Applying Multi-dimensional Analysis to a Russian Webcorpus: Searching for Evidence of Genres

Anisya Katinskaya
Russian State University for the Humanities
Moscow, Russia
a.katinsky@gmail.com

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

Abstract

The paper presents an application of Multi-dimensional (MD) analysis initially developed for the analysis of register variation in English (Biber, 1988) to the investigation of a genre diverse corpus, which was built from modern texts of the Russian Web. The analysis is based on the idea that each linguistic feature has different frequencies in different registers, and statistically stable co-occurrence of linguistic features across texts can be used for automatic identification of texts with similar communicative functions. By using a software tool which counts a set of linguistic features in texts in Russian and by performing factor analysis in R, we identified six dimensions of variation. These dimensions show significant similarities with Biber's original dimensions of variation. We studied the distribution of texts in the space of the dimensions of our factors and investigated their link to 17 externally defined Functional Text Dimensions (Forsyth and Sharoff, 2014), which were assigned to each text of the corpus by a group of annotators. The results show that dimensions of linguistic feature variation can be used for better understanding of the genre structure of the Russian Web.

1 Introduction

Automatic genre classification is an important step in different kinds of text processing tasks and in scientific research of linguists working with corpus data. As Mikhail Bakhtin (1996) said about genres: “Specific function (scientific, technical, journalistic, official, and informal) and specific conditions of each communication field generate specific genres, i.e., thematic, compositional, and stylistic types of utterances”. This idea has special importance for texts from the Web since this communication field is in the process of continuous change, so it is difficult to make a fixed classification of Web genres, so that the annotators normally disagree about the genre labels (Sharoff et al., 2010). For that reason, we will use the Functional Text Dimensions (FTDs) which allow determining the similarity of texts in terms of their functional characteristics (Forsyth and Sharoff, 2014).

Since Biber’s work (Biber, 1988) the idea for classification via a link between genres and their linguistic categories has been implemented by numerous researchers (Nakamura, 1993; Michos et al., 1996; Sigley, 1997; Stamatatos et al., 2001; Finn et al., 2002; Finn and Kushmerick, 2003; Lee and Myaeng, 2004). Linguistic parameters of different genres for Russian have also been studied. Braslavski (2011) investigated genre analysis in the context of Web search. A small set of simple syntactic constructions was used to distinguish fiction, news and scientific texts in (Klyshinsky et al., 2013). These three types of texts were also investigated in the space of 11 low-level frequency parameters, e.g., type/token ratio or verb frequency, in (Yagunova and Pospelova, 2014).

Our idea is to implement the MD analysis for Russian and, firstly, to test whether this approach could be used for finding sets of linguistic features covering a wide range of web texts rather than just three genres. Secondly, unlike (Sharoff et al., 2010) and (Forsyth and Sharoff, 2014) these studies have not investigated the issue of inter-annotator reliability.

MD analysis has not been applied to Russian language before, but it ha been used to analyse texts in English (Biber, 1988; de Mönnink et al.,

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 65–74,
Hissar, Bulgaria, 10–11 September 2015.

2003; Crossley and Louwerse, 2007; Daems et al., 2013), Nukulaelae Tuvaluan (Besnier, 1988), Somali (Biber and Hared, 1994), Korean (Kim and Biber, 1994), Spanish (Biber and Tracy-Ventura, 2007; Parodi, 2007), Gaelic (Lamb, 2008), Brazilian Portuguese (Berber Sardinha et al., 2014).

In spite of variation in the use of terms “genre” and “register” among researchers, this study refers to externally recognized text types, e.g., news or fiction, as “genres” (Lee, 2001).

In Section 2 we shortly describe the methodology of the MD approach. In Section 3 we describe the corpus we used, the principles how we chose a set of linguistic features and the software tool built for extracting these features from texts. In Section 4 we analyse the dimensions of linguistic feature variation resulting from factor analysis and briefly compare them to dimensions from other works. Section 5 shows the distribution the FTDs in the factor space. In Section 6 we analyse the results and discuss possible applications.

2 Short Overview of Multi-dimensional Analysis

The procedure of the MD analysis can be described in several methodological steps (Biber et al., 2007). Firstly, texts are collected as a corpus representing the variety of genres. Then a research is performed to define a set of linguistics features to be found in texts of the corpus along with functions of features.

The third step is to develop a computer program to automatically identify linguistic features. After tagging of the corpus and correcting results by the researcher, additional programs compute frequency counts of each linguistic feature in each text, and the counts are normalized.

The next step is to conduct the procedure of finding latent features (co-occurrence patterns) among the linguistic features using factor analysis of the obtained frequency counts. Each set of co-occurrence patterns is referred to as a factor. The factors are interpreted in terms of their functions as underlying dimensions of linguistic feature variation. Factor scores of each text are calculated with respect to each dimension of variation. Then mean factor scores for each genre are computed and compared to each other to analyse specific linguistic features of each genre.

3 Data acquisition

3.1 Description of the Corpus

The corpus used for the experiment consists of 618 texts (see Table 1). The texts were collected from Open Corpora (Bocharov et al., 2011), as well as from news portals (e.g., chaskor.ru, ru.wikinews.org, ria.ru, lenta.ru), Wikipedia and other online encyclopedias (e.g., krugosvet.ru), online magazines (e.g., vogue.ru, popmech.ru) and text collections (primarily fiction, e.g., lib.ru), blogs (e.g., vk.com, lifejournal.com, habrahabr.ru), forums (e.g., forum.hackersoft.ru, litforum.ru), scientific and popular scientific journals (e.g., cyberleninka.ru, sci-article.ru), promotional web-sites (e.g., mvideo.ru, avito.ru), legal resources (e.g., base.garant.ru, consultant.ru), and other online resources.

Number of texts: 618
Number of words: 741831
Number of sentences: 52031
Length of texts: 88 (min), 10848 (max), 573 (med.)
Number of texts < 200 words: 133
Number of texts > 200 words: 482

Table 1: Annotated corpus used in study.

Noticeable differences in the length of texts are mostly determined by their genre characteristics: it is difficult to find a very long advertisement or a joke and a very short scientific paper or a law.

Despite the fact that we could have used a big collection of texts from the Web, at this stage we decided to settle on a manually built and quite small corpus for several reasons. Firstly, even in texts obtained from the same source, e.g., news or blog posts, we can often find considerable variation in subgenres. For instance, one news text from chaskor.ru expresses the author’s attitude to the topic, whereas the second one is relatively neutral, so these two texts from the same news portal differ in their FTD A17 (evaluation).[1]

Secondly, we tried to obtain maximal variety of Web genres. Thirdly, annotation of texts on 17 parameters is very labour-intensive, while we wanted to ensure a reasonable level of inter-annotator agreement. A significant part of the corpus was annotated by 11 annotators with three annotations per text. Then the full corpus annotation has been revised by 2 annotators.

[1] http://goo.gl/XZdg1t and http://goo.gl/wMkuCL

Class	Main FTDs	Num of texts	Main interpretation
C1	A1, A13	21	Argumentative texts
C2	A11	101	Personal blogs
C3	A8	79	News reports
C4	A9	76	Legal texts
C5	A12	66	Advertisement
C6	A14	59	Scientific texts
C7	A16 (-A14)	186	Encyclopedic texts
C8	A7	33	Instructional texts
C9	A4, A16	10	Fictional texts

Table 2: Classes of the FTDs.

The annotated texts were clustered with scores of 17 FTDs as predictors. Clustering was performed by a variant of kNN, which had additional constraints to limit the size of small clusters; the method is fully described in (Lagutin et al., 2015). After manual analysis of the clustering results, nine stable classes (C1-C9) were revealed and interpreted as reliably annotated genres, which can also be described on the basis of their principal FTDs (see Table 2). In our paper below we will treat these classes as genres for illustrating dimensions of linguistic feature variation.

3.2 Linguistic Features

Sets of grammatical and lexico-grammatical linguistic features identified by Biber's tagger range from 60 to 120+ linguistic variables. The largest inventory (Berber Sardinha et al., 2014) comprises 190 features. For our purposes, we relied on the list presented in the Appendix 2 in (Biber et al., 2007) and the description of features in the manual of Multidimensional Analysis Tagger (Nini, 2014) that replicates Biber's tagger, while adapting the English features to reflect Russian grammar.

There are several reasons why we have chosen a relatively short list of features. Firstly, necessary features should be accessible for extraction from texts by the tools available to us (morphological tagger and our program, which we will describe further). For instance, it is very difficult to specify the difference between phrasal coordination (e.g., coordination of extended noun phrases) and independent clause coordination, using only POS tags and a small window (from 1 to 10 words) for shallow parsing. We plan to add a syntactic module to the next version of the feature tagger.

Secondly, each feature reflected the Russian grammar. For example, researchers disagree with respect to the existence of proforms of verbal phrases in Russian. Preposition stranding (when a preposition with an object occurs somewhere other than adjacent to its object, e.g., *the thing I was thinking of*) does not exist in Russian; therefore, we did not include such linguistic features to the list. The reflexive pronouns in Russian do exist, but their forms are different from the reflexive pronouns in English, which derive from personal pronouns and can be added to the corresponding features as it was done in MAT v.1.2. (*myself* as the first person pronoun, *itself* as the third person pronoun, and so on). For this reason, the Russian reflexive pronouns are considered as an independent linguistic feature.

Under nominalizations we mean verbal nouns like *возрождение*, 'revival', or *вход*, 'entrance'. A feature called 'wh-relative' means relative clause with a wh-element (e.g., *который*, 'which') that is fronted to the beginning of the clause. 'Wh-question' marks interrogative sentences with a wh-element at the left edge (e.g., *кто*, 'who'). A feature 'that-complement' means a complement clause with the complementizer *что* or *чтобы* ('that') at the left edge. More details see in (Bailyn, 2012).

The third reason is that we want to test our hypothesis about appropriateness of the Multi-dimensional approach for the task of automatic genre classification of texts of the Russian Web. The list of features can be extended in the future.

3.3 MD Analysis for Russian

Biber's computational tools have been used to tag lexical, grammatical, and syntactical features and to count their frequencies in each analysed text. Using large-scale dictionaries and context-dependent disambiguation algorithms, the tagger marks word classes and syntactic information. The description of the early version of the tagger is presented in (Biber, 1988), computational methods are outlined in (Biber, 1993a; Biber et al., 2007).

We have developed a program in Python, which uses a morphologically parsed corpus as an input.[2]

[2] https://github.com/Askinkaty/MDRus_analyser

Factors	**PA1**	**PA2**	**PA3**	**PA4**	**PA5**	**PA6**
PA1	1.00	-0.01	-0.16	0.30	0.49	0.17
PA2	-0.01	1.00	-0.28	0.13	0.17	-0.03
PA3	-0.16	-0.28	1.00	-0.53	-0.43	-0.22
PA4	0.3	0.13	-0.53	1.00	0.46	0.48
PA5	0.49	0.17	-0.43	0.46	1.00	0.20
PA6	0.17	-0.03	-0.22	0.48	0.20	1.00

Table 3: Inter-factor correlation.

	PA4	**PA1**	**PA2**	**PA3**	**PA5**	**PA6**
Proportion Variance	0.08	0.09	0.06	0.06	0.05	0.04
Cumulative Variance	0.08	0.17	0.23	0.29	0.34	0.38
Proportion Explained	0.22	0.22	0.15	0.15	0.14	0.11
Cumulative Proportion	0.22	0.45	0.60	0.75	0.89	1.00

Table 4: Output of the factor analysis.

RFTagger (Schmid and Laws, 2008) was used to process the corpus with the accuracy rate close to accuracy of the tools described in (Sharoff and Nivre, 2011), which is near 95-97%.[3] For the lexical features we used dictionaries derived from the Russian National Corpus.[4]

Then we can run our feature analyser for each text to identify and count linguistic variables. We have developed a processing algorithm for each feature, considering the requirements of Russian grammar and possible ambiguity, which we try to resolve by relatively simple methods such as specifying contextual conditions and exceptions. For example, we have to identify time adverbs *весной* ('in spring') or *порой* ('at times'), which might be confused with nouns. In almost every case RFTagger processes them as nouns. Therefore, we should specify the context in which these words cannot be used as adverbs (e.g., if one of these words agrees with an adjective or a pronoun, it is likely to be a noun).

All processing rules were tested on wider outputs obtained from the General Internet Corpus of Russian (GICR) (Piperski et al., 2013). Because we work with texts from the Web, we took into account some possible mistakes. For instance, people often make mistakes with conjunctions like *вследствие того что* ('because of'), *ввиду того что* ('in view of that') and miss commas or white spaces between words in these complex conjunctions. For our practical purposes, we have attempted a unified processing of the most common cases of this sort.

Unlike Biber, we did not edit the results of feature extraction because it is labour-intensive and not consistent with the idea of applying the method to a large-scale corpus in the next step. Accuracy of the most complicated rules (e.g., detection of proforms of noun phrases) is around 67-85%, simple rules have much higher accuracy, mostly above 95%.

Counted frequencies of all features in each text (except for word length, sentence length, and type/token ratio) are divided by the number of words in the texts. As an output of the program we get a matrix of 618 to 40 including the frequencies of 40 linguistic variables for each text.

4 Searching for Dimensions of Variation

4.1 Factor Analysis

Factor analysis is an important part of the MD analysis. It is a useful tool for investigating the underlying structure of complex phenomena and for reducing data to a smaller set of latent variables called factors. Each of the observed variables is assumed to depend on a linear combination of factors, and the coefficients (the strength of relation to a factor) are known as factor loadings. For the justification of factor analysis for genre research we refer the reader to (Biber, 1988).

[3] http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/

[4] http://www.ruscorpora.ru

Linguistic features, which are observed variables in our study, are supposed to co-occur in different texts. We are interested in systematic patterns among this co-occurrence. Patterns of variation reflect an underlying system of factors, with which variables have strong association. A rotated factor analysis was performed in R with Promax rotation since we assume possible correlation among factors (Kabakoff, 2011).

The inter-factor correlation ranges from -0.53 to 0.49, see Table 3. Other output of the factor analysis is presented in Table 4.

Table 5 presents linguistic features with factor loadings over 0.3 or below -0.3 correlating with corresponding factors. Features with lower loadings cannot be considered as informative for interpretation of factors. Large loading means stronger correlation between a feature and a factor. Only three features have been excluded (verbal adverb, concessive subordinate clauses, and pied-piping, which for Russian is interpreted as a preposition moved to the front of its relative clause) due to low factor loadings. Six dimensions of feature variation were selected as optimal for our data. Dimensions 3 and 5 are relatively small: each of them includes only three features.

The factor structure is very stable and does not change significantly if different models (maximum likelihood, iterated principal axis, etc.) or different types of rotation are used.

4.2 Interpretation of Dimensions of Variation

Each factor combines linguistic features that serve related communicative functions. It is also important that a feature can have positive or negative loading in a factor; therefore, features with opposite loadings have a complementary distribution. In our case, only three factors have so called negative features, i.e., features with negative loadings. For convenience, we will call the obtained factors as dimensions and rename PA4, PA1, PA2, PA3, PA5, and PA6 to D1, D2, D3, D4, D5, and D6 correspondingly.

The positive features of Dimension 1 (D1) are 1st person pronouns, 2nd person pronouns, exclamation, and wh-questions what can be associated with interactivity and indicates dialogue. A possible interpretation of place adverbs in D1 is proposed in (Biber, 1988), according to which place and time adverbs are ‘reflecting the description of other people in particular places and times’. Nouns, long words, prepositional phrases, and attributive adjectives mostly relate to the informational purpose (high frequency of nouns and modifiers of noun phrases usually signs high informational saturation). It follows that D1 is very close to the ‘Informational vs. Involved’ dimension in (Biber, 1993b) since it also includes such features as nouns, word length, prepositional phrases, attributive adjectives vs. 1st and 2nd personal pronouns and wh-questions.

Dimension 1: interactive/informative
POSITIVE FEATURES: 1th person pronoun, 2nd person pronoun, place adverb, exclamation, wh-question
NEGATIVE FEATURES: word length, nouns, attributive adjective, all prepositional phrases (total PP)

Dimension 2: presentation of personal view of subject/impersonal
POSITIVE FEATURES: pro-form of noun phrase (pro-form of NP), negation, mental verb, that-complement, speech verb, wh-relative, 3rd person pronoun, indefinite pronoun, predicative adjective, pro-form of adjective phrase (pro-adjective), reflexive pronoun, causative subordinate clause

Dimension 3: narrative/non-narrative
POSITIVE FEATURES: past tense, perfect aspect
NEGATIVE FEATURES: present tense

Dimension 4 : abstract/non-abstract
POSITIVE FEATURES: passive participle clause, agentless passive, nominalization, passive with agent, active participle clause, type/token ratio
NEGATIVE FEATURES: sentence length

Dimension 5
POSITIVE FEATURES: all adverbs, time adverb, indefinite pronoun

Dimension 6: directive/ non-directive/
POSITIVE FEATURES: infinitive, conditional subordinate clause, imperative mood, purpose subordinate clause

Table 5: Result of the factor analysis (factorial structure).

Dimension 2 (D2) combines features that can be interpreted as a report of speech of others (3th person pronouns, speech verbs) and features that can be used to frame a personal attitude towards some topic (mental verbs, that-complements, reflexive pronouns). Some features have a referential meaning: wh-relatives (elaborated reference), pro-forms of noun phrases, pro-adjectives, and indefinite pronouns, which can be interpreted as generalized reference in a shared context of communication between the author and the reader. D2 is somewhat similar to the dimension in (Grieve et al., 2010) called Thematic Variation Dimension and also similar in several features to the argumentative Dimension 2 in (Berber Sardinha et al., 2014). We will interpret D2 as the dimension presenting an informal personal argumentation or personal opinion on something like other's words or a context that is well known to the reader.

Two positive (past tense and perfect aspect) and one negative (present tense) features allow interpreting of Dimension 3 (D3) as narrative vs. non-narrative. A similar dimension in (Biber, 2004) has the label called 'Narrative-focused discourse'.

Dimension 4 (D4) includes a set of features like agentless passive, passive with an agent, many non-repeating words, and nominalizations and can be interpreted as presenting an abstract style of writing. It also correlates with high frequency of active and passive participle clauses. The negative correlation of this dimension with the average sentence length is unexpected. D4 is almost similar to the dimension called 'Abstract vs. Non-Abstract style' in (Biber, 1993b).

It is more difficult to interpret Dimension 5 (D5), which includes only the total number of adverbs, time adverbs (both usually narrative features), and indefinite pronouns. In the next section we will investigate which kind of texts is characterized by D5. This dimension is stable in the space of 40 features. However, having run the analysis with 63 linguistic features (it has not been fully tested at the time of writing), we got that adverbs and indefinite pronouns do not form a separate dimension and correlate with other dimensions along with place adverbs.

The features of Dimension 6 (D6) (infinitives, conditional subordinate clauses, imperative mood, purpose subordinate clauses) reflect the directive function. Purpose subordinate clauses mostly refer not to how some action can be performed but for what purpose. This dimension can be compared to Biber's dimension named 'Overt expression of persuasion' including infinitives, conditional subordination, and different modals (Biber, 1993b).

5 Distribution of Classes of the FTDs in the Space of Dimensions

It is interesting to see how the classes of the FTDs, which we interpret as genres in this study, relate to the six dimensions of feature variation. For this purpose, we counted dimension scores of each class by summation of dimension scores of texts having a value 2 on the corresponding FTD (or FTDs), see Table 2.

Class	D1	D2	D3	D4	D5	D6
C1	3.2	7.7	0.6	-4.1	4.4	2.8
C2	13.7	12.3	3.5	-12.9	12.0	7.7
C3	-4.0	-1.8	2.5	0.4	-1.9	-1.8
C4	-16.8	-15.7	-6.2	17.9	-15.8	-8.5
C5	-4.6	-8.1	-2.8	4.5	-4.8	-2.1
C6	-5.6	-6.0	-3.2	7.0	-5.8	-5.7
C7	-6.5	-7.8	-1.7	6.1	-5.6	-4.5
C8	6.0	-2.4	1.5	-3.5	3.0	9.0
C9	24.8	12.3	9.5	-19.5	13.4	9.8

Table 6: Medians of dimension scores for C1-C9.

Medians of the dimension scores of the classes are presented in Table 6. The difference between scores is statistically significant with p-value < 0.05.

Clusters C1 and C2 are quite close to each other; however, all average factor scores of C2 are considerably stronger than average factor scores of C1. The first class is a class of argumentative texts. Samples from C1 mostly include political articles, blog posts about social situation, and religious texts. Most of the texts are non-informative, slightly interactive, non-narrative, non-abstract, and slightly directive (religious texts are very directive). C1 has one main dimension D2, which means expressing a personal point of view on a particular subject and on positions of other people.

C2 is a big class of different personal blogs. These blogs are non-abstract, highly interactive, and expressing personal positions about a subject. The class has a relatively high value on the narrative dimension D3 because it is heterogeneous to some extent and includes a set of narrative personal stories.

A number of reviews from C2 (blogs with personal reasons about something like a political situation, a tour, a concert, or a book) have especially high scores on the argumentative dimension D2. So, if we want to distinguish reviews from other types of blogs automatically, we should take this feature into account.

C3 is a class of news reports. The texts of this class appear to be informative (D1), not presenting personal thoughts about describing events, not sharing the same context with readers (D2), narrative (D3), mostly neutral on 'abstract vs. non-abstract style' (D4), and non-directive (D6).

All legal texts belong to C4. This class is characterized by the highest value of D4, and other dimension values are low. The texts are very abstract, very informative, non-narrative, non-directive, and not presenting any personal positions about subjects. C4 is the most informative class in the set; its texts are characterized by long words, a large number of noun phrases, and its modifiers.

C5, the class of advertisements, has a large standard deviation for D1, D2 and D4-D6. The analysis of the texts which factor scores are far from the means of the dimensions indicated above showed that C5 includes very different sets of advertisements. It has an impact on the resulting dimension scores of the class. Most of advertising texts are informative, not showing personal argumentation about anything, and not highly directive; however, in the corpus we have several advertisements on dating sites which have appeal to potential partners and strong motivation to write a respond. As opposed to these addressee and personal focused texts, another set of advertisements is highly abstract because it describes technically complicated products (cameras, automobiles, synthesizers, etc.). It is unusual for advertisements in our corpus and more typical for scientific texts. So, we could see that values of dimensions scores could help us to find different subgenres in the genres of advertisement in the present corpus.

The type of texts related to a field of Science and Technology is included in the class C6. All the texts of the class are informative, non-narrative, non-directive, abstract, and not presenting a personal position. It is relevant for the scientific articles presented in the corpus.

C7 (encyclopedic texts) and C8 (instructive texts) have large standard deviations for D1, D2, D4, and D5. C7 includes texts which are highly informative and abstract, but also it contains a set of texts which do define some topics but not encyclopedic at all (interactive, non-abstract, and presenting a personal argumentation), e.g., a description of an episode from The Simpsons, a movie review or an obituary. We suppose that the problem with C7 can be solved by adding to the corpus more variety of texts defining some topic, especially texts not written by academic language. On the other hand, it might be reasonable to suppose that we had some errors in the annotation on the FTD 16 (defining a topic).

The main characteristic of C8 is high values of D6, which has a directive meaning. Looking at the samples from the class, we can understand that large values of a standard deviation are due to the fact that C8 consists of two sets of different instructions. The first small set consists of technical instructions, user's guides, and recipes; they all are informative, non-interactive, and abstract. The second big set includes highly interactive and non-abstract texts with some kind of informal communication with the reader, for example, a blog post advising on how to quit smoking. This spread of dimension values shows two different types (or subgenres) of instructions in our corpus.

Fiction texts of C9 are highly interactive, presenting personal attitude and argumentation, highly narrative, non-abstract, and directive. We undoubtedly should extend the corpus for further research because it contains only 10 fictional texts although they are quite long. Even though it is difficult to analyse the class C9, it shows the highest values on D3 (narrative vs. non-narrative). C9 is also highly marked on D2 which once again shows the close proximity of D2 to a personal side of discourse. Only C9 has as high values on D6 as C8. Analysis of the samples of C9 showed that it is mostly due to specific features of fictional texts in our corpus (high frequency of infinitives, purpose and conditional subordinate clauses).

Concerning D5, which includes only such features as total adverbs, time adverbs, and indefinite pronouns, we have a hypothesis that this dimension is a part of some other dimension, which might be or might be not presented in our current set. After analysing the medians of the scores on D5, we can suppose that D5 is close to D1 or D2. High positive scores on D5 mark mostly personal blogs and fictional texts. Negative scores are typical for legal texts, scientific, encyclopedic texts, and adverts. All texts labeled by the highest values of D5 are personal blog stories, so we assume that D5 is a part of D2. D5 is also very similar to the negative pole of the dimension called 'Elaborated vs. Situated reference' in (Biber, 1993b) including such features as place, time and other adverbs.

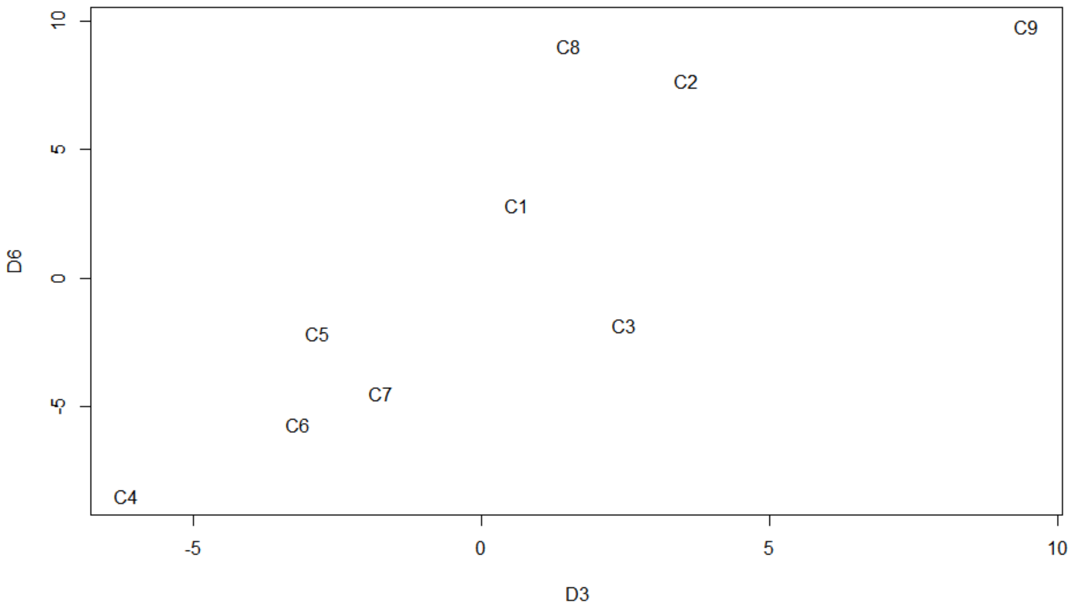

Figure 1: Distribution of 9 classes of the FDTs in D3 (narrative vs. non-narrative) and D6 (directive vs. non-directive).

Figure 3 presents an illustration how the classes are located in the space of D3 (narrative vs. non-narrative) and D6 (directive vs. non-directive).

6 Conclusions and Further Research

We have investigated variations in the linguistic properties of texts from the Russian Web by applying Biber's Multi-dimensional analysis to a Web corpus and have successfully used a much bigger Web corpus (GICR) to build a linguistic tagger.

By using factor analysis, we found six dimensions around which all functionally similar linguistic features are grouped for the presented corpus, and which were interpreted from the point of view of their functions. The dimensions obtained in this study are very similar to (Biber, 1993b) except for D2, which combines features usually found in several other dimensions such as ‘Involved’ (e.g., mental verbs or negation), ‘Elaborated reference’ (e.g., relative clauses), and 'Narrative' (e.g., 3rd personal pronouns, speech verbs). A larger corpus might provide a better match to the classical features.

Russian is not fundamentally different from English with respect to implementation of MD analysis; many features can be mapped, even though more morphological and syntactical features need to be processed.

The results of the MD analysis show that the classes of the FTDs (close to traditional genres) and the dimensions of language variation in Russian have evident connection. Every class has its own place in the multidimensional space of linguistic features. Deviations in dimension values for each text in each cluster allow us to find errors in annotation or functional groups of texts within a cluster (e.g., technical instructions vs. advice in C8). This shows that the MD approach can be used for finding text features specific for different genres and also for detecting fine-grained differences between subgenres.

The FTDs are not genres, but we assume that different genres in big corpora can be described by sets of different FTDs, so we should be able to identify them in texts. Our analysis shows that every major FTD describing a genre corresponds to a set of linguistic features. This could be used for the purpose of the automatic genre classification (the results of the first experiments with classification see in Lagutin et al., 2015).

In further research we intend to examine the FTDs in the space of an extended set of linguistic features, to experiment with a bigger corpus and to add discourse structure features.

Acknowledgements

This study has been supported by a joint project with Skoltech and Moscow State University, ‘Brains, minds and machines’ (Joint Lab. Agr. - No 081-R 1.10.2013).

References

John F. Bailyn. 2012. *The Syntax of Russian*. Cambridge: CUP, pages 84–109.

Mikhail Bakhtin. 1996. The problems of speech genres [Problemy rechevykh zhanrov]. *Russian dictionaries [Risskie slovari]. Collected writings*. Moscow, pages 159–206.

Tony Berber Sardinha, Carlos Kauffmann, and Cristina Mayer Acunzo. 2014. A multi-dimensional analysis of register variation in Brazilian Portuguese. *Corpora*, volume 9(2), pages 239–271.

Niko Besnier. 1988. The linguistic relationships of spoken and written Nukulaelae registers. *Language*, volume 64(4), pages 707–736.

Douglas Biber. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language*, volume 62, pages 384–414.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge: CUP.

Douglas Biber. 1993a. Using register-diversified corpora for general language studies. *Computational Linguistics*, volume 19, pages 219–241.

Douglas Biber. 1993b. The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings. *Computers and the Humanities*, volume 26, pages 331–345.

Douglas Biber. 2004. Conversation text types: a multi-dimensional analysis. In Gérald Purnelle, Cédrick Fairon, and Anne Dister (eds.) *Le poids des mots: Proc. of the 7th International Conference on the Statistical Analysis of Textual Data*, Louvain: Presses universitaires de Louvain, p.15–34.

Douglas Biber and Mohamed Hared. 1994. Linguistic correlates of the transition to literacy in Somali: language adaptation in six press registers. In Douglas Biber and Edward Finegan (eds.) *Sociolinguistic Perspectives on Register*, Oxford, pages 182–216.

Douglas Biber and Nicole Tracy-Ventura. 2007. Dimensions of register variation in Spanish. In Giovanni Parodi (ed.) *Working with Spanish Corpora*, London: Continuum, pages 54–89.

Douglas Biber, Ulla Connor, and Thomas A. Upton. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*, Amsterdam – Philadelphia, pages 261–271.

Victor Bocharov, Svetlana Bichineva, Dmitry Granovsky, et al. 2011. Quality assurance tools in the OpenCorpora project. *Proc. Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, pages 101–110.

Pavel Braslavski. 2011. Marrying relevance and genre rankings: an exploratory study. *Genres on the Web Computational Models and Empirical Studies. Text, Speech and Language Technology*, volume 42, pages 191–208.

Scott A. Crossley and Max Louwerse. 2007. Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics*, volume 12(4), pages 453–478.

Jocelyne Daems, Dirk Speelman, and Tom Ruette. 2013. Register analysis in blogs: correlation between professional sector and functional dimensions. *Leuven Working Papers in Linguistics*, volume 2(1), pages 1–27.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, pages 1–15.

Aidan Finn, Nicholas Kushmerich, and Barry Smyth. 2003. Learning to classify documents according to genre. *IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis*, Acapulco, pages 35–45.

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

Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova. 2010. Variations among blogs: a multi-dimensional analysis. In Alexander Mehler, Serge Sharoff, and Marina Santini (eds.) *Genres on the Web: Computational Models and Empirical Studies*, Berlin – New York: Springer, pages 303–323.

Robert Kabakoff. 2011. *R in Action. Data Analysis and Graphics with R*. New York, pages 347–348.

YouJin Kim and Douglas Biber. 1994. A corpus-based analysis of register variation in Korean. In Douglas Biber and Eric Finegan (eds.) *Sociolinguistic Perspectives on Register*, Oxford, pages 157–181.

Eduard Klyshinsky, Natalia Kochetkova, Oksana Mansurova, Elena Iagounova, Vadim Maximov, and Olesia Karpik. 2013. Development of Russian subcategorization frames and its properties investigation. *Keldysh Institute preprints*, Moscow, no. 41, pages 1–23.

Mikhail Lagutin, Anisya Katinskaya, Vladimir Selegey, Serge Sharoff, and Alexey Sorokin. 2015. Automatic classification of web texts using Functional Text Dimensions. *Proc. Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, volume 1, pages 398–414.

William Lamb. 2008. *Scottish Gaelic Speech and Writing: Register Variation in an Endangered Language*. Belfast: Cló Ollscoil na Banríona.

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*, volume 5(3), pages 37–72.

Yong-Bae Lee and Sung Hyon Myaeng. 2004. Automatic identification of text genres and their roles in subject-based categorization. *Proc. of the 37th Hawaii International Conference on System Science (HICSS ’04)*, pages 1–10.

Stephanos E. Michos, Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. 1996. An empirical text categorizing computational model based on stylistic aspects. *Proc. of the 8 the International Conference on Tools with Artificial Intelligence (TAI’96),* pages 71–77.

Inge de Mönnink, Niek Brom, and Nelleke Oostdijk. 2003. Using the MF/MD method for automatic text classification. In Sylviane Granger and Stephanie Petch Tyson (eds.) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*, Amsterdam, pages 15–25.

Junsaku Nakamura. 1993. Statistical methods and large corpora – a new tool for describing text type. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli (eds.) *Text and technology*, Philadelphia – Amsterdam, pages 291–312.

Andrea Nini. 2014. *Multidimensional Analysis Tagger 1.2, Manual*. Available: http://sites.google.com/site/multidimensionaltagger

Giovanni Parodi. 2007. Variation across registers in Spanish: exploring the El Grial PUCV corpus. In Giovanni Parodi (ed.) *Working with Spanish Corpora,* London: Continuum, pages 11–53.

Alexander Piperski, Vladimir Belikov, Nikolay Kopylov, Vladimir Selegey, and Serge Sharoff. 2013. Big and diverse is beautiful: a large corpus of Russian to study linguistic variation. *Proc. 8th Web as Corpus Workshop (WAC-8),* pages 24–29.

Alexey Sorokin, Anisya Katinskaya, and Serge Sharoff. 2014. Associating symptoms with syndromes: reliable genre annotation for a large Russian webcorpus. *Proc. Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, pages 646–659.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. *Proc. of the 22nd International Conference on Computational Linguistics (COLING’08)*, Manchester, volume 1, pages 777–784.

Serge Sharoff. 2010. In the garden and in the jungle: comparing genres in the BNC and the internet. In Alexander Mehler, Serge Sharoff, and Marina Santini (eds.) *Genres on the Web: Computational Models and Empirical Studies*, Berlin – New York: Springer, pages 149–166.

Serge Sharoff and Joakim Nivre. 2011. The proper place of men and machines in language technology: processing Russian without any linguistic knowledge. *Proc. Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, pages 591–604.

Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics,* volume 26(4), pages 471–495.

Robert Sigley. 1997. Text categories and where you can stick them: a crude formality index. *International Journal of Corpus Linguistics*, volume 2(2), pages 199–237.

Elena Yagunova and Anna Pospelova. 2014. Opyt primeneniya stilevyh i zhanrovyh harakteristik dlya opisaniya stilevyh osobennostej kollekcij tekstov. *Novye informacionnye tekhnologii v avtomatizirovannyh sistemah,* no. 17, pages 347–356.