June 14-16, 2023

Russian Construction 2.0: New Features and New Perspectives of the Biggest Construction Ever Built

Alexander Orlov

HSE University alexander.orlov98@gmail.com

Daria Demidova

UiT The Arctic University of Norway dashademidova1998@gmail.com

Ekaterina Rakhilina

HSE University
Vinogradov Institute for Russian
language (Russian Academy of
Sciences)

rakhilina@gmail.com

Zoia Butenko

HSE University
UiT The Arctic University of Norway
zoiab@uio.no

Vladimir Starchenko

HSE University vsstarchenko@hse.ru

Olga Lyashevskaya

HSE University
Vinogradov Institute for Russian
language (Russian Academy of
Sciences)
olesar@yandex.ru

Abstract

Russian construction is an open-access linguistic database containing detailed descriptions of over 3,800 Russian grammatical constructions. In this paper we present a new, enlarged and updated version of Russian Constructicon (RusCxn) as well as new trajectories of development which were opened for the resource after the update. Since its first release, RusCxn, has undergone many significant changes. Our team has expanded the number of constructions present in the database 1,5 times, introduced new meta-information features such as glosses, significantly reworked the architecture and the design of Russian Construction's website, and improved the search facilities. The above-mentioned changes not only make RusCxn more attractive and convenient-to-use, but they can also greatly facilitate typological research in the field of Construction Grammar and improve the mapping between constructicography-orinented resources for different languages.

Keywords: Construction Grammar; construction; construction.

DOI: 10.28995/2075-7182-2023-22-378-385

Русский Конструктикон 2.0: Новые особенности и новые перспективы развития самого большого в мире конструктикона

Александр Викторович Орлов

Национальный исследовательский университет «Высшая школа экономики»

alexander.orlov98@gmail.com

Зоя Алексеевна Бутенко

Национальный исследовательский университет «Высшая школа экономики»
Университет Тромсё — Норвежский арктический университет zoiab@uio.no

Дарья Александровна Демидова Университет Тромсё — Норвежский арктический университет

dashademidova1998@gmail.com

Екатерина Владимировна Рахилина

Национальный исследовательский университет «Высшая школа экономики» Институт русского языка им. В. В. Виноградова РАН rakhilina@qmail.com

Владимир Миронович Старченко

Национальный исследовательский университет «Высшая школа экономики» vsstarchenko@hse.ru

Ольга Николаевна Ляшевская

Национальный исследовательский университет «Высшая школа экономики» Институт русского языка им. В. В. Виноградова РАН olesar@yandex.ru

Аннотапия

Русский конструктикон — это бесплатная электронная лингвистическая база данных, содержащая подробные описания более 3800 русских грамматических конструкций. В этой статье мы хотим представить новую, расширенную и обновленную версию Русского Конструктикона, а также рассказать о новых перспективах развития ресурса, открывшихся после масштабного обновления. С момента своего первого выпуска Русский Конструктикон претерпел множество значительных изменений. Наша команла в 1.5 раза увеличила количество представленных на ресурсе конструкций, разработала новые типы мета-информации для описываемых конструкций, в частности глоссы, значительно переработала архитектуру и дизайн сайта Русского Конструктикона, а также улучшила механизм поиска. Эти изменения не только делают Русский Конструктикон более привлекательным и удобным в использовании, но также могут значительно облегчить типологические исследования в области грамматики конструкций и улучшить связь между конструктографическими ресурсами для разных языков.

Ключевые слова: грамматика конструкций; конструкция; конструктикон.

Introduction 1

In this paper we introduce a new, enlarged and upgraded version of "The Russian Construction (RusCxn)" as well as discuss the new prospects which became available for the resource after the upgrade.

1.1 **Basic terms**

The term Construction denotes both a system of constructions of a particular language, and a detailed description of this system, normally presented in a form of a searchable database.

Construction is a key term employed by Construction Grammar (CxG), which assumes that constructions are fundamental building blocks of a human language [1], [2], [3], [4]. Under this theory, any linguistic pattern or model can be recognized as a construction "as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist" [1: 5] – in other words, constructions are understood as fixed linguistic patterns lacking compositionality¹. Thus, grammatical structures as small in scope as prefixation and as large in scope as passive voice can be viewed as constructions. While constructions for some languages (i.e., English, Brazilian Portuguese) store all linguistic patterns compatible with the definition above, many prefer to narrow the object of their research, focusing only on grammatical constructions - predominantly multiword linguistic patterns, which (1) lie on the border of lexis and grammar and (2) which are partially schematic $[6], [7].^2$

¹ See more on compositionality in [5].

² Yet it is worth noting that such type of constructions better be called *quasi-grammatical* since, as stated in their definition, they express meanings which combine grammatical elements with the lexical ones. Albeit the term grammatical construction has been used in CxG works to denote such quasi-grammatical constructions since the release of [6], and we are not going to dispute it in this paper.

In the Russian Construction we also mainly focus on grammatical constructions. The examples of such constructions include *NP-Nom Cop umo надо*³ 'NP-Nom Cop what needed' – a construction signifying superior quality of NP, or *без пяти минут NP* 'without five minutes NP' – a construction meaning that NP will in the nearest future experience a change in its (social) status. Please note that both constructions express meanings which are neither fully lexical nor fully grammatical, and that neither meaning is clearly entailed from the component parts of a construction.

For the sake of simplicity, hereby we will use the term *construction* to specifically denote grammatical constructions as defined in [6].

1.2 Russian Constructioon 1.0

First presented to public in 2020 [7], the Russian Construction is a joint project of HSE University and The Arctic University of Norway. Upon its release, it became the largest construction of any language, featuring over 2,000 constructions. It was also among the most functional. In RusCxn, each construction was accompanied by a substantial amount of meta-information, incl. definition, examples, semantic type of a construction, syntactic structure of a construction, etc. (see more in [8]). RusCxn also became one of the two constructions (the other being Swedish) which were specifically targeted not only towards the academics but also towards the L2 students of a particular language, providing CEFR levels and easy-to-read definitions of the constructions stored.

Clearly, with such a swift start, RusCxn had potential for becoming an exemplary construction. Yet, obviously, RusCxn came not without its own flaws. Some of these flaws, e.g., lack of glossing system, inability to form unique URLs for individual constructions, etc., seriously hindered both current usability and future perspectives of the resource. Therefore, after delivering the first version of the database to the public, RusCxn team continued efforts on the project.

Our team, which has expanded over time, implemented a number of fundamental changes in the Russian Construction since its first release: we have greatly increased the number of constructions present in the database, filled the gaps in RusCxn's instrumentarium and addressed some significant faults in the design of RusCxn's website. Hereby, we will refer to the new and upgraded version of RusCxn as 'The Russian Construction 2.0' (as opposed to RusCxn 1.0 version delivered in 2020).

This article aims to describe the biggest changes in the second iteration of the project. Section 2 addresses improvements on the 'theoretical side' of the Russian Construction, with 2.1 reporting on the increase in the number of constructions described in RusCxn and 2.2 - on the introduction of glossing system. Section 3 highlights the improvements on the 'computational side of thing', namely the launch of a new website for the resource. Section 4 describes further perspectives of the Russian Construction which became available after the big update.

2 RusCxn 2.0: contents update

2.1 Expanding the Number of Constructions Described

In the Russian Construction 2.0 the number of featured constructions increased from 2,200 to 3,800.

The new constructions originated from the following sources: (1) a list of phrases to depict manner, retrieved from ruscorpora.ru (~2,500 entries) [9], (2) Thesaurus dictionary of the Russian idioms [(>8,000 entries) [10], (3) a list of constructions collected manually from the Russian fiction books (~600 entries).

All the entries were manually examined by several annotators for compliance with our criteria. To begin with, it was necessary to make sure that the constructions under examination were not already present in our database (i.e., did not match the constructions from RusCxn 1.0). Thus, a significant number of entries (~800) were eliminated on the first stage because of repetition.

On the next stage, the units under consideration had to be checked for compliance with the definition of the construction adopted in RusCxn. A substantial number of entries from sources 1 and 2 (\sim 6,000) were eliminated at this stage. For instance, [10] contains a large number of proverbs and sayings that do

³ In the Russian Construction we developed a special system of notation for construction formulae. NP(-Nom) = noun phrase (with a noun in the nominative case), Cop = copula. See more in [7]. Hereby all examples of constructions are presented in accordance with how they appear on RusCxn's website, i.e., without transliteration.

not form constructions according to our definition, as they lack a free slot. In addition, lexical constructions, the semantics of which go well beyond the framework of quasi-grammatical meanings explored by RusCxn, were also excluded (cf. *VP в чем мать родила* 'lit. VP in which the mother gave birth' signifying *naked* – a fully lexical meaning). The process of annotation of phraseological units from [10] is described in more detail in [11].

Finally, we controlled for the frequency of use and stylistic coloring of the remaining constructions. Thus, we did not include into the final update constructions which have fallen out of use through time or were rather rare (as demonstrated by the data from Russian National Corpus). Because of the formal constraints imposed on us by the pedagogical nature of the resource, we were also unable to include any constructions that might appear rude or explicit to our users⁴.

Subsequently, more than 1,600 constructions were added to the new version of the Russian Construction, increasing its volume by more than 50 per cent. These constructions are annotated in accordance with RusCxn rules and should be available to the public via a new website by the date of publication.

2.2 Introducing the Glossing System⁵

With definitions of constructions in multiple languages and semantic equivalents of Russian constructions in English and Norwegian, the Russian Construction was meant to be an internationally oriented resource since its release. Nevertheless, the usability of RusCxn 1.0 for non-Russian speaking linguists was significantly hindered by the lack of glosses or any other device non-Russophones could use to understand the inner structure of constructions. Therefore, in RusCxn 2.0, we developed a glossing system.

2.2.1 Glossing Format

Our glossing system is based in the Leipzig glossing format [12]. These are some examples of glossed constructions from the Russian Construction:

- (1) ко-му как-ое дел-о Сор до NP-Gen who-DAT.SG which-NOM.SG.N deal-NOM.SG Сор to NP-Gen
- (2) пош-л-и/пойд-ём VP-Pfv.Fut/VP-Ipfv.Inf go-PST-PL/go-FUT.1PL VP-Pfv.Fut/VP-Ipfv.Inf
- (3) больно Adv/Adj/Pred too<painfully ADV/ADJ/PRED

We describe the glossing rules in detail in [13]. In this article, we address only some key features of our system.

To begin with, in our glossing system, we utilize the symbol < for translating roots with multiple meanings, provided the literal/original meaning of a root is distinct from the contextual meaning. Consider word больно in (3): even though it originally means *painfully*, in this particular construction it is used as an intensifier, better translated as *too*. Under our glossing rules literal or original meaning of a root should appear to the right of <, whilst contextual meaning – to the left.

In addition to this, symbol < can be used to convey the origins of some function words. Cf. second *umo* in (4) – a complementizer originating from a word meaning *what*, or *uuub* in (5), – a particle that comes from PRS.2SG of a verb *видеть* 'see'[14].

- (4) ну и что, что XP PTCL and what COMP<what XP
- (5) ишь, как-ой Adj-Nom Cop! PTCL<see{OBSOLETE}.PRS.2SG which-NOM Adj-Nom Cop

In the Russian Construction 2.0 we do not provide translation equivalents for complementizers, interjections, or particles, as it is rarely possible to find an exact equivalent for such words; yet we believe that preserving their source-meanings, where possible, may prove useful for some researchers.

Unlike many other glossing systems, one adopted at RusCxn 2.0 aspires to retain the original stylistics of roots and words glossed through the use of special stylistic labels. We currently have two labels:

⁴ Given that the constructions excluded on stages 2 and 3 might still be of interest to some researchers, we plan to build a separate resource for hosting such entries.

⁵ All the additions described in this section are to appear on the new version of the site by the date of publication.

OBSOLETE, marking usages which sound old-fashioned or which are no longer used in speech on their own, such as *cezo* in (6), and SUBSTAND, marking colloquial and/or inappropriate in a written language usages such as *a nom* in (6).

(6)ни c того ни c се-го C1 NEG from that.GEN.SG NEG this {OBSOLETE}-GEN.SG Cl from (NP-Dat) VP-Inf (7) Cop В лом (NP-DAT) **COP** in unwillingness {SUBSTAND} < break VP-INF

Glosses will occupy a separate section in the description of a construction. Since glosses are of interest to a limited number of users only, they will be displayed in an advanced meta-information section.

2.2.2 The Great Slash Problem – Resolved

In RusCxn slash ('/') is utilized to denote that some elements of a construction are in a free variation with each other, e.g., in *NP-Gen.Pl Cop выше крыш-и/голов-ы* 'lit. NP-Gen.Pl Cop higher than the roof/head' speaker can say either *roof* or *head* without any significant difference in meaning. Such notation, convenient at a first glance, becomes rather problematic when variative parts of a given construction consist of more than one word. Consider (8) where slash suddenly denotes free variation not between neighboring words *ущерб* and *во*, but rather between phrases *в ущерб* and *во вред*.

VPвущерб/вовредNP-DatVPindamage.ACC.SG/inharm.ACC.SGNP-DAT

The extreme example of this is construction (9), where multi-layered variation is found: first, *Pred* interleaves *Adj.Short*, and then the string of words *μe maκ* (y) *Pred/Adj.Short* interleaves *μe maκ* y) *We u Adv VP*.

(9) NP-Nom Cop не так (уж и) Pred/Adj.Short/не так уж и Adv VP NP-Nom Cop NEG so (FOC.PTCL and) Pred/Adj.Short/NEG so (FOC.PTCL and) Adv VP

Even though notations as (8) and (9) appear comprehensible to native speakers, they might be difficult to understand for L2 learners of the language because of the lack of clearly marked borders for variative parts, not to mention that such inconsistent notation is obviously unfit for automatized computational analysis. At the same time, marking borders for variative parts with square brackets or alike devices could pose a problem of its own, since, with too many notation symbols, construction formula risks to become unreadable.

After the introduction of a special section for glossing, this problem was partially resolved. Square brackets to mark the borders of variative parts were introduced in all construction formulae with slashes. Thus, (8) and (9), under new notation will look like this in the database:

- (10) VP [в ущерб]/[во вред] NP-Dat VP [in damage.ACC.SG]/[in harm.ACC.SG] NP-DAT
- [NP-Nom Cop не так (уж и) [Pred]/[Adj.Short]]/[не так (уж и) Adv VP] [NP-Nom Cop NEG so (FOC.PTCL and) [Pred]/[Adj.Short]]/[NEG so (FOC.PTCL and) Adv VP]

To avoid overloading the construction entry with special symbols, the square brackets will not be visible in the main line introducing construction formula on the website, yet they will be visible in the gloss section.

2.2.3 Gloss Search

The Russian Construction 2.0 will feature a search facility for glosses, which will be part of Advanced search. This facility will allow to search constructions containing an individual gloss or a combination of glosses located at a particular distance from each other. A gloss can be represented by a translation (typed manually), a grammatical category (chosen from a pre-existing list), or a combination of a translation and a grammatical category. Translation part also allows for the use of regular expressions (e.g., one can find all constructions containing words glossed with 'put' and 'lay' with one query 'put|lay'). The gloss search facility will be available via *Advanced Search* tab of our website and can be combined with other search facilities present in advanced search (semantic type of construction, syntactic structure of anchor, CEFR level, etc.).

3 RusCxn 2.0: Website Overhaul

The data from RusCxn 1.0 was available to users through a website built upon Github. The resource worked rather slow since the site had to cache all the data from a pre-made Google Spreadsheet with constructions at every opening. In addition to that, since the searching process was carried out at the expense of the front-end, the issuance of results was performed on the same webpage with a single unchanged URL. The descriptions of individual constructions also did not have individual URLs since this information was not pre-stored in any kind of a database. Thus, it was not technically possible to directly link RusCxn's data with the data from its other satellite resources, e.g. Constructesize![15], containing exercises on the constructions, or Pragmaticon[16], containing related discourse formulas [17]. It was also unfeasible to provide a direct link to a particular construction in the description of a different construction, for example, to signal their similarity or synonymity, or to formally depict construction families [18] in the database. To sum up, the RusCxn 1.0's website design was inconvenient for both ordinary users and academics, seriously limiting research possibilities and general perspectives of the resource.

In Russian Construction 2.0 we resolved the above-mentioned inconveniences by developing a totally new web-platform for the project. The new site is based on an SQL database. The resource is currently hosted ruscorpora.ru and is available through https://constructicon.ruscorpora.ru/. New design allows for assignment of unique URLs to each page with the description of a construction and to each search query; it also allows for a more swift and efficient processing of the data.

In addition to the creation of a new back-end for our resource, we introduced significant changes to the front-end. To accommodate new users, we added two sliders (in Russian and English) containing explanatory information about the resource. Each slider answers five basic questions about RusCxn in a simple and vivid language with several examples. The questions are *What is a construction?*, *What is construction?*, *What is the purpose of the Russian Construction?*, *What can you find here?*, and *Who built this resource and how?*. Now we also display sample queries in a search bar (cf. *He говоря o* in Fig.1) to better familiarize new users with the format of queries and the content of the resourse.

We additionally enhanced the appearance and the general usability of the website by changing a color scheme, text font, and a configuration of the plain text and widgets throughout the resource. For instance, the output window on the main page is now located under the search bar and is reduced in size to give way for the slider (yet the results are more readable than before due to the darkened color of the text and the font which prevents amalgamation).

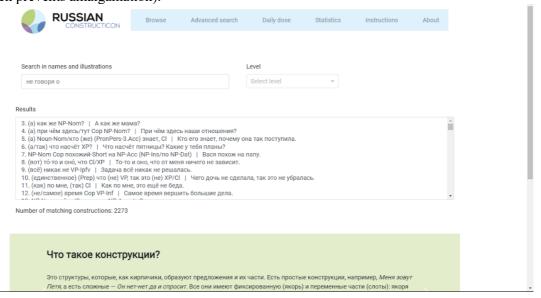


Figure 1: The main page of Russian Construction after the update

On top of that, we fixed several bugs in the searching mechanism, completely redesigned the Advanced search, and significantly changed the display of meta-information for a construction. Additionally, an option to choose language was introduced at the top of relevant pages (previously, the site would

feature duplicative sections like *Instructions Russian* and *Instructions English* in its header). The contents of the text pages were also rewritten to improve their comprehensibility.

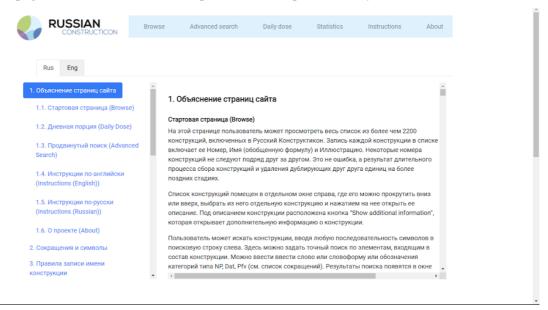


Figure 2: The instructions page of Russian Construction after the update

4 Instead of conclusion: New Perspectives for the Russian Construction

The changes implemented in the second version of RusCxn not only made it more attractive and convenient to use for both researchers and L2 learners, but also opened new prospects for improved integration with the constructions for other languages and with the satellite resources of the Russian Construction.

First, we shall discuss how improvements in RusCxn 2.0 may facilitate typological research in Construction Grammar and improve connectivity between constructions for different languages. The talks about somehow 'aligning' different databases with constructions to foster typological research in the field have been around since the first major conference on constructicography [19], as few cross-linguistic studies of constructions that existed at that time proved to be rather useful for both theoretical and applied linguistics [20]. Yet, up to date, there still exists no device or platform that could facilitate typological study of constructions from different languages. We reckon that such a platform should be based on a universal system of glossing, and we are happy to be the pioneers in this field. Even though currently RusCxn remains the only fully glossed resource of its kind, our team is actively working on Hill-Mari, Persian and Ukrainian constructions, which all have the same architecture and, thus, will be easily mappable to each other, provided they also have glosses. We hope that researcher teams that work on constructions for other languages will also join our endeavor, so that we can create a big typologically oriented platform for conducting constructicography studies at a fundamentally new level.

Besides that, we shall talk about improved cross-connectivity between RusCxn and other resources targeted at Russian constructiography, such as Constructesize![15], Pragmaticon[16], and Diachronicon (in development). These platforms include much data directly connectable to the constructions from RusCxn: exercises on constructions for L2 learners, diachronically related discourse formulas, and history and origins of some Russian constructions respectively. Nevertheless, in RusCxn 1.0 we were unable to easily map these data because of the lack of unique URLS for our constructions. Now, with a new website architecture, we can conduct studies involving these platforms more easily.

All in all, the Russian Construction has been in development for over eight years. In this article we present a second iteration of the resources, enlarged and updated. In the future we shall continue working on the project to remain on the cutting edge of constructicography with the largest and (possibly) the greatest construction ever made.

This work was partially carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15- 2020-793: "Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research".

Funding

This work was partially carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793: "Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research".

References

- [1] Goldberg, A. E. (2006), Constructions at Work: The Nature of Generalizations in Language, Oxford University Press, Oxford.
- [2] Fillmore, Ch. J., Kay, P., O'Connor, M. C. (1988), Regularity and idiomaticity in grammatical constructions: The case of let alone, Language, Vol. 64(3), pp. 501–538.
- [3] Croft, W. (2001), Radical Construction Grammar, Oxford University Press, Oxford.
- [4] Rakhilina, E. V. (ed.) (2010), Linguistics of constructions [Lingvistika konstrukcij], Azbukovnik, Moscow.
- [5] Szabó, Z. G. Compositionality // The Stanford Encyclopedia of Philosophy (Fall 2022 Edition).
- [6] Ehrlemark, A., Johansson, R., Lyngfelt, B. (2016), Retrieving Occurrences of Grammatical Constructions, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp. 815–824.
- [7] Endresen A. A., Zhukova V. A., Mordashova D. D., Rakhilina E. V., Lyashevskaya O. N. (2020), The Russian Constructicon: A New Linguistic Resource, Its Design and Key Characteristics [Russkij Konstruktikon: novyj lingvisticheskij resurs, ego ustrojstvo i specifika] // Proceedings of the International Conference "Dialog 2020" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2020", Moscow. Pp. 241-255.
- [8] Janda, L. A., Lyashevskaya, O., Nesset, T., Rakhilina, E., Tyers, F. M. (2018), A construction for Russian: Filling in the Gaps, Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (eds.), Constructicography: Construction development across languages, John Benjamins, Amsterdam, pp. 165–181
- [9] The Russian National Corpus (ruscorpora.ru). 2003—2023.
- [10] Baranov, A. N., Dobrovol'skij, D. O., Kiseleva, K. L., Kozerenko, A. D., Voznesenskaya, M. M., & Korobova, M. M. (2007), Thesaurus dictionary of the Russian idioms [Slovar'-tezaurus sovremennoj russkoj idiomatiki], Avanta+, Moscow.
- [11] Rakhilina E. V., Zhukova V. A., Demidova D. A., Kudryavceva P. S., Rozovskaya G. P., Endresen A. A., Janda L. A. (2022), Phraseology in the Perspective of Russian Constructicon [Frazeologiya v rakurse «Russkogo konstruktikona»] Vinogradov Institute for Russian language open series, Vol. 2 (32). Pp. 13-44.
- [12] Leipzig Glossing Rules (2008), Conventions for interlinear morpheme-by-morpheme glosses. URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.
- [13] Demidova D., Makarova P. (forth.) Glossing Russian Construction. Arctic University of Norway open series, in prep.
- [14] Vasmer, M. (1953), Russisches etymologisches Wörterbuch/1 A-K.
- [15] Endresen, A., Zhukova V., Lonshakov G., Demidova D., Kalanova N., Bjørgve E., Lavén D. H., Janda L. A., Butenko Z., Perevoshchikova T. (2022), Con-struxercise! Hands-on learning of Russian constructions. A digital educational resource. https://constructicon.github.io/construxercise-rus/
- [16] Buzanov, A., Bychkova, P., Molchanova, A., Postnikova, A., & Ryzhova, D. (2022). Multilingual Pragmaticon: Database of Discourse Formulae. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3331-3336).
- [17] Bychkova, P., & Rakhilina, E. (forth.). Towards pragmatic construction typology: The case of discourse formulae // A. Barotto & S. Mattiola (Eds.), Discourse phenomena in typological perspective. John Benjamins Publishing Company, in press.
- [18] Endresen, A., & Janda, L. A. (2020), Taking construction grammar one step further: Families, clusters, and networks of evaluative constructions in Russian. Frontiers in Psychology, 11.
- [19] Lyngfelt, B., Torrent, T. T., Laviola, A., Bäckström, L., Hannesdóttir, A. H., & da Silva Matos, E. E. (2018), Aligning constructicons across languages: A trilingual comparison between English, Swedish, and Brazilian Portuguese. // Constructicography. John Benjamins, New York. pp. 255-302.
- [20] Boas, H. C. (2010). Contrastive studies in construction grammar, John Benjamins, New York.