Microsyntactic Phenomena as a Computational Linguistics Issue

Leonid Iomdin
A.A.Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia
iomdin@gmail.com

Abstract

Microsyntactic linguistic units, such as syntactic idioms and non-standard syntactic constructions, are poorly represented in linguistic resources, mostly because the former are elements occupying an intermediate position between the lexicon and the grammar and the latter are too specific to be routinely tackled by general grammars. Consequently, many such units produce substantial gaps in systems intended to solve sophisticated computational linguistics tasks, such as parsing, deep semantic analysis, question answering, machine translation, or text generation. They also present obstacles for applying advanced techniques to these tasks, such as machine learning. The paper discusses an approach aimed at bridging such gaps, focusing on the development of monolingual and multilingual corpora where microsyntactic units are to be tagged.

1 Introduction

This work is largely based on the theory of microsyntax developed by the author over the last 15 years (see e.g. L.L.Iomdin 2013, 2014, 2015). In this theory, which has much in common with construction grammar (Fillmore 1988, Goldberg 1995, Rakhilina 2010) two main groups of linguistic units are distinguished: lexically oriented syntactic idioms and lexically independent non-standard syntactic constructions. Throughout the paper, I will be mostly concerned with syntactic idioms, which means that the concepts "microsyntactic units" and "syntactic idioms" are basically synonymous.

Microsyntactic linguistic units, such as syntactic idioms and non-standard syntactic constructions,[1] are poorly represented in linguistic resources, mostly because the former occupy an intermediate position between the lexicon and the grammar and the latter are too specific to be routinely tackled by general grammars. Consequently, many such units produce substantial gaps in systems intended to solve sophisticated computational linguistics tasks, such as parsing, deep semantic analysis, question answering, machine translation, or text generation. They also present obstacles for applying advanced techniques to these tasks, such as machine learning. I will discuss an approach aimed at bridging such gaps, focusing on the development of monolingual and multilingual corpora where microsyntactic units are to be tagged.

One of the difficult issues in dealing with microsyntactic units is the fact that, in many cases, they are extremely difficult to identify, discriminate between, and even nominate adequately. This happens because such units may have an intricate set of senses. In addition, a microsyntactic unit may be homonymous with a free word sequence so that it is unclear which of the two occurs in a given text. To illustrate this assertion, let us consider the English polysemous microsyntactic unit *all the same*, which may (1) refer to something happening despite some fact, as in *She was kind, but all the same she terrified me*, *or* (2) denote an indifference: *It is all the same to me whether you stay or go.* It is not always easy to choose between the two readings, and the task is even more difficult for an NLP system. On top of this, these two interpretations of the idiom have to compete with word sequences of *all*, *the* and *same* occurring in sentences like *Men are all the same* or *If all the cells in our body have the same DNA, then why aren't they all the same cell?*, which have nothing to do with the idiom.

Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces, pages 8–17, Osaka, Japan, December 11 2016.

In what follows, I will focus on Russian, where such idiomatic units are numerous and extremely varied. The situation is especially typical for units formed with functional words: pronouns, adverbs, particles, prepositions, and conjunctions.

By way of illustration, let us look at some instructive microsyntactic elements which require special attention and in-depth research.

(a) ***Tot že*** is a two-word adjective meaning “identical”, as in (1). This phrasemic unit is homonymous with a sequence including the pronominal adjective (as in (2a)) or noun (as in (2b)) ***tot*** ‘that’, used anaphorically, followed by the discourse particle ***že*** (with the meaning close to ‘as concerns’), cf.

(1) *Ja čital* ***tot že*** *rasskaz, čto i ty* 'I read the same story as you did' vs.

(2a) *Prišlos' kupit' novyj xolodil'nik –* ***tot že*** *slomalsja* lit. 'One had to buy a new fridge: as far as that one (= the old fridge) is concerned, it broke down' ('One had to buy a new fridge because the old one had broken down');

(2b) *Ja pozval druga v kino, tot* ***že*** *predložil pojti na futbol* lit. ‘I invited my friend to the movies, as for him, he proposed to go to football’ (I invited my friend to the movies but he suggested that we should go to a football game’).

(b) a two-word phrasemic adjectival quantifier ***ni odin*** *'not a single one'* (3), which may be homonymous with a random juxtaposition of a member of the two-part coordinating conjunction *ni.. ni* 'neither… nor' and the numeral *odin* 'one' (4):

(3) *V komnate ne bylo ni odnogo čeloveka* ‘There was not a single person in the room’ vs.

(4) “*Irka ne žalela dlja Nataši ni duxov, ni odnogo iz svoix ženixov, no eto nikogda ničem ne končalos'* (Victoria Tokareva) *‘Ira did not hesitate to spare Natasha her perfume, or one of her suitors, but that never brought any result’* (the story goes that Ira was willing to agree that one of her suitors should court her friend Natasha instead of her).

In this paper, we will investigate two polysemous microsyntactic units of Russian – *v silu* and *kak by*[2] – in order to find out with what other entities or groups of entities they come into contact. This task will be largely solved by corpus techniques.

2 Microsyntactic Markup in the SynTagRus Corpus

It is well known that lexically annotated text corpora are extremely helpful in lexical ambiguity resolution, especially in computational linguistics tasks. Lexical annotation means that polysemous words occurring in the corpus (ideally, all such words) are tagged for concrete lexical senses, specified by some lexicographic resource, be it a traditional explanatory dictionary or an electronic thesaurus like WordNet. Such lexically annotated corpora play a crucial role in word sense disambiguation (WSD) tasks. These tasks are normally solved by machine learning techniques, which are rapidly developing and improving. Research work in this area performed in varied paradigms for a multitude of languages is immense; recent papers, to cite but a few, include a comprehensive review by Navigly 2009, a paper by Moro et al. 2014, and newest research involving neural networks presented by Dayu Yuan et al. 2016.

It is to be added that text corpora, fully or at least partially tagged for lexical senses, are extremely helpful in disambiguation tasks within theoretical semantics and lexicography not necessarily directly related to computational linguistics or automatic text processing (see e.g. B.Iomdin 2014, B.Iomdin et al. 2016, Lopukhin and Lopukhina 2016).

We may regret that text corpora tagged for senses of «normal» words are not large enough, but they do exist and thus are at researchers' disposal. In contrast, to the best of my knowledge, there have been no resources so far to offer texts annotated for phraseological units of any types, including of course syntactic idioms. We have endeavored to mend the situation by introducing microsyntactic markup in the deeply annotated corpus of Russian texts, SynTagRus. This corpus, created in the Laboratory of Computational Linguistics of A.A.Kharkevich Institute of Information Transmission

[2] We are not glossing the microsyntactic units here because of their polysemy: they will be glossed later when individual senses are discussed.

Problems of the Russian Academy of Sciences in Moscow, contains several types of annotation: morphological, syntactic (in the form of dependency trees), lexico-functional, elliptic, and anaphoric annotation. (For details, see e.g. Dyachenko et al. 2015, where the current state of SynTagRus is presented.)

Microsyntactic annotation of the corpus is far from being an easy task. An important reason for that is the fact that no comprehensive, or even representative, list of microsyntactic units in general, and syntactic idioms in particular, is available to researchers. This is true of any language, including Russian. To overcome this deficiency, we resorted to two different strategies of tagging corpus sentences for microsyntactic elements:

1) continuous examination of a text aimed at finding all candidates to microsyntactic elements;

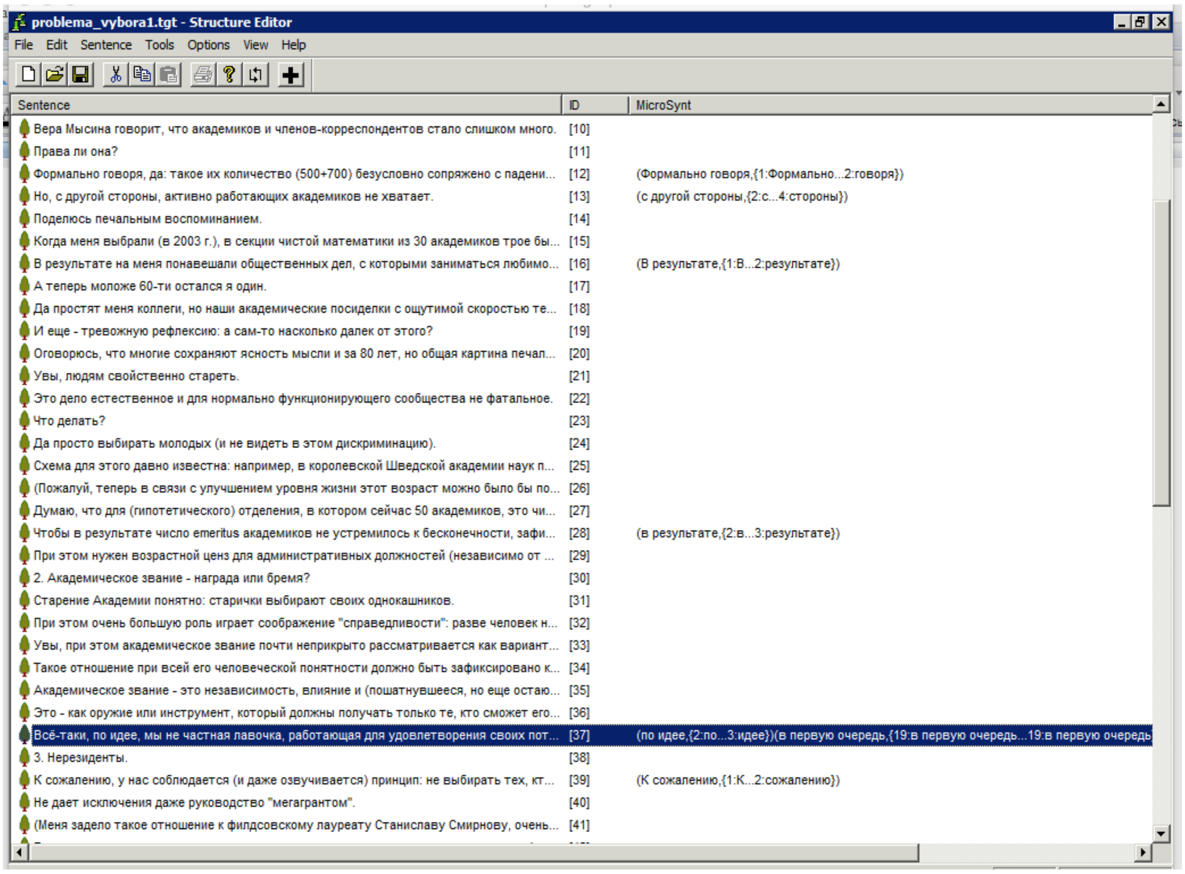
2) preliminary search for linear strings or syntactic subtrees composed of such words about which we have had previous knowledge or reasonable conjecture that they form, or may form, microsyntactic units. To give a few examples, these are strings or subtrees like *vse ravno* 'all the same', *kak budto* 'as though', *kol' skoro* 'since; as long as', *razve čto* 'if only, except that', *poka čto* 'so far; for the time being', *tol'ko liš'* 'nothing but; as soon as', *malo li* 'all sorts of things'; *vo čto by to ni stalo* 'at any cost; whatever happens', *ni razu* 'not once', *to i delo* 'over and over again', *čert znaet* + *interrogative word* 'devil knows (what, where,…)' etc.[3]

Understandably, in both cases only manual annotation of text for microsyntatic elements was possible: even its partial automation is a matter of the future (see however the discussion at the end of Section 3 below).

As a result of continuous scrutiny and targeted search of the material, we were able to obtain a draft version of microsyntactic markup of the corpus, which was later processed in detail. A thorough analysis of post-processed results revealed that the number of microsyntactic elements occurring in the text is quite considerable. In numerous texts, as many as 25% of sentences contain at least one microsyntactic element.

Fig. 1 below represents a connected fragment of a randomly chosen text of the corpus, which was annotated according to the first strategy. It is easy to see that, out of 30 sentences, 6 sentences contain syntactic idioms, whilst one sentence features two such idioms: adverbials *po idee* 'in theory; at face value' and *v pervyju očered'* 'primarily; first and foremost. '

[3] In order to avoid extended discussion, which could lead us far from the main topic, we list only one or two English equivalents for all microsyntactic units cited. Interestingly, in all of these cases Russian microsyntactic units correspond to multiword English microsyntactic units which we use as glosses. It can thus be hypothesized that the number and composition of microsyntactic phenomena in various languages are commensurable.


problema_vybora1.tgt - Structure Editor
File Edit Sentence Tools Options View Help

Sentence	ID	MicroSynt
Вера Мысина говорит, что академиков и членов-корреспондентов стало слишком много.	[10]	
Права ли она?	[11]	
Формально говоря, да: такое их количество (500+700) безусловно сопряжено с падени...	[12]	(Формально говоря,{1:Формально...2:говоря})
Но, с другой стороны, активно работающих академиков не хватает.	[13]	(с другой стороны,{2:с...4:стороны})
Поделюсь печальным воспоминанием.	[14]	
Когда меня выбрали (в 2003 г.), в секции чистой математики из 30 академиков трое бы...	[15]	
В результате на меня понавешали общественных дел, с которыми заниматься любимо...	[16]	(В результате,{1:В...2:результате})
А теперь моложе 60-ти остался я один.	[17]	
Да простят меня коллеги, но наши академические посиделки с ощутимой скоростью те...	[18]	
И еще - тревожную рефлексию: а сам-то насколько далек от этого?	[19]	
Оговорюсь, что многие сохраняют ясность мысли и за 80 лет, но общая картина печал...	[20]	
Увы, людям свойственно стареть.	[21]	
Это дело естественное и для нормально функционирующего сообщества не фатальное.	[22]	
Что делать?	[23]	
Да просто выбирать молодых (и не видеть в этом дискриминацию).	[24]	
Схема для этого давно известна: например, в королевской Шведской академии наук п...	[25]	
(Пожалуй, теперь в связи с улучшением уровня жизни этот возраст можно было бы по...	[26]	
Думаю, что для (гипотетического) отделения, в котором сейчас 50 академиков, это чи...	[27]	
Чтобы в результате число emeritus академиков не устремилось к бесконечности, зафи...	[28]	(в результате,{2:в...3:результате})
При этом нужен возрастной ценз для административных должностей (независимо от ...	[29]	
2. Академическое звание - награда или бремя?	[30]	
Старение Академии понятно: старички выбирают своих однокашников.	[31]	
При этом очень большую роль играет соображение "справедливости": разве человек н...	[32]	
Увы, при этом академическое звание почти неприкрыто рассматривается как вариант...	[33]	
Такое отношение при всей его человеческой понятности должно быть зафиксировано к...	[34]	
Академическое звание - это независимость, влияние и (пошатнувшееся, но еще остаю...	[35]	
Это - как оружие или инструмент, который должны получать только те, кто сможет его...	[36]	
Всё-таки, по идее, мы не частная лавочка, работающая для удовлетворения своих пот...	[37]	(по идее,{2:по...3:идее})(в первую очередь,{19:в первую очередь...19:в первую очередь
3. Нерезиденты.	[38]	
К сожалению, у нас соблюдается (и даже озвучивается) принцип: не выбирать тех, кт...	[39]	(К сожалению,{1:К...2:сожалению})
Не дает исключения даже руководство "мегагрантом".	[40]	
(Меня задело такое отношение к филдсовскому лауреату Станиславу Смирнову, очень...	[41]	

Fig. 1. Annotation of a SynTagRus text with microsyntactic elements.

We will now concentrate on the second strategy of microsyntactic annotation and consider two specific examples, represented in Sections 3 and 4, respectively.

3 A polysemous microsyntactic unit *v silu* and its context

Fig. 2 represent a subcorpus fragment in which syntactic structures of each sentence contains a binary subtree consisting of the preposition *v* '≅ in' and the noun *sila* '≅ force' in the accusative case, singular number, dominated by the preposition *v*.[4] The whole SynTagRus corpus (amounting to 1,000,000 words today) was found to contain 86 such sentences.

The annotation of this fragment performed by linguist experts revealed that the majority of these sentences (in all, 57 of them) contain a compound preposition *v silu* 'because of; by virtue of'[5], as in the sentence

(5) ***V silu*** *specifiki Moskvy daže takie obščepriznannye kriterii, kak mestopoloženie doma i cena kvadratnogo metra, nel'zja sčitat' osnovnymi.* 'By virtue of the specific character of Moscow, even such generally recognized criteria as the location of the house and the price of a square meter cannot be considered to be the main ones'.

Six sentences contain a very different microsyntactic unit, which could reasonably be named «v takuju-to silu» 'to such and such extent', as in

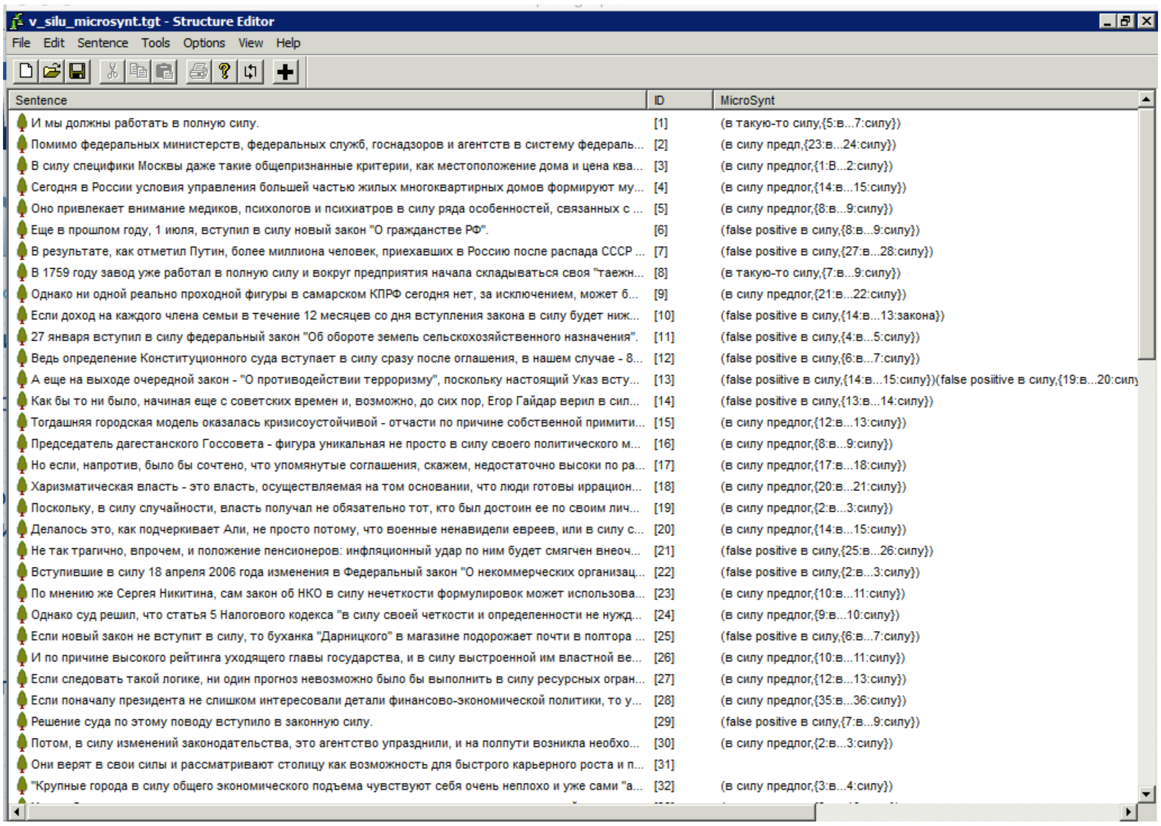
(6) *Eto govorit o tom, čto my ne spravljaemsja s potokom našix doxodov, ne v sostojanii* ***v polnuju silu*** *aktivizirovat' biznes*. 'This says that we are not coping with the flow of our incomes, are unable to activate the business in full swing'.

[4] It must be noted that the syntactic formalism underlying the syntactic annotation of SynTagRus heavily relies on the syntactic component of the Meaning ⇔ Text theory by Igor Mel'čuk (see Mel'čuk 1974), which was later refined and extended by Juri Apresjan and his colleagues during the elaboration of a multipurpose linguistic processor ETAP (see Apresjan et al. 1992, 2010). For the reasons of space, I am skipping the details of this theory.

[5] All compound prepositions are naturally considered to be microsyntactic units.

This unit, an adverbial of degree, consists of two permanent word elements, *v* and *silu*, and one variable element – an adjective modifying the noun *silu*. SynTagRus represents only one option of this unit – *v polnuju silu* 'in full swing' (plus two occurrences of the adverb *vpolsily* 'at half strength', which may be viewed as an option of the «v takuju-to silu» unit). In fact, Russian texts may contain other adjectives that occupy the variable position of the microsyntactic unit – *v nepolnuju silu* 'at reduced strength', *v polovinnuju silu* (a very close synonym of *vpolsily*), quantifier nouns (*v polovinu sily*, another equivalent of *vpolsily*) and even numerals like *tri sily* (something like 'using three strengths'), as in

(7) *Snova vzjalis' za sunduk, teper' uže v tri sily, i snova opustili – ne te sily* [Valentin Rasputin] 'They put their hands to the trunk, this time merging three strengths, but lowered them again – the strengths were wrong'.



Sentence	ID	MicroSynt
И мы должны работать в полную силу.	[1]	(в такую-то силу,{5:в...7:силу})
Помимо федеральных министерств, федеральных служб, госнадзоров и агентств в систему федераль...	[2]	(в силу предл,{23:в...24:силу})
В силу специфики Москвы даже такие общепризнанные критерии, как местоположение дома и цена ква...	[3]	(в силу предлог,{1:В...2:силу})
Сегодня в России условия управления большей частью жилых многоквартирных домов формируют му...	[4]	(в силу предлог,{14:в...15:силу})
Оно привлекает внимание медиков, психологов и психиатров в силу ряда особенностей, связанных с ...	[5]	(в силу предлог,{8:в...9:силу})
Еще в прошлом году, 1 июля, вступил в силу новый закон "О гражданстве РФ".	[6]	(false positive в силу,{8:в...9:силу})
В результате, как отметил Путин, более миллиона человек, приехавших в Россию после распада СССР ...	[7]	(false positive в силу,{27:в...28:силу})
В 1759 году завод уже работал в полную силу и вокруг предприятия начала складываться своя "таежн...	[8]	(в такую-то силу,{7:в...9:силу})
Однако ни одной реально проходной фигуры в самарском КПРФ сегодня нет, за исключением, может б...	[9]	(в силу предлог,{21:в...22:силу})
Если доход на каждого члена семьи в течение 12 месяцев со дня вступления закона в силу будет ниж...	[10]	(false positive в силу,{14:в...13:закона})
27 января вступил в силу федеральный закон "Об обороте земель сельскохозяйственного назначения".	[11]	(false positive в силу,{4:в...5:силу})
Ведь определение Конституционного суда вступает в силу сразу после оглашения, в нашем случае - 8...	[12]	(false positive в силу,{6:в...7:силу})
А еще на выходе очередной закон - "О противодействии терроризму", поскольку настоящий Указ всту...	[13]	(false posiitive в силу,{14:в...15:силу})(false posiitive в силу,{19:в...20:силу
Как бы то ни было, начиная еще с советских времен и, возможно, до сих пор, Егор Гайдар верил в сил...	[14]	(false positive в силу,{13:в...14:силу})
Тогдашняя городская модель оказалась кризисоустойчивой - отчасти по причине собственной примити...	[15]	(в силу предлог,{12:в...13:силу})
Председатель дагестанского Госсовета - фигура уникальная не просто в силу своего политического м...	[16]	(в силу предлог,{8:в...9:силу})
Но если, напротив, было бы сочтено, что упомянутые соглашения, скажем, недостаточно высоки по ра...	[17]	(в силу предлог,{17:в...18:силу})
Харизматическая власть - это власть, осуществляемая на том основании, что люди готовы иррацион...	[18]	(в силу предлог,{20:в...21:силу})
Поскольку, в силу случайности, власть получал не обязательно тот, кто был достоин ее по своим лич...	[19]	(в силу предлог,{2:в...3:силу})
Делалось это, как подчеркивает Али, не просто потому, что военные ненавидели евреев, или в силу с...	[20]	(в силу предлог,{14:в...15:силу})
Не так трагично, впрочем, и положение пенсионеров: инфляционный удар по ним будет смягчен внеоч...	[21]	(false positive в силу,{25:в...26:силу})
Вступившие в силу 18 апреля 2006 года изменения в Федеральный закон "О некоммерческих организац...	[22]	(false positive в силу,{2:в...3:силу})
По мнению же Сергея Никитина, сам закон об НКО в силу нечеткости формулировок может использова...	[23]	(в силу предлог,{10:в...11:силу})
Однако суд решил, что статья 5 Налогового кодекса "в силу своей четкости и определенности не нужд...	[24]	(в силу предлог,{9:в...10:силу})
Если новый закон не вступит в силу, то буханка "Дарницкого" в магазине подорожает почти в полтора ...	[25]	(false positive в силу,{6:в...7:силу})
И по причине высокого рейтинга уходящего главы государства, и в силу выстроенной им властной ве...	[26]	(в силу предлог,{10:в...11:силу})
Если следовать такой логике, ни один прогноз невозможно было бы выполнить в силу ресурсных огран...	[27]	(в силу предлог,{12:в...13:силу})
Если поначалу президента не слишком интересовали детали финансово-экономической политики, то у...	[28]	(в силу предлог,{35:в...36:силу})
Решение суда по этому поводу вступило в законную силу.	[29]	(false positive в силу,{7:в...9:силу})
Потом, в силу изменений законодательства, это агентство упразднили, и на полпути возникла необхо...	[30]	(в силу предлог,{2:в...3:силу})
Они верят в свои силы и рассматривают столицу как возможность для быстрого карьерного роста и п...	[31]	
"Крупные города в силу общего экономического подъема чувствуют себя очень неплохо и уже сами "а...	[32]	(в силу предлог,{3:в...4:силу})

Fig. 2. Annotation of a SynTagRus fragment containing the binary subtree *v silu* with microsyntactic elements.

In 21 sentences of the samplng the subtree *v silu* occurred within expressions like *vstupat' v silu* 'come into effect', as in

(8) *Rešenie suda vstupilo v zakonnuju silu* 'The court's decision came into legal force'.

In my opinion, there is no syntactic idiom in (8); instead, we have to do with a specific meaning of the noun *sila* 'validity'. This meaning is specified in traditional dictionaries. What is important, however, is that the noun *sila* cannot be used freely in this meaning: it can only appear in combination with lexical finctions like IncepOper1 (*vstupat' v silu* 'come into force', FinOper1 (*utračivat' silu* 'be terminated') and with weakly idiomatic variants of Oper1 (*byt' v sile* 'be in force' and *ostavat'sja v sile* 'remain in force'.[6]

Since expressions like *vstupat' v silu* turn out to be very frequent, we believe that it is reasonable to leave them in the microsyntactic annotation of the corpus as false positives so that the respective contexts could be used in automatic disambigution of regular and microsyntactic units (e.g. with the help of machine learning techniques). We believe that such disambiguation will be possible in future.

[6] Lexical functions present an important element of semantic and lexicographical components of the Meaning ⇔Text theory. The discussion of this issue, however, is far beyond the topic of my paper.

It is worth mentioning that in the whole sampling for *v silu* there is only one chance sentence which has nothing to do with any of the two syntactic idioms postulated above and at the same time contains no false positives: this is the sentence

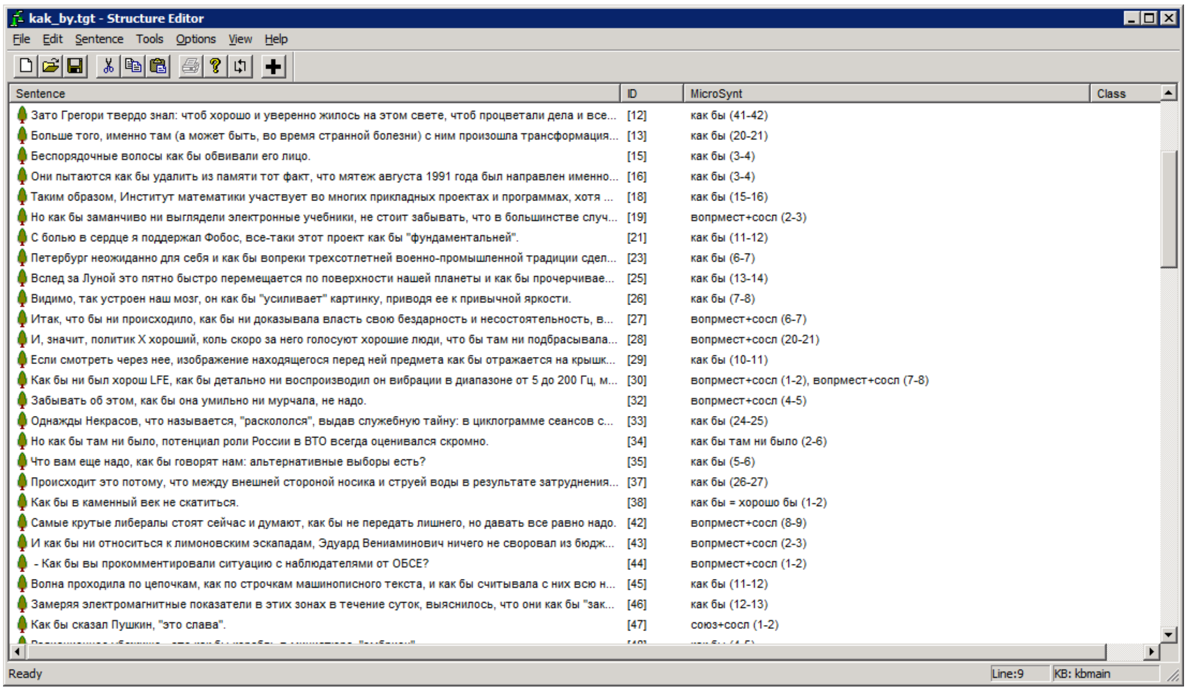
(9) *Vošlo by v silu novoe pokolenie, osoznal by svoi interesy srednij klass* ‘If the new generation gained strength, the middle class would become over of its interests’.

It should be emphasized that in this case SynTagRus provides a fairly satisfactory representation of syntactic idioms and freer collocations with *v silu*: two syntactic idioms and one lexical functional construction are amply covered. The author is aware of only one other syntactic idiom that contains the subtree *v silu* and is absent from SynTagRus*:* this is the adverbial meaning 'at a certain level', occurring in expressions like *igrat' v silu pervogo razrjada <pervorazrjadnika, grossmejstera>* etc. 'play at the level of the higher athletic rank, the first-rank sportsman, the grand master’

It cannot of course be excluded that there are other microsyntactic idioms based on this group of words: so far, however, we are aware of none.

4 A polysemous microsyntactic unit *kak by* and its context

Using the same strategy of the preliminary search of potential syntactic phrasemic units, we obtained a corpus sample of 116 sentences containing the *kak by* string, assuming that at least some of these sentences must contain phrasemic units. Fig. 3 below shows a fragment of this sample.


kak_by.tgt - Structure Editor
File Edit Sentence Tools Options View Help
Ready
Line:9
KB: kbmain

Sentence	ID	MicroSynt	Class
Зато Грегори твердо знал: чтоб хорошо и уверенно жилось на этом свете, чтоб процветали дела и все...	[12]	как бы (41-42)	
Больше того, именно там (а может быть, во время странной болезни) с ним произошла трансформация...	[13]	как бы (20-21)	
Беспорядочные волосы как бы обвивали его лицо.	[15]	как бы (3-4)	
Они пытаются как бы удалить из памяти тот факт, что мятеж августа 1991 года был направлен именно...	[16]	как бы (3-4)	
Таким образом, Институт математики участвует во многих прикладных проектах и программах, хотя ...	[18]	как бы (15-16)	
Но как бы заманчиво ни выглядели электронные учебники, не стоит забывать, что в большинстве случ...	[19]	вопрмест+сосл (2-3)	
С болью в сердце я поддержал Фобос, все-таки этот проект как бы "фундаментальней".	[21]	как бы (11-12)	
Петербург неожиданно для себя и как бы вопреки трехсотлетней военно-промышленной традиции сдел...	[23]	как бы (6-7)	
Вслед за Луной это пятно быстро перемещается по поверхности нашей планеты и как бы прочерчивае...	[25]	как бы (13-14)	
Видимо, так устроен наш мозг, он как бы "усиливает" картинку, приводя ее к привычной яркости.	[26]	как бы (7-8)	
Итак, что бы ни происходило, как бы ни доказывала власть свою бездарность и несостоятельность, в...	[27]	вопрмест+сосл (6-7)	
И, значит, политик X хороший, коль скоро за него голосуют хорошие люди, что бы там ни подбрасывала...	[28]	вопрмест+сосл (20-21)	
Если смотреть через нее, изображение находящегося перед ней предмета как бы отражается на крышк...	[29]	как бы (10-11)	
Как бы ни был хорош LFE, как бы детально ни воспроизводил он вибрации в диапазоне от 5 до 200 Гц, м...	[30]	вопрмест+сосл (1-2), вопрмест+сосл (7-8)	
Забывать об этом, как бы она умильно ни мурчала, не надо.	[32]	вопрмест+сосл (4-5)	
Однажды Некрасов, что называется, "раскололся", выдав служебную тайну: в циклограмме сеансов с...	[33]	как бы (24-25)	
Но как бы там ни было, потенциал роли России в ВТО всегда оценивался скромно.	[34]	как бы там ни было (2-6)	
Что вам еще надо, как бы говорят нам: альтернативные выборы есть?	[35]	как бы (5-6)	
Происходит это потому, что между внешней стороной носика и струей воды в результате затруднения...	[37]	как бы (26-27)	
Как бы в каменный век не скатиться.	[38]	как бы = хорошо бы (1-2)	
Самые крутые либералы стоят сейчас и думают, как бы не передать лишнего, но давать все равно надо.	[42]	вопрмест+сосл (8-9)	
И как бы ни относиться к лимоновским эскападам, Эдуард Вениаминович ничего не своровал из бюдж...	[43]	вопрмест+сосл (2-3)	
- Как бы вы прокомментировали ситуацию с наблюдателями от ОБСЕ?	[44]	вопрмест+сосл (1-2)	
Волна проходила по цепочкам, как по строчкам машинописного текста, и как бы считывала с них всю н...	[45]	как бы (11-12)	
Замеряя электромагнитные показатели в этих зонах в течение суток, выяснилось, что они как бы "зак...	[46]	как бы (12-13)	
Как бы сказал Пушкин, "это слава".	[47]	союз+сосл (1-2)	

Fig. 3. Microsyntactic annotation of a corpus sample with sentences containing the *kak by* string.

A thorough study of these data shows a very interesting and complicated microsyntactic picture.

First of all, many of the sentences contain the phrasemic unit we shall call ***kak by 1*** that can be treated as a discourse particle with the semantics of comparison or uncertainty, as in (10) and (11):

(10) *Gazety, sledovatel'no, imejuščie dejstvitel'no obščestvennoe značenie, sut'* ***kak by*** *akushery obščestvennogo mnenija, pomogajuščie emu javit'sja na svet Božij* (N. Danilevskij) 'Therefore, the newspapers having a true public value are, **in a way**, obstetricians of the public opinion, helping it to be borne'.

Here the author compares newspapers to obstetricians and warns the reader that he exploits a metaphor, by using the *kak by* expression.

(11) *Tolpa sostojala iz ljuda prostogo, i čekistam bylo ne s ruki xvatat'* ***kak by*** *svoix – trudjaščixsja* (A. Tkačenko) 'The crowd were just common people, and the security agents did not think it fit to arrest those who were, **so to say**, of their own kind'.

Here the speaker considers the word *svoix* 'of their kind' not proper enough.

Intrinsically, ***kak by 1*** can almost be viewed as one word (which in internet discussions is even sometimes written without a space: *kakby*). In SynTagRus, the words *kak* and *by* are connected with the auxiliary syntactic relation reserved for syntactically integrated entities (*kak* -> *by*). The author is quite unclear whether *kak* here is a pronoun, a conjunction, or neither. In any case, it is the syntactic head of the unit, governed, in its turn, by the next word by a restrictive syntactic relation; see sentence (12) and its syntactic structure shown in Fig. 4:

(12) *Besporjadočnye volosy kak by obvivali ego lico* ('The unruly hair as if twined itself around his face').

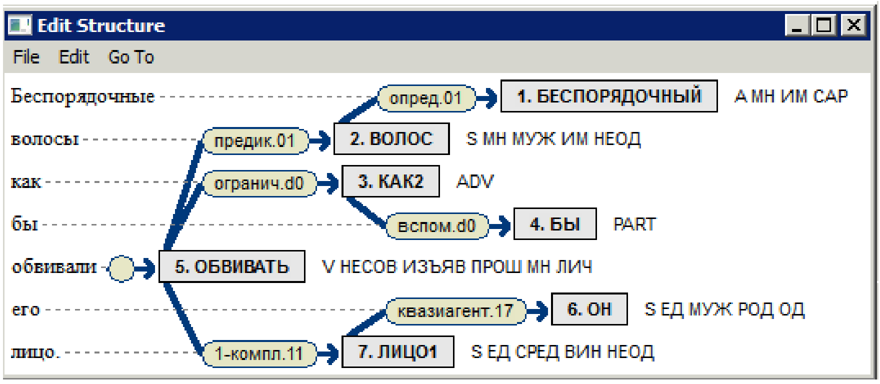


Fig. 4. Syntactic structure of (12) containg the phrasemic unit ***kak by 1***.

Another syntactic phrasemic unit that can be identified in the corpus is the conjunction ***kak by 2*** which is only used as a strongly governed word with many predicates sharing the semantics of apprehension, such as the verbs *bojat'sja* 'to be afraid', *opasat'sja* 'to fear', *ispugat'sja* 'to be scared', *sledit'* 'to make sure', the nouns *bojazn', strax, opasenie* 'fear', and even the predicative adverbs *strashno, bojazno* 'fearful':

(13) *V universitete Galilej poseščal takže lekcii po geometrii… i nastol'ko uvljoksja aetoj naukoj, čto otec stal opasat'sja,* **kak by** *aeto ne pomešalo izučeniju mediciny* (Wikipedia) 'In the university, Galileo also attended lectures in geometry… and became so enthusiastic about this science that his father started fearing lest it could interfere with his learning medicine'.

(14) *Vo Frajburge za nim [Gor'kim] po pjatam xodili špiki: nemeckie, — bojavšiesja, čto on sdelal revoljuciju, i sovetskie, — sledivšie,* **kak by** *on ne sdelal kontrrevoljuciju* (V. Khodasevich) 'In Freiburg, Gorky was closely followed by spies: German ones, who were afraid of him because he organized the revolution, and Soviet ones, who were making sure that he would not organize a counter-revolution'.

(15) *Smotri,* **kak by** *tebe ne požalet', čto smeješ'sja* (A. Herzen) 'Watch out you don't regret that you are laughing';

(16) *Vyvesti aeskadron iz stanicy bylo rešeno iz opasenija,* **kak by** *aeskadron ne vosstal, uznav ob areste Fomina* (M. Šoloxov) 'It was decided to draw out the squadron from the village, in fear that the squadron would rise if it learned about Fomin's arrest'.

(17) *Tolpa očutilas' neožidanno tak blizko k imperatoram, čto Rostovu, stojavšemu v perednix rjadax ejo, stalo strašno,* **kak by** *ego ne uznali* (L. Tolstoy) 'The crowd suddenly got so close to the emperors that Rostov, who was standing in its first rows, started to fear that he **could** be recognized'.

Sometimes the government pattern of this conjunction can be modified by the expletive pronoun *to*:

(18) *S odnoj storony, neobxodimo bylo v celjax samooborony koe-čto pozaimstvovat', koe-čemu poučit'sja u Evropy; s drugoj storony, nado bylo opasat'sja togo,* **kak by** *pri aetom ne popast' v*

kul'turnuju, duxovnuju zavisimost' ot Evropy (N. Trubetskoj) 'On the one hand, it was necessary for self-defence to learn something from Europe; on the other hand, one had to **make sure not to** get into cultural and mental dependence on Europe'.

Semantic, syntactic and collocational properties of the syntactic idiom ***kak by 2*** are very interesting and need individual research. We can make only a few remarks here. First, the conjunction requires the presence of the negative particle *ne* as a direct dependent. Of the head verb of the subordinate close. Second, the verb has to be either a finite form in the past tense, or an infinitive (in the latter case, the implicit subject of the infinitive should coincide with the subject of the head predicate, cf. example 15 above and 19):

(19) *Ona snova pošla, opasajas', kak by ne natknut'sja gde-nibud' na policiju* [Vasil Bykov] 'She$_i$ went once again, fearing lest she$_i$ should run onto the police somewhere'.

As far as the embedment of *kak by 2* into the syntactic structure of the sentence is concerned, I believe that the most natural solution would be to view the first element of the idiom as a conjunction and subordinate it to the head predicate using a respective valency relation; the verb of the subordinate clause should depend on *kak*, and *by* should be linked to this latter verb. Accordingly, the elements of the idioms turn out to be syntactically unlinked with each other, as in Fig. 5:

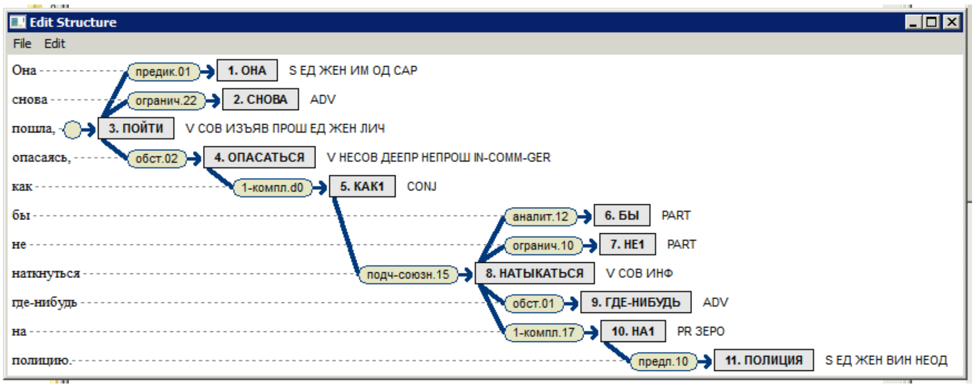


Fig. 5. The syntactic structure of the sentence with the embedded syntactic idiom *kak by 2*.

The next syntactic idiom composed of *kak* and *by* is a modal lexical unit that implicitly expresses the speaker's wish. Let us refer to this idiom as ***kak by 3***. It is represented in such corpus sentences as

(20) ***Kak by*** *v kamennyj vek ne skatit'sja* 'It would be good not to slide back into the stone age'

At first glance, this idiom is close to the microsyntactic unit *kak by 2* described above: in both cases, one has to do with the speaker's wish to avoid some unpleasant situation (and hence, with his fear that such a situation may happen. However the picture seems to me to be drastically different. Note that in (20) the verb *skatit'sja* 'slide back' belongs to the scope of the implicit predicate of the speaker's wish together **with the negation**; the speaker wishes no-slidng back to the stone age. In contrast, sentences with *kak by 2* contain the predicate of fear whose scope does not contain the negation. The approximate equivalence of (20) and

(21) *Bojus', kak by v kamennyj vek be skatit'sja* 'I fear lest we slide back into the stone age'

follows from the correlation of the semantics of wish and fear: *I fear X* is known to mean 'I expect X and I do not want X to happen'. In constructions with *kak by 3* the verbal negation is frequent but not at all obligatory, cf.

(22) *Kak by obojtis' bez etogo, ostaviv samuju sut'* [A.Bitov] 'I wish we could manage without it, leaving only the most crucial thing.

Another syntactic idiom appearing in the subcorpus is a discourse unit *kak by ne tak* ≈ ‘contrary to expectations, the situation is different and much worse'. Normally, this unit forms a separate utterance or a separate clause:

(23) *Vy dumaete, teper' on po krajnej mere ujdet? Kak by ne tak!* [I.S.Turgenev] 'Do you think he will now at least leave?' Like hell he will’.

Finally, the subcorpus includes a very frequent parenthetic expression *kak by to ni bylo* ‘be that as it may’, with a variation *kak by tam ni bylo,* which can also be viewed as a microsyntactic unit:

(24) *Teper' vse eto bylo pozadi, no kak by tam ni bylo, videt' špiona Gardnera bylo emu neprijatno.* [Ju. Dombrovsky]. This was now all over but however that may be he did not like seeing the spy Gardner.

It is naturally rather easy to identify units like *kak by ne tak* or *kak by to ni* bylo due to their length and strict word order; yet, counterexamples are also possible, cf.

(25) *Vse kak by ne tak už ploxo* 'It seems that all is not so bad'

where the «shorter» idiom *kak by 1* can be detected

As in the first subcorpus containing units *v silu*, the present subcorpus has a number of sentences that do not involve microsyntactic units formed with *kak by.* In particular, some sentences contain construction with the emphatic particle *ni*:
(26) *Kak by nam ni xotelos' povysit' kačestvo školnogo obrazovanija, na eto potrebuetsja ešče mnogo let* 'However much we want to improve the quality of school education, this will require many years yet'.

Clearly neither *kak* nor *by* are constituent elements of this construction: *kak* may be replaced by any interrogative word, and *by* may be absent, as in
(27) *Čto on ni predprinimaet, ničego ne menjaetsja* 'Whatever he undertakes, nothing changes'.

Finally, some sentences represent a Wackernagel shift of the particle *by* forming the subjunctive mood into the position after *kak*, as in
(28) *Kak by ty otvetil* 'How would you answer'.

As in the first subcorpus, we leave «false positive» tags in all such cases.

To conclude, we need to note that in this case, too, the corpus is representative enough for the syntactic idioms postulated. Yet, we were able to find an interesting microsyntactic idiom formed with kak and by beyond the material of the corpus. It can be illustrated by a sentence present in the Russian National Corpus:
(29) — Kak by ne burja moskovskaja sobiraetsja, - pokrutil golovoj storož i povernul s pogosta von. [B.Evseev]. 'Isn't it the case that the Moscow tempest is approaching? – The watchman twisted his head and went away from the cemetery’

The first part of (29) means the following: There are signs that the Moscow tempest is approaching, which is undesirable. Importantly, in such cases a semantically void negation must be present – just like in the case with kak by 2. However it is not attached to the verb but immediately follows kak by, ths forming a new syntactic idiom which could be called kak by ne. This idiom has a rather close synonym – už ne (with an obligatory li particle). - ‘Už bne burja li moskovskaja sobiraetsja?

It goes without saying that one cannot discuss all features of the newly developed resource – a corpus with microsyntactic annotation. It is to be hoped that I could demonstrate the fact that this resource is likely to be very helpful.

Acknowledgements

The author is grateful to the Russian Humanitarian Research Foundation for their support of this work (grant No. 15-04-00562).

References

Ju.D.Apresjan, I.M.Boguslavsky, L.L.Iomdin, A.V.Lazursky, L.G.Mitjushin, V.Z.Sannikov, L.L.Tsinman (1992). Lingvističeskij protessor dlja složnyx informatisonnyx sistem. [A linguistic processor for complex information systems.] Moscow, Nauka Publishers. 256 p. [In Russian.]

Ju.D.Apresjan, I.M.Boguslavsky, L.L.Iomdin, V.Z.Sannikov (2010). Teoretičeskie problemy russkogo sintaksisa: Vzaimodejstvie grammatiki i slovarja. [Theoretical Issues of Russian Syntax: Interaction of the Grammar and the Lexicon.] / Ju.D.Apresjan, ed. Moscow, Jazyki slavjanskix kul'tur. 408 p. [In Russian.]

Dayu Yuan, Ryan Doherty, Julian Richardson, Colin Evans, Eric Altendorf (2016). Word Sense Disambiguation with Neural Language Models. https://arxiv.org/pdf/1603.07012v1.pdf.

P.V.Dyachenko, L.L.Iomdin, A.V.Lazursky, L.G.Mityushin, O.Yu.Podlesskaya, V.G.Sizov, T.I.Frolova, L.L.Tsinman (2015). Sovremennoe sostojanie gluboko annotirovannogo korpusa tekstov russkogo jazyka (SynTagRus). [The current state of the deeply annotated corpus of Russian texts (SynTagRus)]. // Nacional'nyj korpus russkogo jazyka. 10 let proektu. Trudy Instituta russkogo jazyka im. V.V. Vinogradova. M. Issue 6, p. 272-299. [In Russian.]

Ch. Fillmore (1988). The Mechanisms of Construction Grammar. // Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society. pp. 35-55.

A. Goldberg (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

B.L.Iomdin (2014). Mnogoznačnye slova v kontekste i vne konteksta. [Polysemous words in context and out of context.] Voprosy jazykoznaniya, No. 4. P. 87-103.

B.L.Iomdin, A.A.Lopukhina, K.A.Lopukhin, G.V.Nosyrev (2016). Word Sense Frequency of Similar Polysemous Words in Different Languages. // Computational Linguistics and Intellectual Technologies. Dialogue 2016, p. 214-225.

L.L.Iomdin (2013). Nekotorye mikrosintaksičeskie konstruktsii v russkom jazyke s učastiem slova *što* v kačestve sostavnogo elementa. [Certain microsyntactic constructions in Russian which contain the word *što* as a constituent element.] Južnoslovenski filolog. Beograd, LXIX, 137-147. [In Russian.]

L.L.Iomdin (2014). Xorošo menja tam ne bylo: sintaksis i semantika odnogo klassa russkix razgovornyx konstruktsij. [Good thing I wasn't there: syntax and semantics of a class of Russian colloquial constructions]. // Grammaticalization and lexicalization in the Slavic languages. Proceedings from the 36th meeting of the commission on the grammatical structure of the Slavic languages of the International committee of Slavists. München-Berlin-Washington/D.C.: Verlag Otto Sagner. 436 p. (Die Welt der Slaven. Bd. 55), p. 423-436. [In Russian.]

L.L.Iomdin (2015). Konstruktsii mikrosintaksisa, obrazovannye russkoj leksemoj *raz*. [Constructions of microsyntax built by the Russian word *raz*.]. SLAVIA, časopis pro slovanskou filologii, ročník 84, sešit 3, s. 291-306. [In Russian.]

K.A.Lopukhin, A.A.Lopukhina (2016). Word Sense Disambiguation for Russian Verbs using Semantic Vectors and Dictionary Entries. // Computational Linguistics and Intellectual Technologies. Dialogue 2016, p. 393-405.

I.A.Mel'čuk (1974). Opyt teorii lingvističeskix modelej «Smysl ⇔ Tekst». [An experience of creating the theory of linguistic models of the Meaning ⇔ Text type.] Moscow, Nauka Publishers. [In Russian.]

R. Navigli (2009). Word sense disambiguation: A survey. ACM Comput. Survey,. 41(2):1–69.

A. Moro, A. Raganato, R. Navigli (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics (TACL), 2, pp. 231-244.

E.V.Rakhilina (ed.) (2010). Lingvistika konstruktsij. [The linguistics of constructions]. Moscow, Azbukovnik Publishers. 584 p. [In Russian.]