EESTI KEELE INSTITUUT

Constructions and their collo-profiles in the Estonian Constructicon: Identifying semantic clusters and CEFR levels

Jelena Kallas · Heete Sahkai · Geda Paulsen · Ene Vainik · Ahto Kiil · Kertu Saul · Raili Pool

23rd EAAL Conference: LANGUAGE (TEACHING & LEARNING) IN A CHANGING WORLD
April 23-24 2026, Tallinn

EESTI KEELE INSTITUUT

Background

Research project (PRG 1978, 2023–2027)

Expanding the scope of a multipurpose lexicographic resource for grammar and L2 competence

Aim

To develop a pedagogically oriented Estonian constructicon (as an extension of the EKI Combined Dictionary)

Current Stage

Workflow design
Inventory
Data model
User interface

Method

Semi-automatic compilation
Corpus-based

EESTI KEELE INSTITUUT

Basic notions

Constructions – conventional, learned form-mearning pairings at varying levels of abstraction and complexity (Goldberg 2013)

Constructicon – a theoretical conception of language as a structured inventory of constructions; a collection of construction desriptions (Lyngfelt 2018)

Collexemes – lexems attracted to a construction **(**Stefanowitsch & Gries 2003)

Collostructions – lexeme-construction collocations **(**Stefanowitsch & Gries 2003; Herbst 2018, 2020)

Collo-profile – set of collexemes associated with a constructional slot (Herbst 2020; Ziem & Feldmüller 2023)

Case study: Nominal Quantifier Construction

EESTI KEELE INSTITUUT

Two nominal slots

[Noun + $Noun_{part}$]
the first noun functions as a quantifier
the noun in the partitive case refers to the quantified entity

tass *kohvi*
cup-NOM coffee-PART
'a cup of coffee'

karp *komme*
box-NOM sweet-PART
'a box of sweets'

Pictures generated by ChatGTP 5.0

Reference: Heete Sahkai, Jelena Kallas, Ahto Kiil, Geda Paulsen, and Kertu Saul. 2025. Using large language models to identify constructional instances in corpus data. In Electronic Lexicography in the21st Century (eLex 2025): Intelligent Lexicography. Lexical Computing CZ s.r.o.

Case study: Nominal Quantifier Construction

Picture generated by ChatGTP 5.0

412 quantifier nouns registered in the EKI Combined Dictionary

At least 665 quantifier nouns in Estonian

Reference: Heete Sahkai, Jelena Kallas, Ahto Kiil, Geda Paulsen, and Kertu Saul. 2025. Using large language models to identify constructional instances in corpus data. In Electronic Lexicography in the21st Century (eLex 2025): Intelligent Lexicography. Lexical Computing CZ s.r.o.

EESTI KEELE INSTITUUT

Focus

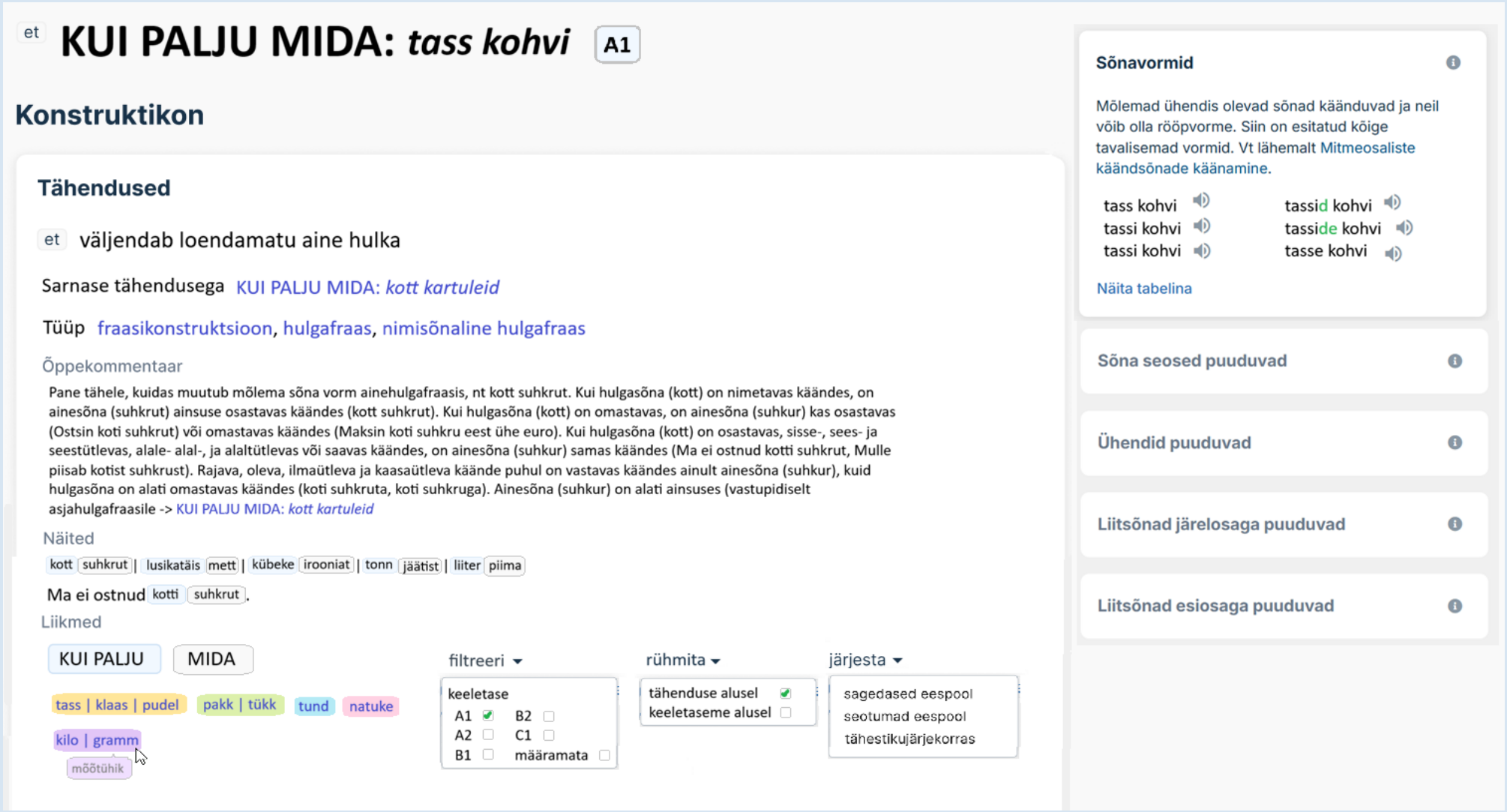
The design of a workflow that combines traditional corpus linguistic methods with the semantic capabilities of LLMs to

- structure collexemes into semantically meaningful clusters, using
  - L1 corpora
  - lexicographic data

- assign CEFR levels to constructions and collostructions, using
  - L2 textbook
  - learner corpora

Practical relevance: presenting the collo-profile in an accessible and structured way

EESTI KEELE INSTITUUT



Nominal Quantifier Construction entry in the dictionary portal Sõnaveeb (prototype)

EESTI KEELE INSTITUUT

Research questions

RQ1: Can LLMs group collexemes into categories with semantically meaningful labels, and which type of input produces the most reliable results?

RQ2: Can LLMs identify instances of constructions in L2 textbooks?

RQ3: Can LLMs identify instances of constructions in L2 learner corpora?

RQ4: Do textbooks and learner corpora correlate in providing CEFR-level evidence for constructions?

EESTI KEELE INSTITUUT

Structuring collo-profiles into semantically meaningful clusters

EESTI KEELE INSTITUUT

Data and procedure

1. Data: 192 nominal quantifiers from EKI Combined Dictionary (CombiDict) and Estonian Balanced Corpus
2. **LLM-based processing** (GPT-5.4)

Prompts using different input data:

a) word
b) word + phrase
c) word + corpus sentence
d) word + CombiDic definition
   preliminary stage: definition selection
e) word + CombiDic examples
   preliminary stage: example selection

Prompt
Based on the [input], identify the **most optimal, logical, and meaningful** classification system.
If necessary, consult online sources for comparison with other languages.
If needed, use www.sõnaveeb.ee to clarify the meanings of the words.

3. Creation of gold standard categorization (13 clusters, two independent annotators, Cohen's κ = 0,7)
4. Evaluation metrics:
   Categorization Similarity: Adjusted Mutual Information (AMI) and Adjusted Rand Index (ARI)
   Labelling similarity: human evaluation

Results

EESTI KEELE INSTITUUT

10—14 clusters depending on input data

The highest agreement (ARI=0.759, AMI=0.766) with gold standard was achieved using word+ CombiDic definition as input (comparable to the agreement between human annotators)

Category labelling shows minimal differences across input conditions (≈70–80% overlap with the gold standard)

GPT-5.4 clustering (using Word+CombiDic definition as input, 14 clusters)

1. Containers and the contained quantity(!): *kann* 'jug, pitcher', *kannutäis* 'a jugful'
2. Words for small quantities: *ivake* 'a tiny bit, a speck', *piisake* 'a small drop', *raas* 'a morsel, a tiny amount'
3. Portions: *lonks* 'a sip, a gulp', *näpuotsäis* 'a pinch', *mahv* 'a puff, a drag'
4. Words for groups and collectives: *grupp* 'group', *rühm* 'group, unit', *parv* 'flock, swarm'
5. Piece-, slice- and part- quantifiers: *käär* 'a slice', *viil* 'a slice', *riba* 'a strip'
6. Standardised units of measure and calculation: *gramm* 'gram', *kraad* 'degree', *liiter* 'liter'
7. Units of time: *päev* 'day', *kuu* 'month', *minut* 'minute'
8. Set- and bundle- words: *kiht* 'layer', *kimp* 'bundle, bunch', *kobar* 'bunch'
9. Part and proportion quantifiers: *enamus* 'majority', *enamik* 'majority'
10. Sequence and series nouns(!): *rida* 'row, line', *rivi* 'row, rank', *rodu* 'a series, a lot (often informal)'
11. Packaging and batch units: *pakk* 'pack, package', *partii* 'batch, lot'
12. Classification and type nouns: *klass* 'class', *liik* 'type, species', *sort* 'kind'
13. Words for big quantities: *hunnik* 'heap, pile', *kuhil* 'heap, mound', *lasu* 'pile'
14. Words for general quantity: *arv* 'number', *hulk* 'amount, quantity', *kogus* 'quantity, amount'

Discussion

EESTI KEELE INSTITUUT

There are several possible meaningful categorisations; ¾ of the words are easily clustered; ¼ is variable

Polysemy of the words → two or more options for categorizing, e.g *toop* (container, unit), *promill* (unit, proportion)

The advantage of using LLMs: unlike other data-driven methods (e.g. using WordNet-based similarity measures or word embeddings), LLMs provide both the clustering and the labels

EESTI KEELE INSTITUUT

Assigning CEFR levels to constructions and collostructions using L2 textbook and learner corpora

EESTI KEELE INSTITUUT

CEFR-level assignment

1. **Construction as a whole**

   The lowest level at which a schematic construction appears with different collexemes in L2 data

2. **Collostructions**

   The lowest level at which a specific collexeme appears in the construction more than once in L2 data

Corpora

L2 textbook corpora

- Estonian as a Second Language Coursebook Sentences Corpus 2021
- Estonian as a Second Language School Coursebook Sentences Corpus 2021
- 500 000 words, 57 964 sentences
- Divided into subcorpora by CEFR proficiency levels (A1–C1)

L2 learner corpus

- EMMA corpus (EKI subsection)
- 12 076 texts from tests, assignments, and exams, 134 551 sentences
- Divided into subcorpora by CEFR proficiency levels (A1–C1)

RQ2: Can LLMs identify instances of constructions in L2 textbooks?

EESTI KEELE INSTITUUT

Model	Precision	Recall	F1
Est-RoBERTa	**0.9747**	**0.9167**	**0.9448**
Claude-Sonnet-4	**0.9394**	0.7381	0.8255
o3-mini	0.8721	**0.8929**	**0.8814**
GPT-4.1	0.7293	0.7976	0.7613

Method

- EstRoBERTa: 8,500 positive + 17,000 negative examples
- 3 commercial LLMs: 10 positive + 5 negative examples

Key conclusions

- LLMs can successfully identify constructional instances **with few-shot fine-tuning**
- LLM recall results comparable to EstRoBERTa
- Can be applied to any L1 corpora

Reference: Heete Sahkai, Jelena Kallas, Ahto Kiil, Geda Paulsen, and Kertu Saul. 2025. Using large language models to identify constructional instances in corpus data. In Electronic Lexicography in the21st Century (eLex 2025): Intelligent Lexicography. Lexical Computing CZ s.r.o.

RQ3: Can LLMs identify instances of constructions in L2 learner corpora?

EESTI KEELE INSTITUUT

- **Models**: 7 commercial LLMs (Claude Sonnet 4.5, Claude Opus 4.1, Gemini Flash 2.5, o3, **o3-mini**, GPT-5-mini, **GPT-5**)
- **Prompt types**
  - **baseline** (9 rules, 15 examples)
  - **extended** (9 rules, 15 examples, role, stopwords, edge-cases)
  - **reduced** (9 rules, 3 examples, role, stopwords)
  - **minimal** (9 rules, 3 examples, role)
  - **zero-shot** (9 rules, role)
- **Results**: a very small number of examples can still support strong construction retrieval

Results for GPT-5

Prompt	Precision	Recall	F1
baseline	0.9034	0.9589	0.9302
extended	**0.9295**	**0.9932**	**0.9603**
reduced	0.8813	0.9658	0.9216
minimal	**0.8968**	**0.9521**	**0.9236**
zero-shot	0.9214	0.8836	0.9021

Reference: Heete Sahkai, Jelena Kallas, Ahto Kiil, Geda Paulsen, and Kertu Saul. 2025. Using large language models to identify constructional instances in corpus data. In Electronic Lexicography in the21st Century (eLex 2025): Intelligent Lexicography. Lexical Computing CZ s.r.o.

RQ4: Do the textbook and learner corpora correlate in providing CEFR-level evidence for constructions and collostructions?

Goal: to evaluate and compare textbook and learner corpora as sources for assigning CEFR levels

Method:

- extraction of the instances of the construction from the textbook and learner corpus (see RQ2 and RQ3)
- identification of the collexemes of the construction in the two corpora
- assignment of CEFR level to the construction based on each corpus (i.e. the lowest level at which the construction appears with more than one collexeme in the corpus)
- assignment of CEFR level to each collostruction based on each corpus (i.e. the lowest level at which the collostruction appears more than once in the corpus)
- comparison of the collexemes and CEFR levels identified based on the two corpora

RQ4 results 1: Collexemes of the NQC in textbook and learner copora

EESTI KEELE INSTITUUT

CEFR level	Collexemes in textbook corpus	Collexemes in learner corpus
A1	*tass, klaas, pudel, pakk; tükk, natuke; kilo, gramm, tund*	–
A2	*teelusikatäis, hulk, enamik, paar; kilogramm, tonn, liiter; minut, tunnike, päev, nädal, kuu*	*pudel, kilo, pakk, tund*
B1	*tassike, kann, keedukann, tuub, tünder; kimp, kuhi, rida; osa, arv, jagu, suutäis, korvitäis, limonaadipudelitäis; protsent*	*aasta, enamik, enamus, grupp, hulk, hunnik, kuu, nädal, osa, paar, peotäis, tükk, valik, raas*
B2	*kamp, koorem, põlvkond, valik; konteineritäis, hetk; kilomeeter*	*arv, minut*
C1	*näpuotsatäis*	–

NQC collexemes in textbook (both school and adult textbooks) and learner corpus

- Twice as many collexemes in the textbook vs. learner corpus: 40 vs. 20.
- The construction appears one level lower in textbook vs. learner corpus: A1 vs. A2.

RQ4 results 2: Correspondence of the CEFR levels of collostructions in textbook and learner copora

Collexeme	Level in learner corpus	Level in textbook corpus	Collexeme	Level in learner corpus	Level in textbook corpus
tund ‘hour’	A2	A1	nädal ‘week’	B1	A2
pudel ‘bottle’	A2	A1	enamus ‘majority’	B1	-
kilo ‘kilogram’	A2	A1	grupp ‘group’	B1	-
pakk ‘package’	A2	A1	hunnik ‘heap’	B1	-
hulk ‘amount’	B1	A2	valik ‘selection’	B1	B2
enamik ‘majority’	B1	A2	paar ‘pair’	B1	A2
tükk ‘piece’	B1	A1	peotäis ‘handful’	B1	-
osa ‘part’	B1	B1	raas ‘crumb’	B1	-
kuu ‘month’	B1	A2	arv ‘number’	B2	B1
aasta ‘year’	B1	-	minut ‘minute’	B2	A2

CEFR levels of collexeme-NQC pairings in learner and textbook corpus

- The majority of the collexemes in the learner corpus also appear in the textbook corpus (14 out of 20, or 70%).
- Of the 14 words that appear in both corpora, 12 appear one or two levels lower in the textbook vs. learner corpus, just like the construction as a whole appears at a lower level in the textbook corpus.

EESTI KEELE INSTITUUT

RQ4 Conclusions

- There is a systematic correspondence between the CEFR levels identified based on textbook and learner corpora: in the textbook corpus, constructions and collostructions appear one or two levels lower. The correspondence may indicate a natural learning order.
- The collexemes in the learner corpus also appear in the textbook corpus, but not vice versa.

Likely reasons:

- different corpus size
- different size of active vs. passive vocabulary
- topic restrictions in assignments

EESTI KEELE INSTITUUT

Overall conclusions

- LLMs can be used to organize constructional collexemes into semantic categories with meaningful labels
- LLMs can be used to identify instances of constructions both in textbook and learner corpora
- Textbook and learner corpora correlate in terms of the CEFR levels of constructions and collostructions
  - The level based on textbook corpus indicates the level at which a construction or collostruction is included in the teaching materials – useful for teaching
  - The level based on learner corpus indicates the level at which a construction or collostruction has been acquired – useful for assessment

References

Goldberg, A. (2013). Constructionist approaches. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 15–31). Oxford University Press. doi.org/10.1093/oxfordhb/9780195396683.013.0002

Herbst, T. (2018). Is language a collostructicon? A proposal for looking at collocations, valency, argument structure and other constructions. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical collocation analysis: Advances and applications* (pp. 1–22). Springer.

Herbst, T. (2020). Constructions, generalizations, and the unpredictability of language: Moving towards colloconstruction grammar. *Constructions and Frames, 12*(1), 56–96.

Lyngfelt, B. (2018). Introduction: Constructicons and constructicography. In B. Lyngfelt, L. Borin, K. Ohara, & T. T. Torrent (Eds.), *Constructicography: Constructicon development across languages* (pp. 1–18). John Benjamins Publishing Company. doi.org/10.1075/cal.22.01lyn

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243. doi.org/10.1075/ijcl.8.2.03ste

Ziem, A., & Feldmüller, T. (2023). Dimensions of constructional meanings in the German constructicon: Why collo-profiles matter. *Yearbook of the German Cognitive Linguistics Association, 11*(1), 203–226. doi.org/10.1515/gcla-2023-0010

EESTI KEELE INSTITUUT

Thank you!