Associating symptoms with syndromes
Reliable genre annotation for a large Russian webcorpus

Alexey Sorokin, Anisya Katinskaya, Serge Sharoff
Moscow State University, Russian State University for the Humanities, University of Leeds
alexey.sorokin@list.ru, a.katinsky@gmail.com, s.sharoff@leeds.ac.uk

Abstract

The paper describes several experiments aimed at establishing the parameters for genre annotation of potentially any text which can be collected from the Russian web. We started with a set of text classification parameters, refined them iteratively in several studies and established a reliable framework, which was further subjected to clustering analysis. Overall, we obtained the level of agreement for Krippendorff's α to be in the range of $0.51 < \alpha < 0.84$. We have also discovered the most common combinations of parameters in the test corpus, which should form the basis for classifying very large samples of the Russian web.

1 Introduction

Genre classification is often referred to as "jungle", this metaphor has been used by numerous researchers (Kilgarriff, 2001; Lee, 2001; Sharoff, 2010). While the web users see the differences between individual texts, it is difficult to agree upon a set of labels, which can cover the majority of webpages and which can be, at the same time, applied reliably by annotators (Sharoff et al., 2010). One reason for this is the sheer number of possible genre labels, up to 6,000 according to some studies (Adamzik, 1995). Another reason is a high degree of genre hybridism, especially on the web where many texts are not controlled by the institutional gate-keepers (Santini et al., 2010).

This study continued a line of genre classification experiments which started from adaptation of John Sinclair's typology of communicative aims to the needs of the Russian National Corpus (Sinclair, 2003; Sharoff, 2005), which in turn led to the Functional Genre Classes used in (Sharoff, 2010). The lack of reliability in classification of random web texts, which we investigated in the TTC project,[1] led to a proposal for introducing Functional Text Dimensions (FTDs), which can be used to judge the similarity of texts (Forsyth and Sharoff, 2014). It proposed such FTDs as:[2]

A1: Argumentative To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view?

A7: Instructive To what extent does the text aim at teaching the reader how to do something?

A12: Promotional To what extent does the text promote a commercial product or service?

[1] http://www.ttc-project.eu/

[2] See the full FTD list which we use in Appendix 1.

The annotators can express their opinions concerning each of these parameters on a scale from 0 (absent) to 2 (strongly present). From earlier studies we know that the annotators tend to achieve better agreement using such scales in comparison to atomic genre labels (Forsyth and Sharoff, 2014; Sharoff et al., 2010). The higher values in FTDs then serve as “symptoms”, from which we can infer a number of “syndromes”, i.e. traditional look'n'feel genres.

In this study we try to investigate the FTD values obtained from a diverse range of Russian texts to improve the FTD set and also to compare it against a recent project aimed at wide-ranging genre classification of the web (Egbert and Biber, 2013). The goal of this study is to make genre annotation reliable, i.e. the annotators should not struggle with the ambiguity of their choices, interpretable, i.e. the investigators understand the ranges of their choices, and applicable to any web page containing running text.

In Section 2 we introduce the corpus and the annotation procedure. In Section 3 we present the level of the interannotator agreement achieved on our corpus. In Section 4 we describe experiments at clustering our annotated corpus using the FTD values as an attempt to map the “symptoms” observed in our annotation to “syndromes”. Finally in Section 5 we analyse the results by comparing the clusters against the register list from (Egbert and Biber, 2013), which is in turn based on an earlier study (Rehm et al., 2008).

2 Corpus and the annotation procedure

Corpus selection and annotation went through three stages.

Stage 1 We selected 226 texts from Open Corpora (Bocharov et al., 2011) with the intention of enriching this collection with genre labels. Using texts with permissible licenses should have also helped in distributing the results of annotation freely. We requested two annotations per text to determine the degree of inter-annotator agreement. Another goal was to “debug” the definitions of the Functional Text Dimensions. For independent attribution of texts we used the Brown Corpus categories.

Stage 2 Since we found that the genre range of the Open Corpora collection is quite limited, we enriched the corpus by requesting *diverse* genre sets from a new round of annotators, who were also asked to annotate their own texts. The main sources of the texts were blogs, news portals, Wikipedia and other online encyclopedias, forums, online magazines, web libraries, promotional web-sites and legal resources. This iteration led to further feedback on the FTDs. We decided to add three more FTDs such as A16 (defining a topic), A17 (expressing judgments) and A18 (interaction between participants). This was primarily due to the fact that some texts in Corpus 2, e.g. Wikipedia articles, were not distinguished from other texts by the fifteen original FTDs.

Stage 3 Corpus 2 consisting of 514 texts was annotated by 11 annotators, overall 3 annotators per text. Table 1 describes quantitative characteristics of Corpus 1 and Corpus 2. Given that the Brown Corpus categories were found to be unsuitable for a large number of webtexts, we asked the annotators to attribute each text by using one of the registers used in Egbert and Biber (2013), e.g. ‘News report/blog’, ‘Description with intent to sell’, ‘Review’. This was needed to get some attribution which is independent from the FTDs and to evaluate the results. In addition,

Corpus 1			Corpus 2		
Documents	Words	Sentences	Documents	Words	Sentences
226	273319	16587	514	640476	41961

Table 1: Annotated corpora used in this study

the annotators were asked to give comments about the annotation procedure, the FTDs and the registers which caused the greatest difficulty for annotation.

As we are primarily interested in investigating *language* of the Web, we deliberately focused on the webpages with running text, leaving out such types of webcontent as social network profiles, e-commerce pages, forms, etc.

The annotation results show considerable correlation between several functional dimensions, such as A10 and A18, A14 and A15. Strong correlation between A10 and A18 gives reasons for removing duplicate FTDs. In the case of A14 and A15, we can assume that correlation was influenced by a small number of A15 documents that do *not* belong to the field of Science and Technology. The last annotation round also suggested a new "poetic" FTD, which roughly corresponds to the poetic function of language according to Jakobson (1960), covering various kinds of texts concerned with making an aesthetic impression.

3 Interannotator agreement

We measured the interannotator agreement using several methods, primarily using Krippendorff's α (Krippendorff, 2004), which treats the annotators are interchangeable and measures the difference between disagreement expected by chance vs observed disagreement:

$$\alpha = 1 - \frac{D_{observed}}{D_{chance}}$$

The corresponding results are shown in Table 2. Overall, the agreement is above 70%, but for some FTDs, such as newly added $A16, A17$ and $A18$ the level of agreement was lower. Some FTDs tested in previous studies, in particular $A5$ (flippant) or $A6$ (informal) also demonstrated considerable disagreement between annotators. However, in comparison to achieving agreement on atomic genre labels (Sharoff et al., 2010), FTDs overall offer better acceptable agreement judgments.[3] We also measured Krippendorff's alpha for the Biber register labels, the agreement was 53.0% for the register labels, 65.8% for the general registers and 58.7% using a compound distance function, assigning 1 for different labels in the same general register and 2 for a pair of elements in diverse registers.

Given systematic errors from some annotators for some FTDs (caused by their misunderstanding of the instructions), as well as possible issues with fatigue (coming from annotation of more than 100 texts), we need to exclude badly annotated texts from future research since the characteristics of such documents cannot be considered as "predictive" values of the FTDs. The second methodological reason for reducing the number of texts is that it is worse in general to have noisy data than to have less data, especially when the nature of noise is unclear. In the clusterisation task the presence of noise may potentially deform the shape of clusters or create spurious clusters which do not

[3] $\alpha \geq 60\%$ is usually treated as the acceptability threshold (Krippendorff, 2004).

	A1	A3	A4	A5	A6	A7	A8	A9
Overall	56.39	58.44	71.80	51.39	55.53	69.58	78.00	84.22
Annotator 1	67.31	51.16	83.86	58.42	35.33	69.15	84.62	79.66
Annotator 2	54.67	68.94	71.71	35.06	61.41	68.86	64.68	80.56
Annotator 3	28.81	53.22	68.60	58.50	52.36	64.87	77.14	81.71
Annotator 4	66.34	54.51	72.39	60.72	53.66	80.03	58.48	87.19
Annotator 5	58.25	67.92	82.74	36.04	69.28	84.89	77.65	81.16
Annotator 6	50.62	57.37	60.14	40.69	60.92	69.69	61.25	69.63
Annotator 7	66.13	72.92	84.04	41.73	59.13	83.87	82.17	81.01
Annotator 8	50.35	64.07	77.38	57.04	56.40	52.30	85.09	91.08
Annotator 9	58.82	50.67	69.86	62.38	33.24	46.23	38.87	92.96
Annotator 10	38.26	34.99	61.49	46.11	52.11	80.19	88.85	81.30
Annotator 11	51.64	58.51	52.51	45.02	64.61	74.49	74.83	90.38
	A12	A13	A14	A15	A16	A17	A18	Total
Overall	62.09	58.71	52.14	56.80	63.55	51.43	49.32	62.95
Annotator 1	62.68	51.42	59.24	61.53	42.80	64.51	34.72	62.99
Annotator 2	46.58	57.11	50.02	62.89	69.89	60.04	34.66	62.78
Annotator 3	86.29	60.07	12.59	57.23	59.54	-12.33	66.03	54.99
Annotator 4	75.61	60.92	66.03	67.73	53.15	50.44	59.79	65.72
Annotator 5	42.25	49.72	60.59	64.84	70.52	48.29	58.08	66.19
Annotator 6	52.29	65.94	53.21	46.59	54.84	46.84	40.27	59.27
Annotator 7	63.86	47.40	57.57	64.07	56.01	62.10	36.93	67.14
Annotator 8	59.46	52.94	46.10	54.98	74.17	57.94	53.34	67.26
Annotator 9	64.96	72.58	10.62	32.36	67.45	58.04	55.90	57.85
Annotator 10	63.00	62.21	61.90	59.55	65.49	33.79	50.67	59.93
Annotator 11	61.35	66.78	40.09	38.53	60.98	48.12	47.08	62.08

Table 2: Krippendorff's α values for FTDs and annotators

correspond to any pattern in the data. However, we cannot simply exclude from the sample the documents for which there was no agreement since two or more annotators agreed on all the FTDs only for 199 of 500 initial texts. Moreover, we found that the more informal the text is the lower the probability of their agreement. It implies that the distribution of the FTD values over the reliable texts considerably differs from the same distribution over the set of all documents.

Therefore, we tried to preserve as many documents as possible unless they can be considered as reliable. The key idea is to remove the *ratings* of annotators when they disagree with other annotators and keep their ratings otherwise. For every FTD we evaluated the quality of experts using Krippendorff's α: for any text annotated by k annotators ($k = 3$ in our case) we created $k - 1$ pairs of ratings assigned by the current annotator together with the rating given by another annotator. Then for every FTD we selected the worst annotator and removed his/her ratings for all documents where s/he disagreed with other annotators. Then individual annotation qualities were recalculated and a new worst annotator was selected until the agreement quality for the worst annotator reached a fixed threshold or became close to the mean value of individual agreement for all annotators. We used the α value of 0.75 for the threshold, also we tried the values from 0.6 to 0.7.

Selection strategy	# documents
Alpha threshold 0.75	283
Alpha threshold 0.75, 0.5 disagreement allowed	315
Alpha threshold 0.6	238
Alpha threshold 0.6, 0.5 disagreement allowed	263
2 annotators agreement	199
No selection	500

Table 3: The number of remaining documents for different selection strategies

	A1	A3	A4	A5	A6	A7	A8	A9
Overall	86.73	81.34	87.64	76.21	89.88	87.97	90.06	86.28
Best annotator	97.09	92.81	94.46	96.98	96.77	93.93	99.67	93.96
Worst annotator	75.07	73.32	80.06	65.51	78.50	77.75	78.84	80.20
	A12	A13	A14	A15	A16	A17	A18	Total
Overall	88.47	88.70	53.24	58.50	88.47	58.62	85.26	80.97
Best annotator	98.45	93.09	66.89	68.94	96.21	71.56	94.68	81.79
Worst annotator	80.47	75.17	12.57	39.74	75.77	21.81	77.48	77.53

Table 4: Krippendorff's α values for FTDs after removing unreliable annotations

Since these thresholds produced a lower number of reliable texts, we allowed different annotators to disagree by 0.5. The sizes of obtained collections of documents are shown in table 3. Table 4 contains the values of Krippendorff's α for FTDs after the removal procedure with threshold 0.75 and 0.5 disagreement allowed.

4 Clustering

The ultimate goal of our project is to provide automatic genre classification using the FTDs, which can be detected reliably, and also to map them to genre syndromes. Therefore, we are interested in exploring frequent combinations of FTD values corresponding to stable patterns, which may be considered as similar to genres in a traditional genre system like Egbert and Biber (2013). Our primary task in this section is to detect clusters in the FTDs space and to investigate whether these clusters are linguistically relevant.

We used several methods to cluster the texts by the values of their FTDs. Since the FTD values are discrete, we used the taxicab (also known as Manhattan) metric as the distance function. The discreteness of our feature space makes the clustering task problematic. The general difficulty is that we do not know the probability distribution on the sample set, especially on its subset selected for clusterisation. Also it is difficult to perform feature weighting for discrete features with a small number of feature values. Both agglomerative and iterative methods have their own drawbacks complicating their usage for our task.

The main difficulty with the agglomerative methods is their sensitivity to the order of objects. This sensitivity becomes more dramatic since the discreteness of our feature space creates many ties. Also it is impossible to correct the wrong choices made during early clusterisation steps. However,

Selection method	# documents	k-means	k-medoids	Hierarchical
0.75	283	0.864	0.766	0.769 (R)
		0.468	0.425	0.423 (S)
0.75, 0.5 disagreement	315	0.810	0.757	0.714 (R)
		0.444	0.431	0.380 (S)
0.6	238	0.844	0.805	0.871 (R)
		0.480	0.440	0.383 (S)
0.6, 0.5 disagreement	263	0.783	0.745	0.719 (R)
		0.427	0.440	0.365 (S)
2 annotators agreement	199	0.890	0.863	0.827 (R)
		0.508	0.505	0.450 (S)
No selection	500	0.691	0.531	0.490 (R)
		0.329	0.318	0.168 (S)

Table 5: Robustness and silhouette for different selection strategies

the usage of iterative methods is even more problematic. First of all, the most common methods such as EM or K-means are not suitable for discrete spaces. We can use k-medoids instead but this does not fix the problem of detecting the number of clusters. Also iterative methods are sensitive to initial cluster approximations.

Therefore we decided to combine hierarchical and iterative methods. First of all we performed agglomerative clustering using the weighted linkage. We detected the number of clusters k searching for the knee position of the evaluation graph by the well-known L-method (Salvador and Chan, 2004). Then we used these clusters as initial approximations for the k-medoids algorithm. We also tried k-means, even though it has limitations when applied to discrete data.

The proposed method of clusterisation is unstable, since it is sensitive to the initial order of objects. We measured the robustness of clusterisation using the adjusted Rand index (ARI), which measures the degree of intersection between different clusterisation runs. Precisely, we ran the clusterisation algorithm for 20 different random orderings of data and calculated the average value of ARI between the clusters obtained in this way. To estimate the internal quality of clustering we applied the silhouette score (Rousseeuw, 1987), which assesses how well all objects lie within their clusters.

The values of the indices for different selection strategies and different clusterisation methods are shown in Table 5, each vertical section corresponds to a particular data selection method and consists of two rows: the first row contains the values of the robustness index (R), the second one contains the silhouette score (S).

Unfortunately, the scores are not particulary high because of the annotation noise. It means the additional procedure of noise removal must be performed. We determine and remove the outliers using the silhouette score. To be removed an object should have the silhouette score lower then the minimum of the predefined threshold t (we choose $t = 0.25$) and current threshold $t' = \mu - \sigma$ with μ and σ being, respectively, the mean and the deviation of the individual silhouette scores. After excluding the unreliable objects, the scores were recalculated, and the process was repeated until the Rand score exceeds 90%. The convergence of the algorithm was rather fast which means that the core elements of clusters are clustered independently of data ordering.

Selection method	Initial # documents	k-means	k-medoids	Hierarchical
0.75	283	0.924	0.984	0.972
		0.670	0.660	0.625
		206	194	203
		16	14	15
0.75, 0.5 disagreement	315	0.936	0.993	0.927
		0.566	0.686	0.530
		248	209	219
		8	16	8
0.6	238	0.954	0.925	0.931
		0.631	0.605	0.486
		188	196	206
		9	10	8
0.6, 0.5 disagreement	263	0.959	0.956	0.948
		0.574	0.550	0.579
		215	216	214
		11	11	11
2 annotators agreement	199	0.991	0.991	0.968
		0.696	0.654	0.659
		153	162	160
		13	9	12
No selection	500	0.915	0.953	0.998
		0.414	0.570	0.641
		358	295	250
		13	12	15

Table 6: Different methods for cleaning clusters

The values of the scores are presented in Table 6. Each vertical section contains four rows, containing the values of the adjusted Rand index, the silhouette score, the number of objects attached to clusters and the final number of clusters.

So different clusterisation algorithms produce different number of clusters in the final clusterisation experiment. Though these values are very important for further usage of obtained clusters as classes for document classification, they cannot be considered as absolutely reliable. To uncover better the cluster structure we made an additional experiment when the number of clusters was fixed through the algorithm. We varied the number of clusters from 8 to 15. In this experiment we used the initial set of 315 documents, obtained in the run of selection algorithm with 0.75 threshold and allowed 0.5 disagreement between annotators.

Some basic clusters were found independently of the number of clusters. The documents in each cluster usually share principal FTDs, i.e. FTDs, which are equal to 2. Below we list these clusters together with their principal dimensions.

1. “Instructions” (21 texts), principal dimension $A7$.
2. News (64 texts), principal dimension $A8$.
3. “Legalese”(11 texts), principal dimension $A9$.

4. “Specialised technical texts” (13 texts), principal dimensions $A14, A15$.
5. Descriptive and encyclopedic texts (49 texts), principal dimension $A16$.
6. Adverts (13 texts), principal dimensions $A1, A12, A17$.
7. Argumentative propaganda texts (13 texts), principal dimensions $A1, A13, A17$.

5 Interpretation of clustering results

We were interested in comparing the clusters to Biber's registers. We identify the clusters with the set of their principal FTDs, while indicating other FTDs which are significant only for a portion of our texts.

The cluster with the principal dimension A9 is the only cluster which contains documents annotated within a single Biber's register (‘Legal terms and conditions’). Some texts of the A9 cluster also contain higher values of the A7 FTD (instructions), describing a proper legal procedure to achieve something, as well as, the A16 FTD (a text defining a topic), e.g., definitions of the kinds of taxation or of land property.

The cluster with the principal dimension A7 includes several kinds of instructional texts according to Biber (’How-to’, ’Instructions’, ’Recipe’), as well as a smaller amount of ’Technical support’, ’Advice’, ’Self-help’, where ’Advice’ and ’Self-help’ belong to the general register of ’Opinion’, while the instructional texts belong to the general register of ’Non-opinion’ (Egbert and Biber, 2013). The degree of recommendation is represented by variation in the A17 FTD.

The A8 (news) cluster combines the narrative registers which report an event (’News report/blog’, ’Sports report’), as well as a smaller amount of ’Magazine article’ and ’Obituary’. The presence of other registers in this cluster might be due to errors the annotators made in assigning these registers. ’Magazine article’ is a very general category, which can contain more or less of argumentation (A1), entertainment (A5), information (A8) or evaluation (A17).

The A16 (definitions) cluster includes texts annotated with descriptive registers (’Encyclopedia article’, ’Legal terms and conditions’, ’Description of a person’, a smaller amount of ’Description of a thing’, ’Research article’, ’Abstract’), as well as one narrative register (’Biographical story/history’). It is necessary to emphasise that all the texts annotated as belonging to these registers and to the A16 cluster also have a property of defining a topic, e.g., life of Peter the Great, a kind of cheese, a national holiday in India. We suppose that from the point of view of the annotators the distinction between biographical, descriptive and historical texts is less important. In this cluster a small number of texts (mostly ’Encyclopedia article’) also have the secondary dimension A15 (text requiring specialist knowledge).

The A14A15 cluster contains texts belonging to the field of Science and Technology (’Encyclopedia article’, ’Research article’), along with a small amount of ’Technical support’, ’Technical report’, ’How-to’ and ’Instructions’), at this stage of experiment all texts of this cluster require readers to have background knowledge. The texts annotated as ’How-to’ and ’Instructions’ are technical, e.g. texts about how to write a compiler. The secondary dimension A16 in this cluster is also prominent. Our clustering procedure probably separated sci vs non-sci texts of the descriptive kind. Sci texts also tend to require more specialist knowledge in comparison to biographies.

The A1A12A17 cluster combines advertisements (’Description with intent to sell’, ’Persuasive article or essay’, ’Opinion blog’, ’Review’). The A1A13A17 cluster mainly includes the texts which correspond to the Biber's ’Persuasive article or essay’, but also a small number of other registers with persuasion and argumentation (’Prayer’, ’Religious blog/sermon’, ’Research article’).

Even though the results of clustering are described in terms of the FTDs, the number of possible clusters will be much more extensive, when we include more texts in the analysis. Many existing clusters are better to be described with a set of FTDs, such as the A1A12A17 or the A1A13A17 clusters.

Having the preliminary results we suppose that some of our clusters are similar to Biber's general registers, e.g., the A7 cluster corresponds to Instructional texts, the A1A12A17 cluster to 'Intent to sell' and the A1A13A17 cluster to 'Persuasive articles, although there are some differences, e.g. we do not have lyrical or opinion clusters in this corpus. However, this is primarily because of the composition of our corpus.

'Persuasive article' and 'Review' are widespread across our clusters. The texts annotated as 'Persuasive article' are mostly presented in the A1A12A17 cluster and in the A1A13A17 cluster (adverts and argumentative propaganda texts), which is reasonable. There are only isolated cases of this register in the A7 cluster and the A8 cluster. Similar situation is with 'Review', which mostly occurs in the A1A12A17 cluster and sporadically occurs in the A8 cluster and the A16 cluster. We can expect a lot of variation in terms of the functional dimensions in texts of such registers, and this leads to the lack of internal stability of these registers in terms of the clusters they have been assigned to.

6 Conclusions

We have tested the Functional Text Dimensions framework on a wide variety of Russian texts and suggested changes to these dimensions in comparison to previous studies. For example, we removed largely duplicate dimensions with strong pairwise correlation and suggested new dimensions to make distinctions between important text types (see Appendix 1). Finally, we analysed the reliability of FTD annotation and obtained data to improve the annotation quality. The results of annotation are available from: `http://corpus.leeds.ac.uk/serge/webgenres/`

We have also experimented with various clustering techniques and detected patterns in annotated data, which lead to the possibility of uncovering "syndromes" of features corresponding to look'n'feel genres, such as 'News', 'Legal texts', 'Research papers'. However, identification of the clusters defined by several principal FTDs requires further research. In the initial steps for this research, we compared these clusters to other genre classification frameworks, in particular to registers from (Egbert and Biber, 2013). We detected cases of good agreement, as well as disagreement between the two annotation approaches, which can potentially lead to enrichment of both methods. In particular, some registers, e.g., 'Persuasive articles', exhibit internal variation, which is best explained by using the FTDs.

One of the important outcomes of the annotation experiment is that it demonstrates the possibility to achieve acceptable interannotator agreement on the FTDs, while the annotators often disagree with respect to atomic labels Sharoff et al. (2010). 'News' as a cluster corresponds nicely to the A8 principal dimension. However, this does not make 'News' a reliable label. On the contrary, a text with a high A8 value can be more or less argumentative (A1), light-hearted (A5), or can contain an overview of a topic (A16). A traditional genre palette forces the annotators to choose one label, which can be 'News report', 'Short story', 'Magazine article', 'Opinion', 'Persuasive article', etc. It is natural for different annotators to consider a text from different viewpoints, thus reducing reliability of their annotation. If linguistically similar texts receive different labels, this in turn reduces accuracy of an automatic classifier. At the same, a combination of several FTDs is more likely to achieve both reliable annotation and reliable classification.

In the next step, we would like to train classifiers for each of these dimensions and apply them to a large Russian webcorpus of about 50 billion words Piperski et al. (2013). This should provide us with a genre map of the entire space of Russian web texts, so that the linguistic researchers can select a corpus subset according to their interests, e.g., personal narratives (A11) or evaluative texts (A17). To develop the classifiers we will need more research into the linguistic features associated with particular FTDs.

Acknowledgements

Research reported in this paper was partly funded by a grant provided by the Skolkovo Institute of Science and Technology. We are grateful to Alexander Piperski for his help in organising the evaluation procedure in the Russian State University for the Humanities and to Vladimir Selegey for general support.

References

Adamzik K. Textsorten – Texttypologie. Eine kommentierte Bibliographie. — Münster : Nodus, 1995.

Bocharov V. , Bichineva S., Granovsky D. et al. Quality assurance tools in the OpenCorpora project // Proc. Dialogue, Russian International Conference on Computational Linguistics. — Bekasovo, 2011.

Egbert J., Biber D. Developing a user-based method of register classification // Proc. 8th Web as Corpus Workshop. — Lancaster, 2013. — July.

Forsyth R., Sharoff S. Document dissimilarity within and across languages: a benchmarking study // Literary and Linguistic Computing. — 2014. — Vol. 29. — P. 6–22.

Jakobson R. Linguistics and poetics // Style in Language / Ed. by T. A. Sebeok. — The M.I.T. Press, 1960. — P. 350–377.

Kilgarriff A. The web as corpus // Proc. of Corpus Linguistics 2001. — Lancaster, 2001. — URL: `http://www.itri.bton.ac.uk/techreports/ITRI-01-14.abs.html`.

Krippendorff K. Reliability in content analysis: Some common misconceptions and recommendations // Human Communication Research. — 2004. — Vol. 30, no. 3. — URL: `http://faculty.washington.edu/jwilker/559/PAP/krippendorf-reliability.pdf`.

Lee D. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle // Language Learning and Technology. — 2001. — Vol. 5, no. 3. — P. 37–72. — URL: `http://llt.msu.edu/vol5num3/pdf/lee.pdf`.

Piperski A., Belikov V., Kopylov N et al, Big and diverse is beautiful: A large corpus of Russian to study linguistic variation // Proc. 8^{th} Web as Corpus Workshop (WAC-8). — 2013.

Rehm G., Santini M., Mehler A. et al. Towards a reference corpus of web genres for the evaluation of genre identification systems // Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008. — Marrakech, 2008.

Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of computational and applied mathematics. — 1987. — Vol. 20. — P. 53–65.

Salvador S., Chan P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms // Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on. — 2004. — P. 576–584.

Santini M., Mehler A., Sharoff S. Riding the rough waves of genre on the web // Genres on the Web: Computational Models and Empirical Studies / Ed. by Alexander Mehler, Serge Sharoff, Marina Santini. — Berlin/New York : Springer, 2010.

Sharoff S. Methods and tools for development of the Russian Reference Corpus // Corpus Linguistics Around the World / Ed. by D. Archer, A. Wilson, P. Rayson. — Amsterdam : Rodopi, 2005. — P. 167–180.

Sharoff S. In the garden and in the jungle: Comparing genres in the BNC and 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.

Sharoff S., Wu Z., Markert K. The Web library of Babel: evaluating genre collections // Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010. — Malta, 2010. — URL: `http://corpus.leeds.ac.uk/serge/publications/lrec2010.pdf`.

Sinclair J. Corpora for lexicography // A Practical Guide to Lexicography / Ed. by P. van Sterkenberg. — Amsterdam : Benjamins, 2003. — P. 167–178.

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, using colloquialism and/or slang (as opposed to the "standard" or "prestige" variety of language)?
A7.	instructive	To what extent does the text aim at teaching the reader how to do something? (e.g. a tutorial)
A8.	news	To what extent does the text appear to be a news report such as might be found in a newspaper, i.e. an informative report of recent events? (recent at the time of writing. None if a news source does not provide new information about what happened, while analysing 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	Does the text report a first-person point of view?
A12.	compuff	To what extent does the text promote a commercial 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? (i.e. any promotion of not-for-profit causes)
A14.	scitech	To what extent would you consider the text as belonging in the field of Science, Technology and/or Engineering? (As opposed to the Arts, Humanities &/or Social Studies. This is not necessarily a research paper. A newswire text can include scientific contents, so it can be judged as Strongly or Partly.)
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.	encyc	To what extent does the text provide information to define a topic? (For example, encyclopedic articles or text books).
A17.	eval	To what extent do you judge the text to evaluate something? (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 dialogue lines coming from theatre plays).
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)

Rating	**Levels:**
0	none or hardly at all;
0.5	slightly;
1	somewhat or partly;
2	strongly or very much so.