Automatic Identification and Classification of Ambiguous Toponyms *

Kirill Boyarsky [1]
Boyarin9@yandex.ru

Eugeny Kanevsky [2]
eak300@mail.ru

Daria Butorina [1]
daanareevna20@gmail.com

[1] ITMO University
[2] Institute for Regional Economy Studies RAS
Saint Petersburg, Russian Federation

Abstract

Currently, various versions of TextMining technology are widely used, among other tasks, to allow retrieving various information from documents. In particular, the information can refer to certain named entities: personalities, geographical objects, etc. In this case, there inevitably arise problems associated with the ambiguity of names. The paper studies the potentialities of disambiguation for Russian texts of political, artistic, and highly specialized nature using the toponyms. We identified words that form collocations with toponyms (marker words). It is shown that the use of short lists of such words can significantly increase the reliability of identifying a toponym and determining its type.

Keywords: *toponyms, ambiguity, text analysis, parser, marker words*

1 Introduction

The mass media produce daily a plethora of data, in particular, the news that describe various events taking place in one or another region of the globe. More often than not, of importance in regional news are the so-called **named entities**. The discovery of named entities is a key problem of information retrieval (structured data) from unstructured or semistructured documents [Starostin et al, 2016]. Its main point is to find the names or identifiers of objects of a certain type in the text. For the first time ever, the problem was stated as far back as 1996 at the Message Understanding Conference, where the entities under consideration were organizations, places, people, and some numeric expressions. Later, it was examined

at the conferences on computational natural language learning (CoNLL) CoNLL-2002 and CoNLL-2003 [Tjong, Sang and De Meulder, 2003].

The **named entity** is today understood to be an object of a certain type, which has a name or an identifier. Which types (classes) the system distinguishes is defined within a particular task. Typical for news texts are PERSON, ORGANIZATION, LOCATION, and MISCELLANEOUS (numeric expressions, dates, events, slogans etc.). In a few cases, the number of classes may be greater: studies are known to involve 29 [Brunstein, 2002] or even 200 [Shinyama and Sekine, 2004] of those, including, among others, food, books, events, animals, plants, airports, and so on.

The discovery of named entities can be used for text classification, sentiment analysis, paraphrase identification etc. The importance of the problems (as well as the complexity of solving them) is substantiated by holding regular contests of the corresponding software solutions [Starostin et al, 2016, Loukachevitch and Rubtsova, 2016, Panchenko et al, 2018].

Of great importance in solving the problems is the correct extraction from the texts of those geographical entity names (toponyms), to which the given message may bear relation. As noted in [Lieberman snd Samet, 2011], the toponym retrieval problem is rather involved. Because of the homonymy, one and the same word may designate a geographical entity or that completely unrelated to toponyms.

These particular features cause the analysis to use more accurate models than just a "bag of words", where the word order in the text is disregarded altogether [Barsegyan et al, 2007]. A method of the kind is the construction of a subordination tree [Testelec, 2001]. Yet, even following this way, one has to resolve issues arising in connection with the high degree of homonymity of Russian texts.

2 Problem Statement

The means used to analyze texts is parser SemSin [Boyarsky and Kanevsky, 2015] that performs the in-depth syntactic analysis of Russian texts. It employs the expanded version of the dictionary by V. A. Tuzov [Tuzov, 2004], its size exceeding 194,000 lexemes (about 170,000 words). The classifier has been expanded to 1,700 classes. The text analysis proceeds under control of production rules [Boyarsky and Kanevsky, 2011a, Boyarsky and Kanevsky, 2011b]. Accomplished along with the syntactic analysis of a sentence are the grammar and POS disambiguation, phrase segmentation, the construction of syntactic dependency tree, and, fairly often, the semantic disambiguation as well.

The parser allows adjusting the semantics to a certain subject domain. To do this, it suffices to pretrain the parser: process several texts for the selected subject domain and identify a number of homonymic lexemes that distort the parsing semantics. The lexemes along with their classes are placed into a special file, which ensures the automatic elimination in all subsequent analysis sessions. As evidenced in practice, the number of such lexemes is not large: selecting 35 lexemes to adjust the parser to analyze texts from a rather specific domain of historical shipbuilding proved to be enough [Artemova et al., 2015].

The parser operation result, presented as an xml file, contains the lemma and complete information on grammatic (part of speech, gender, number, case . . .), syntactic (relation types), and semantic (classifier class) properties of words in sentences. The data are a "semi-finished product" of a kind and can be used for further analysis and retrieval of any information of interest. This is why making the ambiguity as low as possible as well as raising the parsing

accuracy is important.

The present work aims to study the feasibility to establish the following:

– is a capitalized word a toponym?

– which toponym does the word belong to?

– which syntactic relation connects the word to the subordination tree?

When analyzing toponyms, it must be borne in mind that no reference book is capable of covering all geographic names. So, one should be ready to encounter an unknown name in the text. It can be a word either absent in the dictionary at all or having another meaning. The present work disregards the situation when a toponym coincides with a common vocabulary word (*Belukha* [the mountain] and *Belukha* [a white porpoise], *Tigr* [the river Tigris] and *tigr*[a tiger]). It typically suffices in such case to carry an elementary graphematic analysis as to the word capitalization.

In addition, no task is set to find the geographic location of an object. Apparently, the problem in the general case has no solution without a detailed analysis of large text fragments. Now, there is *Verblyud* [camel] mountain both in Pyatigorsk and in Kamchatka, while there are at least three *Chyornaya rechka* [black river] within 30 km distance from St. Petersburg. In any case, this is not a task for a parser.

About 90 classes in the semantic classifier we use contain words that are proper names, with 30 of them being allocated for identities of toponyms. Table 1 lists merged classes that contain toponyms or proper names, homonymic to toponyms, and the number of lexemes in the classes.

Table 1: Semantic classes of proper names

Semantic classes	Number of lexemes
Terrestrial natural objects (mountains, deserts, islands...)	700
Natural water bodies (rivers, sees, gulfs...)	1200
«Cities» — settlements (cities, towns, streets...)	4600
«Counties» — administrative units (countries, regions, provinces, states...)	800
Astronomical objects (stars, constellations, planets)	140
Last names	13700
First names	4400
Mythological creatures	120
Institutions (industry, research, sports...)	2700
Enginery (automobiles, ships, weapons...)	300

When parsing, an attempt was made to establish uniquely, if possible, to which class the entity named with a capitalized word belongs.

To assess the parsing quality, a comparison with the performance of parser ETAP-4 was undertaken [ETAP-4, 2019].

3 Semantic homonymy

The text analysis starts, before anything else, with preprocessing, when the text is split into sentences and words and the morphological properties of words are found. Then, the text is, word by word, input to the syntactic-semantic analyzer. The module of rules for processing toponyms is enabled upon detection of capitalized words. Further on, the word will be called a **target** one.

Termed as semantic homonymy is the situation when the target word names objects of different nature: settlements, rivers, islands, planets etc. Some cases of semantic homonymy are presented in Table 2. The selection used the Great Encyclopedic Dictionary [BES, 1997]. The coincidences of names are grouped into conditional classes with somewhat greater degree of detail than in Table 1.

An asterisk in Table 2 marks the cases when not more than 10 coincident names fall into two classes, two asterisks—from 11 to 20, and three asterisks—more than 20.

Note that one and the same word may mean both a toponym and a "non-geographical" entity (company name, first/last personal name etc.), with graphematic analysis yielding nothing of use here.

Table 2. Toponym homonymy

	Mountain	City	Gulf	Star	Lake	Island	Planet	Region	River	Country	County	Last name	First name	Institution	Misc
Mountain								*	*		*		*	*	*
City			*	*	*	**		*	***	**	**	***	**	**	**
Gulf						*					*	*			*
Star										*		*	*	*	*
Lake						*				*		*	*	*	*
Island										**	*	*	*	*	*
Planet													*	*	*
Region														*	*
River										*	**	*	**	**	*
Country												*	*	*	*
County												*	*	*	*

The most frequent situations are those, when a settlement name coincides with a last personal name, there are more than 100 such cases. Here, the city may be named after someone (*Kirov, Korolyov, Houston*), or there can be a mere coincidence (*London, Khotyn*). The coincidence city—first name is typically of a random nature (*Lyon, Olympia, Milan*). The coincidences of city names with those of rivers are, as a rule, related to the geographical location of settlements (*Aldan, Kabul, Narva*).

There are many cases of coinciding names of islands and island states situated on them (*Haiti, Ireland, Cuba*). The same goes for a number of city names (*Dikson, Kodiak, Zanzibar*), though the city of Vancouver, e. g., is not situated on the Vancouver island.

To find the target word semantics, the closer context is analyzed.

4 Abbreviations

First of all, the presence is checked of a left-contiguous abbreviation, which may mean a geographical entity (*g.* [mountain or sity], *o.* [island], *oz.* [lake], *pos.* [settlement], *r.* [river]), an address (*per.* [alley], *pl.* [square], *pr./prosp.* [avenue], *ul.* [street]), or a person (*g.* [Mr.], *gg.* [Messrs.]). If an abbreviation for a geographical entity is detected, the capitalized word is thought a toponym. For instance, *r. Volga* [the Volga river] will be expressly a river rather than a car brand or a sports club.

As typical, the situation is complicated by homonymy, this time, of abbreviations. The abbreviation "*o.*" is easy enough to differentiate from a preposition (this has no dot) and from an initial (this is capitalized), leaving the options "*ostrov*" [island] or "*otets*" [father]. Parser ETAP-4 chooses "father" in all cases, even in the sentence *Ya posetil o. Madagaskar* [I visited the Magadascar island]. Our parser treats "*o.*" as an island, the target word takes on the island class independently of the presence in the dictionary and its dictionary meaning. For instance, *o. Sumatra* [the Sumatra island] (present in the dictionary as an island), *o. Vozneseniya* [the Ascension island] (present in the dictionary, but with a different meaning), *o. Baratang* [the Baratang island] (absent from the dictionary). A problem remains how to interpret the abbreviation in combinations of *o. Ivan* type. It turns out that the islands named after masculine names are quite few, so one can reasonably assume "*o.*" to have meaning "*otets*" [father] in that case only, if the target word is a dictionary word of the masculine gender.

Even more confused is the situation with the abbreviation "*g.*", which may mean "*gora*" [mountain], "*gorod*" [city], "*gospodin*" [mister], "*god*" [year]. ETAP-4 produces interpretation "*gorod*" [city] for target words, present in the dictionary, and "*god*" [year]—otherwise. Upon detection of the left-contiguous abbreviation "*g.*" for the target word, our system goes on analyzing the left context, and the interpretation "*gorod*" [city] is selected only if a numerical token is detected. The interpretation in the rest of cases is decided by the target word class: "*g.*" in the combination *g. Novgorod* means "*gorod*" [city] and takes on the settlement class, while "*g.*" in the combination *g. Everest* is already "*gora*" [mountain]. All options are retained for words that are absent from the dictionary. In individual cases, one succeeds to refine the class during later parsing.

5 Determiners

Determiners will be understood as the words that directly name a toponym class: river, island, city etc. It should be noted that the number and case of a determiner is exactly the same as those of a toponym.

If there is a determiner and the toponym is in the dictionary, the semantic disambiguation presents no difficulty. So, *Baikal* in combination *ozero Baikal* [the Baikal lake] is the lake rather than the drink; the relation is "Name" (its counterpart in ETAP-4 is the appositive relation).

If there is a determiner, but the target word is not in the dictionary as a toponym (*Ya priyekhal na ostrov Zub* [I came to the Zub island]), the parsing is correct both in SemSin and ETAP-4. Note that the subordination tree, constructed in ETAP-4, has no semantic labels, so that the correctness of the semantic class cannot be verified. If the word is absent from the dictionary altogether (*ostrov Aogasima* [the Aogashima island])—the class and relation type

settings in SemSin are correct, while ETAP-4 yields the relation "quasi-agent", i. e. fails to identify the word as a name.

If the toponym is in the dictionary, but there is no determiner, the semantic disambiguation is attended with great difficulties that one succeeds to overcome in individual cases only. Consider the sentence *Ya priyekhal na Lenu* [I came to the Lena]. Upon the morphological parsing, the word form *Lenu* has four meanings: the dative of lexeme *Len* (either a land allotment, or an administrative unit in Sweden) and the accusative of lexeme *Lena* (either a river, or a personal name). The two first options are discarded by the results of the graphematic analysis (they should have been capitalized), the last two remain.

In the course of construction of the full subordination tree, however, the analysis of the semantic classes that are required by the verb "*priyekhal*" [to come] allows disambiguation of the homonymy, leaving the river name only. ETAP-4 yields label "animate" for lexeme *Lena*. Understandably, both options remain in the sentence *Ya uvidel Lenu* [I saw Lena].

6 Marker Words

In some cases when the determiner is absent, one can try to ascertain the semantics of a proper name by marker words. The latter will be understood as the words often used with toponyms of a certain type. For instance, such words can be *techeniye* [flow], *ruslo* [bed], *ust'ye* [mouth] for rivers, *vershina* [summit], *sklon* [slope] for mountains, *bereg* [shore], *ostrov* [island] for lakes and so on. A feature of marker words is that they augment a proper name using the genitive (this is the relation "quasi-agent" in ETAP-4) rather than as a name.

We have studied the issue how accurately this or that marker word allows unique identification of the semantic class of a toponym. The results of the expert analysis of several thousands of sentences are presented in Tables 3–5.

Table 3. Marker words for rivers

Lexeme	Total in RNC	Capitalized word on the right	Selected for analysis	Percentage of co-occurrence with a river name
Izluchina [Loop]	543	118	124	100 %
Nizov'e [Lower reach]	704	391	229	100 %
Mezhdurech'e [Interstream]	211	81	95	99 %
Pritok [Tributary]	3534	941	480	99 %
Verhov'e [Upper reach]	1537	671	603	98 %
Ust'ye [Mouth]	6020	2242	743	97 %
Del'ta [Delta]	905	291	243	90 %
Pojma [Flood land]	2473	103	70	84 %
Istok [Waterhead]	3677	490	180	83 %
Ruslo [Bed]	4567	239	104	78 %
Techeniye [Flow]	51572	1225	224	52 %
Vodorazdel [Watershed]	658	53	50	44 %
Bassejn [Basin]	7137	820	422	38 %
Bereg [Bank]	75107	14275	770	23 %
Voda [Water]	156992	3185	280	22 %

It shall be noted that our study used much greater number of candidates to marker words, many of which are out of the tables because of their excessive ambiguousness and low percentage of co-occurrence with the corresponding toponym class. So, Table 3 contains no words *glu-*

bina [depth], *dlina* [length], *dno* [bottom], *naberezhnaya* [quay], *omut* [whirlpool], *stremnina* [quickwater], and *shirina* [width], though they are occasionally found in front of a river name (. . . *stol' nepokhozhem ni na Ligovku, ni na naberezhnyie Nevy* [. . . so unlike either *Ligovka,* or the *Neva* quays]). Table 4 contains no word naberezhnyie [quays], though it may stand in front of a city name (*Mazankovyie i kamennyie dvortsy zapolnili naberezhnyie Peterburga* [wattle and daub houses and stone palaces filled the quays of St. Petersburg]), while Table 5 contains no words *dolina* [valley], *obryv* [precipice], *peshchera* [cave], *pik* [pinnacle], *poverkhnost'* [surface], *ushchel'e* [ravine], *khrebet* [ridge] (*Prakticheski vse gornyie ushchel'ya Issyk-Kulya prigodny dlya trekov* [Practically all mountain gorges of Issyk Kul are fit for trekking]). First, selected as marker words were those of class "landscape", which often occur together with names of rivers, cities, and mountains. Next, the Russian National Corpus (RNC) [NKRYa, 2019] was sampled under an additional condition of the right-contiguous capitalized word (a candidate to toponyms). Further, eliminated were the cases when there was a punctuation mark between the marker and the candidate. The resulting list was appended with sentences, retrieved under the same rules from news, literary, and sports texts of total size 55 mln. of words.

Consider the example: *On vozglavil ekspeditsiyu v Man'zhuriyu s tsel'yu izyskaniya tropy, kotoraya spryamlyala by izluchinu Amura* [He headed an expedition to Manchuria for the purpose of finding the path, which would straighten the Amur loop]. The word *Amur* [Amur] as a proper name has at least five meanings: a river, a planet, a firm, a sporting club, a god. With consideration of the marker word *izluchina* [loop], a single meaning is left, a river.

As seen from Table 3, there is probability 95 % and greater for the co-occurrence of some markers with river names, the percentage for others being much lower. The latter include, e. g., the word *vodorazdel* [watershed]. The word in Russian has a very versatile semantics, and, as the table shows, there are just 44 % cases of co-occurrence with a river name. The marker occurs along with other toponyms and proper names in 50 % of cases:

- *Akademik Gmelin. . . opredelil vodorazdel Indiyskogo i Ledovitogo okeanov.* [Academician Gmelin. . . identified the watershed of the Indian and Arctic oceans]
- *Topograf . . . zasnyal vodorazdel Ural'skogo khrebta.* [The surveyor mapped the watershed of the Ural range]
- *"Pamyatnik" zastyl vodorazdelom Pushkina myortvogo i Pushkina zhivogo v nashem soznanii.* ["The Monument" froze as a watershed of dead Pushkin and alive Pushkin in our mind]

However, given that practically all names of mountain systems and sees are found in dictionaries, one can confidently predict that the words "*Baranikhi*" and "*Pogindeny*" in the sentence . . . *Dolinoy reki spustilis' vniz i v neskol'kikh verstakh ot khrebta, sluzhashchego vodorazdelom Baranikhi i Pogindeny, ostanovilis' na nochleg* [We came down by the river valley and put up for the night in a few miles from the ridge that served as a watershed of Baranikhi and Pogindeny] are river names. In the course of parsing, this allows predicting the semantics of the capitalized words that are not in the dictionary.

The word *bereg* [bank, shore] occurs typically with names of rivers, lakes, islands, sees, countries, regions. For this reason, a river name rather than a personal one is univocally left in the sentence *Lager' razbili na beregu Yany* [The camp was pitched on the Yana bank]. The parsing can be luckily refined in some cases by expanding the context and analyzing the

adjectives, left-contiguous to the marker word. Thus, one succeeds to distinguish *mineral'nyie vody* [mineral waters] from *vody Volgi* [Volga waters], and recognize that what is meant in the sentence *Gorod vyros na oboikh beregakh El'by* [The city grew on both Elba banks] is a river rather than an island. Typical for the marker bereg in the meaning "river bank" are also the adjectives l*evyi* [left] and *pravyi* [right]; for the marker *basseyn* in the meaning "swimming pool"—*detskii* [wading], *krytyi* [indoor], *otkrytyi* [outdoor], *plavatel'nyi* [swimming], and *sobstvennyi* [private]; for the marker *techeniye* [flow] in the meaning "river flow"—*verkhniy* [upper], *nizhniy* [lower], *sredniy* [middle].

Table 4. Marker words for cities

Lexeme	Total in RNC	Capitalized word on the right	Selected for analysis	Percentage of co-occurrence with a city name
Predmest'e [Outskirt]	1375	455	274	86 %
Prigorod [Suburb]	1458	493	454	94 %
Centr [Downtown]	51990	6377	476	41 %
Plural only				
Kvartaly [Blocks]	2067	180	224	74 %
Bul'vary [Boulevards]	1203	74	46	56 %
Ulitsy [Streets]	37093	3599	498	55 %
Vokzaly [Terminals]	1156	144	39	46 %
ZHiteli [Inhabitants]	27593	4607	520	46 %
Parki [Parks]	2319	144	176	43 %
Pereulki [Alleys]	2424	146	176	43 %
Prospekty [Avenues]	609	50	38	37 %

It shall be noted that even here the parsing can be refined in some cases by expanding the context and analyzing the adjectives, left-contiguous to the marker words *prigorod* [suburb] or *predmest'e* [outskirt]. If the words are preceded by an adjective derived from a city name, the target word is a name of a suburb or an outskirt as such: *Tak, v pechal'no izvestnom svoyey tyur'moy bagdadskom prigorode Abu-Graib vplot' do posledney voyny v Irake raspolagalsya natsional'nyi bank semyan* [So, Bagdad suburb Abu Ghraib, notorious for its prison, hosted the National Seedbank continuing until the last war in Iraq] (suburb name) vs *Po etomu povodu v prigorode Khabarovska sostoyalas' ofitsial'naya tseremoniya* [This was the occasion for an official ceremony to be held in a suburb of Khabarovsk] (relation by the genitive). In both cases, ETAP-4 yields the relation "quasi-agent". In some cases, one succeeds even to disambiguate the toponym semantically: *Ni doma, ni mashiny zdes' v prigorode Vankuvera pochti nye zapirayutsya* [Here, in the Vancouver suburb, neither houses, nor cars are hardly ever locked] (Vankuver [Vancouver] is a city and an island, the analysis leaves a city).

Table 5. Marker words for mountains

Lexeme	Total in RNC	Capitalized word on the right	Selected for analysis	Percentage of co-occurrence with a mountain name
Vershina [Peak]	17801	1739	500	62 %
Sklon [Slope]	8796	739	395	81 %

Iz izvestnyakov pochti polnost'yu slozheny zapadnyie sklony Urala [The western slopes of

the Urals are almost completely composed of limestone] (Ural is a river, mountains, region, and a sporting club; the analysis leaves mountains).

In some cases, the correct classification of a toponym is quite difficult. So, the river names *Ona* and *Talaya* in the sentence *Za nim v istokakh Ony and Taloy nachinalas' yego tayga* [Beyond it, in the waterheads of Ona and Talaya, his taiga started] are detected by neither our parser, nor ETAP-4.

7 Conclusion

The analysis of the usage of toponyms in Russian texts shows that the degree of their homonymity is rather high. Automatic classification of toponyms is complicated also by the existence of a vast number of geographical names that are absent from the common vocabulary.

To resolve the problem, it has been proposed to use marker words, which allow inferring the presence of a toponym circumstantially, in addition to the determiners that indicate the toponym class directly. So, the markers *izluchina* [loop], *nizov'ye* [lower reach], *mezhdurech'ye* [interstream], *pritok* [tributary], *verkhovye* [upper reach], and *ust'ye* [mouth] precede a river name in the overwhelming majority of cases. Their presence secures the accuracy of at least 95 % of interpreting a word that is absent from the dictionary, but is capitalized, as a river name.

It has been noted that even recognition of a capitalized word as a toponym fails often to determine the type of a geographical entity unequivocally (Table 2). "Microdictionaries" of marker words have been compiled for toponyms of the most homonymic classes: rivers, cities, and mountains. Their use has been shownto allow raising the accuracy of determination of toponyms and, in some cases, to allow establishing their class.

The results hold out the hope that marker words are capable of disambiguating not only toponyms, but also other semantic classes. For instance, serving as markers for last names could be such words as *mneniye* [opinion], *vizit* [visit], *otstavka* [retirement], and so on. Compilation of lists of such words could be the subject of further research.

References

[Starostin et al, 2016] Starostin A. S. et al (2016) Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian / Starostin A. S., Bocharov V. V., Alexeeva S. V., Bodrova A. A., Chuchunkov A. S., Dzhumaev S. S., Efimenko I. V., Granovsky D. V., Khoroshevsky V. F., Krylova I. V., Nikolaeva M. A., Smurov I. M., Toldova S. Y. // FactRuEval 2016: Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 1–4 iyulya 2016 g.). Vyp. 15 (22).—M.: Izd-vo RGGU.

[Tjong, Sang and De Meulder, 2003] Erik F. Tjong, Kim Sang and Fien De Meulder (2003) Introduction to the conll-2003 shared task: language-independent named entity recognition. // In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, CONLL '03, Stroudsburg, PA, USA, 2003. P. 142—147.

[Brunstein, 2002] Brunstein A (2002) Annotation Guidelines for Answer Types // BBN technologies (2002). URL: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.

[Shinyama and Sekine, 2004] Yusuke Shinyama and Satoshi Sekine (2004) Named entity discovery using comparable news articles. // In Proceedings of the 20th international conference on Computational Linguistics, COLING '04, Stroudsburg, PA, USA, 2004. P. 848–853.

[Loukachevitch and Rubtsova, 2016] Loukachevitch N. V., Rubtsova Y. V (2016) SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis. // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 1–4 iyulya 2016 g.). Vyp. 15 (22).—M.: Izd-vo RGGU, 2016. S. 416-426.

[Panchenko A. et al, 2018] Panchenko A. et al (2018) RUSSE2018: a Shared Task on Word Sense Induction for the Russian Language / Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Arefyev N., Leontyev A., Loukachevitch N. // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 30 maya–2 iyunya 2018 g.). Vyp. 17 (24), 2018. C. 546-564.

[Lieberman snd Samet, 2011] Michael D. Lieberman, Hanan Samet (2011) Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11. Beijing, China, 2011. P. 1—36.

[[Barsegyan et al, 2007] Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Holod I.I. (2007) Tekhnologii analiza dannyh: Data Mining, Visual Mining, Text Mining, OLAP. 2-e izd., pererab. i dop. SPb.: BHV-Peterburg, 2007. (In Russian).

[Testelec, 2001] Testelec YA.G. (2001) Vvedenie v obshchij sintaksis. M.: RGGU, 2001. (In Russian).

[Boyarsky and Kanevsky, 2015] Boyarsky K.K., Kanevsky E.A. (2015) Semantiko-sintaksicheskiy parser SEMSIN // Nauchno-tekhnicheskiy vestnik informatsionnyh tekhnologiy, mekhaniki i optiki. 2015, T. 15, 5. – S. 869-–876. (In Russian).

[Tuzov, 2004] Tuzov V.A. (2004) Komp'yuternaya semantika russkogo yazyka. SPb. Izd-vo S.-Peterb. un-ta, 2004. (In Russian). “”

[Boyarsky and Kanevsky, 2011a] Boyarsky K.K., Kanevsky E.A. (2011) Yazyk pravil dlya postroeniya sintaksicheskogo dereva // Internet i sovremennoe obschestvo: Materialy XIV Vserossiyskoy ob"edinennoy konferentsii "Internet i sovremennoe obschestvo". – SPb.: OOO "Mul'tiProdzhektSistemServis", 2011. S. 233–237. (In Russian).

[Boyarsky and Kanevsky, 2011b] Boyarsky K.K., Kanevsky E.A. (2011) Sistema produktsionnyh pravil dlya postroeniya sintaksicheskogo dereva predlozheniya. Prikladna lingvistika ta lingvistichni tekhnologii: MegaLing-2011. K.: Dovira, 2012. S. 73–80. (In Russian).

[Artemova et al., 2015] Artemova G. et al. (2015) Text Categorization for Generation of Historical Shipbuilding Ontology / G. Artemova, K. Boyarsky, D. Gouzévitch, N. Gusarova, N. Dobrenko, E. Kanevsky, D. Petrova. // Communications in Computer and Information Science, 2015, v. 468. P. 1—14.

[ETAP-4, 2019] Lingvisticheskiy protsessor ETAP-4 [Elektronnyy resurs] // URL: http://www.http://proling.iitp.ru/ru/etap4 (data obrascheniya: 22.02.2019). (In Russian).

[BES, 1997] Bol'shoy entsiklopedicheskiy slovar' (1997) / Gl. red. A.M. Prohorov. – 2-e izd., pererab. i dop. – M.: Bol'shaya Rossiyskaya Entsikl., 1997. (In Russian).

[NKRYa, 2019] Natsional'nyy corpus russkogo yazyka [Elektronnyy resurs] //URL:http://www.ruscorpora.ru/ (data obrascheniya: 19.01.2019). (In Russian).