AUTOMATIC CLASSIFICATION OF WEB TEXTS USING FUNCTIONAL TEXT DIMENSIONS

Lagutin M. B. (lagutinmb@mail.ru)*,[1],
Katinskaya A. Y. (a.katinsky@gmail.com)[2],
Selegey V. P. (Vladimir_S@abbyy.com)[2,3,4],
Sharoff S. (s.sharoff@leeds.ac.uk)[2,5],
Sorokin A. A. (alexey.sorokin@list.ru)*,[1,2,3]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Russian State University of Humanities, Moscow, Russia
[3]Moscow Institute of Physics and Technology, Moscow, Russia
[4]ABBYY, Moscow, Russia
[5]Leeds University, Leeds, UK

The work addresses automatic genre classification of Web texts. We show that functional text dimensions could be used for this tasks, with their stable combinations (clusters) corresponding to genres. Basing on a gold standard corpus, we construct a list of such genres. We also show that functional dimensions values can be automatically extracted from language features. In the conclusion we discuss the application of our results for automatic annotation of large Web corpora.

Introduction

It is well-known that an additional genre annotation could be very useful for corpus studies. Genre obviously affects lexical, syntactical and other text parameters. Using of genre information seems promising in different tasks of computational linguistic, e.g. for language models refinement. Therefore a reliable genre annotation is very desirable and ideally it should be done automatically since manual annotation even of medium-size corpora is an extremely labour-intensive task. It is especially true for corpora of texts from the Internet, in particular for General Internet-Corpora of Russian Language (GICR) which is currently under development (Belikov et al., 2012).

Standard systems cannot fit genre structure of Internet perfectly since they lack such concepts as “blog” or “forum discussion”. Mechanical extension of the system with new categories does not correct this deficiency: numerous annotation experiments have shown that inter-annotator agreement for Internet texts is rather decent. It is not due to a bad qualification of annotators or unclear instructions: most Internet texts demonstrate a real mixture of various genres without any clear border between them. That leads to an objective disagreement between annotators and low quality of automatic genre classification.

* Corresponding author

One possible way to avoid this problem is to replace genres with Functional Text Dimensions (FTDs). The system of FTDs of S. Sharoff (Forsyth and Sharoff, 2014) showed better inter-rater agreement both for Russian and English languages than genres of D. Biber (Egbert and Biber, 2013). The drawback of this approach is that assigned rates are uninformative: it is not very clear, for example, what means that a particular text has a value of 1 for the feature A7. Therefore, there is a need to establish the correspondence between the genre of a text and its FTDs.. Since the dimensions are not independent, some combinations of FTDs are more stable than others. Such stable and frequent combinations form analogues of traditional genres. Hence clustering of FTD values is an unavoidable preliminary stage of genre classification for web texts.

The present work extends our previous study (Sorokin, Katinskaya, Sharoff, 2014) addressing similar questions. We discovered several natural and stable clusters in the space of FTDs, however, their detection suffers from a serious noise. The noise originates both from imperfect annotation and an unbalanced corpus structure. We have tried to improve both the homogeneity of the corpus and the reliability of its annotation. We also made a preliminary experiment on automatic FTD detection. The achieved percentage was quite high (about 70%) which gives us hope for the further automatic genre annotation of the whole corpus or at least a large segment of it. We start the paper with describing our corpus and its preprocessing. Afterwards we make a statistical analysis of the FTD space and describe the algorithm of automatic genre classification. In the conclusion we discuss how to use our method for creating automatic genre annotation.

Corpus and functional dimensions

Our word is devoted to automatic genre classification. Since, in the strict sense, the notion of genre is not defined for texts from the Web, beforehand we analyze the space of Functional Text Dimensions (Forsyth, Sharoff, 2014) and how texts are located in this space. We used 17 FTDs given in Appendix I. In our previous studies we picked out a benchmark corpus of 618 texts. It contains most of the texts from our previous study (Sorokin, Katinskaya, Sharoff, 2014) except for the ones which do not permit reliable annotation by the FTDs. The corpus was enlarged by 90 texts from each of three popular platforms: blogs.mail.ru, vkontakte.ru and livejournal.com (its Russian segment). When collecting this corpus, we tried to cover as much various combinations of FTD values as possible. Each text was carefully annotated by two raters by 17 FTDs. Annotation scheme and guidelines were thoroughly examined during our previous studies, so we tried to make the annotation process as objective as possible. The presence of each dimension was rated on the following scale:

- 0—absent;
- 0,5—slightly;
- 1—partially;
- 2—present at most part.

The inter-rater agreement (Krippendorff α) achieved 90–95%, which is unattainable when using traditional genre systems. The annotation procedure was extensively tested in our previous research, so its results form a “gold standard” for future annotation studies.

Statistical analysis of annotation results

The intermediate goal of our study is to reveal stable combinations (or clusters) of FTD values ("pseudogenres"). Let us analyze the annotation results using the histogram below. The distribution of FTD values vary between dimensions. The features **A4**, **A5**, **A7**, **A18** are more rare than others and **A19** was detected only in 13 texts. So there is little hope to find any clusters in the subspace of these features. Even such clusters been found, their size would be too small to allow reliable automatic detection.

On the opposite side, the dimension **A16** is the most frequent. Usually it is found together with other FTDs. The explanation is abundance of encyclopedic texts in our corpus. The features also vary by the fraction of '2' scores: the dimensions **A4**, **A7**, **A9**, **A12**, **A14** were scored very categorically with the fraction of '2'-s above 70%. On the contrary, the annotators were not so confident in assigning features **A1**, **A3**, **A6**, **A15**, **A17**, **A19** (30–40%). The extreme uncertainty was **A17** ("evaluation") with only 22% of '2' between positive values.

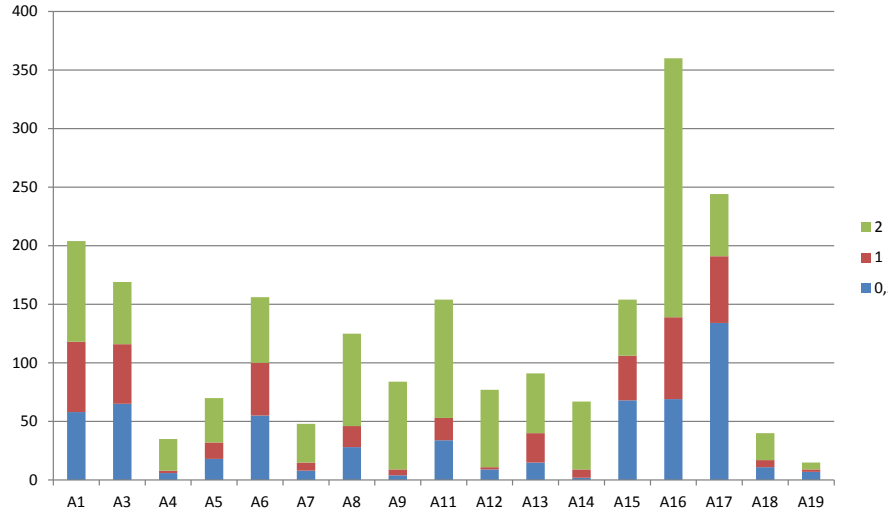

Figure 1. The distribution of FTD values

Geometrically, the annotations are the points of 17-dimensional cube located on the grid with 0.5 interval. We are looking for clusters in this cube defining a cluster as a dense and isolated subset of points. Since borderline texts are ubiquitous, there is no hope for isolation, so we care only about density. Formally, our task is to find the vertexes of 17-dimensional cube with the densest clouds of points around them. Such points will be the centers of clusters.

Clusterization: method and results

To facilitate clustering we binarize the values of the FTDs. Extreme values 0 and 2 are naturally mapped to themselves. Since the difference between '0' (absent) and '0.5'

(slight) is rather vague, we also map 0.5 to 0. Processing 1s in the same manner would be a crude oversimplification, so we use a twofold recoding. We create two datasets, the first one with 1s replaced by 0s, and the second with 1s replaced by 2s. We perform clustering for both datasets and then compare the results. By clustering in this context we mean calculating the frequencies of cube vertexes and detecting the most stable combinations.

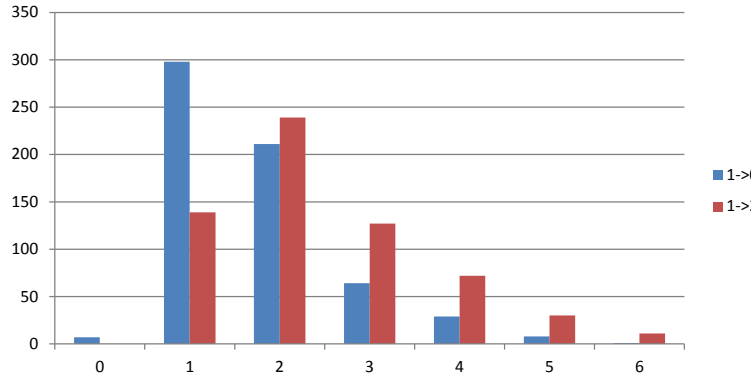

Figure 2. Histogram of the number of positive features for FTD combinations

For a vector of FTD values we call its rank the number of positive features in that vector. Figure 2 shows the frequency distribution of dataset points after various recodings. After 1→0 recoding the most frequent rank is 1 (298 times) and for 1→2 recoding such rank is 2 (239 times). Weights greater than 4 are rare: there are 5 such vertexes in the case of 1→0 recoding and 41 in case of 1→2 transformation. Both this quantities are negligible as compared to the size of the whole corpus so we may restrict our attention to combinations with 4 or less nonzero values. That gives us 3214 possible points in 17-dimensional cube.

Table 1 shows 24 top vertexes according to their frequencies after 1→0 and 1→2 recodings. For every prototype (cluster center) **v** we also measure **CW(v)**—the number of initial vertexes (before recoding) for which v is the closest prototype between the 24 selected with respect to standard Euclidean distance. In case of ties the frequency is divided by the number of closest prototypes. The prototypes are sorted in the descending order by their **CW**. Most of the cluster centers are of rank 1 and 2 except for the two prototypes of higher rank: **A14+A15+A16** and **A3+A6+A11+A17**. For the majority (352) of annotation vectors the number of closest prototypes is 1. In 92 cases the point had 2 prototypes on the same distance (most of the time the distance of 1). In 169 cases (27,3%) the points had no prototype on the distance less than 2 and were considered as noise.

Table 1. The most frequent combinations of FTD values

A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15	A16	A17	A18	A19	Weight1->0	Weight1->2	CW	Rank
0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	71	53	68.20	1
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	75	37	59.75	1
0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	37	46	42.00	2
0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	25	29	31.00	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	10	26.15	1
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	30	20	26.00	1
0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	24	22	25.50	2
0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	21	4	17.50	1
0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	12	19	17.00	3
2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	11	14	16.00	2
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	13	6	13.00	1
0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	7	15	11.25	2
0	2	0	0	2	0	0	0	2	0	0	0	0	0	2	0	0	6	4	11.00	4
0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	14	2	10.70	1
0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	3	10	10.50	2
0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	16	2	10.00	1
0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	10	5	9.00	2
2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	8	3	9.00	2
0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	5	3	8.50	2
0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	7	1	8.00	2
2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	11	6.70	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	7	1	5.40	1
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	7	3	5.00	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	1.15	0

Clusters and their classes

Most of the elicited clusters are too small to be used as pseudogenres: only 2 clusters from 24 are of size 60 or greater (>10% of the collection) and 5 more clusters contain 25–40 texts (5–7%).Hence reliable automatic classification for these clusters is a hopeless task since no algorithm can detect such small classes. Fortunately, the clusters themselves possess hierarchical structure and can be grouped into higher order classes. When grouping the clusters we used the correlation between the features **A1**–**A18** and the diagram of cluster proximity (Figure 3). The sizes of the rectangles are proportional to the size of the corresponding cluster. Note that the dimension **A19** is excluded from the future consideration due to its scarcity.

We made the following decisions on the basis of this diagram:

1) The most inhabited clusters **A8** ("news texts") and **A16** ("encyclopedic texts") form single classes.
2) Any cluster containing the FTDs **A14**, **A15** belong to the same class due to a high weight of clusters **A14+A15+A16** and **A14+A16**, and considerable correlation between **A14** and **A15** ($\rho = 0{,}42$). Note that the features **A14** ("scientific character") and **A15** ("texts to specialists")' have a lot in common already by their definition.
3) The clusters containing **A3**, **A6**, **A11** and **A17** are joined together since this features rarely appear severally and their correlation (ρ(**A3**, **A6**) = 0,5; ρ(**A3**, **A11**) = 0,54; ρ(**A3**, **A17**) = 0,4) is high (e.g., significant on $\rho = 0{,}001$ level).

4) Since the weight of cluster **A1+A13** is high (16), and their correlation ρ(**A1**, **A13**) = 0,32 is significant on the level $\rho < 0{,}0001$, we unite all the clusters containing **A1** and **A13** (except for **A1+A11** which is already in the class for **A11** dimension).
5) Join the clusters **A9**, **A9+A16** together, as well as **A12** and **A12+A16**.
6) The features **A4**, **A5**, **A7**, **A18** do not have enough positive values to form classes but their total weight (104 texts) is too high to consider them as noise. We sort these features according to their frequencies (from high to low) and attribute yet unannotated text to the first FTD with value of 1 or 2. For example, the FTD **A7** ("instructive text") occurred in this sense in 36 texts. Remaining dimensions are tightly correlated (ρ(**A4**, **A18**) = 0,44; ρ(**A4**, **A5**) = 0,28) and it is natural to join them together. Certainly, these decisions are justified only for this collection. We call this method the method of presence.

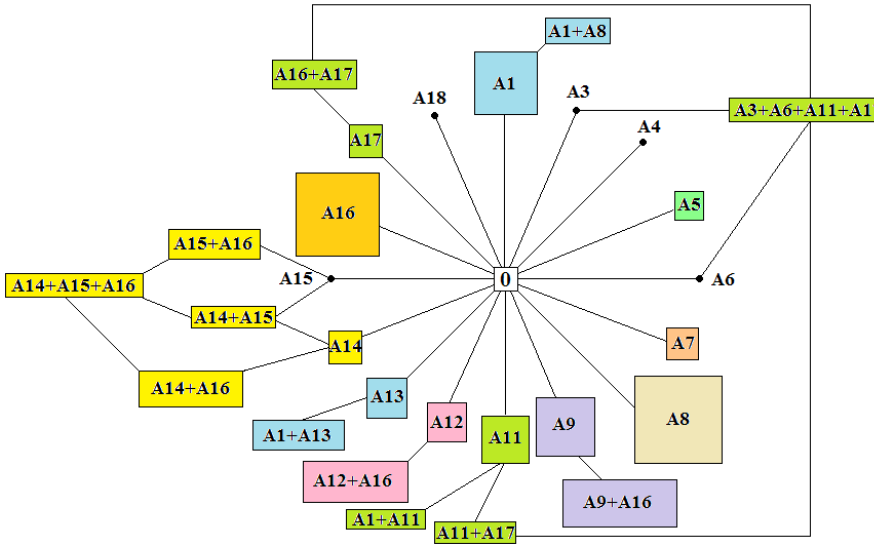

Figure 3. Diagram of cluster proximity

We obtain a list of 9 clusters **C1**–**C9** with their rounded weights as superscripts:

Table 2. Joining clusters to classes

Class	Clusters in the class
C1	$A1^{26}$, $(A1+A13)^{17}$, $A13^{11}$, $(A1+A8)^{7}$
C2	$A11^{18}$, $(A3+A6+A11+A17)^{11}$, $(A11+A1)^{9}$, $(A11+A17)^{8}$
C3	$A8^{68}$
C4	$(A9+A16)^{42}$, $A9^{26}$
C5	$(A12+A16)^{31}$, $A12^{10}$

Class	Clusters in the class
C6	(A14+A16)[26], (A14+A15+A16)[17], (A15+A16)[11], (A14+A15)[9], A14[5]
C7	A16[60]
C8	A7[35] (*weight based on the presence method*)
C9	(A4 or A5 or A18)[69] (*weight based on the presence method)*

86 of 618 texts (13,9%) were not attributed to any class and were considered as noise. For comparison we give the results of clusterization in our previous work.

Table 3. Clusters of FTDs according to [Sorokin et al., 2014]

Class	Principal dimensions	Size
"instructive texts"	A7	21
"news texts"	A8	64
"legal texts"	A9	11
"scientific or technical texts"	A14, A15	13
"encyclopedic texts"	A16	49
"advertising texts"	A1, A12, A17	13
"propaganda texts"	A1, A13, A17	13
"noise"	—	131

Thus, our results roughly correspond to our previous work. Slight differences arose from corpus modifications and changing of the clustering procedure. It is worth noting that whereas clusterization of the present work is in some sense "manual" and "ad hoc", in (Sorokin, Katinskaya, Sharoff, 2014) we used a fully automatic procedure. The compatibility of results justifies the claim that clusters are linguistically relevant pseudogenres, not just the mathematical curiosity, and correspond to some observable language features. In what follows we study this problem for some predefined set of language features.

Language features

To address automatic genre classification we use the feature space of 40 language features **B1**–**B40** given in Appendix 2. The inventory of features was built for the experiment with applying multi-dimensional analysis (see Biber, 1988) to Russian and was based on the set of English linguistic features presented in (Biber et al, 2007). We adapted the features for Russian grammar and slightly restricted their set basing on the accessibility of features for automatic extraction without involving external complicated tools. We calculated the absolute frequencies of features using special program taking the morphologically tagged text as input. All the counts except for word length, sentence length and type/token ratio were normalized by the text length. As a result we obtained a 618x40 matrix containing the frequencies of 40 linguistic variables for each text. Before dealing with automatic classification we study the distribution of these features. Figure 4 contains the scatter diagram for the first 5 features.

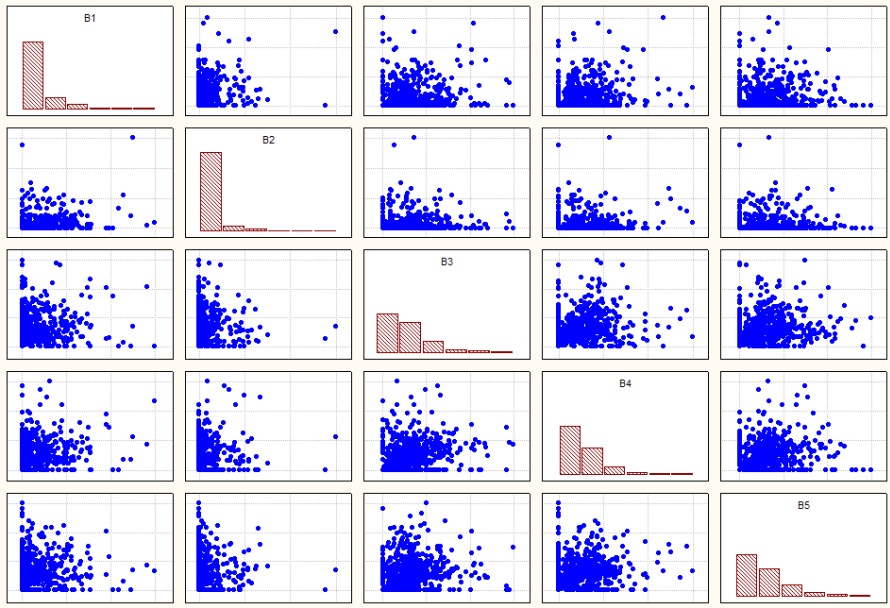

Figure 4. Scatter diagram for the features **B1–B5**

Analysis of Figure 4 leads to the following observations:
a) there are some outliers (for example, for the feature **B2**);
b) zero values are rather frequent, about 24% of all texts;
c) the distribution tails are quite heavy (see the histograms).

To prevent the distortion of regression coefficients by outliers we replace them by the closest "inlying" value. Heavy distribution tails make the model unrobust; we applied the Box-Cox normalizing transformation (Kutner et al., 2004) to reduce their effect. The value B is mapped to a new value B′

$$\mathbf{B}' = \frac{\mathbf{B}^{\lambda} - 1}{\lambda}$$

with λ estimated automatically using maximal likelihood.

We used the following scheme of transformation:
a) all zeros were temporarily referred as missing values;
b) the Box-Cox transformation was applied to the features with missing values;
c) the features were standardized to have mean 0 and deviation 1.

$$\mathbf{X} = \frac{\mathbf{B}' - \mu}{\sigma}$$

After this transformation the absolute values of the coefficients correspond to their effect

d) the missing values were replaced by -3 (an actual minimal value of a standardized normal distribution according to "3 sigma rule").

The scatter diagram for transformed features is given on Figure 5. The tails became less heavy and the shape better fits two-dimensional normal distribution, which makes the estimation of regression coefficients more robust.

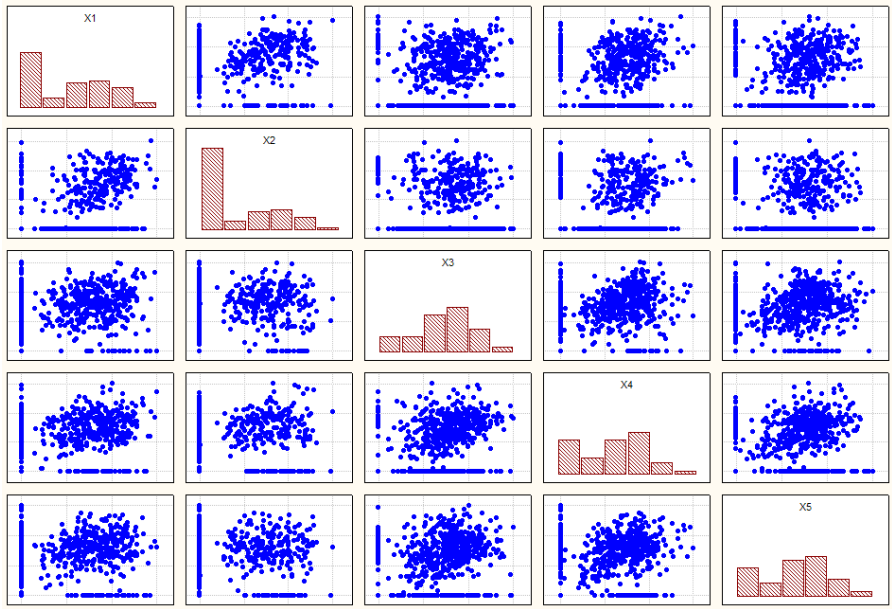

Figure 5. Scatter diagram for transformed features

Spearman coefficient matrix for features **B1–B40** is given in Table 4. The coefficients significant at 0,001 level are marked as red. Most of the features are tightly interconnected. Consequently, there is no reason to use factor analysis of principal components analysis since they are appropriate for correlation matrixes with block structure. In the further we apply logistic regression which is invariant under linear transformations so we will use the features themselves, not the principal components or factors.

Table 4. Spearman correlation coefficients for features **B1–B40**

Variable	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24	B25	B26	B27	B28	B29	B30	B31	B32	B33	B34	B35	B36	B37	B38	B39	B40
B1	0,18	0,35	0,23	0,25	0,35	0,13	0,22	0,15	0,28	0,34	0,46	0,22	-0,41	-0,63	-0,49	-0,52	0,30	0,32	0,39	0,16	0,17	-0,10	-0,25	0,33	-0,55	-0,24	-0,05	0,12	-0,50	-0,31	0,29	0,21	-0,44	0,14	0,10	0,05	-0,02	0,44
B2	0,15	0,19	0,13	0,12	0,21	-0,10	0,20	0,19	0,18	0,13	0,30	0,26	-0,28	-0,49	-0,33	-0,34	0,25	0,24	0,28	0,08	0,10	-0,05	-0,20	0,18	-0,42	-0,23	-0,07	0,14	-0,30	-0,29	0,47	0,17	-0,32	0,03	0,02	0,22	0,09	0,32
B3	1,00	0,29	0,26	0,30	0,16	0,31	0,25	0,12	0,06	0,19	0,28	0,16	-0,29	-0,29	-0,30	-0,29	0,20	0,35	0,30	0,23	0,21	0,09	-0,13	-0,12	-0,28	-0,11	-0,06	0,13	-0,33	-0,10	0,07	0,27	-0,37	0,19	0,01	-0,04	0,06	0,39
B4	0,29	1,00	0,27	0,20	0,16	0,13	0,21	0,10	0,00	0,12	0,24	0,00	-0,22	-0,28	-0,27	-0,26	0,26	0,23	0,25	0,07	0,18	0,03	-0,12	0,03	0,18	0,02	-0,06	0,17	-0,25	-0,07	0,09	0,21	-0,22	-0,01	0,07	-0,04	0,02	0,23
B5	0,26	0,27	1,00	0,30	0,12	0,15	0,10	0,11	-0,02	0,18	0,24	-0,11	-0,15	-0,23	-0,15	-0,25	0,19	0,13	0,25	0,16	0,31	0,06	-0,01	-0,13	-0,13	0,02	-0,06	0,04	-0,22	-0,11	-0,06	0,24	-0,19	0,11	0,07	0,04	-0,01	0,23
B6	0,30	0,20	0,30	1,00	0,40	0,21	0,29	0,14	0,22	0,38	0,40	0,27	-0,25	-0,35	-0,26	-0,20	0,35	0,39	0,44	0,41	0,52	0,24	-0,11	-0,03	-0,29	-0,04	0,29	0,22	-0,29	-0,16	0,24	0,38	-0,36	0,34	0,26	0,27	0,28	0,46
B7	0,16	0,16	0,12	0,40	1,00	0,15	0,21	0,20	0,40	0,42	0,44	0,33	-0,34	-0,42	-0,24	-0,06	0,38	0,45	0,44	0,36	0,31	0,14	-0,12	0,14	0,40	-0,09	0,31	0,20	-0,19	-0,13	0,29	0,28	-0,46	0,35	0,33	0,35	0,22	0,57
B8	0,31	0,13	0,15	0,21	0,15	1,00	0,65	-0,36	0,09	0,36	0,21	0,07	-0,33	-0,06	0,03	-0,07	0,03	0,42	0,30	0,35	0,21	0,07	0,11	0,08	-0,26	-0,05	0,13	-0,10	-0,11	-0,10	-0,01	0,10	-0,30	0,01	-0,13	0,00	-0,21	0,15
B9	0,25	0,21	0,10	0,29	0,21	0,65	1,00	-0,23	0,23	0,42	0,32	0,25	-0,31	-0,20	-0,08	-0,03	0,45	0,45	0,44	0,39	0,26	0,13	0,11	0,13	-0,32	-0,10	0,23	0,30	-0,14	-0,21	0,42	0,19	-0,46	0,12	0,09	0,25	0,40	0,32
B10	0,12	0,10	0,11	0,14	0,20	-0,36	-0,23	1,00	0,15	-0,01	0,16	0,16	-0,06	-0,21	-0,10	0,01	0,17	0,09	0,18	0,11	0,20	0,20	-0,14	0,14	-0,00	-0,01	0,06	0,19	-0,03	0,18	0,09	0,20	-0,09	0,23	0,12	0,13	0,01	0,23
B11	0,06	-0,00	-0,02	0,22	0,40	0,09	0,23	0,15	1,00	0,43	0,44	0,37	-0,46	-0,35	-0,24	-0,00	0,20	0,21	0,33	0,32	0,21	0,17	-0,15	0,34	-0,47	-0,23	0,48	0,17	-0,20	-0,13	0,41	0,22	-0,37	0,30	0,28	0,22	0,36	0,28
B12	0,19	0,12	0,18	0,38	0,42	0,36	0,42	-0,01	0,43	1,00	0,65	0,15	-0,41	-0,36	-0,30	-0,19	0,30	0,43	0,42	0,36	0,30	0,11	-0,02	0,18	-0,45	-0,08	0,41	0,27	-0,36	-0,21	0,28	0,16	-0,43	0,25	0,28	0,13	0,30	0,41
B13	0,28	0,24	0,24	0,40	0,44	0,21	0,32	0,16	0,44	0,65	1,00	0,08	-0,51	-0,61	-0,41	-0,36	0,44	0,30	0,46	0,19	0,28	0,05	-0,20	0,07	-0,47	-0,24	0,21	0,23	-0,41	-0,28	0,25	0,35	-0,43	0,18	0,29	0,15	0,06	0,52
B14	0,16	0,00	-0,11	0,27	0,33	0,07	0,25	0,16	0,37	0,15	0,08	1,00	-0,29	-0,27	-0,33	-0,06	0,20	0,20	0,30	0,26	0,13	0,19	-0,09	0,33	-0,32	-0,27	0,43	0,06	-0,23	-0,19	0,35	0,12	-0,35	0,32	0,30	0,36	0,34	0,24
B15	-0,29	-0,22	-0,15	-0,25	-0,34	-0,33	-0,31	-0,06	-0,40	-0,41	-0,51	-0,29	1,00	0,34	0,44	0,35	-0,20	-0,38	-0,37	-0,24	-0,27	0,03	0,25	-0,16	0,63	0,31	-0,22	-0,19	0,46	0,34	-0,29	-0,25	0,48	-0,22	-0,18	-0,06	-0,09	-0,42
B16	-0,29	-0,28	-0,23	-0,35	-0,42	-0,06	-0,20	-0,21	-0,35	-0,36	-0,61	-0,27	0,34	1,00	0,42	0,45	-0,39	-0,33	-0,43	-0,16	-0,19	0,05	0,32	-0,20	0,62	0,27	-0,01	-0,16	0,41	0,34	-0,34	-0,34	0,43	-0,18	-0,19	-0,26	-0,08	-0,54
B17	-0,30	-0,27	-0,15	-0,26	-0,24	0,03	-0,08	-0,10	-0,24	-0,30	-0,41	-0,28	0,44	0,42	1,00	0,49	-0,26	-0,25	-0,28	-0,22	-0,29	0,12	0,31	-0,16	0,48	0,28	-0,04	-0,16	0,88	0,33	-0,24	-0,19	0,39	-0,13	-0,19	-0,03	-0,03	-0,36
B18	-0,29	-0,26	-0,25	-0,20	-0,06	-0,07	-0,03	0,01	-0,00	-0,19	-0,36	0,06	0,35	0,45	0,49	1,00	-0,12	-0,11	-0,09	0,00	-0,09	0,42	0,34	0,25	0,38	0,29	0,29	-0,02	0,60	0,38	-0,01	-0,09	0,24	0,22	0,24	0,28	0,30	-0,15
B19	0,20	0,26	0,19	0,35	0,38	0,03	0,45	0,17	0,20	0,30	0,44	0,20	-0,20	-0,39	-0,26	-0,12	1,00	0,29	0,39	0,25	0,25	0,05	-0,03	0,07	-0,27	-0,06	0,16	0,23	-0,26	-0,14	0,30	0,38	-0,40	0,17	0,11	0,43	0,46	0,45
B20	0,35	0,23	0,13	0,39	0,45	0,42	0,45	0,09	0,31	0,43	0,30	0,38	-0,38	-0,33	-0,25	-0,11	0,29	1,00	0,45	0,47	0,30	0,14	-0,06	0,33	-0,41	-0,20	0,36	0,31	-0,29	-0,25	0,36	0,28	-0,55	0,36	0,19	0,23	0,18	0,45
B21	0,30	0,25	0,25	0,44	0,44	0,30	0,44	0,18	0,33	0,42	0,46	0,30	-0,37	-0,43	-0,28	-0,09	0,39	0,45	1,00	0,49	0,33	0,15	-0,11	0,22	-0,42	-0,21	0,30	0,28	-0,29	-0,16	0,37	0,31	-0,47	0,23	0,10	0,27	0,24	0,44
B22	0,23	0,07	0,16	0,41	0,36	0,35	0,39	0,11	0,32	0,36	0,19	0,26	-0,24	-0,16	-0,22	0,00	0,25	0,47	0,49	1,00	0,35	0,27	-0,00	0,16	-0,28	-0,09	0,41	0,28	-0,24	-0,13	0,26	0,29	-0,40	0,39	0,19	0,31	0,45	0,34
B23	0,21	0,18	0,13	0,21	0,26	0,20	0,23	0,30	0,26	0,13	0,27	-0,19	-0,20	-0,09	0,25	0,30	0,33	0,35	1,00	0,28	-0,06	-0,07	-0,19	0,01	0,36	0,24	-0,23	-0,15	0,16	0,28	-0,30	0,33	0,19	0,12	0,25	0,31		
B24	0,09	0,03	0,06	0,24	0,14	0,07	0,13	0,20	0,17	0,11	0,05	0,19	0,03	0,05	0,12	0,42	0,06	0,14	0,15	0,27	0,28	1,00	0,22	0,08	0,01	0,13	0,48	0,21	0,13	0,15	0,26	0,11	-0,14	0,35	0,28	0,22	0,37	0,10
B25	-0,13	-0,12	-0,01	-0,11	-0,12	0,11	0,11	-0,14	-0,15	-0,02	-0,28	-0,09	0,25	0,32	0,31	0,34	-0,03	-0,06	-0,11	-0,00	-0,06	0,22	1,00	-0,08	0,17	0,21	0,01	-0,06	0,29	0,20	-0,16	-0,19	0,06	-0,03	0,05	-0,04	0,06	0,20
B26	-0,12	0,03	-0,13	-0,03	0,14	-0,08	0,13	0,14	0,34	0,18	0,07	0,33	-0,16	-0,20	-0,16	0,25	0,30	0,33	0,35	1,00	-0,07	0,08	0,08	1,00	-0,36	-0,15	0,46	0,35	-0,13	-0,01	0,48	-0,00	-0,20	0,35	0,07	0,22	0,24	0,20
B27	-0,28	-0,18	-0,13	-0,29	-0,40	-0,26	-0,32	-0,00	-0,47	-0,45	-0,47	-0,32	0,63	0,62	0,48	0,38	-0,27	-0,41	-0,42	-0,28	-0,19	0,01	0,17	-0,36	1,00	0,36	-0,13	-0,20	0,51	0,46	-0,41	-0,23	0,70	-0,22	-0,21	-0,24	-0,18	-0,45
B28	0,11	0,02	0,02	-0,04	-0,09	-0,05	0,10	-0,01	0,23	-0,08	-0,24	-0,27	0,31	0,27	0,28	0,25	0,06	-0,20	-0,25	-0,09	0,01	0,13	0,21	-0,15	0,36	1,00	-0,16	-0,03	0,30	0,34	-0,28	-0,14	0,22	0,05	0,16	0,06	0,10	0,18
B29	-0,06	-0,06	-0,06	0,29	0,31	0,13	0,23	0,06	0,48	0,41	0,21	0,43	-0,22	-0,01	-0,04	0,29	0,16	0,36	0,30	0,41	0,36	0,48	0,01	0,46	-0,13	-0,16	1,00	0,48	-0,04	0,08	0,48	0,32	-0,09	0,60	0,63	0,33	0,59	0,20
B30	0,13	0,17	0,04	0,22	0,20	0,16	0,30	0,19	0,17	0,27	0,23	0,06	-0,19	-0,19	-0,16	-0,02	0,23	0,31	0,28	0,29	0,24	0,21	-0,06	0,35	-0,20	-0,03	0,48	1,00	-0,16	0,05	0,31	0,17	-0,19	0,18	0,17	0,10	0,28	0,22
B31	-0,33	-0,25	-0,22	-0,29	-0,19	-0,11	0,14	-0,03	-0,20	-0,36	-0,41	-0,23	0,46	0,45	0,88	0,60	-0,25	-0,29	-0,29	-0,24	-0,23	0,13	0,29	-0,13	0,51	0,30	-0,04	-0,16	1,00	0,40	-0,20	-0,21	0,44	-0,09	-0,21	0,00	-0,06	-0,34
B32	-0,10	-0,07	-0,11	-0,16	-0,13	-0,18	-0,21	0,18	-0,13	-0,21	-0,28	-0,19	0,34	0,34	0,33	0,38	-0,14	-0,25	-0,16	-0,13	-0,15	0,15	0,20	-0,01	0,46	0,34	0,08	0,05	0,40	1,00	-0,18	-0,09	0,34	0,03	0,04	-0,04	-0,08	-0,25
B33	0,07	0,09	-0,06	0,24	0,29	-0,01	0,42	0,09	0,41	0,28	0,25	0,35	-0,29	-0,34	-0,24	-0,01	0,30	0,36	0,37	0,26	0,16	0,26	-0,16	0,48	-0,41	-0,26	0,48	0,31	-0,20	-0,18	1,00	0,18	-0,34	0,36	0,23	0,29	0,34	0,33
B34	0,27	0,21	0,24	0,38	0,28	0,10	0,19	0,20	0,22	0,16	0,35	0,12	-0,25	-0,34	-0,16	-0,09	0,38	0,28	0,31	0,29	0,28	0,11	-0,18	-0,09	-0,23	-0,14	0,32	0,17	-0,21	-0,09	0,18	1,00	-0,27	0,39	0,21	0,32	0,27	0,35
B35	-0,37	-0,22	-0,19	-0,36	-0,46	-0,38	-0,46	-0,09	-0,37	-0,43	-0,43	-0,35	0,48	0,43	0,39	0,24	-0,40	-0,55	-0,47	-0,40	-0,30	-0,14	0,06	-0,20	0,70	0,22	-0,09	-0,19	0,44	0,34	-0,34	-0,27	1,00	-0,29	-0,23	-0,29	-0,23	-0,48
B36	0,19	-0,01	0,11	0,34	0,35	0,07	0,12	0,23	0,30	0,25	0,18	0,32	-0,22	-0,18	-0,13	0,22	0,17	0,36	0,23	0,39	0,33	0,35	-0,03	0,35	-0,22	0,05	0,60	0,18	-0,09	0,03	0,36	0,39	-0,29	1,00	0,54	0,36	0,35	0,32
B37	0,01	0,07	0,07	0,26	0,33	0,01	0,09	0,12	0,28	0,28	0,29	0,30	-0,18	-0,19	-0,19	0,24	0,11	0,19	0,10	0,19	0,19	0,28	0,05	0,07	-0,21	0,16	0,63	0,17	-0,21	0,04	0,23	0,21	-0,23	0,54	1,00	0,29	0,30	0,31
B38	-0,04	-0,04	0,04	0,27	0,33	-0,13	0,25	0,13	0,22	0,13	0,15	0,36	-0,06	-0,26	-0,03	0,28	0,43	0,23	0,27	0,31	0,12	0,22	-0,04	0,22	-0,24	-0,06	0,33	0,10	-0,00	-0,04	0,29	0,32	-0,29	0,36	0,29	1,00	0,45	0,30
B39	0,06	0,02	-0,01	0,28	0,22	0,09	0,40	0,01	0,36	0,30	0,06	0,34	-0,09	-0,08	-0,03	0,30	0,46	0,18	0,24	0,45	0,25	0,37	0,06	0,24	-0,18	0,10	0,59	0,28	-0,06	-0,08	0,34	0,27	-0,23	0,35	0,30	0,45	1,00	0,12
B40	0,39	0,23	0,23	0,46	0,57	0,21	0,32	0,23	0,28	0,41	0,52	0,24	-0,42	-0,54	-0,36	-0,15	0,45	0,45	0,44	0,34	0,31	0,10	-0,20	0,20	-0,45	-0,18	0,20	0,22	-0,34	-0,25	0,33	0,35	-0,48	0,32	0,31	0,30	0,12	1,00

Class prediction

To achieve better robustness we removed 86 (13,9%) texts of gold standard, which we not assigned to any cluster and consequently were considered as noise. For every class among **C1–C9** we constructed a single logistic model (Bishop, 2006) separating it from the other classes. For every class we used its corpus frequency to set classification threshold. The values of thresholds, sensitivity and specificity are given in Table 5.

Table 5. Threshold, sensitivity and specificity for classes **C1–C9**

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Threshold τ	0,11	0,13	0,14	0,13	0,08	0,14	0,09	0,07	0,13
Sensitivity	81%	81%	92%	94%	94%	85%	81%	94%	85%
Specificity	86%	88%	93%	97%	90%	88%	85%	94%	83%

For multiclass classification we used one-versus-all (OVA) approach: a point was assigned to the class whose separation hyperplane was on the largest signed distance. Distance was measured according to the formula below; negative value means that logistic model does not assign this point to a class under consideration.

$$d = \ln \frac{\left(\frac{1}{\tau} - 1\right) / \left(\frac{1}{\rho} - 1\right)}{\sqrt{b_1^2 + \ldots + b_1^2}},\ b_1, \ldots, b_9 \text{ — } \textit{feature weights in regression model.}$$

We obtained the following table of cross-classification:

Table 6. Table of cross-classification for classes **C1–C9**

Class	C1	C2	C3	C4	C5	C6	C7	C8	C9	Total
C1	37	9	3	0	3	1	1	1	1	**56**
C2	9	41	0	0	2	3	2	1	9	**67**
C3	6	0	59	0	0	0	7	0	1	**73**
C4	0	0	1	64	0	3	0	1	0	**69**
C5	0	0	0	0	34	0	4	2	1	**41**
C6	1	0	1	4	3	57	9	0	0	**75**
C7	4	3	4	1	0	1	30	1	3	**47**
C8	1	0	0	1	1	0	0	30	2	**35**
C9	4	6	1	0	2	1	2	3	50	**69**
Accuracy	66%	61%	81%	93%	83%	76%	64%	86%	72%	**76%**
	78%		81%	93%	83%	80%		86%	72%	**81%**

The pairs **C1, C2** and **C6, C7** are the most difficult to separate. If we join the elements of each pair together, accuracy grows from 76% to 81%. Unification of **C6** and C7 is justified from linguistical point of view also: the class **C7** contains texts with principal dimension **A16** (at first, encyclopedic), whereas **C6** comprises documents

with principal FTDs **A14** (scientific or technical) and **A15** (texts for specialists). For example, scientific encyclopedic articles are on the border; therefore, these classes cannot be reliably distinguished not only automatically but by human annotator also. Classification accuracy for all classes lies in the diapason **72–93%**. Thus, logistic regression model permits to perform automatic genre classification with high accuracy at least for the texts of gold standard.

Predicting functional text dimensions

Almost surely, new clusters will emerge for a corpus of other size or origin. Therefore an automatic procedure should be designed to detect new clusters in data and assign texts to these clusters, as well as to separate dense regions from ubiquitous noise. To address these tasks the values of the very FTDs should be known. Again we used logistic regression, but in its weighted variant: when predicting a feature, the texts with rate 2 for this feature had weight 2. Inversed backward procedure was used to increase stability. As earlier, we used the frequency of a feature to determine classification threshold. The values of threshold, sensitivity and specificity for 16 dimensions are given in Table 7. The test values of performance measures were calculated using 70%/30% train-test split.

Table 7. Accuracy of automatic FTD prediction

	A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15	A16	A17	A18
Threshold	0,33	0,23	0,09	0,14	0,23	0,11	0,25	0,22	0,31	0,20	0,19	0,18	0,20	0,61	0,24	0,08
Sensitivity full	76%	88%	92%	79%	82%	94%	86%	96%	83%	89%	74%	86%	71%	76%	72%	91%
Specificity full	78%	91%	96%	80%	83%	93%	86%	97%	86%	90%	80%	89%	75%	77%	83%	96%
Sensitivity test	79%	94%	92%	76%	81%	91%	80%	98%	79%	90%	72%	91%	73%	72%	71%	89%
Specificity test	67%	76%	100%	81%	72%	95%	73%	83%	90%	80%	64%	71%	75%	66%	88%	100%

Classification accuracy is quite high. Dimensions **A1**, **A13**, **A16** were predicted a bit worse than others on the test sample. Average specificity on the full data and test sample was 83% and average sensitivity was 86% and 80%, respectively. To make the model even more reliable we selected only the features which are significant at p-level=0,05. Since features are standartized to have the same mean and variance, the absolute values of weights reflect the importance of the corresponding features for a predicted dimension. Logistic model also predicts class probabilities. For example, the probabilities below confidently attribute text to the class **C4** (**A9**+**A16**).

A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15	A16	A17	A18
0,024	0,001	0,000	0,011	0,021	0,008	0,011	0,962	0,003	0,535	0,005	0,000	0,247	0,728	0,003	0,000

FTD probability scores being known, the easiest way to classify text is the following: multiply all the probabilities by 2 and assign text to the closest prototype in the sense of Euclidean distance. When this distance exceeds 2, a text is considered as noise. We refer to such classification mechanism as crude classification. The results of crude classification of FTDs for gold standard are given below. They are quite more decent than before, which justifies the term “crude classification”.

Table 8. Results of crude classification

	Noise	C1	C2	C3	C4	C5	C6	C7	C8	C9	Total
Accuracy	43%	45%	31%	64%	94%	66%	57%	55%	80%	57%	**58%**
	43%	48%		64%	94%	66%	80%		80%	57%	**65%**

Such method allows us to detect 43% of noise. Investigating its structure more thoroughly, we see that most of the noise texts are rare combinations of 2 or 3 functional dimensions. Such combinations may potentially become new cluster centers for a more representative corpus.

Rank	0	1	2	3	4	5	6
Count	3	5	44	23	8	2	1
Percentage	3%	6%	51%	27%	9%	2%	1%

Application to a large corpus

The ultimate goal of our research is to prepare the corpus for automatic genre annotation. The results of automatic classification are rather optimistic in this sense, showing that FTD values could in principle be predicted reliably. However, we need to check whether the model learnt on the gold standard would be correct on another corpus. For this task we should manually annotate another corpus and compare the values of FTDs with those predicted by the logistic models. Since this comparison is time and labour-consuming, we just explore the combinations of features more frequent on a new corpus.

1) We took 1,000,000 LJ-posts from the current version of GICR and applied the logistic models learnt on the gold standard. This yields 16 probability scores for every text. For the sake of simplicity we performed binary classification without intermediate values.
2) We binarized the scores for every feature using the same threshold 0,75. Thus every text is mapped to a 16-dimensional binary vector, which naturally corresponds to a set of FTDs of this text. The vectors which occur more than 1000 times in the corpora were considered as the prototypes.
3) We attached the texts to the closest prototype by the Euclidean distance between the probability scores of the text and the prototype. The texts with no prototype within the distance 1 were not attached to any cluster. Here a cluster is a set of texts with the same prototype.

We detected 120 prototypes with ranks varying from 0 to 5 (rank is the number of positive FTDs). For 208,533 of 1,000,000 texts no prototype was found. Using the "knee method" of (Salvador, 2004) we discovered the 9 most frequent clusters which are the prototypes for 266,635 texts.

Table 9. The most frequent combinations of FTD values after crude classification

Feature combination	Texts number
A8	49,738
A3+A6+A11	35,506
A16	34,578
A3+A6+A11+A17	31,214
A8+A16	30,219
A3+A5+A6+A11+A17	23,036
A1+A3+A6+A11	20,917
A12+A16	20,874
A3+A4+A6+A11	20,553

Some of the combinations (**A8**, **A16**, **A12+A16** etc.) occur already in the gold standard; while the combinations of features **A3**, **A6**, **A11** increased their frequency. This property is the most noticeable for FTD A17 "evaluation". This results are quite expectable since argumentative (**A1**), informal (**A6**) and evaluative (**A17**) texts are usual in blogosphere. However, additional verification of assigned rates is necessary.

Conclusions

While straightforward classification of a large Web corpus into genres is problematic because of the disagreement between the human annotations, we managed to provide an overview of the most frequent genre options in the corpus. Overall, our research leads to the following conclusions:

1) There exist well-formed dense clusters in the FTD space.
2) Language features allow prediction of the FTD values with high accuracy (about 75%).
3) We can detect very similar clusters when the model learnt on the gold standard are applied to a new corpus.

The achieved results are rather promising, but there is still a long way to go before achieving reliable genre annotation of large real-world corpora. Our plan is to replace traditional genres with FTD clusters, and the present research demonstrates, that such clusters can be detected automatically. However, the list of pseudogenres observed in the "gold standard" is obviously incomplete, so that we need to uncover the cluster structure of bigger corpora in future research.

In future genre annotation experiments we will match the texts in the corpus with the label of the closest cluster in the FTD space. For some texts the classifier can be confident, assigning scores about 1.0 to the principal dimensions and near zero scores for other FTDs. Hence we can assign a cluster label (therefore, genre label) for such texts reliably. For other texts the classifier can be less confident, and the best strategy would be to refuse attributing any genre label or to give a list of nearby clusters. Therefore, we plan to continue our research in the following directions:

1. To collect a near-exhaustive list of possible clusters in the FTD space using a bigger corpus.
2. Check the accuracy of automatic detection of these clusters
3. Consider the centers of such clusters as prototypes
4. Assign the documents of the corpora to their closest prototype provided the prototype is indeed close in the FTD space

Though this procedure definitively would not allow to detect genre for arbitrary Web text, we plan to assign labels to a large fraction of the corpus with sufficient confidence. Even incomplete annotation gives a possibility to study e.g genre-dependent linguistic parameters and the linguistic correlates of genre labels.

References

1. *Biber, D.* Variation across speech and writing. Cambridge: CUP, 1988
2. *Biber D., Connor, U. and Upton, T.* Discourse on the move: using corpus analysis to describe discourse structure. Amsterdam—Philadelphia, 2007.
3. *Bishop C.* Pattern recognition and machine learning. Springer, New York, 2006.
4. *Egbert J., Biber D.* Developing a user-based method of register classification // Proc. 8th Web as Corpus Workshop. Lancaster, 2013. July.
5. *Forsyth R., Sharoff S.* (2014). Document dissimilarity within and across languages: a benchmarking study // Literary and Linguistic Computing. — 2014. — 29(1):6–22.
6. *Kilgariff A.* The Web as corpus // Proc. of corpus linguistics 2001. Lancaster, 2001.
7. *Kutner M., Nachtsheim C., Neter J., and Li W.* (2004). Applied Linear Statistical Models, McGraw-Hill/Irwin, Homewood, IL, 2004.
8. *Salvador S., Chan P.* // Proc. of 16 IEEE International Conference on Tools with Artificial Intelligence (2004). P. 576–584.
9. *Sharoff S.* In the garden and in the jungle: Comparing genres in the BNC and the Internet // Genres on the Web: Computational Models and Empirical Studies / Ed. by Alexander Mehler, Serge Sharoff, Marina Santini. Berlin/New York : Springer, 2010. P. 149–166.
10. *Sorokin A., Katinskaya A., Sharoff S.* Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus // Proc. Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2014.
11. *Беликов В. И., Селегей В. П., Шаров С. А.* Пролегомены к проекту Генерального интернет-корпуса русского языка. // Труды конференции Диалог 2012.

Appendix 1. Functional Text Dimensions

Code	Label	Question to be answered
A1	argum	To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', if argumentation is obvious)
A3	emotive	To what extent is the text concerned with expressing feelings or emotions? ('None' for neutral explanations, descriptions and/or reportage.)
A4	fictive	To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.)
A5	flippant	To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader? ('None' if it appears earnest or serious; even when it tries to keep the reader interested and involved)
A6	informal	To what extent is the text's content written in an informal style? (as opposed to the «standard» or «prestige» variety of language)
A7	instruct	To what extent does the text aim at teaching the reader how to do something? (For example, a tutorial or an FAQ)
A8	hardnews	To what extent does the text appear to be an informative report of recent events? (Recent at the time of writing. Announcements of future events can be considered hardnews too. 'None' if a news article only analyses information from other sources).
A9	legal	To what extent does the text lay down a contract or specify a set of regulations? (This includes copyright notices.)
A11	personal	To what extent does the text report from a first-person point of view? (For example, a diary-like blog entry.)
A12	compuff	To what extent does the text promote a product or service?
A13	ideopuff	To what extent is the text intended to promote a political movement, party, religious faith or other non-commercial cause?
A14	scitech	To what extent would you consider the text as representing research? (It does not have to be a research paper. For example, 'Strongly' or 'Partly' if a newswire text has scientific contents.)
A15	specialist	To what extent does the text require background knowledge or access to a reference source of a specialised subject area in order to be comprehensible? (such as wouldn't be expected of the so-called "general reader")
A16	info	To what extent does the text provide information to define a topic? (For example, encyclopedic articles or text books).
A17	eval	To what extent does the text evaluate a specific entity by endorsing or criticising it? (For example, by providing a product review).
A18	dialogue	To what extent does the text contain active interaction between several participants? (For example, forums or scripted dialogues).
A19	poetic	To what extent does the author of the text pay attention to its aesthetic appearance? ('Strongly' for poetry, language experiments, uses of language for art purposes).

Appendix 2. Linguistic features for automatic classification

B1	first_person_pronoun
B2	second_person_pronoun
B3	third_person_pronoun
B4	reflexive_pronoun
B5	adjective_pronoun
B6	nom_pronoun
B7	indefinite_pron
B8	past_tense
B9	perf_aspect
B10	present_tense
B11	place_adverb
B12	time_adverb
B13	total_adverb
B14	wh_questions
B15	nominalization
B16	nouns
B17	passive
B18	by_passive
B19	infinitive
B20	speech_verb

B21	mental_verb
B22	that_compl
B23	wh_relative
B24	pied_piping
B25	total_PP
B26	exclamation
B27	word_length
B28	type_token_ratio
B29	sentence_length
B30	verbal_adverb
B31	passive_participial_clauses
B32	active_participial_clauses
B33	imperative_mood
B34	predicative_adjectives
B35	attributive_adjective
B36	causative_subordinate
B37	concessive_subordinate
B38	conditional_subordinate
B39	purpose_subordinate
B40	Negation