Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2016”

Moscow, June 1–4, 2016

RUSSIAN MINORITY LANGUAGES ON THE WEB: DESCRIPTIVE STATISTICS

Orekhov B. (nevmenandr@gmail.com),
Krylova I. (krylova93@gmail.com),
Popov I. (imvanya@gmail.com),
Stepanova E. (stepanovayekaterina@gmail.com),
Zaydelman L. (luda.zaidelman@yandex.ru)

National Research University Higher School of Economics,
Moscow, Russia

The paper presents quantitative data about the web segments in minority languages of Russia. An ad-hoc search procedure allows to locate sites and pages on social networks that contain texts in a certain language of Russia. According to our data, there are texts in at least 48 of the examined languages on the Internet. We compared the gathered statistical data with the data from Wikipedia and the number of native speakers and found out that none of the “live” online data has a good correlation with the offline-life of language.

Keywords: minority languages, web as a resource, social networks, sociolinguistics

Acknowledgements: we thank Kirill Reshetnikov and Daria Ignatenko for collecting lexical markers for some of the languages and Dmitry Granovsky for his assistance in proofreading the text.

МИНОРИТАРНЫЕ ЯЗЫКИ РОССИИ В ИНТЕРНЕТЕ: ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Орехов Б. (nevmenandr@gmail.com),
Зайдельман Л. (luda.zaidelman@yandex.ru),
Крылова И. (krylova93@gmail.com),
Попов И. (imvanya@gmail.com),
Степанова Е. (stepanovayekaterina@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

В работе представлены количественные данные об интернет-сегментах на миноритарных языках России. Специальная технология поиска позволяет находить сайты и страницы в социальных сетях, на которых присутствуют тексты на одном из языков России. По нашим данным, в интернете присутствуют тексты по крайней мере на 48 языках из тех, которые мы обследовали. Мы сравнили собранные статистические данные с данными Википедии и числом носителей и обнаружили, что ни один параметр онлайн-данных не коррелирует с оффлайн-данными по языку (числом носителей).

Ключевые слова: миноритарные языки, малые языки, интернет как ресурс, социальные сети, социолингвистика

Благодарность: мы благодарим Кирилла Решетников и Дарью Игнатенко, которые собрали лексические маркеры для части языков, а также Дмитрия Грановского за помощь в работе над текстом.

1. Introduction

There are over a hundred national languages in Russia, excluding Russian and languages that are official in other countries. Some of these count more than a million native speakers. However, linguistic tools for all of these minority languages are equally scarce. The lack of tools—first and foremost, the corpora—results from the lack of digitized texts in said languages. The goal of our work is to form full collections of texts in Russian national languages, which can then be used to create text datasets (like annotated corpora, sets of n-grams and so on), and to count the number of sites where they are present.

We use several Internet sources to gather corpora—Yandex web-search as a word index store of the Internet, and the API of the most popular Russian social network Vkontakte to download texts from communities consisting of enthusiasts eager to speak a certain language.

2. Related works

Although papers dedicated to minority languages show a wide variety of approaches taken to the topic, they can generally be classified as quantitative vs. non-quantitative. Paper [4] presents a quantitative analysis of the Bashkir Internet, while such papers as [7], [8] display mostly qualitative analysis of Udmurt. There are also papers somewhere in between with little manual quantitative analysis [6].

There is an evident interest in investigating the Internet of minority languages, collecting corpora and developing NLP tools for such languages. In 2007 a huge project was launched by Scannell [9] for gathering texts of many minority and under-resourced languages. Even though are no particular corpora available, the website provides a lot of useful information about more than 2000 minority languages, including word n-grams and lists of urls to find texts on them.

Another topic worth mentioning is the contribution of modern technology to the well-being of minority languages. Mass media [12] argue that the Internet grants a new life to minority languages, anticipating an opportunity for the native speakers to talk to each other despite the distance. NLP researchers can also benefit from this as they gain access to new sources of texts in minority languages.

3. Methods

We use an almost automatic method for gathering corpora widely known as "seed words" and described in Wacky-papers [1], [2]. The method consists of seven stages, the first four of which were described in detail in [10]:

1. Search for lexical markers in grammars and phrasebooks.
   - Lexical markers are words that uniquely define the language that they belong to.

2. Search with Yandex.XML [15] for domains (websites) that contain the words obtained in stage 1.
   - We expect that texts found on pages in these domains are in the language to which the lexical marker belongs.

3. In the domains found in stage 2 find all pages that contain the lexical markers from stage 1 (send queries like "site: example.com 'marker'" to the Yandex search engine).

4. For each domain from stage 2 count the number of pages found in it on stage 3 and sort the domains into four groups: 1) download the whole domain, 2) download certain pages, 3) pages with files, 4) pages of social networks.
   - At the moment, we do not work with the third domain group.

5. Remove ambiguous domains from further processing (domains that were found several times using markers from different languages).

6. Download texts from the Vkontakte social network via its API [14], download texts from the Internet using a web-crawler (scrapy [13], Beautiful soup [11]).

7. For each text identify its language using the letter n-gram method.

Our methods have certain limitations. First of all, it is quite difficult to find a graphically unique word for a language. We gather and check lexical markers manually by running a search query that consists of a marker word and checking whether all found pages are in the expected language. However, considering that for some languages it might be very difficult to check every found page, there is no full confidence that several pages from the search result will not match another language.

Apart from that, we have query limits from the Yandex search engine—we can send only 1000 queries per day. As a result, the third stage of our method may take about 10–20 days for a widely spoken language with many lexical markers. The Yakut language with 450 140 natives and 22 lexical markers has the longest search history of 48 days. Some languages, on the other hand, can be searched in less than a day. The median search time over all 49 processed languages is 1,5 days.

Having collected all the urls, we also need to clean the lists and exclude, using a semi-automatic procedure, various music and video sites that contain only titles in the minority language. We also exclude the sites that were found by markers of different languages to consider them later.

When we download texts from a certain domain or a community, we already know the expected language of the obtained texts. However, the majority of these texts, especially when it comes to the texts from Vkontakte, are in fact in Russian or are useless and contain only links or pictures (see Fig. 2). Since our primary focus is compiling a collection of texts, we are only interested in texts in our target languages, so it is important for us to be able to identify the language of a text.

A common approach to identifying the language of a text is the letter n-gram method. The main idea of the method is to compare the list of the most frequent n-grams for the given text with the n-gram standards for the expected languages. The result of each comparison is a value called “distance” that represents the difference between the language of the text and the language with which it was compared. Once all the distances are calculated, the least one is determined and the corresponding language is considered to be the language of the text [3]. This method has proved to be very effective and quite accurate. However, the downside of the n-gram method is that in order to get the initial language standards it requires corpora, which, ironically, are exactly what we do not have.

Nevertheless, with a few modifications it is possible to apply the letter n-gram method to our problem. First of all, instead of creating n-gram standards for all languages, we only create a n-gram standard for Russian (more specifically, trigrams, which proved to provide the best results). Why do we assume that using a single standard would be enough? As already mentioned, we always know the expected language of our texts since they are downloaded from communities and domains already assigned to certain languages. As a result, there are only three options for each text: it may be written in the expected language, it may be written in Russian or it may be a “garbage” text, i.e. a text that consists entirely of Latin symbols (hyperlinks) or emoticons. Garbage texts can be easily identified and discarded by checking whether they contain cyrillic symbols. If a text contains cyrillic symbols and, therefore, is not garbage, we create its letter trigram standard and calculate the distance between Russian and the language of the text based on their standards. Finally, the distance value is compared with the empirically obtained threshold value, and in case the distance value is greater than the threshold, we can say that the text is not written in Russian or, in other words, is written in the expected language.

Unfortunately, it is very common for texts found on social networks to be quite short (like “спасибо” or “удачи”). These texts often cannot be correctly identified as Russian using just the letter n-gram method due to the skewed frequency data. Nevertheless, we were able to improve language identification by additionally comparing texts against the list of Russian word unigrams. The modified method proved to be more effective when dealing with short texts.

4. Results

We are aware of 96 different minor languages that exist in the Russian Federation. So far, we have found lexical markers and searched for 49 of them, and have got some preliminary results for about 40 of them.

4.1. Social networks

For 30 minor languages we were able to find and download at least one Vkontakte community. A total of 1735 communities were downloaded with 1633 of them containing at least one text, where by text we mean either a post on the community wall or a comment under a post. Fig. 1 shows the distribution of communities by language.

In Fig. 2 you can see the distribution of texts by language. Similarly to the previous figure, bar heights represent the total number of posts extracted from communities that “speak” a corresponding language. A base-10 logarithmic scale is used to account for an order-of-magnitude difference between the figures for some of the languages; the black section of each bar represents the fraction of the posts that are actually written in this language. Clearly, for all languages the majority of texts are not written in said languages, but, instead, are either in Russian or are “garbage” texts.

A question then arises of whether there is any functional relationship between the total number of texts in a community and the number of texts that are actually written in the language to which the community is assigned. Applying linear regression to the data reveals that this might be the case: with the linear regression coefficient at 0.383 (p-value $< 2.2e^{-16}$), we can say that the number of texts that are written in the target language is 2.5 times less than the total number of texts in the community (see fig. 3).

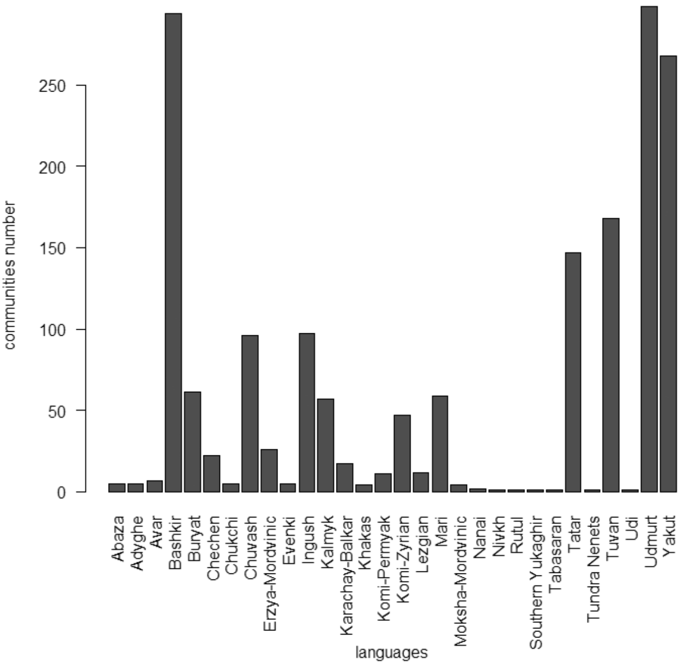

Fig. 1. Communities distribution by language

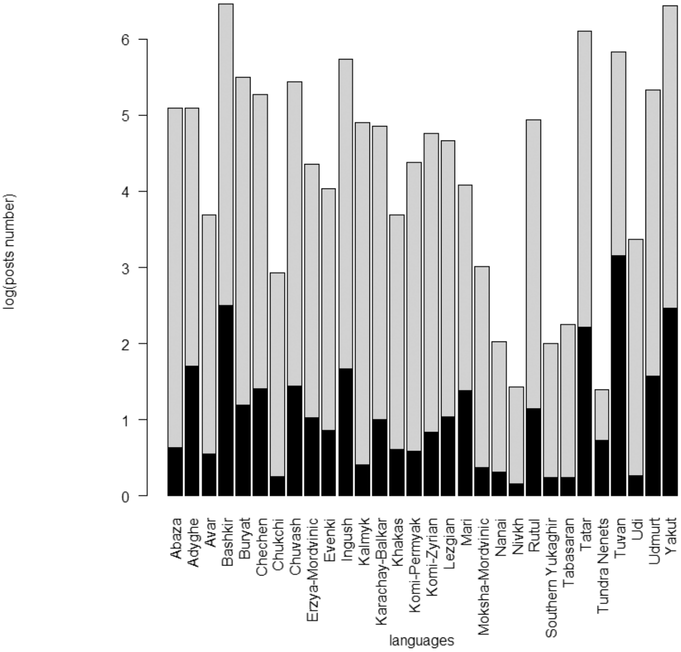

Fig. 2. Texts distribution by language

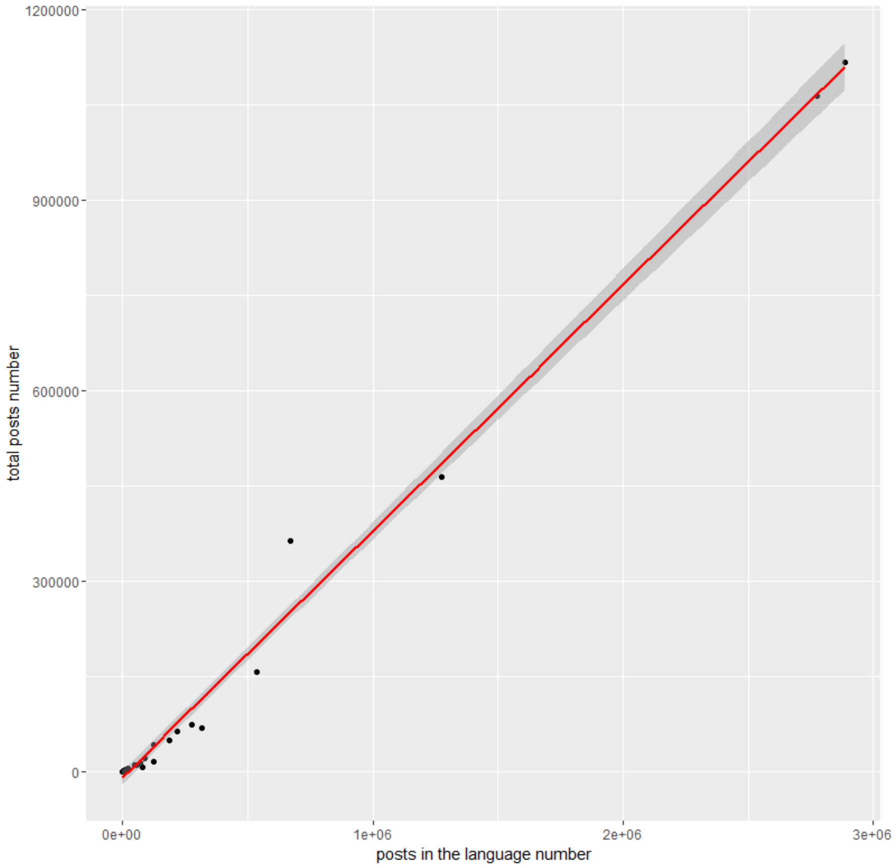

Fig. 3. Linear regression plot of total number of texts in communities attributed to a certain language vs. number of texts in said language

4.2. Internet

Currently, we have clean lists of urls (*example.com/page*) and domains (*example.com*) gathered for 49 minor languages. By clean lists we mean lists without non-informative sites such as music and video sites where a minor language might only be present in titles. It took us 239 search days to find 379 different "download-whole" language domains. We pay great attention to these domains, because, in our opinion, the number of such domains reflects language's self-sufficiency and its online development level.

In Table 1 you can see domain information for several groups of languages: the most active ones (Bashkir, Tatar and Yakut), the middle ones (Chuvash, Buryat, Kalmyk) and the least represented on the Internet (Rutul, Shor, Moksha-Mordvinic, Itelmen).

Table 1. Minor language domains information: ISO code, population, lexical markers found, domains found and days for search

ISO-639	Language	Natives	Markers	Domains	Days
ba	Bashkir	1,150,000	8	74	19
tt	Tatar	4,280,000	3	59	12
sah	Yakut	450,140	22	52	48
cv	Chuvash	1,152,404	5	20	8
bxr	Buryat	283,000	9	19	9
xal	Kalmyk	80,546	13	5	6
rut	Rutul	30,360	6	1	0.5
cjs	Shor	2,839	2	0	0.2
mdf	Moksha-Mordvinic	2,025	8	0	3.5
itl	Itelmen	7	2	0	0.5

Below you can see the histogram for all the processed minor languages. Please note that 19 languages with zero domains were excluded from the list for readability purpose. The exclude languages are: Aleut, Archi, Even, Forest Yukaghir, Gorno-Altai, Itelmen, Koryak, Kubachi-Ashtin, Mansi, Moksha-Mordvinic, Nanai, Nivkh, Nogai, Shor, Tabasaran, Tofa, Tsakhur, Tundra Yukaghir, Udi.

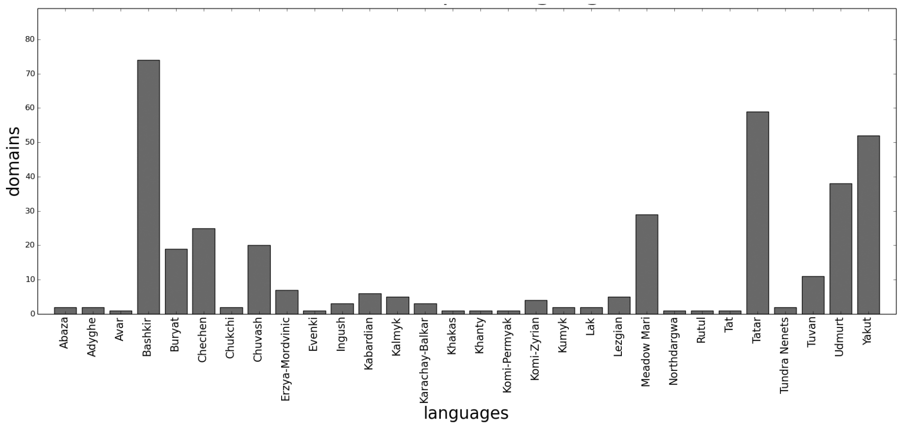

Fig. 4. Domains distribution by language

4.3. Comparison with Wikipedia data

When you set out to collect texts for a corpora, Wikipedia seems to be an obvious choice as a source for texts. Indeed, Wikipedia is basically a very large collection of texts, and many languages of Russia have their own language-specific Wikipedia. However, as was shown in [5], Wikipedia can sometimes be an inadequate representation of a language.

With actual data at our disposal, we were able to compare languages based on their parameters derived from the data. We created a dataset where rows corresponded to the languages and columns to the following parameters:

- the number of articles on the Wikipedia in the given language in 2014
- the number of articles on the Wikipedia in the given language in 2015
- the number of communities on Vkontakte in the given language
- the number of texts in this language in Vkontakte communities
- the number of tokens in this language in Vkontakte texts
- the number of whole-download domains for this language
- the number of webpages (urls) with at least a word in the minor language
- the number of tokens in this language on downloaded webpages
- the number of native speakers

We used this dataset to check if there were correlations between the parameters. We removed all the languages for which we do not have full information, e.g. for Bashkir we have not downloaded or processed texts from the web, and for Kumyk we have not processed Vkontakte communities. With such languages removed, there are 33 languages for which we have data both from Vkontakte and the web. We used this data as input to calculate linear correlations, the results are presented in Table 2, in which each cell contains a correlation coefficient for the corresponding parameters, and cells with values greater than 0.7 are grayed out.

Some of the correlations are rather obvious, e.g. the number of Wikipedia articles in 2014 vs. 2015 or the number of Vkontakte communities that "speak" a certain language vs. the number of texts in this language on Vkontakte, number of tokens on the web vs number of urls. However, the table also reveals something less trivial: neither the number of Vkontakte communities using a language nor the number of domains for this language correlates with the number of the language's native speakers. This means that the relationship between an actual language and its online representation is at least nonlinear, and that a language is used differently online.

Table 2. Pearson correlation coefficients for language parameters

	wiki 2014	wiki 2015	communities	posts in language	tokens on vk	domains number	urls	tokens in web	native speakers
wiki 2014	1	0.978	0.263	0.008	0.024	0.469	0.674	0.612	0.465
wiki 2015	0.978	1	0.376	0.131	0.142	0.592	0.668	0.660	0.617
communities	0.263	0.376	1	0.925	0.895	0.825	0.380	0.525	0.401
posts in language	0.008	0.131	0.925	1	0.982	0.714	0.064	0.346	0.299
tokens on vk	0.024	0.142	0.895	0.982	1	0.726	0.068	0.357	0.270
domains number	0.469	0.592	0.825	0.714	0.726	1	0.540	0.684	0.587
urls	0.674	0.668	0.380	0.064	0.068	0.540	1	0.765	0.270
tokens in web	0.612	0.660	0.525	0.346	0.357	0.684	0.765	1	0.470
native speakers	0.465	0.617	0.401	0.299	0.270	0.587	0.270	0.470	1

5. Conclusion and future plans

So far, we have briefly analyzed about 40 minor languages. We found out the most active languages in the Vkontakte social network, as well as the ones which are most represented on the Internet in general (based on the number of domains written predominantly in the minor language). We compared the gathered statistical data with the data from Wikipedia and the number of native speakers and found out that none of the "live" online data has a good correlation with the offline-life of language.

That means that, according to our data, the offline life of a language is completely different from its online existence. Obviously, representation of the language on the Web, affected by various factors, is not limited to the number of its speakers. We have found possible indicators of language activity on the Internet—the number of webpages and the number of communities on social networks, which demonstrate the degree of the web-vitality of a language. However, these assumptions require further research.

We plan to continue and deepen our analysis. As of today, we have already downloaded texts from all the domains and communities for 41 languages, tagged the downloaded texts with the language identification tool and counted the number of distinct webpages and posts per language, assessed the amount of minority-language text in them and counted token information for them. We plan to perform thematic analysis of websites and Vkontakte communities and compare our experience and results with the works of foreign minor language researchers and with those of social linguistics researchers.

All text collections and additional data are available on our website http://web-corpora.net/minorlangs/ (as of today, the website only has a Russian interface).

References

1. *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, Vol. 43(3), pp. 209–226.
2. *Boleda, G., Bott, S., Meza, R., Castillo, C., Badia, T., López, V.* (2006), CUCWeb: a Catalan corpus built from the Web, Proceedings of the 2nd International Workshop on Web as Corpus, Trento, pp. 19–26.
3. *Cavnar, W. B., Trenkle, J. M.* (1994), N-gram-based text categorization. Ann Arbor MI, Vol. 48113(2), pp. 161–175.
4. *Orekhov B. V., Gallyamov A. A.* (2013), Bashkir internet: lexis and pragmatics in a quantitative aspect [Bashkirskiy internet: leksika i pragmatika v kolichestvennom aspekte], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2012"], Bekasovo, pp. 502–509.

5. *Orekhov B. V., Reshetnikov K. Yu.* (2014), To the assessment of Wikipedia as a linguistic source [K otsenke Vikipedii kak lingvisticheskogo istochnika], Contemporary Russian on the Internet [Sovremennyy russkiy yazyk v internete], Moscow, Languages of slavic culture [Jazyki slavjanskoy kul'tury], pp. 310–321.
6. *Pischlöger C.* (2014), Notes from Murjol Underground: super udmurts in cyberspace [Zapis(k)i iz Murzhol Undeground: Super udmurty v Cyberspace], Proceedings of IV international science-practical conference "Florov's readings" [Trudy IV Mezhdunarodnoy nauchno-prakticheskoy konferentsii "Frolovskie chteniya"], Glazov, pp. 56–59.
7. *Sacharnych D. M.* (2006), National Udmurt Internet: what interferes with the development [Udmurtskiy natsional'nyy internet: chto meshaet razvitiyu?], available at: http://udmurt.info/pdf/texts/udmint.pdf
8. *Sakharnykh D. M.* (2008), New in national Udmurt Internet development: winter-autumn 2008 [Novoe v razvitii udmurtskogo natsional'nogo interneta: Zima-osen' 2008], available at: http://udmurt.info/pdf/texts/udmurtnet-zima-osenj-2008.pdf
9. *Scannell, K. P.* (2007), The Crúbadán Project: Corpus building for under-resourced languages. Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, Vol. 4, pp. 5–15.
10. *Zaydelman L., Krylova I., Orekhov B., Stepanova E.* (2015), Languages of Russia: Using Social Networks to Collect Texts, (in print) In Proceedings of the 9th Summer School in Information Retrieval and Young Scientist Conference (RuSSIR 2015)—Revised and Selected Papers, Communications in Computer and Information Science, St. Petersburg, Vol. XXX, pp. 1–8.
11. *Beautiful Soup*—http://www.crummy.com/software/BeautifulSoup/bs4/doc/
12. New technologies contribute to the preservation of minor languages [Novye tekhnologii sposobstvuyut sokhraneniyu malykh yazykov]—http://www.polit.ru/news/2014/01/15/ps_lang/
13. *Scrapy*—http://doc.scrapy.org/en/latest/
14. *VK API*—https://vk.com/dev/api_requests
15. *Yandex.XML*—https://tech.yandex.ru/xml/