How to build a constructicon in five years

The Russian example

Laura A. Janda[1] Anna Endresen[1] Valentina Zhukova[2]
Daria Mordashova[3,4] and Ekaterina Rakhilina[2,5]
[1] UiT The Arctic University of Norway | [2] National Research University Higher School of Economics | [3] Institute of Linguistics of the Russian Academy of Sciences | [4] Lomonosov Moscow State University | [5] Vinogradov Institute for Russian language of the Russian Academy of Sciences

We provide a practical step-by-step methodology of how to build a full-scale constructicon resource for a natural language, sharing our experience from the nearly completed project of the Russian Constructicon, an open-access searchable database of over 2,200 Russian constructions (https://site.uit.no/russian-constructicon/). The constructions are organized in families, clusters, and networks based on their semantic and syntactic properties, illustrated with corpus examples, and tagged for the CEFR level of language proficiency. The resource is designed for both researchers and L2 learners of Russian and offers the largest electronic database of constructions built for any language. We explain what makes the Russian Constructicon different from other constructicons, report on the major stages of our work, and share the methods used to systematically expand the inventory of constructions. Our objective is to encourage colleagues to build constructicon resources for additional natural languages, thus taking Construction Grammar to a new quantitative and qualitative level, facilitating cross-linguistic comparison.

Keywords: constructicon, construction grammar, Russian, corpus

1. Why build a constructicon?

If you are a linguist working on individual constructions in a language X, you might wonder why one should bother building a constructicon resource, and even if you accept this challenge, you might wonder where to start, how to proceed, and how to organize this endeavor.

https://doi.org/10.1075/bjl.00043.jan
Belgian Journal of Linguistics Volume 34 (2020), pp. 161–173. issn 0774-5141 | e-issn 1569-9676

The primary objective of this article is to address linguists working in the framework of Construction Grammar in order to inspire and motivate them to build constructicon resources for their languages, by presenting the ideas and tools we utilized in building a constructicon for Russian.

Constructions are the elements that structure languages (Fillmore, Kay, and O'Connor 1988; Croft 2001; Goldberg 2006). In essence, each language is a structured inventory of constructions, and thus it is theoretically possible to model an entire language as a constructicon. The term *constructicon* refers to both a structured inventory of grammatical constrtuctions and a description of this inventory. Today, constructicon resources are under development for only a handful of languages, namely English, Swedish, German, Brazilian Portuguese, Japanese, and Russian (Lyngfelt et al. 2018).

The growth of this emergent sub-discipline of Construction Grammar, termed 'constructicography', promises crucial benefits both for linguists and for language learners. Our understanding of how networks of constructions work largely depends on the amount of publicly available data on constructions. Moreover, thoroughly annotated and searchable databases of constructions can serve the needs of Natural Language Processing (NLP). Recognizing semi-compositional constructions in running text is crucial for machine translation, extraction of information and other applications (Dunietz, Levin, and Petruck 2017). It is now high time to build comparable constructicon resources for additional natural languages.

In what follows, we provide a practical guide for how to build a full-scale constructicon resource for a natural language, sharing our experience from the Russian Constructicon project (https://site.uit.no/russian-constructicon/). We report on a group project carried out over a five-year period (2016–2020) that succeeded to collect, describe and illustrate an inventory of over 2200 multi-word constructions of Contemporary Standard Russian (Janda et al. 2018; Endresen et al. 2020).

We start with a brief overview of characteristics of the Russian Constructicon resource (Section 2), then outline the major stages of our work, focusing on methods for expanding and structuring the inventory of constructions (Sections 3 and 4). Section 5 presents an illustration of our method. The article concludes with recommendations based on our experience.

2. Features of the Russian constructicon resource

The Russian Constructicon resource provides a large-scale model of the system of Russian constructions for the benefit of linguists, second language learners, and NLP. The goal of modelling a language as a constructicon and the needs of users have motivated the design of the project. The scope and organization of the project are detailed in this section.

2.1 The scope of the project

In the broadest sense, a construction is any recurrent form-meaning pairing in a language, at any level of complexity, from morpheme through lexeme through phrase to discourse structure (Goldberg 2006, 5). The constructicon of a language is an open-class inventory that is potentially limitless. Therefore it would be unrealistic to expect to produce a comprehensive constructicon resource. Furthermore, many items that a comprehensive constructicon should contain are already available in existing reference works, such as dictionaries (that contain lexeme-level constructions), phraseological dictionaries (that contain idioms where all the slots are fixed), and grammars (that explain basic schematic types of sentences and use of function words).

What remains are entrenched multi-word expressions that contain at least one open (not fixed) slot, and these are the strategic target of the Russian Constructicon resource. More precisely we have collected partially schematic phrases that are repeatedly used in Russian to convey meanings that range along a scale from fully transparent (compositional) to opaque. A salient feature of such constructions is the fact that their form, while motivated, is also to some extent arbitrary.

The following examples illustrate the type of constructions targeted in the Russian Constructicon resource, namely constructions that are neither merely schematic sentence types nor fully fixed idioms. A typical construction in the resource includes a fixed part, called the 'anchor' and one or more slots that can be filled with a restricted set of lexemes. This type of construction is partially schematic because part of it (the anchor) is fixed, while the rest is variable. Partially schematic constructions are likewise the focus of the Swedish constructicon resource (Lyngfelt et al. 2018, 42), and are referred to as 'constructions of microsyntax' in the Russian linguistic literature. For example, in the construction *net čtoby VP-Inf, Cl* [instead of X-ing, Y] illustrated in (1), the anchor is *net čtoby* literally 'no in-order', and the open slots are the infinitive verb (here filled by *podoždat'* 'wait') and the following clause.

(1) *Net čtoby podoždat', on uše-l-ø bez nas!*
no in.order wait-INF he leave-PST-M.SG without we.GEN
'Instead of having waited for us, he just left!'

This example strongly illustrates non-compositionality since it is not possible to predict the meaning of this construction based on its components. Linguists, learners, and NLP specialists face challenges in accounting for such constructions.

In addition to non-compositional constructions like the one above, high-frequency compositional constructions are targeted in our project, such as *(NP-Dat) Cop možno VP-Inf* [possible to X] illustrated in (2), where the adverb *možno* 'possible' is added to an infinitive to mean 'it is possible to X'.

(2) *Do Moskv-y iz London-a možno dolete-t' za četyr-e*
to Moscow-GEN from London-GEN possible fly-INF behind four-ACC
čas-a.
hour-GEN.SG
'It is possible to fly from London to Moscow in four hours.'

Even such a construction is somewhat arbitrary, since it would be theoretically possible to use a different adverb or a different form of the verb (perhaps a gerund or a deverbal noun), however in Russian the usual way to express this meaning is with precisely this construction.

Further types of compositional but arbitrary constructions targeted in the Russian Constructicon resource include constructions where the anchor is a verb with a specific argument structure or where a derivational morpheme serves as part of the anchor. For example, in *NP-Nom načinat' NP-Ins* [X begin as Y] illustrated in (3), the conventionalized choice of the instrumental case with the verb *načinat'* 'begin' indicates the status of the person as a salient and temporary property.

(3) *On načina-l-ø učitel-em.*
he begin-PST-M.SG teacher-INS.SG
'He began his career as a teacher.'

An example of a derivational morpheme embedded in a construction is *NP-Nom pere-Verb vse NP-Acc.Pl*[re-X all Ys] as in (4), where the prefix *pere-* specifies distributive semantics.

(4) *Ja pere-my-l-ø vs-e tarelk-i v dom-e.*
I PERE-wash-PST-M.SG all-ACC.PL dish-ACC.PL in house-LOC.SG
'I washed all of the dishes in the house.'

While the instrumental case and the prefix *pere-* are motivated from the perspective of Russian grammar, their use in these constructions is also an arbitrary language-specific fact that must be accounted for by linguists and mastered by learners.

In sum, the Russian Constructicon resource targets recurrent linguistic patterns that 'fall between the cracks' of dictionaries and grammars, yet are essential to full mastery of the language.

Some constructicons are connected to a FrameNet resource, based on Fillmore's work on frames. According to Fillmore and Atkins (1992, 75), a frame is a cognitive structure, the knowledge of which is presupposed for the concepts encoded by constructional constituents. Though Russian lacks a fully developed FrameNet resource, there exists a FrameBank (https://github.com/olesar/framebank) that focuses primarily on verbs and their argument structure. The data of FrameBank and the Russian Constructicon partially overlap. In the future, we might add cross-references to frames described in the Russian FrameBank where appropriate.

2.2 The presentation of constructions

The presentation of constructions in the Russian Constructicon resource is tailored to the needs of the projected users: linguists, second language learners, and NLP researchers. To this end, we provide both detailed linguistic classification and user-friendly guidance. Each construction is supplied with:

- a *name*, which is a schematic description of the construction; such as *net čtoby VP-Inf, Cl* [instead of X-ing, Y]
- a *brief illustration*; such as (1)
- a *definition* stated in non-technical language in Russian (with translations into English and Norwegian); in this case: "The construction indicates that the speaker expresses dissatisfaction with the fact that the interlocutor has not taken a given action or is undertaking or has undertaken a different action."
- a *CEFR language proficiency level* (from A1 to C2) to help learners target appropriate constructions; in this case C1
- a series of *semantic and syntactic tags*
- a list of *common fillers* for the open slot(s)
- a *usage label* specifying the type of speech (Neutral, Colloquial, Formal, Obsolete)
- a *structure* in terms of Universal Dependencies (https://universaldependencies.org)
- three to five *corpus examples* from the Russian National Corpus (www.ruscorpora.ru)

In addition, both the definition and the corpus examples are tagged for semantic roles (Agent, Experiencer, etc.). All of the information about each construction is searchable. For example, linguists can search for semantic and syntactic parameters, learners can search for constructions at a given proficiency level, and both types of users can enter strings (for example, of anchor words) to search for specific constructions. The Universal Dependency structure, the glossing system, and the lists of common fillers of the slots serve the purposes of Natural Language Processing, facilitating automatic recognition of constructions in authentic Russian texts. The system of semantic tags is based on terminology from typological literature (cf. the "universal grammatical set of meanings", Plungian 2011, 65). Taken together, these features make the Russian Constructicon a multi-functional resource, designed for language pedagogy, language research, and language technology. Among other constructicon projects, only the Swedish Constructicon (Lyngfelt et al. 2018, 41, 94) pursues pedagogical goals and has been created not only for linguists but also for learners of Swedish.

3. Reaching and exceeding a critical mass of constructions

Linguistically, we can classify constructions according to their semantics and their formal structures. However, the classification becomes reliable only after a representative sample has been obtained. A critical mass of constructions is needed in order to establish their classification, which is uncertain prior to that point. In other words, we had to repeatedly cycle through the tasks of collecting and classifying constructions in order to arrive at a stable system which could then be exploited for further expansion of the constructicon with only minor adjustments. Our process proceeded in three stages, visualized in Figure 1 as the Initial inventory, Corpus-based expansion, and System-based expansion. Numbers inside the bars reflect the quantity of constructions added in each stage, and dates indicate the approximate timing of the stages.

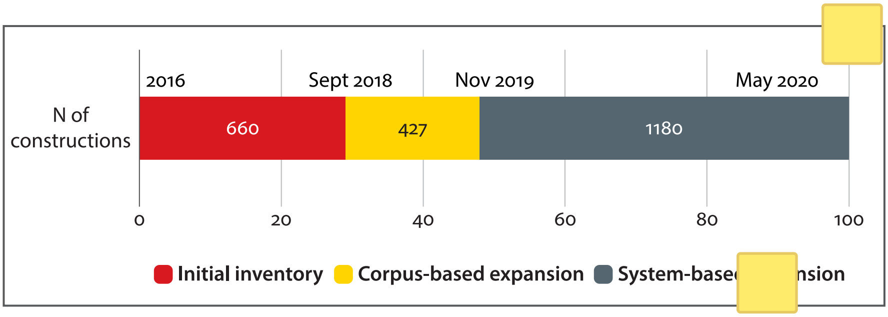

Figure 1. Stages of the Russian constructicon project

The Initial inventory of 660 constructions was amassed manually in Stage 1 from a variety of sources including textbooks for learners of Russian (especially Janda and Clancy 2002) and scholarly literature on Russian constructions (especially Rakhilina 2010), as well as a crowd-sourced Google spreadsheet. At this stage we decided what kinds of constructions to focus on in our project (see Section 2.1), established most of the conventions that would be used in the presentation of constructions (see Section 2.2) and began to explore the semantic and syntactic system of the constructicon (see Section 4). This stage involved continuous revisions in our procedure as we grappled with the dimensions of the project.

The Corpus-based expansion in Stage 2 continued the manual heterogeneous collection of constructions, at this stage culled from running texts of various kinds, particularly those that contain dialogues and spoken discourse, as well as an automatically extracted list of highly frequent collocations attested in the Russian National Corpus. In this stage, we added 427 constructions to the Initial inventory. In addition to adding constructions, we continued the work on classification of semantic and syntactic types, using the new constructions to verify and refine the classification. Once we had reached a critical mass of over one thousand constructions, the classification became stable and robust enough to facilitate the identification of 'families' of constructions (see Section 4). In other words, on the basis of our semantic and syntactic tags we were able to discover groups of constructions that were internally relatively homogeneous.

Families of constructions served as the basis for the more rapid and extensive System-based expansion of the constructicon in Stage 3, which more than doubled the size of the inventory to over 2,200 items. We examined semantic families of constructions found in the database and searched for their synonyms, antonyms, and related constructions containing the same or similar anchor words in order to fill gaps in each family. Thus the classification system facilitated addition of constructions in a significantly more efficient manner. This stage yielded not only quantitative but also qualitative change in the constructicon: semantic classification of constructions turned what initially was a list of unrelated items into a structured system of constructions.

4. Identifying families: Theoretical motivation and methodology

4.1 Theoretical motivation

One of the tenets of Construction Grammar is the idea that constructions are related to each other. Following the example of Goldberg (2006) and her analysis of the English Subject Auxiliary Inversion family of constructions, we have developed the means to transform the inventory of constructions into a structured system. One of the crucial challenges of a constructicon resource is to reveal and represent this system, that is, the complex relationships (both hierarchical and lateral) between constructions. One strategy is to focus on the relationship of parent vs daughter constructions, i.e. a more abstract schema vs its specific instantiation. In addition, we identify meaningful groupings: *families* that form *clusters*, and ultimately *networks*.

We define a family of constructions as a relatively homogeneous group of about two to nine constructions that exhibit family resemblance in that they share some semantic, syntactic (function in a clause and structure of the fixed part), and structural properties (e.g. reduplication, negation, inversion, etc.). Family resemblance means that the constructions in a family share various subsets of these properties. The families within a cluster in turn share properties in a prototypical vs. peripheral distribution. We have elaborated a multi-level set of semantic and syntactic tags that facilitate identification of families and clusters.

4.2 Methodology

Annotation was undertaken by a panel of three native speakers who worked to achieve consensus on the tagging of each construction. A number of semantic and syntactic tags were assigned to each construction by the panel. The annotation was continuously refined and cross-checked by the entire panel, minimizing subjectivity and guaranteeing consistency. In this process we took into account existing scholarship relevant to semantic and syntactic classification, from both Russian and typological scholarly traditions (Plungian 2011).

In all we employ 53 general semantic tags, many of which have subtags, yielding an overall inventory of 173 subtags. Over 40% of the constructions bear more than one semantic tag. Figure 2 displays the distribution of the most frequent general semantic tags. Figure 3 displays the distribution of constructions across eleven syntactic tags.

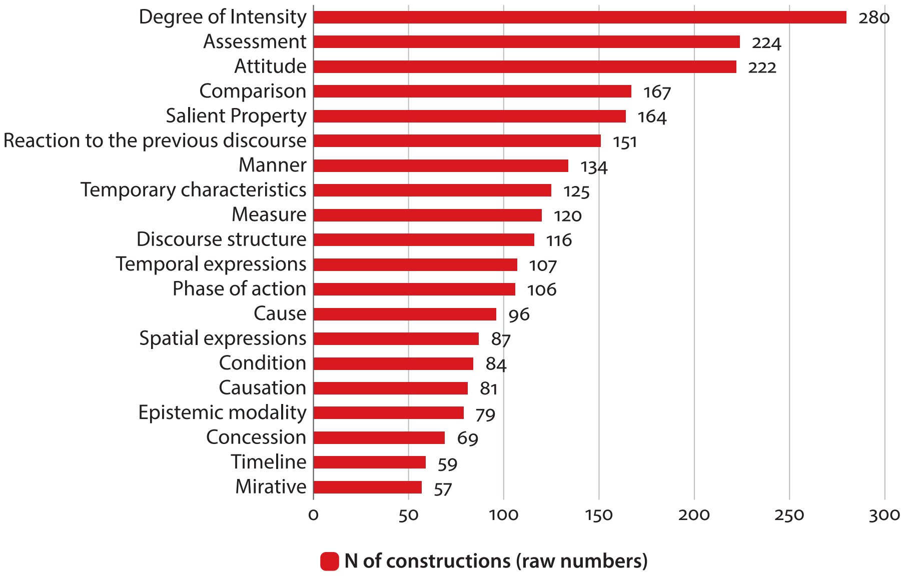

Figure 2. Distribution of constructions across twenty most frequent general semantic tags

We investigated the intersection of semantic and syntactic classifications to identify meaningful groupings of constructions. Among the constructions that received each general semantic tag, we examined syntactic patterns in order to find more homogeneous groups of constructions. Thus, we arrived at smaller groups of 2–9 constructions that shared more or less the same syntactic structure and more narrowly specified semantics. These smallest groups we call families. We furthermore

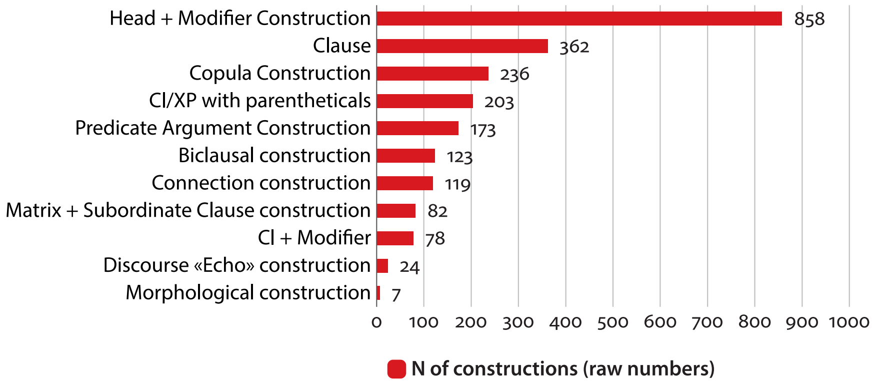

Figure 3. Distribution of constructions across general syntactic tags

examined how families are related to each other within clusters and how clusters comprise networks. As a rule, our general semantic tags correspond to networks, and the subtypes correspond to clusters. An illustration of this approach is presented in Section 5.

5. Turning a list into a structured inventory

We illustrate the method outlined in Section 4 with the network of Prohibitive constructions diagrammed in Figure 4, consisting of two clusters and a total of eleven families.

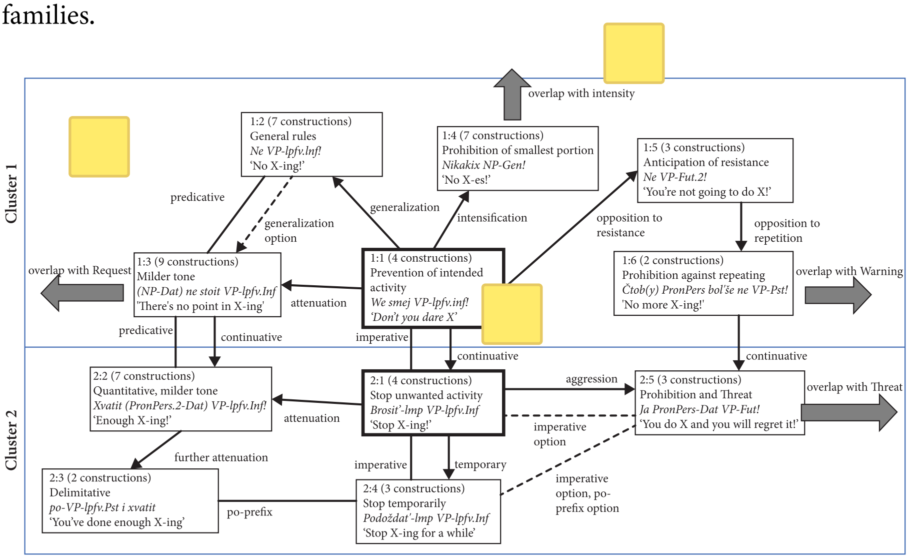

Figure 4. Network of prohibitive constructions

In Figure 4, boxes represent families indexed as cluster:family, followed by a brief description and illustrative example. Thick boxes indicate prototypes. Lines with arrows indicate semantic transitions. Lines without arrows indicate syntactic/formal similarities. Dotted lines and arrows indicate weaker relationships. Thick arrows indicate overlap with other networks of constructions.

Whereas constructions in Cluster 1 ask a hearer to refrain from doing something, constructions in Cluster 2 express 'continuative prohibition', asking a hearer to stop doing something. All constructions in Cluster 1 contain overt markers of negation; such markers are absent from Cluster 2. Cluster 1 is centered around its prototype, family 1:1, containing negated imperative constructions. Lines represent the relationships that hold among families and are tagged for semantic transitions and shared formal properties. A semantic transition to generalized prohibitions connects 1:1 to 1:2, with transitions to the remaining families in Cluster 1 labeled in Figure 4. Prohibitions in 1:3 can be either generalized or individual, indicated by a dotted arrow, and 1:2 shares the syntactic form of predicative with 1:3. Three families in Cluster 1 (1:3, 1:4, and 1:6) share constructions across other networks (Request, Intensity, and Warning), indicated by the thick arrows.

Cluster 1 is connected to Cluster 2 through three pairs of families. In each pair, the semantic transition is from standard prohibition in Cluster 1 to continuative prohibition in Cluster 2. In both clusters, families to the left represent generalization and attenuation, as opposed to more combative prohibitions on the right. In addition, the two prototypical families (1:1 and 2:1) share the syntactic form of imperative (also shared by 2:4) and families 1:3 and 2:2 share the form of predicative. The *po-* prefix is a necessary feature of 2:3 and 2:4, and optionally found in 2:5, where there is also some use of imperative forms.

The Prohibitive network demonstrates the complex of semantic and formal properties that structure the constructicon.

6. Conclusion

We hope that this article will encourage the building of constructicons for a wider variety of languages to serve both language learners and linguists. While the Russian Constructicon represents just one possible model, we can share lessons that from our experience can be valuable to other similar projects. This is not a project for an individual; it is essential to build a team of researchers because a constructicon requires a variety of skills and a long-term commitment. As with any collaborative project, funding is essential. We found that it was possible to 'package' funding for the Russian Constructicon under the umbrella of grant projects primarily aimed at language pedagogy and international cooperation. A strategic focus on constructions that are otherwise underrepresented in pedagogical and reference works helps to keep the project manageable and also makes it easier to ‘sell’ in grant proposals. A further ‘selling point’ is a user-friendly design that addresses the needs of multiple audiences: the Russian Constructicon is a resource both for learners and for linguists. In terms of presentation, we started by ‘piggybacking’ on an existing architecture (the Swedish Constructicon), making it possible to work through the first two stages of our project without having to start from scratch with the design of an interface. We are grateful for the big advantage this gave us, which ultimately made it possible to envision something that would better represent the Russian Constructicon. Once we began to uncover the relationships among constructions (illustrated in Section 4), we had something that was no longer an inventory, but a system, and we needed a new interface that could do justice to that structure. We look forward to further expanding and refining the Russian Constructicon in its new design and welcome comments and critique.

Funding

We acknowledge funding from the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education: grants NCM-RU-2016/10025 and CPRU-2017/10027.

References

Croft, William. 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001

Dunietz, Jesse, Lori Levin, and Miriam R. L. Petruck. 2017. “Construction Detection in a Conventional NLP Pipeline.” In *The papers from the 2017 AAAI Spring Symposium on Computational Construction Grammar and Natural Language Understanding. Technical Report SS-17-02*, 178–184.

Endresen, Anna, Valentina Zhukova, Daria Mordashova, Ekaterina Rakhilina, and Olga Lyashevskaya. 2020. “Russkij konstruktikon: Novyj lingvističeskij resurs, ego ustrojstvo i specifika [The Russian Constructicon: A new linguistic resource, its design and key characteristics].” In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue-2020”*, 226–241. Published on-line.

Fillmore, Charles J., and Beryl T. Atkins. 1992. “Toward a Frame-Based Lexicon: The Semantics of RISK and Its Neighbors.” In *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization*, ed. by Adrienne Lehrer, and Eva Kittay, 75–102. Hillsdale, NJ: Lawrence Erlbaum.

Fillmore, Charles J., Paul Kay, and Mary C. O'Connor. 1988. "Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*." *Language* 64(3): 501–538. https://doi.org/10.2307/414531

Goldberg, Adele. 2006. *Constructions at Work. The Nature of Generalization in Language.* Oxford: Oxford University Press.

Janda, Laura A., and Steven J. Clancy. 2002. *The Case Book for Russian*. Bloomington, IN: Slavica Publishers.

Janda, Laura A., Olga Lyashevskaya, Tore Nesset, Ekaterina Rakhilina, Francis M. Tyers. 2018. "A Constructicon for Russian: Filling in the Gaps." In *Constructicography: Constructicon Development Across Languages*, ed. by Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago T. Torrent, 165–181. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.22.06jan

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark, and Rudolf Rydstedt. 2018. "Constructicography at Work: Theory Meets Practice in the Swedish Constructicon." In *Constructicography: Constructicon Development Across Languages*, ed. by Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago T. Torrent, 41–106. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.22.03lyn

Plungian, Vladimir A. 2011. *Vvedenie v grammatičeskuju semantiku: Grammatičeskie značenija i grammatičeskie sistemy jazykov mira* [An introduction to grammatical semantics: Grammatical meanings and grammatical systems in the languages of the world]. Moscow: Russian State University for the Humanities Press.

Rakhilina, Ekaterina V. (ed.) 2010. *Lingvistika konstrukcij* [Linguistics of constructions]. Moscow: Izdatel'stvo Azbukovnik.

Abbreviations

2	2nd person	LOC	locative
ACC	accusative	NP	noun phrase
DAT	dative	M	masculine
FUT	future	PL	plural
GEN	genitive	PronPers	personal pronoun
IMP	imperative	PST	past
INF	infinitive	SG	singular
INS	instrumental	VP	verb phrase
IPFV	imperfective	()	optional element

Authors' addresses

Laura A. Janda
Department of Language and Culture
UiT The Arctic University of Norway
Hansine Hansens veg 18
N-9037 Tromsø
Norway

laura.janda@uit.no

Anna Endresen
Department of Language and Culture
UiT The Arctic University of Norway
Hansine Hansens veg 18
N-9037 Tromsø
Norway

anna.endresen@uit.no

Valentina Zhukova
School of Linguistics
National Research University Higher School of Economics
20 Myasnitskaya Street
101000 Moscow
Russia

valentina.zh96@gmail.com

Daria Mordashova
Minority Language Research and Preservation Lab
Institute of Linguistics, Russian Academy of Sciences
1 bld. 1 Bolshoy Kislovsky Lane
125009 Moscow
Russia

mordashova.d@yandex.ru

Ekaterina Rakhilina
School of Linguistics
National Research University Higher School of Economics
20 Myasnitskaya Street
101000 Moscow
Russia

rakhilina@gmail.com