# Differential Approach to Web-Corpus Construction

Tatiana Shavrina
rybolos@gmail.com
NRU Higher School of Economics
Moscow, Russia

**Abstract**
A wisely constructed web corpus has a lot more potential applications. The “web as corpus” paradigm which has had its natural continuation as a formulation “web as train set”, provide ample opportunities for NLP-developers and computational linguists, who nevertheless often have to gather all the corresponding data by themselves. In this paper "Taiga" corpus project is described - a new web-corpus of Russian with open text sources, deduplicated, automatically tagged in Universal Dependencies format and enriched with wide specter of metadata. Corpus is constructed in the way its segments provide the most up to date NLP problems with big data. The goal of the project is to provide a growing Russian NLP community with large and accurately collected corpus materials, as well as to reflect the shift in the global paradigm of corpora studies, by which only open source data makes it possible to investigate the distribution of interesting linguistic phenomena independently and completely, refusing the “black box” interface usage.
**Keywords:** corpus construction, web corpus, web corpus construction, machine learning, corpus representativity

# Дифференциальный подход к построению веб-корпусов

Татьяна Олеговна Шаврина
rybolos@gmail.com
Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия

**Аннотация**
Правильно собранный веб-корпус потенциально имеет множество применений. Парадигма “веб как корпус” с ее логическим продолжением “веб как обучающая выборка” породила широкий спектр возможностей для разработчиков в области обработки естественного языка и компьютерных лингвистов, которым, тем не менее, часто приходится собирать все необходимые массивы интернет-текстов самостоятельно. В данной статье описывается новый веб-корпус русского языка “Тайга”, с открытыми исходными текстами, дедубликацией, метатекстовой разметкой и автоматической разметкой морфологии и синтаксиса, который собран таким образом, чтобы его сегменты обеспечивали большими данными самые актуальные задачи компьютерной лингвистики. Целью проекта является обеспечение растущего русского NLP-сообщества большими и аккуратно собранными корпусными данными, а также отражение сдвига мировой парадигмы корпусного исследования, в которой только открытые исходные данные дают возможность изучить распределение интересующих языковых явлений самостоятельно и полно, отказавшись от “черного ящика”.
**Ключевые слова:** построение корпусов, веб-корпуса, корпус русского языка, машинное обучение

## *1. Introduction*

In the last decade corpus projects, whose volume exceeds a billion words, are rapidly developing. For the Russian language there are, for example, ruWAC (Baroni et al., 2009), RuTenTen (Jakubicek et al., 2013), General Internet-Corpus of Russian (Belikov et al., 2013), Aranea Russicum (Benko 2014), Google Ngrams. Their volume is their undeniable dignity, but all these projects are united by one more feature - they provide access to the search, but their materials cannot be obtainable for independent development or autonomous statistical analysis. Corpus linguistics is positioned as a more accurate and computational approach to language, but in fact, a researcher studying a specific linguistic phenomenon, wanting to understand what the absolute frequency or IPM of a phenomenon obtained on corpus does mean, either relies on comparing the search results on different resources or on comparing the relative frequencies of various phenomena (which is laborious in both cases), or, more often and worse, just relies on his own intuition.

Interpretation of the corpus search result is a non-trivial task when the distribution of frequencies of words, parts of speech, text sources and other parameters is hidden from the user – first steps in this field are made by (Lagutin et al., 2016), introducing automatic statistical analysis in the search interface.

Another approach which can be proposed to escape statistical black box challenge, is that only fully open data can be used. In this case, every user is responsible for interpretation of his search results, but can perform any analysis on them. The only point then is that this data should be open-source and enough to represent the a certain part of the language.

Coming up with the problem of representativity on web corpora and big linguistic data, we can reformulate it, basing on the statistical idea of “language is a large set of rare events” and sufficient data amount on every non-hapax linguistic event. Coverage of language variation (firstly introduced by (Belikov et al., 2013)) then is a new main big corpus characteristics.

Summarizing, we can distinguish 3 modern approaches to the collection of large web corpora:

1) classic – crawling every resource, using crawlers, which address to search engines and walk through the pages; and after the material is cleaned from spam and is deduplicated. This approach does not allow saving maximum of metatext information, since all the boilerplate is deleted, but it allows to gather a lot in a short time (example – Common Crawl).

2) fitted – all the materials from the listed thousands of URLs are crawled. Sometimes a fit-function is used, which decodes whereas URL is suitable or not, while crawler addresses to the search engines, like in the first approach (example – Aranea corpora).

3) differential – a small number of large resources are crawled, but they are downloaded as completely as possible, entirely, if possible. This download allows us to state that linguistics variation of the resource is covered entirety (example – General Internet-Corpus of Russian).

The author adheres to the idea of coverage of language variation, creating a web-corpus “Taiga” which meets the interests of both linguists and NLP-developers. Corpus is build by collecting a small number of segments with a lot of homogeneous in each, each segment suitable for solving a different set of NLP tasks.

## *2. Web as train set*

In the methodological preparation for creating a new resource (Shavrina, Shapovalova, 2017), 5 main principles were postulated, which allow a modern web-corpus to be relevant for the engineering approach to language:

1) open source

A corpus should be available to use the material at one's discretion, to modify it, to add new data to it, and publish it. Developing a good vector model on the data, composing new genre-specific frequency dictionaries, etc. then becomes a faster side products by a community of corpus users.

Texts of the corpus can be collected from open sources only, which, nevertheless, can be quite sufficient, since now there are a lot of such resources, and compiling their complete list for each language is a separate, voluminous task.

2) big data

A corpus should provide sufficient volume of the data (more than 10 million words) - thus it is possible to collect implicit information, for example about rare words and their compatibility, syntactic behavior of individual words and different meanings of homonyms. This principle is also very important for users, as for many applications doubling the training data improves the quality more than developing a more complicated algorithm (Banko , Brill, 2001).

3) clear data

There should be minimum share of errors embedded in the data itself - this includes both the sufficient quality of linguistic markup, and the completeness of meta-text information, the ability to uniquely separate one segment or genre from another, to find out their balance for each research.

4) coverage of linguistic variation

Corpus data should represent all possible variability in unbiased proportions for each separate resource.

5) solvability in a given metric

Adequacy of data composition and its' features to the applications is one of the most important principles - otherwise it is useless. Meta-information, such as author's gender, age, text theme, etc (depending on the task) should be sufficient for the research, and on the part of the authors, it is necessary to provide maximum opportunities for users.

With these principles, we believe that a corpus product that meets modern requirements of corpus linguistics can be created - it will not be a black box, it will be reflecting modern language and its features, not biased and capable of encouraging more cooperation between developers and linguists. Refusal of the search interface, in our opinion, can be justified in this case, since obtaining examples for the concrete hypothesis, obtaining an IPM for a specific comparison on an unknown volume of data with an unknown distribution is a long-standing problem, which we wish to escape. As an alternative, a python-compatible API can be provided for community, which will include functions for statistic analysis.

### *3. Possibilities of corpora for machine learning*

By now, our corpus contains data from 14 sites, 44 sources and is of 4,2 billion words or 5,4 billion tokens, all documenting modern Russian language (we considered "modern Russian" as a language of speakers younger than 60 years old).

All of the sources are conditionally divided into 6 segments: news, fiction, poetry, subtitles, social networks, special datasets. First 5 of the segments are homogeneous, whereas the last one is containing different text sources with specific tagging – dataset with text readability annotation, dataset with authors profession and, etc – all coming from magazines. See table 1 for data distribution between segments:

Table 1: Taiga corpus segments

| Segments | Tokens, millions | % |
|---|---|---|
| Fiction (2 sites) | 4605 | 75,60 |
| News (4 sites) | 92 | 1,51 |
| Poems (1 site) | 1211 | 19,88 |
| Social media (4 sites) | 80 | 1,31 |
| Special datasets (2 sites) | 2,5 | 0,04 |
| Subtitles (1 site) | 101 | 1,66 |
| Total | 6091,5 | 100,00 |

These data includes materials, tagged by genre (3,4 billion words from fiction and 483 million words from poems), texts, represented with the time each line is pronounced (78 million words from subtitles), parallel corpus data (including Russian, English, German, Italian – subtitles also), texts tagged with text difficulty (1,7 million words from Special Media), key-words.
All of these resources are firstly brought "as is" to Russian community, except social media data, which is depersonalized and was published under CC license in 2017 during MorphoRuEval initiative (Sorokin et al. 2017) (plain texts) – this data is now included with annotation. Some of the news and fiction sites we have crawled also can be found in the search interface of General Internet Corpus of Russian, although there they are only available for searching examples.

We have gathered these segments with respect to popular NLP-problems:

1) thematic modeling - news with theme tags, all the sites which provide rubrication (news, poems, prose)
2) text classification and clusterization – by genre of fiction, rubrics of news
3) readability of texts – magazine texts with a readability metric for each text, provided by editor.
4) NER and fact extraction - news with references to mentioned person's page or wiki-information, news with personalia tags
5) key-words extraction - news with key-word tags, hashtags on social media
6) authorship attribution - all the texts with author information - magazines, news, fiction
7) chat-bot training - open-source film subtitles
8) text generation - any resource depending on genre
9) rare words studying, frequency dictionaries - literary magazines, social media
10) morphological and syntactic parsers - any resource with respect to the genre for language model creation

The "fiction" segment can be considered the most general one, and as is a few billions of words, it is suitable for word and ngram embeddings creation for general purposes for processing Russian. The

other sources are more specific and are recommended to use according to the problem – for example, word vectors obtained from poetry can be used for gathering only poetic synonyms, and will show completely different behavior than embeddings on subtitles.
In the next chapter you can find examples of solving such problems on the material of our corpus.

## *4. Corpus usage examples*

### *4.1 Embeddings*

Word embeddings are very useful for semantic similarity matching, text classification, etc. Yet a smaller share of attention is given to the comparison of embeddings obtained on homogeneous different corpora. Using RusVectores (Kutuzov, Andreev, 2015) interface one can see, that most common synonyms for 'роза' (rose) normally are 'цветок' (flower), 'орхидея' (orchid), 'лилия' (lily), etc. But using embeddings made on poetic subcorpora of Taiga (with fasttext, dimension 300, window - 5, 10 epochs) we can see, how poetic collocations of Russian behave differently from general word usage situation:

Table 2: Most similar words to 'rose' according to vector models on different corpora

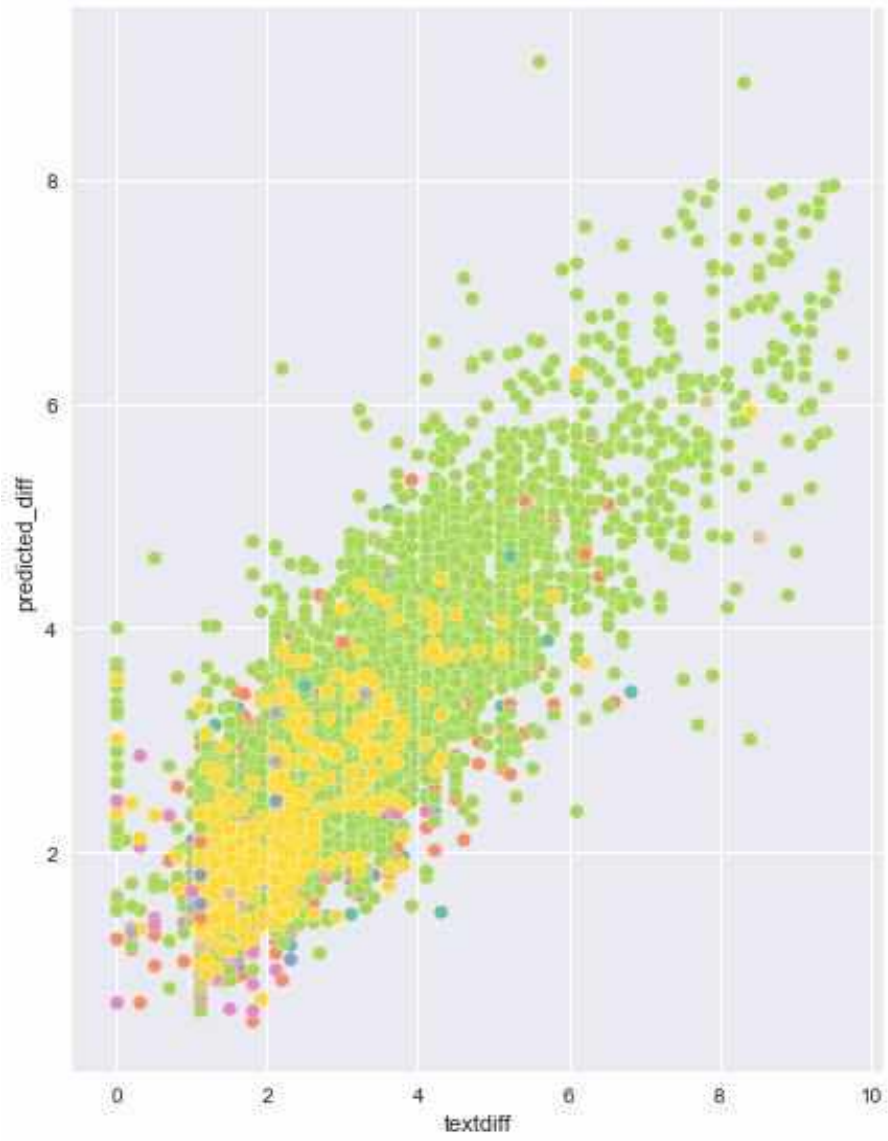
| **News** | | **Araneum fastText** | | **RNC** | | **Taiga poems** | |
|---|---|---|---|---|---|---|---|
| фрезия | freesia | орхидея | orchid | цветы | flowers | оброза | icons |
| гвоздика | carnation | гортензия | hydrangea | гиацинт | hyacinth | береза | birch |
| сябитова | syabit | фрезия | freesia | хризантема | chrysanthemum | проза | prose |
| хризантема | chrysanthemum | хризантема | chrysanthemum | лилия | lily | фроза | frost |
| ирис | iris | гербера | gerbera | маргаритка | daisy | эмброза | embrose |
| рымбаева | rymbayeva | маргаритка | daisy | нарцисс | narcissus | розалия | rosalia |
| гербера | gerbera | фиалка | violet | гелиотроп | heliotrope | гроза | storm |
| яснення | clarification | ирис | iris | букет | bouquet | лукроза | lucrose |
| сирень | lilac | лилия | lily | астр | astern | бяроза | birch |
| яснювальнуть | **error** | цветок | flower | цветок | flower | розали | Rosalie |

### *4.2 Text readability*

By training a simple Ridge regressor from sklearn on 7 000 texts with readability tags (from 0 to 10, continuous scale), we can train a model makes a readability estimation depending on words used in the text.

Testing the model gave a 1.05 RMSE, which is quite a good hit– see picture 1 - there is a linear relationship between the system estimation and the real annotation.

Picture 1 – distribution of text rubrics by their tagged readability (by X) and predicted readability (by Y). Rubrics colored: green – science, yellow – machine learning, orange – technologies, blue – robots, pink – gadgets, violet – transport




*4.3 Text generation*

Due to a great amount of fiction texts, annotated by genre, we can create a text generator almost of every genre. One of the genres, which is underrepresented compared to the others, is an anecdote. Still, I have made a simple HMM generator, which takes word forms as states and also treats sentence end and text end as state. Here are the examples generated: (in Russian, translations given on the right)

(Example 1)

| | |
|---|---|
| *Однажды, когда мне понадобилась ступенька, чтобы положить пришедший товар на стеллаж, я пошёл в фасовочную.* | *Once, when I needed a step to put the goods on the shelf, I went to the package room.* |
| *Там же был плакат "у Кавказа есть вопрос".* | *There was also a poster saying "the Caucasus has a question"* |
| *- Какой у вас красивый половичок.* | *- What a beautiful mat you have.* |
| *- Это не мой.* | *- It is not mine.* |
| *- Талантливый.* | *- Very talented!* |

(Example 2)

| | |
|---|---|
| *Кинорежиссер Никита Михалков сегодня впервые посетил в Екатеринбурге, который ранее резко раскритиковал в Совете Федерации, того что в нем нового - давно забытое старое .* | *Film director Nikita Mikhalkov today visited for the first time in Yekaterinburg, which previously sharply criticized the Council of the Federation of what was new in it – was the long-forgotten old.* |

*4.4 Constraints*

Users are informed, that corpus data is for personal and research purposes, as the open resources that are collected mostly allow their texts to be used for these purposes.

## *5. Corpus format and pipeline*

The corpus is stored in text format in UTF-8 encoding with all the relevant meta-information tags, duplicated as a sqlite database (for each segment). For each text, indent and paragraph structure is kept as in source. All the texts from each source separately have gone deduplication by URL, and are also filtered for non-UTF symbols, html-tags, non-breaking space, etc. by the BeautifulSoup Python package.

Each text can be found both as a plain text and with morphological and syntactic annotation – tagged by Udpipe parser (homonymy resolved) with the models trained on SynTagRus corpus (Boguslavsky 2014). See (3) for annotation example:

Example (3)

```
# newdoc

# newpar
# sent_id = 1
# 2003Armeniya.xml 1
# text = В советский период времени число ИТ- специалистов в Армении составляло около десяти тысяч.
# sent_id = 1
1	В	в	ADP	_	_	3	case	3:case	_
2	советский	советский	ADJ	_	Animacy=Inan|Case=Acc|Degree=Pos|Gender=Masc|Number=Sing
3	amod	3:amod	_
3	период	период	NOUN	_	Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing	11	obl	11:obl	_
4	времени	время	NOUN	_	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing	3	nmod	3:nmod	_
5	число	число	NOUN	_	Animacy=Inan|Case=Acc|Gender=Neut|Number=Sing	11	obj	11:obj	_
```

*6 ИТ ит PROPN _ Animacy=Inan|Case=Nom|Gender=Neut|Number=Sing 8 compound 8:compound SpaceAfter=No*
*7 - - PUNCT _ _ 6 punct 6:punct _*
*8 специалистов специалист NOUN _ Animacy=Anim|Case=Gen|Gender=Masc|Number=Plur 5 nmod 5:nmod _*
*9 в в ADP _ _ 10 case 10:case _*
*10 Армении армения PROPN _ Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing 5 nmod 5:nmod _*
*11 составляло составлять VERB _ Aspect=Imp|Gender=Neut|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act 0 root 0:root _*
*12 около около ADP _ _ 14 case 14:case _*
*13 десяти десять NUM _ Case=Gen 14 nummod 14:nummod _*
*14 тысяч тысяча NOUN _ Animacy=Inan|Case=Gen|Gender=Fem|Number=Plur 11 nsubj 11:nsubj SpaceAfter=No*
*15 . . PUNCT _ _ 14 punct 14:punct _*

Each text has a unique id and all meta-attributes, which can be derived for it, in a unified way: for example, a whole line of attributes by now consists of: corpus segment, text id, title, region, rubric, readability, author's name, author's texts amount, author's profession, author's readers amount, magazine, date, time, tags and URL. And the amount of the attributes and their distribution vary from source to source.

### *6. Results and discussion*

We have created a new corpus resource for machine learning, rich with text attributes, big and open-source. We hope that our work will be useful for Russian natural language processing and will help developing new tools and projects.

In the near future, the main goals are to create a community of users[1] and obtain feedback, bug reports, suggestions for new segments, etc. During winter 2017-2018 we have already collected a lot of new ideas and fruitful criticism, so that now we are presenting a 2.0 release of Taiga corpus, debugged and with much more fiction.

The second goal is to increase the volume of our corpus to 10 billion words through other resources and provide our users more datasets for easy training of the models, testing, organizing tracks, etc. We also acknowledge, that for some purposes there are still list of constraints:

- for example, for creation of differential vector-space models and word embeddings on special segments, our data is still not enough; obtaining data it is not solely a question of time - some of the open data sources on the web are not scalable up to the amount of words that are needed for some tasks, like texts with manual readability rating tags are very few of;
- for some purposes like POS-tagger training, it is not good enough as we have automatic POS and syntactic annotation; it is unclear whether there will be an opportunity in the near future to correct this situation.

We welcome linguistic community and developers to obtain our corpus and express their feedback and questions on pipeline, documentation, sources[2].

---

1 https://groups.google.com/forum/#!forum/taigacorpus

2 https://github.com/TatianaShavrina/taiga_site

**7. Acknowledgments**
Author expresses sincere thanks to Yana Kurmachova, who directly participates in the development of improved syntactic tagging in the corpus and contributed to the development of the project; to Olga Lyashevskaya, whose advice was always to inspire. I also express my respect for the previous group of students of NRU HSE computer linguistics, who have granted us the data on subtitles they collected from open sources.

***References***


Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation, 43*, 209–226.

Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S., (2013) Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In Web as Corpus Workshop (WAC-8).

Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S., (2013) Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In Web as Corpus Workshop (WAC-8).

Benko, Vladimir (2014) Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Ales Horak, Ivan Kopecek and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655.Springer International Publishing Switzerland, 2014. pp. 257-264. ISBN: 978-3-319-10815-5 (Print), 978-3-319-10816-2 (Online).

Boguslavsky, I. (2014). SynTagRus–a Deeply Annotated Corpus of Russian. In Blumenthal, P., Novakova, I., Siepmann, D. (eds.), *Les émotions dans le discours-Emotions in Discourse*, pages 367-380, Peter Lang.

Ginter, Filip; Hajič, Jan; Luotolahti, Juhani; et al., 2017, *CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1989.

Jakubicek, M., A. Kilgarriff, V. Kovar, P. Rychly, and V. Suchomel (2013) The TenTen corpus family. Lancaster: In Proc. Int. Conf. on Corpus Linguistics.

John Walker (2015) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Evaluating Non-Expert Annotations for Natural Language Tasks http://www.tandfonline.com/doi/abs/10.2501/IJA-33-1-181-183?journalCode=rina20

Kilgarriff, A. (2001). The Web as corpus. Proceedings of Corpus Linguistics 2001

Kutuzov, Andrey and Andreev, Igor (2015) Texts in, meaning out: neural language models in semantic similarity task for Russian. Proceedings of the Dialog 2015 Conference, Moscow, Russia

Lagutin M., Kuratov Y., Kopylov N. (2016) Statistical processing of Search results in differential corpora. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016” Moscow, June 1–4, 2016

Lyashevskaya, O., Droganova K., Zeman, D., Alexeeva, M., Gavrilova, T., Mustafina, N., Shakurova, E. (2016). Universal Dependencies for Russian: a New Syntactic Dependencies Tagset. In *Series: Linguistics, WP BRP 44/LNG/2016.*

Michele Banko and Eric Brill (2001) Scaling to Very Very Large Corpora for Natural Language Disambiguation. Microsoft Research, In Proc. of ACL-2001. http://aclweb.org/anthology/P/P01/P01-1005.pdf

Shavrina T., Shapovalova O. (2017) To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser. In proc. of "CORPORA2017", international conference , Saint-Petersbourg, 2017.

Snow, R., O’Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In M. Lapata & H. T. Ng (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 254–263). New York: ACM www.aclweb.org/anthology/D08-1027

Sorokin, A., Shavrina, T., Lyashevskaya, O., Bocharov, V., Alexeeva, S., Droganova, K., Fenogenova, A. (2017). MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2017*, Moscow.