MICROSYNTACTIC ANNOTATION OF CORPORA AND ITS USE IN COMPUTATIONAL LINGUISTICS TASKS[1]

LEONID IOMDIN
A. A. Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

IOMDIN, Leonid: Microsyntactic Annotation of Corpora and Its Use in Computational Linguistics Tasks. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 169 – 178.

Abstract: Microsyntax is a linguistic discipline dealing with idiomatic elements whose important properties are strongly related to syntax. In a way, these elements may be viewed as transitional entities between the lexicon and the grammar, which explains why they are often underrepresented in both of these resource types: the lexicographer fails to see such elements as full-fledged lexical units, while the grammarian finds them too specific to justify the creation of individual well-developed rules. As a result, such elements are poorly covered by linguistic models used in advanced modern computational linguistic tasks like high-quality machine translation or deep semantic analysis. A possible way to mend the situation and improve the coverage and adequate treatment of microsyntactic units in linguistic resources is to develop corpora with microsyntactic annotation, closely linked to specially designed lexicons. The paper shows how this task is solved in the deeply annotated corpus of Russian, SynTagRus.

Keywords: Text corpora, Russian syntactically tagged corpus SynTagRus, syntactic idioms, microsyntactic annotation, microsyntactic dictionary

1 INTRODUCTORY REMARKS

The theory of microsyntax has been developed by the author over the last 15 years (recent publications include [1], [2], [3], [4]). In this theory, which has much in common with construction grammar (see e.g. [5], [6], [7] and [8][2], two main groups of linguistic units are distinguished: lexically centered syntactic idioms and lexically unrestricted non-standard syntactic constructions.[3] Throughout the paper, I will be mostly concerned with these units, which will be referred to as microsyntactic units. Primarily, I will consider syntactic idioms.

[1] The author is grateful to the Russian Humanitarian Scientific Foundation for their support of this work with a grant (No. 15-04-00562). Special thanks also go to anonymous reviewers of the submitted version of the paper, who provided some valuable remarks.

[2] Interestingly, the last paper by P. Lauwers and N, Van Wettere introduces the term "micro-constructionalization", which is an additional evidence of the proximity (but not the identity!) of the approaches.

[3] In fact, some non-standard syntactic constructions are lexically constrained in the sense that they contain two or even more occurrences of the same word. Russian has a great variety of such constructions, each having unique syntactic peculiarities and subtle semantic features, as e.g. ***rabota rabotoj****, no nado otdoxnut' »* 'work is work but one needs a rest' or ***videt' ja ne videl****, no slyshal ob etom. »* ' I didn't really see it but I heard about it' (lit. 'to see I saw not but heard about it'). Probably Russian has many more constructions with lexical repetitions than e.g. English (cf. a relatively full list of English tautological constructions in [9]).

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0027

Microsyntactic units are poorly represented even in traditional linguistic resources, such as monolingual or bilingual dictionaries or descriptive grammars. The reason for this is obvious: syntactic idioms are difficult to attach to a particular lexical entry (so they are often just mentioned and briefly commented on in an entry for one of the words constituting the idiom), while non-standard constructions are too specific to find a place for themselves in general grammars. In computational linguistic resources, microsyntactic elements are even less visible (as are idiomatic entities in general). As a result, they are often disregarded in high-end computational linguistics tasks, such as deep semantic analysis, quality parsing, question answering, or machine translation – or, at best, treated with *ad hoc* solutions.[4]

The project outlined below is an attempt to improve the state of affairs at least partially. The idea is twofold: 1) to create a special dictionary of microsyntactic units of Russian, which should provide comprehensive information on such units and ensure their effective use in computational linguistics applications, and 2) to develop a text corpus which should incorporate annotation of such units. The former type of resource, the Microsyntactic dictionary of Russian, has been described in detail in [4] and [10]. In what follows, I will focus on the second goal, i.e. the development of the corpus with microsyntactic annotation, which, so far, has been only briefly reported in [4] and [11].

2 MICROSYNTACTIC ANNOTATION IN SYNTAGRUS

Rather than create a new corpus with microsyntactic annotation from scratch, we decided to enhance the existing SynTagRus corpus of Russian texts, developed by our Laboratory of computational linguistics at the A. A. Kharkevich Institute for Information Transmission Problems in Moscow. For the recent state-of-the-art of SynTagRus, see [12]. Even though this corpus is not too large (it now contains about 1 million word tokens), it has several layers of annotation, including markup for (1) morphology, (2) syntax (in the formalism of dependency trees), (3) lexical senses (for words whose ambiguity is reflected in the underlying Combinatorial dictionary of Russian),[5](4) parametric lexical functions (in the sense of Meaning Û Text by Igor Meľčuk [14]), (5) certain types of ellipsis and, recently, (6) anaphoric relations: the latter are currently traceable beyond the sentence level so that the antecedents of pronouns can be found either in the same sentence or in a text fragment comprising two preceding sentences (see [15], [16].)

Microsyntactic tagging is thus the seventh layer of SynTagRus markup.

2.1 Purpose of Microsyntactic Tagging

What is the purpose of creating this markup? It is a commonly known fact that a corpus annotated for lexical senses of words is a valuable linguistic resource

[4] A typical *ad hoc* solution is representing a multiword microsyntactic element as a single word, e.g. represent the sequence *in fact* as an unsegmented unit, ignoring cases where it is not, as in *in fact checking* or where it is part of a longer set phrase like *in fact or spirit*.

[5] We also held experiments of supplying SynTagRus with semantic markup on the basis of Juri Apresjan's system of fundamental classification of predicates (see [13]), but this markup is not maintained now.

instrument in solving many sophisticated theoretical and practical tasks, including those associated with theoretical semantics, monolingual and bilingual lexicography, WSD, and deep semantic analysis. In many cases, microsyntactic elements are polysemous, so, in a way, microsyntactic markup is close to lexical sense annotation.

Text corpora annotated for senses of words are few for many languages, including Russian, and they are seldom large; see e.g. [17] for the Russian equivalent of the SemCor corpus annotated with WordNet word senses (see [18], [19] for details).

We may be disappointed with the fact that such corpora are scarce and small, but at least they exist for standard words and are available for researchers. However, there have been no corpus resources at all so far that could provide markup for syntactically challenging phraseological units, including, of course, microsyntactic units. This means that the reported resource is, in all probability, the first one of its kind.

It must be noted that, over the last couple of years, considerable time and effort has been devoted by corpus developers to annotate text corpora of a variety of languages for multiword expressions (MWE) (see e.g. a recent overview [20], with extensive bibliography, and a comprehensive paper [21] on corpus annotation with verbal MWEs – specifically, light verb constructions of various types). It may seem, at first glance, that our research exactly falls within MWE annotation framework. Yet our goal is more specific and, in a way, more ambitious: we focus on linguistic units that have considerable syntactic specificity and strive to present their internal syntactic arrangement and determine how these units are incorporated into the sentence structure.

As a matter of fact, microsyntactic markup of the corpus is not an easy task. On the one hand, it is difficult to discriminate between a microsyntactic element and an arbitrary sequence of words, which may even span over different syntactic chunks. On the other hand, there exist no ready lists of microsyntactic units that could be viewed as exhaustive, or even representative. The available phraseological dictionaries provide no good approximation: most of the traditional idioms present in such dictionaries have no distinctive syntactic properties and cannot be considered as microsyntactic units, while many such units do not appear in such dictionaries.

2.2 Two Strategies

To make up for this lack of initial data, we used two different tactics of tagging SynTagRus for microsyntactic elements:

1) continuous analysis of whole individual texts, aimed at finding all candidates to microsyntactic elements within this text;
2) targeted search for linear strings and/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’, *kak by* ‘sort of’, *vse že* ‘yet’, *kak raz* ‘exactly, namely’, *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* ‘one never knows; 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,…)', *to i delo* 'ever so often', *to li delo* 'how much better', etc.[6]

Sure enough, in both cases only manual annotation of text for microsyntactic elements was possible: partial automation of microsyntactic elements could be done at the first stage of tagging in cases where strings of words constituting such elements had no gaps in between, with subsequent careful editing.

Using both tactics, we were able to obtain draft versions of microsyntactic markup of the corpus fragments, which were later subjected to thorough expert linguistic analysis, which revealed, among other things, that the number of microsyntactic elements occurring in the text is quite considerable. In a considerable number of texts, as many as a quarter of sentences were found to contain at least one microsyntactic element, so the microsyntactic markup turns out to be a frequent corpus feature.

Fig. 1 and Fig. 3 below are screenshots presenting the results of microsyntactic markup obtained by the two tactics. Fig. 1 shows the annotation for a fragment of a running journalistic text called *Kul'turnye olimpijtsy* 'Cultural Olympians'. The text, published by the popular Moscow *Novaya gazeta* newspaper in 2013, is a typical sample of SynTagRus material. It consists of 132 sentences, of which 33 (exactly 25%) were found to contain at least one microsyntactic element.

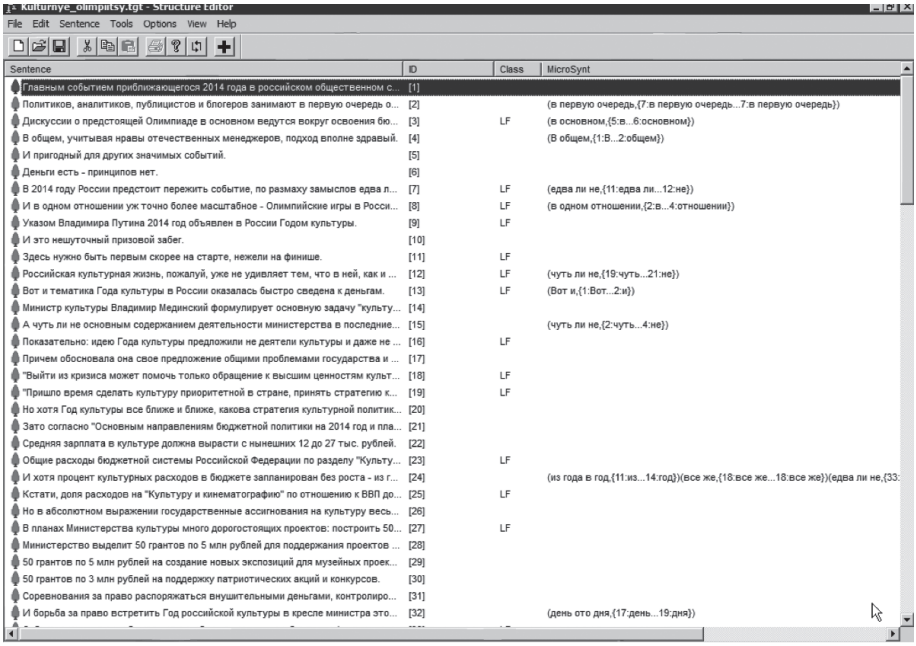

Fig. 1. Microsyntactic markup of a running text

[6] In order to avoid extended discussion, we list only one or two English equivalents for any microsyntactic units cited. Interestingly, in almost all of the above cases Russian microsyntactic units correspond to multiword English microsyntactic units which we use as glosses. It can thus be hypothesized that the number of microsyntactic phenomena and their typology in various languages may be quite comparable.

Currently, the markup looks as follows: a special field in the XML file representing the text cites the name of the microsyntactic element (in the case of syntactic idioms, it is normally a string of words, possibly with a figure attached to it if the syntactic idiom happens to be ambiguous) and the linear segment containing this element. For instance, a rather long sentence (24) of this text

I xotja procent kul'turnyx rasxodov v bjudžete zaplanirovan bez rosta ***- iz goda v god*** *1,5% – on* ***vse že*** *vdvoe vyshe, naprimer, čem procent rasxodov na fizkul'turu i sport, kotorye kažutsja nekotorym publicistam* ***edva li ne*** *glavnym prioritetom sovremennoj Rossii* 'And although the percentage of cultural expenditure in the budget is planned without growth – 1.5% from year to year – it is still twice as high, for example, than the percentage of spending on physical education and sports, which seem to some publicists to be almost the main priority of modern Russia'

contains three microsyntactic units (shown in boldface) – *iz goda v god* 'from year to year', *vse že* 'yet' and *edva li ne* 'almost'. In order to see how these units are incorporated into the syntactic structure, one needs to see the syntactic tree and identify the elements of the syntactic idioms as part of this tree.

Fig. 2 below shows a fragment of the syntactic tree for the above sentence with the first of the syntactic idioms discussed – *iz goda v god:*

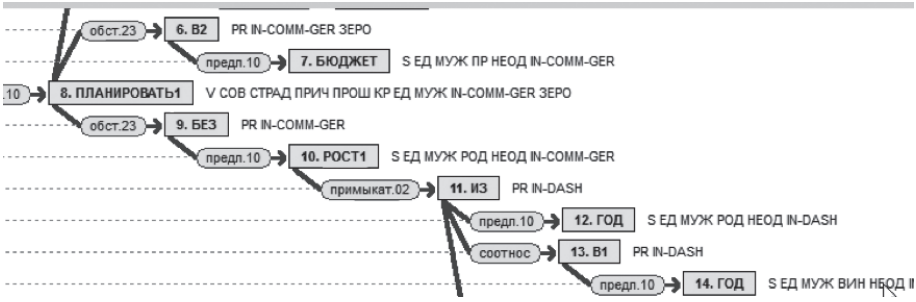

Fig. 2. A fragment of the syntactic tree structure containing a microsyntactic unit

It can be seen that the syntactic idiom occupies the nodes from 11 to 14, its local head, the preposition *iz* 'from', is dominated by the noun *rost* 'growth' and subordinates the noun *god* 'year' using the prepositional syntactic relation. The other prepositional phrase of this idiom (*v god* 'to year') is dominated by the first preposition with the correlative syntactic relation. So, the internal arrangement of the syntactic idiom within the structure has to be determined additionally: if the two prepositional phrases formed no such idiom, both prepositions would be most likely dominated in parallel by a verb or other predicate word.[7]

Fig. 3 below represents the second approach to microsyntactic annotation – the targeted search for possible microsyntactic units. In this case, we searched for sentences that are likely to contain a syntactic idiom *stalo byt'* 'hence, consequently'. The query for this unit (functioning as a parenthetical adverb despite being composed

[7] The syntactic representation of SynTagRus follows the conventions of the ETAP-3 parser (see [22], [23]), which in its turn heavily relies on the syntactic component of the Meaning ⇔ Text theory [14].

of two broad semantics verbs) was simple: find sentences with the wordform *stalo* (‘which is the neuter gender singular of the past tense of the verb *stat’* ‘begin’) followed by the wordform *byt’* (the infinitive of the verb *byt’* ‘be’).

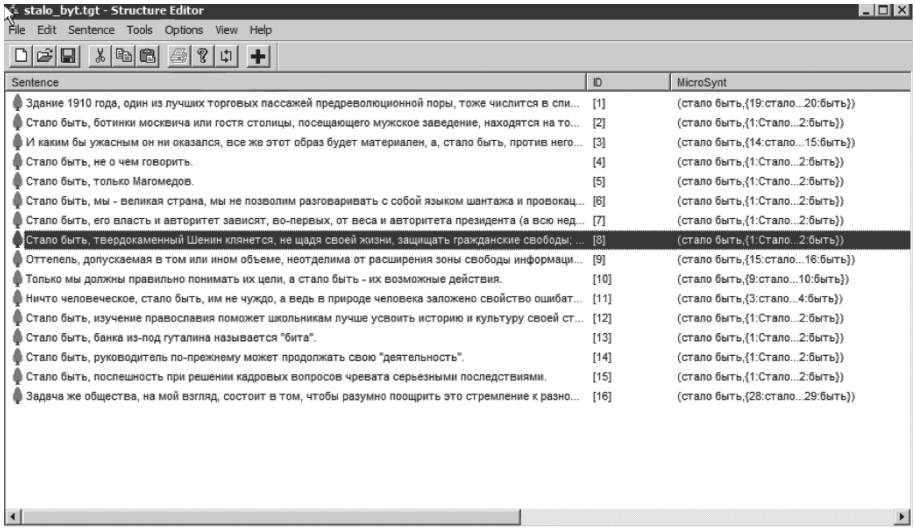

Fig. 3. Microsyntactic Markup of SynTagRus sentences with the unit *stalo byt’*

As seen from the screenshot, all 16 sentences satisfying the search query were tagged for the unambiguous microsyntactic unit *stalo byt’*. This means that, within the corpus, no sentences could be found in which the string *stalo byt’* meant something different (a random juxtaposition of the two wordforms, or a different phrase). It can be conjectured that this binary unit is very stable in the language, effectively excluding other lexical competitors. This hypothesis is easily confirmed by a search for the same string in a much larger corpus (the Russian National Corpus at `www.ruscorpora.ru`): we could find, using rather sophisticated contexts, only a very few sentences in which this string proved to be unrelated to our syntactic idiom. One such sentence, *No kogda by ni žil, nado vo čto by to ni stalo byt’ čestnym čelovekom* (Venedikt Erofeev) ‘Whenever one lives one needs by any means to be an honest man’ happened to have a phrase boundary between *stalo* (which, amusingly, was part of a different syntactic idiom – *vo čto by to ni stalo* ‘at whatever cost, by any means’) and *byt’*. Actually, all other occurrences of the string in the large corpus followed the same pattern as found in SynTagRus.

3 FIRST RESULTS

Even though regular microsyntactic tagging of the SynTagRus corpus was started only a few months ago, a number of linguistically interesting results could already be found.

1. Despite the fact that SynTagRus has a relatively small size, it proved to be quite representative of microsyntactic phenomena. Most microsyntactic elements tagged according to the second tactics of preliminary search for promising occurren-

ces could actually be detected (although some of them could naturally be represented by several examples only).

2. The extent of ambiguity of microsyntactic elements was found to vary significantly from one unit to the other.

Some elements proved to be quite homogenous. In addition to the case considered above (with *stalo byt'*), another microsyntactic unit, *kak byt' (s chem-libo)* 'what to do (about something)' shared the same property of being (almost) unambiguous, and never occurring in extraneous contexts in the SynTagRus corpus (in fact, it requires a lot of linguistic inventiveness to find relevant examples of *kak byt'* falling outside of the syntactic idiom considered (see [10] for more detail).

At the same time, other microsyntactic units proved to be highly ambiguous within the corpus. Moreover, words constituting them occurred in contexts totally unrelated to any of the unit's senses, providing many false positives during the markup. An illustrative example is the ramified set of microsyntactic units *kak by*, which had a number of senses and generated a host of false positives (see the screenshot of Fig. 4 below).

On the one hand, there is a microsyntactic unit which we will refer to as *kak by 1* ≈'sort of': this is a discourse particle with the semantics of comparison or uncertainty, as in sentence (97) from the screenshot in Fig.4: *Takim obrazom, nastupalo kak by ravnovesie* 'Thus, a kind of balance was established'.

On the other hand, there is an entirely different microsyntactic unit, 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 the predicative adverbs *strashno, bojazno* 'fearful', as in sentence (109) from the same screenshot: *Potom ja zamatyvalas' šal'yu i uxodila ne oboračivayas', boyas', kak by mne ne predložili deneg za niščiy vzgljad* (I.Grekova) 'Then I wrapped myself in a shawl and left without turning around, being afraid that I would be offered money for my beggarly look'.

Yet another syntactic idiom composed of *kak* and *by* is a modal sentential adverb that implicitly expresses the speaker's wish – *kak by 3*. It is represented in such corpus sentences as *Kak by v kamennyj vek ne skatit'sja* 'It would be good not to slide back into the stone age' or *Kak by obojtis' bez etogo, ostaviv samuju sut'* [A.Bitov] 'I wish we could manage without it, leaving only the most crucial thing'.

In addition to these senses (plus a set of microsyntactc units which are longer than *kak by* and have to be distinguished from the above units), SynTagRus has a number of sentences that do not involve microsyntactic units formed with *kak by* despite the fact that physically this string is present. In particular, some sentences contain the construction with the emphatic particle *ni*: *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'. We believe that in such cases a good solution is to leave the sentence marked-up, introducing a "false positive" tag. Such a solution may seem a controversial one as it is not routinely applied in corpus annotation. I believe, however, that it may be very helpful not only as a provisional step at preparatory

stages of corpus annotation but as a clear indication of the fact that the respective string does not form an idiomatic unit and represents a free juxtaposition of words otherwise belonging to such a unit. It may be viewed as a sort of negative linguistic material (in the sense of the Russian scholar Lev Shcherba), which can provide interesting linguistic insight for the grammarian and the lexicographer alike.

3. Normally, SynTagRus is representative enough of the most syntactic idioms having the same "lemma" name. However, to be sure that we have not missed anything, additional search is recommended for really ambiguous entities. For the *kak by* host of idioms, we were able to find one more 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:

— *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 meaning of this idiom (*kak by 4*) can be explained as follows: '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*.

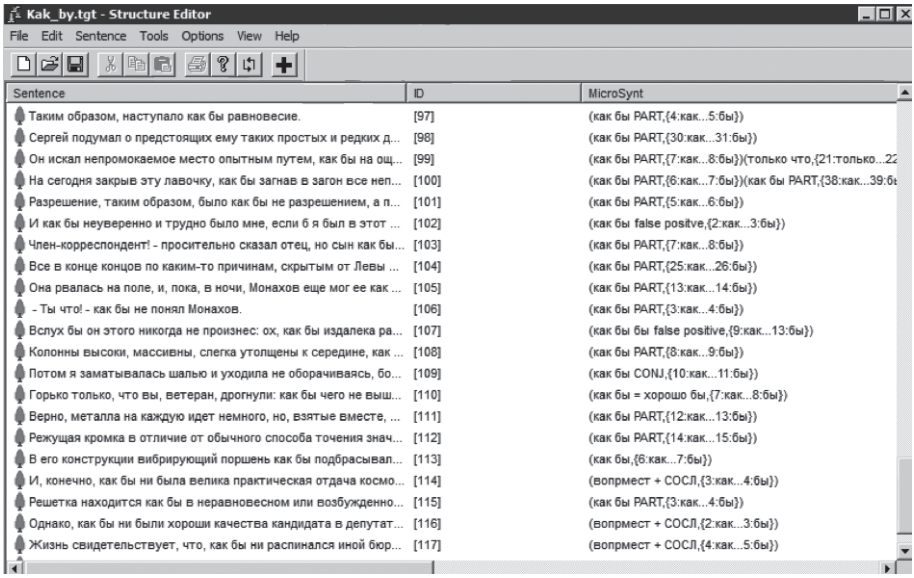

Fig. 4. Microsyntactic markup of a SynTagRus fragment containing sentences with the string *kak by*

4. The material of syntactic idioms present in SynTagRus provides us with valuable data on linear variations of these idioms, their syntactic structure, their obligatory and optional valencies, and most importantly, their unique semantic features, which should be thoroughly accounted for in the resources like Microsyntactic dictionary. We intend to use this opportunity to the fullest extent possible.

References

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

[2] Iomdin, L. L. (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.] In *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*, pages 423–436, Verlag Otto Sagner, München/ Berlin/Washington D.C. [In Russian.]

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

[4] Iomdin, L. (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. In *Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop*, pages 8–18, Osaka, Japan. Accesible at: http://aclweb.org/anthology/W/W16/W16-38.pdf.

[5] Fillmore, Ch. (1988). The Mechanisms of Construction Grammar. In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

[6] Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

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

[8] Lauwers, P. and Wettere, van N. (2017). La Micro-constructionnalisation En Tandem: La Copularisation De Tourner/virer. *Langue française*, 194(2):85–103.

[9] Rhodes, R. (2009). Tautological constructions in English … and beyond. Presented to the Syntax and Semantics Circle, UCB. Accessible at: http://linguistics.berkeley.edu/~russellrhodes/pdfs/syntax_circle_taut_qp.pdf.

[10] Iomdin, L. (2017). Kak nam byt' s konstruktsijami tipa *kak byt*? [What to do about constructions like *what to do*?] *Computational Linguistics and Intellectual Technologies. Dialogue 2017*, 16 (23)(2):150–161. [In Russian, Engl. Abstract.]

[11] Marakasova, A. A. and Iomdin, L. L. (2016). Mikrosintaksičeskaja razmetka v korpuse russkix tekstov SynTagRus [Microsyntactic tagging in the SynTagRus corpus of Russian texts.] In *Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj meždisciplinarnoj školy-konferencii IPPI RAN*, pages 445–449, Repino, Saint Petersburg, Russia. [In Russian.] Accessible at: http://itas2016.iitp.ru/pdf/1570285171.pdf .

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

[13] Apresjan, Ju., D., Iomdin, L. L., Sannikov, A. V., and Sizov, V. G. (2004). Semantičeskaja razmetka v gluboko annotirovannom korpuse russkogo jazyka. [Semantic Tagging in a deeply annotated corpus of Russian.] In *Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika – 2004»,* pages 41–54, Izd-vo Sankt-Peterburgskogo universiteta, Saint Petersburg, Russia. [In Russian.]

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

[15] Inshakova, E. S. (2016). Razrešenie sintaksičeskoj mestoimennoj anafory v sisteme «ETAP-3». [Resolution of syntactic pronominal anaphora in the ETAP-3 system.] In *Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj meždisciplinarnoj školy-konferencii IPPI RAN*, pages 420–429, Repino, Saint Petersburg, Russia. [In Russian.] Accessible at: http://itas2016.iitp.ru/pdf/1570282678.pdf.

[16] Marakasova, A. A. (2016). Avtomatičeskoe razrešenie anafory v russkom tekste: slučaj nulevogo sub"ekta. [Automatic resolution og anaphora in a Russian text: the case of a zero subject.] In *Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj meždisciplinarnoj školy-konferencii IPPI RAN*, pages 431–436, Repino, Saint Petersburg, Russia. [In Russian.] Accessible at: `http://itas2016.iitp.ru/pdf/1570285121.pdf`.

[17] Dikonov, V. G. and Poritski, V. V. (2014). A Virtual Russian Sense Tagged Corpus and Catching Errors In A Russian Û Semantic Pivot Dictionary. *Computational Linguistics and Intellectual Technologies. Dialogue 2014*, 13(20):128–137.

[18] Mihalcea, R. (1998). SemCor semantically tagged corpus, SenseEval 2 & 3 data in SemCor format. Accessible at: `http://www.cse.unt.edu/~rada/downloads.html`.

[19] Petrolito, T. and Bond, F. (2014). A survey of WordNet Annotated Corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–243, Tartu, Estonia.

[20] Rosén, V., Smedt, K. de, Smørdal Losnegaard, G., Bejček, E., Savary, A. and Osenova, P. (2016). MWEs in Treebanks: From Survey to Guidelines. In *Proceedings, LREC 2016, Tenth International Conference on Language Resources and Evaluation*, pages 2323–2330, Portorož, Slovenia.

[21] Savary, A., Sangati, F., Candito, M. et al. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain.

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

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