features to functions

Table 9 lists the weights of the most important features for each FTD in ukWac, which have been selected using the likelihood ratio test.

The results of fitting the LR models to the Web corpora provide further validation for the classic MDA studies (Biber, 1988; Biber, 1995), but but this time on the scale of millions of Web pages. One of the most important MDA dimension, which arises in many studies, is the *Involved-Informational* dimension. The *Involved* end is mostly associated with spontaneous face-to-face interaction, see Figure 5.1 in (Biber, 1995). It contrasts with the *Informational* end which is characteristic of texts concerned with condensed information presentation. On the Web, more *Involved* texts primarily express the function of personal reporting (A11), which is similar to *Involved* spoken texts studied in (Biber, 1988) by being associated with the higher rate of the first person pronouns and *do* acting as a proverb. In contrast, the *Informational* end of this MDA dimension is well detected in texts which are predicted by the neural model as news reporting (A8), instruction (A7) and academic prose (A14). The linguistic features characteristic for these functions are the rates of nouns, prepositions, attributive adjectives, TTR, which all indicate high informational density and exact informational content (Biber, 1988).

Argumentative texts on the Web offer an interesting case study, as Web corpora contain several kinds of argumentation, from formal editorials to informal Web forums. Very much like the *Persuasion* dimension (Biber, 1988, p. 111), it is expressed via modal, private and suasive verbs, causation and conjuncts. The majority of argumentative texts on the Web present information carefully integrated into sentences, so that such texts contain longer words, more attributive adjectives, nominalisations and prepositions, similar to observations in (Biber, 1995, p. 115). Another relevant dimension, which is occasionally produced in MDA studies is the *Argumentative-Reporting* dimension (Biber, 1995, p. 219). The framework proposed in our study reproduces groups of features at both ends of this opposition with the *Argumentative* (A1) texts characterised by wh-questions, private and suasive verbs and demonstrative pronouns, while the *Reporting* end has characteristics of other narrative texts, such as past tense and time adverbials in A4 and A8.

Even though Web pages aimed at commercial promotion (A12) often aim at being appealing and informal, some of their characteristic features take them closer to formal texts, for example, they have more noun phrases, more attributive adjectives, higher TTR and longer words:

(4) *The 'Baronet Supreme' is upholstered using a needle teased hair pad to provide extra support, making it as comfortable as it is affordable. Each individual pocketed spring is enclosed in the finest quality cotton calico, used for its durability and 'breathing' properties. Pictured here in an attractive Mulberry ticking on a turned wood leg...*

Table 10: Text-external features in ruWac

A1.Argument		A8.Newswire		A11.Personal		A12.Promotion		A16.Information	
I40.attrAdj	0.288	J44.wordLength	0.777	C06.1persProns	0.797	E16.Nouns	0.397	E14.nominalizations	0.137
C10.demonstrProns	0.280	A01.pastVerbs	0.256	K50.discoursePart	0.236	I42.ADV	0.308	J43.TTR	0.102
D13.whQuestions	0.246	I39.preposn	0.242	B05.timeAdverbials	0.133	C07.2persProns	0.303	E16.Nouns	0.088
E14.nominalizations	0.236	K45.conjuncts	0.186	K58.seemappear	0.077	I40.attrAdj	0.200	B05.timeAdverbials	0.074
K45.conjuncts	0.171	B05.timeAdverbials	0.171	C11.indefProns	0.071	J44.wordLength	0.199	C07.2persProns	-0.133
P67.analNegn	0.137	C07.2persProns	-0.198	K48.amplifiers	0.062	H34.sentenceRel	0.102	D13.whQuestions	-0.154
H34.sentenceRel	0.113	K56.privateVerbs	-0.271	J44.wordLength	-0.160	K50.discoursePart	-0.292	K55.publicVerbs	-0.157
L53.necessModals	0.104	I40.attrAdj	-0.289	E16.Nouns	-0.213	C08.3persProns	-0.295	I42.ADV	-0.196
H35.causative	0.097	C08.3persProns	-0.350	I40.attrAdj	-0.227	P67.analNegn	-0.321	P67.analNegn	-0.216
K49.generalEmphatics	0.080	I42.ADV	-0.492	C08.3persProns	-0.338	K56.privateVerbs	-0.371	K50.discoursePart	-0.282
C06.1persProns	-0.212	C06.1persProns	-0.653	E14.nominalizations	-0.434	A01.pastVerbs	-0.417	C06.1persProns	-0.575
A4.Fiction		A7.Instruction		A9.Legal		A14.Academic		A17.Review	
A01.pastVerbs	0.683	E16.Nouns	0.331	E14.nominalizations	0.604	E14.nominalizations	0.229	I42.ADV	0.368
C08.3persProns	0.492	H37.conditional	0.290	A03.presVerbs	0.243	A03.presVerbs	0.089	P67.analNegn	0.103
A03.presVerbs	0.237	C07.2persProns	0.238	A01.pastVerbs	0.229	C11.indefProns	0.003	K48.amplifiers	0.100
K55.publicVerbs	0.215	C12.doAsProVerb	0.162	P67.analNegn	0.227	J43.TTR	-0.075	J43.TTR	0.069
C07.2persProns	0.146	B04.placeAdverbials	0.139	K57.suasiveVerbs	0.199	D13.whQuestions	-0.075	H36.concessives	0.062
E16.Nouns	0.142	L53.necessModals	0.133	I39.preposn	0.195	K50.discoursePart	-0.084	K50.discoursePart	0.056
B04.placeAdverbials	0.093	H36.concessives	-0.125	E16.Nouns	0.176	K57.suasiveVerbs	-0.111	K49.generalEmphatics	0.055
K50.discoursePart	-0.106	J43.TTR	-0.133	H37.conditional	0.120	I39.preposn	-0.159	K46.downtoners	0.050
K45.conjuncts	-0.123	K49.generalEmphatics	-0.188	K47.generalHedges	0.099	C08.3persProns	-0.167	A03.presVerbs	-0.081
K48.amplifiers	-0.128	E14.nominalizations	-0.213	B04.placeAdverbials	-0.249	A01.pastVerbs	-0.206	L53.necessModals	-0.096
H34.sentenceRel	-0.129	I40.attrAdj	-0.235	C10.demonstrProns	-0.267	I42.ADV	-0.208	K55.publicVerbs	-0.109
C06.1persProns	-0.154	K50.discoursePart	-0.258	D13.whQuestions	-0.287	P67.analNegn	-0.217	B05.timeAdverbials	-0.111
I39.preposn	-0.184	D13.whQuestions	-0.334	B05.timeAdverbials	-0.317	E16.Nouns	-0.263	E16.Nouns	-0.113
E14.nominalizations	-0.406	A01.pastVerbs	-0.512	C06.1persProns	-0.317	C07.2persProns	-0.265	K56.privateVerbs	-0.119
J44.wordLength	-0.630	C06.1persProns	-0.701	J43.TTR	-0.402	C06.1persProns	-0.437	E14.nominalizations	-0.623

The communicative function of providing reference information (A16) is defined through negative weights of linguistic features, for example, such texts have fewer pronouns, public verbs, etc, than texts expressing other functions. The fact that LR has not detected statistically significant features with positive weights for this function is likely to be related to the presence of two kinds of informative texts on the Web:

1. narrative texts such as encyclopedic articles on historical events or biographies, which contain more verbs in the past tense, as well as time and place adverbials;

2. non-narrative specifications, such as encyclopedic definitions, meeting agendas or reference lists.

Informative reports, which also serve this function, can be either narrative or non-narrative. In the end, even though the communicative function of providing reference information is well recognised, the LR classifier has not described this function in terms of its salient positive features.

Contrary to expectations, texts performing regulatory functions (A9) have markedly lower TTR, which indicates that they contain a large number of formulaic constructions, though they do retain many properties of texts with information carefully integrated into sentences, such as nominalisations, wh-clauses and prepositions.

Table 8 only shows the proportion of the dominant communicative function for each text. If hybrids are defined in the same way as in Table 7 (the likelihood of the second FTD is at least half of the dominant one), 17% of texts in ukWac are detected as hybrids, which is similar to the values obtained from human annotation (Biber and Egbert, 2016). The most common cases of hybridisation are between argumentation and news reporting, argumentation and promotion as well as promotion and reference information.

4.3 Linguistic features across languages

Section 3 established the overall similarity between the composition of ukWac and ruWac in two languages with respect to their communicative functions. Comparison of Tables 9 and 10 shows that across English and Russian these functions are often expressed by similar linguistic features in a way which is similar to MDA studies. For example, features related to the *Informational* dimension, such as attributive adjectives, noun phrases and nominalisations, as well as features related to the *Persuasion* dimension, such as wh-questions, negation and public verbs, are characteristic for argumentation (A01), news reporting (A8) and commercial promotion (A12) in either language. The communicative functions related to narration (A8, A4) have typical features from the *Narrative* dimension in MDA studies, such as the past tense verbs and time and place adverbials. There is some difference in the relative importance of the features across the two languages, such as TTR and word length for News reporting (A8). Nevertheless the fitted logistic regression model assigns positive weights to these features in either language.

Logically there are three factors which impact compatibility of functions and features across languages:

available communicative functions The communicative functions selected as representative for the modern Web are not entirely universal, as also discussed by Biber taking Nukulaelae Tuvaluan as an example. Language use in institutional domains, such as government or academia, is more typical in highly literate societies (Biber, 1995, p. 48). In the end, it might become difficult to obtain a large sample of language use across all functions listed in our study, while other dominant functions might need to be added for annotation.

standard ways for expressing communicative functions Even when a function is common in a given culture, it can lack codified lexicogrammatical features for realising this function as register. For example, it is known that some specific features of academic writing developed over the last two centuries

(Biber and Gray, 2016). In particular, the rate of nouns in academic writing increased, while the rate of verbs decreased. In the end, these features became one of the characteristic properties of scientific writing as also detected in this study (Tables 9 and 10).

language-specific linguistic features Finally, the functions can be similar and well codified, while they can be expressed via language-specific mechanisms in each language, either because of the typological properties or different traditions. For example, the grammatical cases are used in Russian for expressing grammatical functions similar to prepositions in English (Cienki, 1989). In the end, the rate of prepositions in Russian is less likely to be a reliable marker of certain registers. There can also be stylistic preferences in expressing similar communicative functions across the two languages. For example, the argumentative texts in either language are characterised by the higher rate of explicit causation markers and emphatics. However, this study finds that they are more important for identifying argumentative texts in Russian in comparison to English. Similarly, the greater number of formulaic expressions used in academic writing in Russian leads to decreased TTR, so that the TTR weight is positive for detecting academic writing in English, while the TTR weight in Russian is negative (compare A14 in Tables 9 and 10).

5 Related studies on computational analysis of genres

This section describes the key differences of this study from other studies in Automatic Genre Identification (AGI) and Multi-Dimensional Analysis (MDA).

The main aim of AGI studies is to predict stylistic properties using extractable features (Santini et al., 2010; Argamon, 2019). Experiments have been carried out with such features as POS tags (Karlgren and Cutting, 1994), broader linguistic features (Kessler et al., 1997), including theoretically motivated features from Systemic-Functional Linguistics (Argamon et al., 2007). It has been shown that some surface-level features, in particular, most frequent words (Stamatatos et al., 2000) and character-level n-grams (Kanaris and Stamatatos, 2007; Sharoff et al., 2010) are often more efficient than complex feature extraction methods. In spite of their simplicity, the character n-gram features are capable of limited generalisation, such as lexical classes (for example, by providing a single feature *..day* for any day of the week as well as for *yesterday*), part-of-speech tags (*..sly* for adverbs, or *..tion* for nominalisations) or even some syntactic constructions (*..ed by* for passives). However, the use of surface-level features often leads to less robust classification, as the results are influenced by topical biases in genre collections (Petrenz and Webber, 2010). Our study produced a more robust model by replacing the less common words with their POS tags.

Overall, feature-counting approaches to AGI suffer from two issues. First, they do not link similar features together. For example, if the word *good* is available in the training corpus, while *excellent* only occurs in the test corpus, the latter is considered as an unknown feature. Second, feature-counting approaches merely record the frequencies of the individual features (typically words) without considering their use in constructions. For example, the word *shown* is fairly common in texts of different types, while its specific constructions, such as *they* ***have been shown to*** *contain enzymes* vs *bravado* ***he had shown*** *only moments earlier*, are likely to be good indicators of particular text types. The approach used for classification in this study takes into account the similarity between the words via embeddings and takes into account their contexts via biLSTM. As of now, there have been no studies experimenting with AGI on general-purpose corpora using neural networks.

Prediction via logistic regression discussed in Section 4 aims at detecting important features, which is similar to feature selection methods, for example, the Information Gain measure (Yang and Pedersen, 1997),

which estimates the entropy of a feature over all classes:

$$IG(f) = p(x) \sum p(c_i|x) * logp(c_i|x) + p(\bar{x}) \sum p(c_i|\bar{x}) * logp(c_i|\bar{x})$$

where x is a feature, $\bar{x}$ is its absence, and c_i are class labels. If the presence or absence of a feature has less impact on the probability of class labels, it has less Information Gain, so it is less likely to be important for classification. However, in this study we are specifically interested in how useful a feature is for a given class, while feature selection methods report how useful a feature is across all classes.

The goal of the proposed framework is similar to the goal of MDA, which is also aimed at investigating parameters of variation across different registers. Our study relies on the features from the original MDA study (Biber, 1988). Similarly, a cross-lingual study in the MDA paradigm (Biber, 1995) investigated feature distributions across several languages and also compared the MDA text dimensions across those languages. More recently, the MDA setup has been also applied to a large sample of manually annotated Web texts (Biber and Egbert, 2016). The functional dimensions of variation discovered in that study demonstrate considerable similarities of the Web pages to corpora from the earlier MDA studies.

The key difference is that the outcomes of register analysis in this study are explained in terms of communicative functions, as functions remain the same across text collections, they can be predicted via supervised classification, and they can be also used at the annotation stage to compare linguistic features with human perception. The MDA procedure is unsupervised, so its outcomes depend on a combination of the chosen features and the specific corpus. This does not fit the goals of this study, because we want to compare corpora across data sources and across languages. For example, the *Argumentation* dimension has not been produced in the study of (Biber and Egbert, 2016), even though it has been produced in other MDA studies. From the FTD viewpoint *Argumentation* is one of the most common text functions (Table 8), so its features in Web pages need to be described, see its features in Table 9. Also MDA dimensions for different corpora receive the same name, even if they combine different groups of features, for example, the highest score for the *Argumentation* dimension in the respective MDA studies came from infinitives (Biber, 1995, p. 160) vs present tense verbs (Biber, 1995, p. 219). The reason for treating these features as belonging to the *Argumentation* dimension comes from the texts which score higher in this dimension, for example, editorials, so this is the same argument as using human perception in providing the prototypes for the *Argumentation* FTD. The disadvantage of the proposed FTD framework is that its results depend on the accuracy of text classification produced by the neural model, while MDA does not need an external classifier. Our earlier attempt of using MDA features for describing the Web registers (Katinskaya and Sharoff, 2015) was limited by its accuracy.

6 Conclusions and further work

This study presents the results of statistical analysis of the communicative functions as observed in several large Web corpora. This is achieved through reliable automatic genre annotation by means of neural networks trained using a mixed feature representation.

First, the results of digital curation help in understanding the composition of a large corpus, so that a corpus linguist can use the annotations to select suitable subsets of a Web corpus, for example, argumentative texts vs informative reporting vs academic writing to produce contrastive analysis with respect to a particular linguistic phenomenon instead of merely relying on "lots of data" from the Web.

Second, the approach proposed in this paper provides methods for comparing the composition of allegedly similar corpora collected using different pipelines, such as ukWac vs Aranea-En vs enTenTen, or for comparing corpora of different languages using the same pipeline, such as ukWac vs ruWac. Knowing their

composition in terms of functions can give a clue to the reasons behind frequency differences. As a simple example, the frequency of the verb *yell* is nearly twice as common in enTenTen in comparison to ukWac (7.9 vs 4.0 instances per million words). Also, the frequency of *to yell things* is higher than the frequency of *to yell obscenities* in enTenTen, while this is the other way around in ukWac, where the frequency of *to yell obscenities* is higher. The classifier shows a statistically significant prevalence of personal diaries in enTenTen compared to promotional texts in ukWac. This difference in their composition can explain some differences in the frequencies. The majority of uses of *to yell things* in these corpora are found in texts classified as personal reporting, which are more common in enTenTen.

Finally, this paper describes a method for register analysis using text-internal linguistic features through their contribution to text functions within a corpus, across corpora and across languages, for example, to describe typical linguistic features of Web texts aimed at personal reporting or academic writing.

The tools for annotating texts and for extracting linguistic features are available under open-source licenses.[5] The tools can perform:

- automatic detection of text functions using a trainable neural model (with annotated data available for English and Russian);
- automatic tagging of texts with respect to their linguistic features for four languages (English, French, Russian and Spanish);
- fitting a logistic regression model to link the text functions and the linguistic features.

Lists of Web pages from the major Web corpora with their annotations in terms of communicative functions are also available.[6]

One of the interesting directions for further research is to use this framework to probe translation universals, i.e., systematic similarities and differences arising from the process of translation (Baker, 1996). Translated texts are expected to use the same set of linguistic features as texts originally written in the target language, because they need to adhere to its genre conventions. At the same time, the very process of translation can introduce distortions, which are partly caused by the linguistic features and genre conventions of the source language (Kunilovskaya and Sharoff, 2019). If we have the same set of linguistic features for two languages and a classifier which predicts the same set of text functions, this makes it possible to compare the linguistic features of translated texts to texts originally written in this language, because the automatic annotation procedure can select texts having the same function to compare their linguistic features.

Most recently a number of studies in Natural Language Processing suggested replacing the traditional neural networks, such as LSTMs, with pretrained contextualised models, such as BERT (Devlin et al., 2018), which can be fine-tuned to a specific task. BERT-like models are pretrained on very large corpora, so they can make generalisations beyond the limitations of the training set, similarly to the embedding models used in the traditional neural networks. At the same time, the BERT representations are sensitive to the surrounding context, so that its embeddings for *table* are different for *across the dining table* vs *across the periodic table*, unlike word-level embeddings, which can only represent individual words. Also the BERT embeddings are inherently multilingual (Conneau et al., 2020), so that the representations for the *dining table* vs *statistical table* examples will be close to their respective translations *table à manger* vs *tableau périodique*, making such representations suitable for applications across languages without the need of a large training corpus. This makes the pretrained contextualised models attractive as the AGI step to replace the model presented in Section 2.1. However, a BERT-like model will be biased towards the topics of the training corpus, as they depend on the exact word forms. A BERT-like model for AGI needs to be made robust with respect to thematic biases.

In this study we use the same set of linguistic features for all languages. This set might need modifications for more distant languages. For example, Chinese lacks nominalisations or past tense verbs in the same sense as they are used in English, even though the functions, such as news reporting or academic writting, are likely to be similar. An interesting direction of research is to investigate how different sets of linguistic features correspond to the same communicative function provided that we have a multilingual genre-annotated corpus for training.

Finally, this research can be extended through studying homogeneity of texts with respect to their functions, i.e. to study how the functions are distributed over a text. One case is that a hybrid text is fairly homogeneous in performing two functions at the same time, such as news reporting and argumentation, as in Example (1) above. Our neural classifier assigns two functions to this text as expected. However, automatic genre classification also needs to deal with the other case when a hybrid text consists of several functionally separate parts. Even in traditional corpora, such as the BNC, some texts are stored as concatenation of texts of different kinds, e.g., a set of newspaper articles, each of which can serve different functions. Similarly, a Web page can contain a reporting news article and argumentative user comments. Existing genre classification methods treat each text as a homogeneous whole by making predictions at the text level. Instead, the FTD classifier can be applied to detect stylistic shifts within a text and to obtain its genre segmentation profile. This approach resembles TextTiling, which aims at detecting topical shifts in documents (Hearst, 1997). If estimates of text homogeneity are known, they can in turn contribute to better genre identification. Studies in Generic Structure Potential (GSP) have already exposed the link between text structure and registers (Matthiessen, 2015). Automatic recognition of the internal structure of documents can provide new features for the prediction models, as well as statistical grounds for GSP studies.

References

Adamzik, K. (1995). *Textsorten – Texttypologie. Eine kommentierte Bibliographie*. Nodus, Münster.

Argamon, S. (2019). Computational register analysis and synthesis. *Register Studies*, 1.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. In Somers, H., editor, *Terminology, LSP and Translation: Studies in language engineering*. John Benjamins.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Benko, V. (2016). Two years of Aranea: Increasing counts and tuning the pipeline. In *Proc LREC*, Portorož, Slovenia.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, D. and Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.

Biber, D. and Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Cienki, A. J. (1989). *Spatial cognition and the semantics of prepositions in English, Polish, and Russian*, volume 237. Sagner Munich.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (2020). Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Crowston, K., Kwasnik, B., and Rubleske, J. (2010). Problems in the use-centered development of a taxonomy of web genres. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54(2):177–190.

Ferraresi, A., Zanchetta, E., Bernardini, S., and Baroni, M. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *The 4th Web as Corpus Workshop: Can we beat Google? (at LREC 2008)*, Marrakech.

Forsyth, R. and Sharoff, S. (2014). Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing*, 29:6–22.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Görlach, M. (2004). *Text types and the history of English*. Walter de Gruyter, Berlin.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., and Suchomel, V. (2013). The tenten corpus family. In *Proc Corpus Linguistics Conference*, pages 125–127, Lancaster.

Kanaris, I. and Stamatatos, E. (2007). Webpage genre identification using variable-length character n-grams. In *Proc ICTAI*.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *COLING '94: Proc. of the 15th. International Conference on Computational Linguistics*, pages 1071 – 1075, Kyoto, Japan.

Katinskaya, A. and Sharoff, S. (2015). Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In *Proc BSNLP*, Sofia.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35^{th} ACL/8^{th} EACL*, pages 32–38.

Kilgarriff, A. (2001). The web as corpus. In *Proc Corpus Linguistics 2001*, Lancaster.

Kilgarriff, A. and Suchomel, V. (2013). Web spam. In *Proc Web as Corpus workshop (WAC8) at Corpus Linguistics Conference*, Lancaster.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Kunilovskaya, M. and Sharoff, S. (2019). Building functionally similar corpus resources for translation studies. In *Proc RANLP*, pages 583–592, Varna.

Lee, D. (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.

Liu, B. and Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Matthiessen, C. M. (2015). Register in the round: registerial cartography. *Functional Linguistics*, 2(1):1–48.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proc. Workshop at ICLR'13*.

Nesi, H. and Gardner, S. (2012). *Genres across the Disciplines: Student writing in higher education*. Cambridge University Press, Cambridge.

Petrenz, P. and Webber, B. (2010). Stable classification of text genres. *Computational Linguistics*, 34(4):285–293.

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web. In Mehler, A., Sharoff, S., and Santini, M., editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Sharoff, S. (2018). Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Sharoff, S., Goldhahn, D., and Quasthoff, U. (2017). *Frequency Dictionary: Russian*, volume 9 of *Frequency Dictionaries*, chapter Corpus, pages 9–14. Leipziger Universitätsverlag. Uwe Quasthoff, Sabine Fiedler, Erla Hallsteindóttir (editors).

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Sinclair, J. (1991). *Corpus, Concordance and Collocation*. OUP, Oxford.

Sinclair, J. and Ball, J. (1996). Preliminary recommendations on text typology. Technical Report EAG-TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document.

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. Technical report, Oregon State University.

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proc CoNLL 2017 Shared Task*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, 21(3):319–353.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In Fisher, D. H., editor, *Proc ICML*, pages 412–420, Nashville, US.

Yogatama, D., Dyer, C., Ling, W., and Blunsom, P. (2017). Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*.

7 Appendix 1: Linguistic features

The order of the linguistic features and their codes are taken from (Biber, 1988). The conditions for detecting the features for English replicate the published procedures from (Biber, 1988), many of them are expressed via lists of lexical items or via POS annotations, which in this study are provided by UDPIPE (Straka and Straková, 2017). The Russian features are either based on translating the English word lists or on using identical or functionally similar constructions. For example, detecting `C12` (*do* as pro-verb in English) in Russian is based on detecting ellipsis in conditions similar to those used for detecting `C12` in English. Our tool can also detect the same set of linguistic features in French and Spanish. However, they have not been used in Section 4 due to the absence of annotated corpora to train the neural classifier.

Code	Label	Condition
A01	past verbs	VERB, Tense=Past
A03	present verbs	VERB, Tense=Pres
B04	place adverbials	ADV, lex in (*aboard,above,abroad,across...*)
B05	time adverbials	ADV, lex in (*afterwards,again,earlier...*)
C06	first person pronouns	PRON, lex in (*I,we,me,us,my..*)
C07	second person pronouns	PRON, lex in (*you,your,yourself,yourselves*)
C08	third person pronouns	PRON, lex in (*she,he,they,her,him,them,his...*)
C09	impersonal pronouns	Conditions from (Biber, 1988)
C10	demonstrative pronouns	Conditions from (Biber, 1988)
C11	indefinite pronouns	PRON, lex in (*anybody,anyone,anything,everybody...*)
C12	*do* as pro-verb	Conditions from (Biber, 1988)
D13	wh-questions	Conditions from (Biber, 1988)
E14	nominalizations	lex ends with ('tion','ment','ness', 'ism')
E16	nouns	Conditions from (Biber, 1988)
G19	*be* as main verb	Conditions from (Biber, 1988)

Continued on next page

Continued from previous page

Code	Label	Condition
H23	wh-clauses	Conditions from (Biber, 1988)
H34	sentence relatives	Conditions from (Biber, 1988)
H35	causatives	CONJ, lex in (*because*)
H36	concessives	CONJ, lex in (*although,though,tho*)
H37	conditionals	CONJ, lex in (*if, unless*)
H38	other subordination	Conditions from (Biber, 1988)
I39	prepositions	ADP
I40	attributive adjectives	Conditions from (Biber, 1988)
I41	predicative adjectives	Conditions from (Biber, 1988)
I42	adverbs	ADV
J43	type-token ratio	Using 400 words as in (Biber, 1988)
J44	word length	Average length of orthographic words
K45	conjuncts	Conditions from (Biber, 1988)
K46	downtoners	lex in (*almost,barely,hardly,merely..*)
K47	general hedges	lex in (*maybe, at about, something like..*)
K48	amplifiers	lex in (*absolutely,altogether,completely,enormously...*)
K49	general emphatics	Conditions from (Biber, 1988)
K50	discourse particles	Conditions from (Biber, 1988)
K55	public verbs	VERB, lex in (*acknowledge,admit,agree...*)
K56	private verbs	VERB, lex in (*anticipate,assume,believe...*)
K57	suasive verbs	VERB, lex in (*agree,arrange,ask...*)
K58	seem/appear	VERB, lex in (*appear, seem*)
L52	possibility modals	VERB, lex in (*can,may,might,could*)
L53	necessity modals	VERB, lex in (*ought,should,must*)
L54	prediction modals	VERB, lex in (*shall,will,would*), excluding future tense
N59	contractions	Conditions from (Biber, 1988)
N60	that deletion	Conditions from (Biber, 1988)
P66	synthetic negation	Conditions from (Biber, 1988)
P67	analytic negation	Conditions from (Biber, 1988)

8 Appendix 2: Definitions of FTDs

The order and the definitions of the FTDs are taken from (Sharoff, 2018).

Code	Label	Definition	Prototypes
A1	argument	To what extent does the text argue to persuade the reader to support an opinion?	Argumentative blogs, editorials or opinion pieces
A4	fiction	To what extent does the text narrates a fictional story?	Novels, poetry, myths, film plot summaries
A7	instruct	To what extent does the text aim at teaching the reader how something works or at giving advice?	Tutorials, FAQs, manuals
A8	news	To what extent does the text appear to be an informative report of events recent at the time of writing?	Newswires

Continued on next page

Continued from previous page

Code	Label	Definition	Prototypes
A9	legal	To what extent does the text specify a set of regulations?	Laws, contracts, copyright notices, terms&conditions
A11	personal	To what extent does the text report a first-person story?	Diary entries, travel blogs
A12	promotion	To what extent does the text promote a product or service?	Adverts, promotional postings
A14	adacemic	To what extent does the text present academic research?	Academic research papers
A16	info	To what extent does the text provide reference information?	Encyclopedic articles, definitions, specifications
A17	review	To what extent does the text evaluate a specific entity by endorsing or criticising it?	Reviews of a product, location or performance

Address for correspondence
Serge Sharoff
e-mail: s.sharoff@leeds.ac.uk
Centre for Translation Studies
University of Leeds, LS2 9JT, Leeds
United Kingdom