How to build a constructicon in five years: The Russian Example

Abstract

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 2200 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.

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.

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 et al. 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. However, only a small number of constructicon resources presenting constructions that are analyzed, explained and illustrated are under development, namely for English, Swedish, German, Brazilian Portuguese, Japanese, and Russian (Lyngfelt et al. 2018).

The growth of this emergent sub-discipline of Construction Grammar, termed "constructicography" (Lyngfelt et al. 2018), 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. 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 (Co-author et al. Year, Co-author et al. Year).

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), and present the design of a searchable interface (Section 5). 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 both linguists and second language learners. 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. 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. In this sense the (mostly) arbitrary form-meaning relationship we observe for morphemes and lexemes is relevant also for multi-word constructions.

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 (Iomdin 2015). For example: *net čtoby/by VP-Inf, Cl* as in *Net čtoby podoždat', on ušel bez nas!* 'Instead of having waited for us, he just left!' The anchor is *net čtoby/by* literally 'no in-order', and the open slots are the infinitive verb (here filled by *podoždat'* 'wait') and the following clause. This example strongly illustrates non-compositionality since it is not possible to predict the meaning of this construction based on its components, and both

linguists and learners face challenges in explaining and comprehending such constructions.

In addition to non-compositional constructions like the one above, high-frequency compositional constructions are targeted in our project, such as *možno VP-Inf* as in *Do Moskvy iz N'ju-Jorka možno doletet' za desjat' časov* 'It is possible to fly from Moscow to New York in ten hours', where the adverb *možno* 'possible' is added to an infinitive to mean 'it is possible to X'. 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, *NP-Nom načinat' NP-Ins* as in *On načinal učitelem* 'he began his career as a teacher', where the conventionalized choice of the instrumental case with the verb *načinat'* 'begin' indicates the status of the person as a salient and temporary property. An example of a derivational morpheme embedded in a construction is *pere-V vse NP-Acc.Plur* as in *peremyt' vse tarelki v dome* 'wash all of the dishes in the house' where the prefix *pere-* specifies distributive semantics. 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. For example, the Berkeley Constructicon of English is built on the basis of FrameNet (Fillmore et al. 2003), linking each construction to its frame. For example, the construction *be_recip* (e.g. *She is good friends with her mother*) in the Berkeley Constructicon corresponds to the *Reciprocality* frame (Lee-Goldman & Petruck 2018, 32). Constructions receive a FrameNet ID that connects them with semantic resources like WordNet, PropBank, SemLink+ etc. (Palmer et al. 2014). However, this is not the only possible strategy. The Swedish Constructicon relies not on the Swedish FrameNet++, but on SALDO, a system that represents lexicographic resources as a hierarchical network. 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 (Lyashevskaya and Kashkin 2015). The data of FrameBank and the Russian Constructicon partially overlap and the teams working on the two resources overlap as well. 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 and second language learners. 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/by VP-Inf, Cl*
- a *brief illustration*; such as *Net čtoby podoždat', on ušel bez nas!*
- 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
- 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 system of semantic tags is based on terminology from typological literature (cf. the "universal

grammatical set of meanings" Plungian 2011, 65). 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. 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.

2.3 The team

The Russian Constructicon has been made possible by the dedicated efforts of a large multi-national team of collaborators from UiT The Arctic University of Norway (the CLEAR: Cognitive Linguistics Empirical Approaches to Russian research group) and the National Research University Higher School of Economics (School of Linguistics) in Moscow, as well as Indiana University. The team includes both native and non-native speakers of Russian, thus bringing both the intuitions of native linguists and expertise in second language pedagogy to the project. Quantitative and qualitative analysis of corpus data add to the balance. Additional expertise includes computational linguistics, natural language processing, and programming. Both faculty and students at all levels (BA through PhD) contribute to the project.

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.

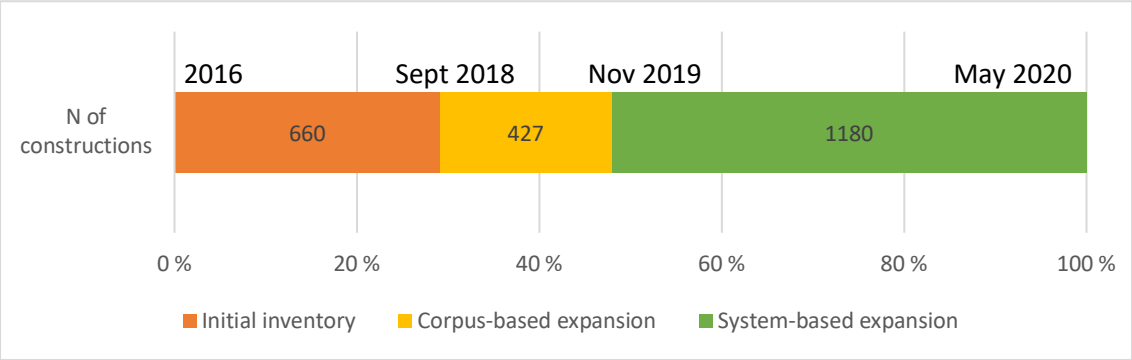

Figure 1: Stages of the Russian Constructicon project

The Initial inventory of 660 constructions was amassed manually in Stage 1 from a variety of sources including textbooks for learners of Russian (especially Co-author, second author Year) and scholarly literature on Russian constructions (especially Co-author Year), 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 heterogenous collection of constructions, at this stage culled from running texts of various kinds, particularly those that contain dialogs 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. 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 2200 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. We gained new insights

regarding hierarchical relations among constructions that are grouped as families, families that form clusters, and clusters that form networks, as well as how all these units can also overlap conceptually and/or by sharing some of the same constructions (see Section 4).

At present descriptions of the constructions added in Stage 3 are being filled out according to the items listed in Section 2.2. Further new constructions will certainly be added in the future, although in principle this is a project that will never be definitively finished given the scope and dynamic nature of language.

4. Turning a list into a structured inventory

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. Our multi-level semantic and syntactic classification makes it possible to identify meaningful groupings of constructions that we term as families, clusters, and networks.

We define a family of constructions as a relatively homogeneous group of about two to ten 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 illustrate this method with the network of Prohibitive constructions diagrammed in Figure 2, consisting of two clusters and a total of eleven families.

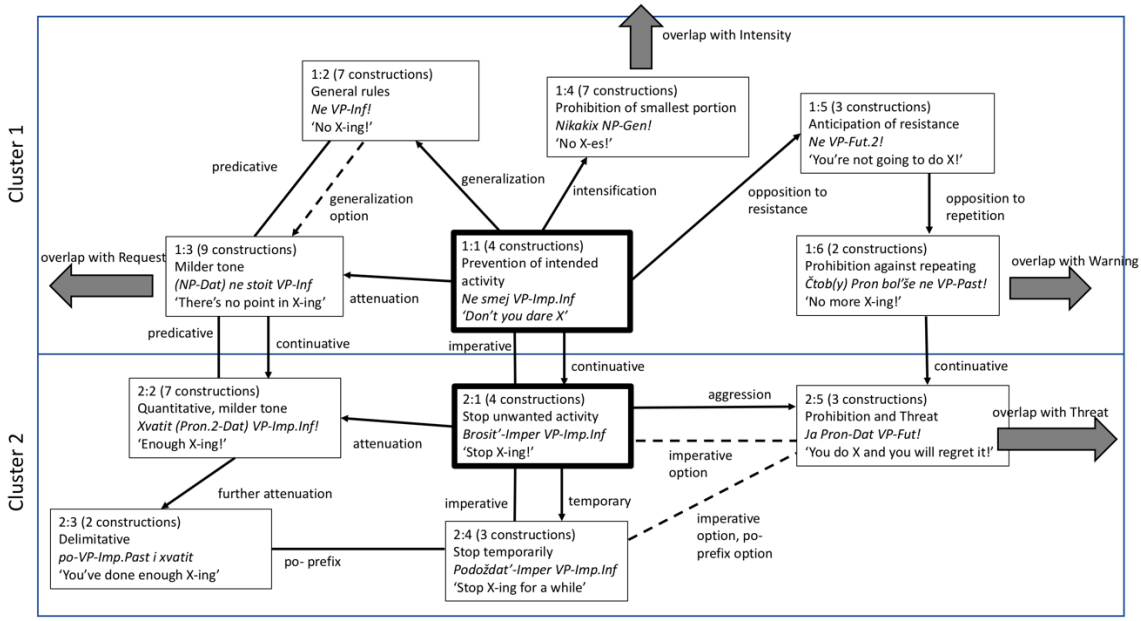

Figure 2: Network of Prohibitive constructions. Boxes represent families indexed as cluster:family, followed by a brief description and illustrative example. Thick boxes indicate prototypes. Lines with arrows indicate semantic transitions. Lines without arrows indicate syntactic/formal similarities. 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 2. 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.

5. A searchable interface

The Russian Constructicon is a multi-purpose, free and open internet resource that can be used without password or login. The initial interface, which is still accessible, shared the same architecture and features as the Swedish Constructicon: http://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus. At present we are designing a new open-source and more user-friendly interface with added options to search by language proficiency level (for learners of Russian) and by various

linguistic properties (for linguists). The new website is under construction and the link will appear at the project website https://site.uit.no/russian-constructicon/ by the end of 2020. The new site will have the following additional features:

Browse. The user sees a list of over 2200 constructions that they can scroll through and search for any string that is part of the name of a construction.

Daily Dose. The user can choose their proficiency level and receive a randomly generated list of five constructions matching their level.

Advanced Search. The user can filter constructions according to various parameters that can be combined: Morphology, Syntax, Semantics, Semantic roles, and Level. It is possible to apply multiple filters on the same search.

Instructions and *About*. These pages explain terminology and search options and provide information about the resource and its developers.

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 Information

The Russian Constructicon has been supported by two grants from the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (Diku, https://diku.no/en):

- "Constructing a Russian Constructicon" (NCM-RU-2016/10025) in 2016 and
- "Targeting Wordforms in Russian Language Learning" (CPRU-2017/10027) in 2017-2020.

Anonymized References

Co-author (ed.) Year. *Book title.* Publisher.

Co-author, co-author, co-author, co-author, co-author. Year. "Article title". In *Papers from the International Conference*, page numbers.

Co-author, co-author, co-author, co-author, co-author. Year. "Article title". In *Book title*, page numbers.

Co-author, second author. Year. *Book title.* Publisher.

References

Croft, William. 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective.* Oxford: Oxford University Press.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R. Petruck. 2003. "Background to FrameNet". In *International Journal of Lexicography* 16(3): 235-250.

Fillmore, Charles J., Paul Kay & Mary C. O'Connor. 1988. "Regularity and idiomaticity in grammatical constructions: The case of *let alone*." *Language* 64(3): 501-538.

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

Iomdin, Leonid Lejbovič. 2015. "Konstrukcii mikrosintaksisa, obrazovannye russkoj leksemoj *raz* [Microsyntactic constructions formed by the Russian lexeme *raz*]." In *Slavia: Časopis pro slovanskou filologii* 84(3): 291-306.

Lee-Goldman R., Petruck, M. 2018. "The FrameNet constructicon in action." In *Lyngfelt et al 2018*, 19-40.

Lyashevskaya, Olga, and Egor Kashkin. 2015. “FrameBank: a database of Russian lexical constructions.” In *Analysis of Images, Social Networks and Texts. Fourth International Conference, AIST 2015, Yekaterinburg, Russia, April 9-11, 2015, Revised Selected Papers. Communications in Computer and Information Science* 542, 337-348. Springer.

Lyngfelt, Benjamin, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent (eds.). 2018. *Constructicography: Constructicon Development Across Languages*. Amsterdam: John Benjamins.

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark, and Rudolf Rydstedt. 2018. “Constructicography at work: Theory meets practice in the Swedish constructicon.” In *Lyngfelt et al 2018*, 41-106.

Palmer, Martha, Claire Bonial, and Diana McCarthy. 2014. “Semlink+: FrameNet, VerbNet and event ontologies.” In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, 13-17.

Plungian, Vladimir Aleksandrovič. 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]. Russian State University for the Humanities, Moscow.