Ene Vainik, Geda Paulsen, Heete Sahkai, Jelena Kallas, Arvi Tavast, and Kristina Koppel

FROM A DICTIONARY TO A CONSTRUCTICON

Putting the Basics on the Map

Abstract This study discusses the possibilities of expanding the scope of the largest Estonian dictionary – the EKI Combined Dictionary – with various types of constructional information. Designing a representation of constructions essentially means building a constructicon. The study starts with a short overview of existing constructicons and the main challenges their creators have faced so far. We address these issues from the point of view of data model reorganisation and database restructuring. Extending the lexicographic resource with constructicographic information is twofold: the existing constructional information must be migrated into a new model and then complemented with additional constructions extracted from a corpus.

Keywords electronic lexicography; grammatical constructions; constructicography; Estonian

1. A New Emerging Field in Electronic Lexicography

As electronic lexicography continues to evolve, a new subfield called constructicography has emerged (Lyngfelt, 2018, p. 1). This field focuses on compiling a resource known as a "constructicon", which parallels dictionaries by presenting complex grammatical constructions as pairs of linguistic forms and meanings. Unlike traditional lexicography, constructicography is theory-driven and rooted in the construction-based view of language (e.g., Croft, 2001; Goldberg, 2003; Hoffmann & Trousdale, 2013), asserting that there is a continuum of linguistic units instead of a lexicon vs. grammar dichotomy and that networks of constructions are fundamental to language (Diessel, 2023). Constructicons aim to not only present the form and meaning of their units but also their interrelations. The use of relational databases facilitates this endeavour. Though approximately ten initiatives worldwide are underway to create such resources for various languages, there is currently no established tradition in this field (Borin & Lyngfelt, in press).

This paper aims to contribute to constructicographic theory and practice by analysing the decisions and challenges of current initiatives (i.e., mapping the basics; see Section 2). In Section 3, we formulate our initial plan to enrich the largest Estonian dictionary, the EKI Combined Dictionary (CombiDic)[1] (Koppel et al., 2019), with information about constructions and present the main points of restructuring the data model to meet the needs of a constructicographic resource.

[1] Accessible via the language portal sõnaveeb.ee

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

2. Mapping the Basics

This section provides an overview of current constructicographic initiatives (Fillmore et al., 2012; Janda et al., 2020; Lyngfelt, Bäckström, et al., 2018; Ohara, 2013; Perek & Patten, 2019; Sass, 2023; Torrent et al., 2014; Ziem et al., 2019), focusing on their target groups, size, types of relations, database types, and connected resources (see Appendix 1). We also list the main challenges reported by contributors and those we have identified.

2.1 The Current Practices

The analysis of current practices showed that most of the initiatives have grown out of ongoing work on a FrameNet database and are linked to it. The reported coverage of constructions varies from 73 to approx. 13 000 entries. The target users can be language experts, L2 learners, language technology or some combination of those. The primary focus varies from semi-schematic and/or idiosyncratic units to abstract valency patterns (argument structure constructions) and, to a somewhat lesser degree, to idiomatic units residing on the "lexical end" of the grammar-lexicon continuum. An essential part of constructicon building is presenting the relations of inheritance (predicted by the theory, e.g., Ziem et al., 2023) and, to a lesser degree, part-whole relationships (e.g., Sass, 2023).

2.2 The Challenges

The process of creating constructicons is challenging and time-consuming, involving various hurdles:

1. Defining what qualifies as a construction for inclusion. While all meaning-form pairs in a language are theoretically constructions, it's crucial to identify suitable constructicon units. This relates to determining the complete inventory of constructions to be included.

2. Presenting constructions at different degrees of schematicity. The data model and description format must be suitable for describing units at various points along the lexicon-grammar continuum, from idioms and phraseological units to intermediate forms and purely schematic grammatical constructions (Ziem et al., 2023).

3. What should a constructicon entry look like? Language learners require simplified descriptions, meta-language and only typical usage examples, while experts seek detailed information and annotated corpus sentences. Ensuring machine readability is another aspect related to the granularity of description. Therefore, a constructicon entry must be adequate, concise, user-friendly, and at the same time formalised (Lyngfelt, Borin, et al., 2018).

4. Naming constructions and organising constructicon units. Since alphabetical order may not suffice, determining organisation criteria, such as taxonomies or other typological systems, is necessary. Should the presentation of

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

categories mimic that of grammar? Should entries be allowed to appear in multiple types?

5. Representing relationships between constructions. Different constructicons adopt varied approaches, reflecting the dataset's purpose: what is suitable for a pedagogically oriented constructicon may not be suitable for an NLP-oriented solution. The task of presenting slot-filler relations is a challenge because the relations must be established first.

6. How is it possible to make information about constructions accessible to users with different linguistic backgrounds? While lexical units can be given translational equivalents, how should a user find the most natural way to express desired meanings in the target language from the constructicon? (see e.g., Lyngfelt et al., 2022).

Borin & Lyngfelt (in press) emphasise that in addition to structural elements derived from theory, it is important to consider whether the constructicon is associated with another lexicographic dataset (such as FrameNet) and what is the overall purpose of the constructicon. The structure must be flexible and adaptable so that new (and possibly different types of) constructions can be added to the collection.

3. Towards a Data Model

The Estonian project primarily targets L2 learners and teachers, along with L1 speakers, researchers, and NLP applications. Due to the lack of a FrameNet for Estonian and the need for a comprehensive lexical and grammatical resource, we are developing the Estonian constructicon as an extension of CombiDic (Koppel et al., 2019). CombiDic is the most extensive and up-to-date Estonian lexicographic data collection, containing a full lexical inventory, morphological paradigms, definitions, syntactic information, semantic types and CEFR proficiency level markers. It has been compiled using the Ekilex Dictionary Writing System (Tavast et al., 2018) on a PostgreSQL database.

3.1 The Data Model

To describe the representation of constructions, we first need to outline the Ekilex data model[2], which is based on a many-to-many relationship between word and meaning. In database terms, such relations are implemented using junction tables. The junction table between word and meaning is called a lexeme, and the data it contains can be described as "this word in this meaning as defined by this dictionary".

The data model already includes the following syntactic data elements: i) government patterns: currently a plain text field within the lexeme; ii) compounds and derivatives: relations between words, e.g., compounds containing information about their components; iii) collocations: until recently separate units closely duplicating the representation of words.

[2] https://github.com/keeleinstituut/ekilex/wiki/Andmemudel

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The first stage of extending Ekilex to accommodate a constructicon involves redesigning the data model for collocations and migrating existing collocations to this new model, along with a more robust representation of government. Eventually, compounds and derivatives will also be unified into the same representation.

As discussed above, a consensus about what exactly is a constructicon is yet to emerge. Our current approach is to start with the corpus, i.e., actualisations of the constructions, and move towards abstraction. As a result, instead of being standalone, i.e., unrelated to a dictionary (e.g., Janda et al., 2020), linked to a dictionary (e.g., Fillmore et al., 2012; Lyngfelt, Bäckström, et al., 2018; Ohara, 2013; Perek & Patten, 2019; Torrent et al., 2014; Ziem et al., 2019), or a replacement for a dictionary (Sass, 2024), we conceptualise the constructicon as something that a dictionary can grow into when expanded by syntactic information.

At the heart of the redesign is the idea of expanding the concept of the headword to include units of language of any length, notably including units that have so far been presented separately as collocations or compounds. Although this increases the number of headwords, it is not a big change conceptually, as many headwords have always been multi-word units. Each expanded headword has parameters controlling its presentation: it can be displayed as a separate headword with its own entry and/or as a collocation and/or compound within the entry of one or more of its components. Just like lexemes, these relations are meaningful data structures on their own rather than simple links between the whole and its parts. They are the data elements with which we implement abstraction from individual realisations of a construction towards its schematic representation. Relations can be tagged using multiple, mutually independent tagging systems, aiming to represent existing morphological and syntactic hierarchies and support ongoing construction grammar research.

A major challenge is the rich morphology of Estonian, including pervasive paradigmatic homonymy, the disambiguation of which has not yet been reliably solved. For sustainability and scalability, we aim for fully structured data in Ekilex, connecting each MWE component to its form, not just the lexeme. This has required manual disambiguation and minor changes to the morphology section of the data model.

3.2 Meeting the Challenges

In the following, we summarise our approach to address the challenges outlined in Section 2.2.

1. In terms of determining what qualifies as a construction to be included, we will start by reorganising and generalising the existing syntactic information in the CombiDic, specifically government patterns and collocational patterns.

2. To represent constructions at different degrees of schematicity, our current approach is to detect phrase structure and argument structure constructions from the morphologically and dependency syntactically annotated corpus and move towards abstraction, thereby achieving a hierarchical representation.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

3. The structure of constructicon entries should cater to the main user groups: concise and simplified for language learners, detailed for experts, and formalised for NLP. This will be achieved by designing a comprehensive entry format with customised interfaces for different users.

4. Regarding the challenge of naming and organising the constructions, we plan to establish a typology of constructions and tag the entries with a set of labels (e.g., parallel names for different user groups, as well as taxonomic/typological and semantic affiliation). Multiple tags per construction will be allowed to facilitate recall by different criteria.

5. Hierarchical relations between entries will be crucial in our data model. The dictionary-based approach makes it possible to establish relations with lexemes as slot-fillers in semi-schematic and schematic constructions. Existing word class and semantic type labels can represent these filler-slot relations.

6. To make constructional information accessible to different user groups, we should integrate it with the general search, potentially including free text search (cf. Sass, 2023), drop-down menus for semantic categories (cf. Janda et al., 2020), a filterable list of all the constructions, and access via a layer of syntactic information in lexical entries.

4. Conclusion

Each constructicon project is unique, but most face similar challenges. This article summarises current constructicographic practices and compares them to our approach for extending a lexicographic resource with constructicographic information. Challenges include defining and naming constructions, and how to represent and display them. The data model must be flexible enough to cover the entire lexicon-grammar continuum. Solutions to these challenges inform the design for representing constructions in the EKI Combined Dictionary and its Ekilex database. Our goal is to redefine existing lexical and grammatical information in Ekilex and enrich it with (semi-)schematic constructions, while addressing the additional complexity of Estonian morphology.

References

Borin, L., & Lyngfelt, B. (in press). Framenets and constructiCons. In *The Cambridge Handbook of Construction Grammar*. Cambridge University Press. https://www.academia.edu/95301779/Framenets_and_constructiCons

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective* (1st ed.). Oxford University PressOxford. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001

Diessel, H. (2023). The Constructicon: Taxonomies and Networks. In *Elements in Construction Grammar*. https://doi.org/10.1017/9781009327848

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Fillmore, C. J., Lee-Goldman, R., & Rhomieux, R. (2012). *The FrameNet Constructicon*. https://www1.icsi.berkeley.edu/pubs/ai/framenetconstructicon11.pdf

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Hoffmann, T., & Trousdale, G. (Eds.). (2013). *The Oxford handbook of construction grammar*. Oxford University Press.

Janda, L. A., Endresen, A., Zhukova, V., Mordashova, D., & Rakhilina, E. (2020). How to build a constructicon in five years. The Russian example. *Belgian Journal of Linguistics*, *34*(1), 161–173. https://doi.org/10.1075/bjl.00043.jan

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. *Proceedings of the eLex 2019 Conference. 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 434–452.* https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_24.pdf

Lyngfelt, B. (2018). Introduction: Constructicons and constructicography. In *Constructicography: Constructicon development across languages* (pp. 1–18). John Benjamins Publishing Company. https://doi.org/10.1075/cal.22.01lyn

Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A., & Rydstedt, R. (2018). Constructicography at work. Theory meets practice in the Swedish constructicon. In *Constructicography: Constructicon development across languages* (pp. 41–106). John Benjamins Publishing Company. https://doi.org/10.1075/cal.22.01lyn

Lyngfelt, B., Borin, L., Ohara, K., & Torrent, T. T. (Eds.). (2018). *Constructicography: Constructicon development across languages* (Vol. 22). John Benjamins Publishing Company. https://doi.org/10.1075/cal.22

Lyngfelt, B., Torrent, T. T., Eli, E. da S. M., & Bäckström, L. (2022). Comparative Concepts as a resource for a multilingual constructicography. In *Valency and constructions: Perspectives on combining words*. Meijerbergs institut för svensk etymologisk forskning.

Ohara, K. (2013). Toward Constructicon Building for Japanese in Japanese FrameNet. *Veredas - Revista de Estudos Linguisticos*, *17*(1), 11–27.

Perek, F., & Patten, A. L. (2019). Towards an English Constructicon using patterns and frames. *International Journal of Corpus Linguistics*, *24*(3), 354–384. https://doi.org/10.1075/ijcl.00016.per

Sass, B. (2023). From a dictionary towards the Hungarian Constructicon. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o.* https://elex.link/elex2023/wp-content/uploads/105.pdf

Sass, B. (2024). We need everything: the "ultimate lexical resource" approach to build a constructicon. Book of Abstracts of EAAL 21st Annual Conference, April 18.-19, 2024, Tallinn, Estonia. https://www.rakenduslingvistika.ee/wp-content/uploads/2024/04/2024-teesid.pdf

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 749–761). Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Torrent, T. T., Lage, L. M., Sampaio, T. F., Tavares, T. da S., & Matos, E. E. da S. (2014). Revisiting border conflicts between FrameNet and Construction Grammar: Annotation policies for the Brazilian Portuguese Constructicon. *Constructions and Frames*, *6*(1), 34–51. https://doi.org/10.1075/cf.6.1.03tor

Ziem, A., Böbel, N., & Willich, A. (2023). What's in the constructicon? Relating constructional forms and constructional meanings on the full range of the lexicon-grammar continuum. In *Book of abstracts of 16th International Cognitive Linguistics Conference.* https://iclc16.github.io/abstracts/ICLC16_BoA.pdf

Ziem, A., Flick, J., & Sandkühler, P. (2019). The German Constructicon Project: Framework, methodology, resources. *Lexicographica*, *35*(2019), 15–40. https://doi.org/10.1515/lex-2019-0003

Acknowledgements

This work was supported by the Estonian Research Council grant (PRG 1978).

Contact information

Ene Vainik
Institute for the Estonian Language
ene.vainik@eki.ee

Geda Paulsen
Institute for the Estonian Language, Uppsala University
geda.paulsen@eki.ee. geda.paulsen@moderna.uu.se

Heete Sahkai
Institute for the Estonian Language
heete.sahkai@eki.ee

Jelena Kallas
Institute for the Estonian Language
jelena.kallas@eki.ee

Arvi Tavast
Institute for the Estonian Language
arvi.tavast@eki.ee

Kristina Koppel
Institute for the Estonian Language
kristina.koppel@eki.ee

Appendix 1: Overview of the current initiatives

Language	Coverage	Target users	Focus	Relations	Related resources	Database format	Reference
English	73	Experts	The frame-based description of valency restrictions of lexicographic units	inheritance	Berkeley FrameNet for English	relational	Fillmore et al., 2012
Japanese	ca 50	Experts	The frame-based description of valency restrictions of lexicographic units	inheritance	Japanese FrameNet	relational	Ohara, 2018
Swedish	394	Learners Experts Language technology L2 Teachers	Semischematic constructions; the aim is to cover grammar as whole	inheritance	Swedish FrameNet; SALDO; Karp, and Korp, applications: pedagogical, LT, lexicographic; crosslingual constructicon	relational	Lyngfelt, Bäckström, et al., 2018
Brasilian Portugese	220	Experts Language technology	The frame-based description of valency restrictions of lexicographic units	(multiple) inheritance	Part of FrameNet for Brasilian Portugese; the data model has been enchanced	relational	Torrent et al., 2014
Russian	3600	Learners Experts Language technology	Semischematic constructions; the aim is to cover grammar as a whole (from morphemes to the discourse units)		Deduced from reference grammars, corpora and pedagogical materials for teachers and students	relational	Janda et al., 2018
German	225	Experts	Semischematic and semi-idiomatic constructions; the aim is to cover grammar as a whole	various horizontal and vertical	Partly linked to German frame-based dictionary on web	relational	Ziem et al., 2019
English	101	General public Learners	Patterns of extending the relational words (verbs, nouns, adjectives): the phraseology will be treated separately	inheritance	Grammar patterns from COBUILD and semantics from FrameNet	XML	Perek & Patten, 2019
Hungarian	13000	Learners	Morphemes, words, and continuous multiword expressions (without slots)	part-whole	The general dictionary will be turned into constructicon by splitting the existing complex unit and presenting them upon request	XML	Sass, 2023

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.