Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

TOPOGRAPHY OF INTERNET CORPORA

Sharoff, Serge
University of Leeds

DOI: 10.28995/2075-7182-2020-19-1134-1137

1. Rationale

Modern linguistics is unthinkable without the use of corpora, while the Internet brought to the service of linguists really Big Data [7]. This made collecting large corpora from the Web much easier than it was in the 1990s. **Table 1** compares popular Russian corpora with respect to their basic parameters such as the number of their texts and the word count. The Russian National Corpus is much smaller than the Web corpora, so it is difficult to use it for making statistically reliable predictions for phenomena occurring less frequently than words or constructions like *парковать, ябеда* or *глубокая пропасть* (all occurring 40 times in the modern part of the RNC). Yet the RNC is considerably more popular with Russian corpus linguists. This is partly because there is much better understanding of its composition in terms of topics and genres in comparison to the Web. The General Internet Corpus of Russian (GICR) tried to address the issue of composition by collecting data from specific sources, so that the researcher knows where the corpus values are coming from. However, even this restriction does not guarantee homogeneity in data.

Table 1: Large Russian corpora

Corpus	#Texts	#Words	#Genres,Domains
RNC'09 [5]	0.04M	100M	177 genres (тост), 96 topics (медицина)
Aranea'16 [2]	2.4M	13995M	fskn.gov.ru, bookap.info, blogspot.co.il
rutenten'13 [4]	38M	20715M	fas.gov.ru, gorodovoy.spb.ru, news.yandex.ru
ruWac'10 [10]	1.3M	2006M	livejournal.com, narod.ru, hiblogger.net
GICR, news'15 [1]	2.2M	599M	lenta.ru, ria.ru, rosbalt.ru
GICR, livejournal'15 [1]	11M	3619M	livejournal.com
GICR, magazines'15 [1]	0.06M	314M	magazines.russ.ru

2. Automatic annotation of texts

In my presentation I will explore Machine Learning methods to understand the topography of Web corpora with the aim of comparing them to each other and to the RNC. It is clearly impossible to apply any established set of genres to the totality of Web pages, partly because of the sheer number of genres of everyday life, and partly because of the hybrid nature of many pages [6], for example, a Web page can contain a combination of news-like reporting with expressing personal opinions, similarly, a Web page with the intention of selling a product can include a very elaborate specification.

A more successful approach is based on the Functional Text Dimensions [8], which allow ranking each text with respect to common communicative functions along with their prototypes, for example:

- **news** To what extent does the text provide an informative report of recent events? (For example, a newswire item).
- **argument** To what extent does the text try to persuade the reader? (For example, an argumentative blog entry or a newspaper opinion column).
- **promotion** To what extent does the text promote a product or service? (For example, adverts, spam).
- **information** To what extent does the text provide reference information to describe something? (For example, encyclopedic articles, dictionary definitions, specifications)

A hybrid text in this scheme can feature on two or three dimensions at the same time, for example, commercial promotion with reference information, or an online news publication as a proportion of proper news reporting mixed with argumentation.

A sufficiently large sample of Web texts annotated with the set of functions as proposed in [8] has been used for training an automatic classifier, which can apply a consistent annotation scheme to both traditional and Web-derived corpora. The accuracy of existing genre classifiers based on neural networks is around 75% per FTD [9], which is acceptable for predicting the basic landscape of large corpora.

3. Analysis of corpus topography

Table 2: Composition of large Russian corpora

FTD	RNC		ruWac		rutenten		Aranea	
Argument	19.76%	6,576	18.20%	222,741	8.30%	1,881,900	9.36%	1,136,912
Fiction	**21.25%**	7,070	1.55%	18,919	3.16%	716,808	1.87%	227,526
Instruction	0.52%	172	1.02%	12,446	5.13%	1,162,261	1.41%	171,095
News	**22.57%**	7,511	5.77%	70,689	**23.83%**	5,401,615	17.66%	2,144,759
Legal	1.89%	628	1.23%	15,088	2.94%	666,964	0.98%	118,500
Personal	4.75%	1,581	**44.29%**	542,111	9.55%	2,165,051	12.84%	1,559,338
Promotion	4.43%	1,474	5.59%	68,432	17.34%	3,929,776	**36.58%**	4,441,845
Academic	3.89%	1,295	4.77%	58,410	5.98%	1,354,483	7.79%	946,241
Information	8.10%	2,696	11.71%	143,324	11.10%	2,514,720	8.07%	979,607
Review	12.84%	4,272	5.87%	71,905	12.67%	2,871,100	3.43%	416,830

Table 2 shows the distribution of predicted communicative functions in the major Russian corpora in terms of the number of texts classified with the predominant FTD. The composition of these corpora is so different that it is difficult to generalise the studies obtained on one corpus, such as the RNC to other corpora. Douglas Biber mentioned that "language may vary across genres even more markedly than across languages" [3]. A study based on a corpus dominated by news reporting (RNC or rutenten) is likely to be different from a study of the same phenomenon in a corpus dominated by promotional texts (Aranea), which are not well represented in the RNC.

What is more, a corpus investigation needs to be sensitive to specific text types (news reporting, promotional texts or academic writing) instead of broad generalisations. A small preliminary study which is based on Biber's features [3] shows some parameters of variation, such as significant differences in the rate of past tense verbs in promotional texts vs news reporting and fiction, since the discourse structure of the latter is based on narration, which is predominantly reported in the past tense. Some of the less expected differences concern the differences in private blogs vs promotional texts. While the latter aim at being informal and engaging, some of their features, such as text-level statistics (lexical density or word length) and lexicogrammatical properties (the rate of nouns, subordinate clauses or attributive adjectives), make them closer to more formal argumentative texts. Differences of this kind become apparent only when texts in large Web corpora receive sensible metadata annotations.

Even though GICR is more careful than other Web corpora by allocating its sources to separate segments, analysis of their composition (Table 3) also demonstrates internal variation within each segment. For example, the most important variation within news texts concerns news reporting vs argumentative texts with considerable differences in their lexicogrammatical properties. As expected, the Livejournal component is dominated by personal stories from private blogs, but the structure of the collection of literary magazines presents considerable unexpected variation. In addition to the expected fiction stories and critical essays, they were found to contain a considerable amount of reference information texts, such as biographical profiles of authors or descriptions of other publications. Another unexpected communicative function concerns personal stories, which are either memoirs or first-person fiction stories which are nearly indistinguishable from private blogs.

Table 3: Composition of GICR subcorpora

FTD	news		livejournal		magazines	
Argument	27.59%	534,212	3.61%	254,830	**26.74%**	11,675
Fiction	0.08%	1,603	1.30%	91,831	17.96%	7,842
Instruction	0.03%	557	0.44%	31,116	0.08%	33
News	**67.03%**	1,298,003	3.56%	251,328	1.15%	500
Legal	0.44%	8,603	0.12%	8,436	0.04%	19
Personal	2.03%	39,280	**85.85%**	6,055,026	**15.64%**	6,830
Promotion	0.44%	8,517	0.89%	62,442	0.09%	41
Academic	0.24%	4,590	0.24%	16,650	2.78%	1,214
Information	1.69%	32,774	2.02%	142,494	**20.81%**	9,086
Review	0.43%	8,390	1.97%	138,656	14.71%	6,424

The tools for annotation and the annotated corpora have been made available to facilitate further research in genre studies.[1]

References

1. *Belikov Vladimir, Kopylov Nikolay, Selegey Vladimir, Sharoff Serge.* Variational Corpus Statistics Using Author Profiles // Proc Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2014.
2. *Benko Vladimír.* Two Years of Aranea: Increasing Counts and Tuning the Pipeline // Proc LREC. Portorož, Slovenia, 2016.
3. *Biber Douglas.* Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge University Press, 1995.
4. *Jakubíček Miloš, Kilgarriff Adam, Kovář Vojtěch, Rychlý Pavel, Suchomel Vít.* The tenten corpus family // Proc Corpus Linguistics Conference. Lancaster, 2013. P. 125–127.
5. *Lyashevskaya Olga, Sharoff Serge.* Chastotny slovar sovremennogo russkogo yazyka. Moscow : Azbukovnik, 2009.
6. *Santini Marina, Mehler Alexander, Sharoff Serge.* 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.
7. *Sharoff Serge.* Open-source Corpora: using the net to fish for linguistic data // International Journal of Corpus Linguistics. 2006. Vol. 11, no. 4. P. 435–462.
8. *Sharoff Serge.* Functional Text Dimensions for the annotation of Web corpora // Corpora. 2018. Vol. 13, no. 1. P. 65–95.
9. *Sharoff Serge.* Finding next of kin: Cross-lingual embedding spaces for related languages // Journal of Natural Language Engineering. 2020. Vol. 26. P. 163–182.
10. *Sharoff Serge, Goldhahn Dirk, Quasthoff Uwe.* Frequency Dictionary: Russian. Leipziger Universitätsverlag, 2017. Vol. 9 of Frequency Dictionaries. P. 9–14. Uwe Quasthoff, Sabine Fiedler, Erla Hallsteindóttir (editors).

[1] https://github.com/ssharoff/genre-keras