Exploring the constructicographic potential of lexicographic data and language models: The case of the Estonian Nominal Quantifier Construction

Heete Sahkai[1], Geda Paulsen[1,2], Ene Vainik[1], Jelena Kallas[1], Ahto Kiil[3], Katrin Tsepelina[1,3], Kertu Saul[1,3] and Arvi Tavast[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia
[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden
[3] University of Tartu, Ülikooli 18, Tartu 50090, Estonia

E-mail: heete.sahkai@eki.ee, geda.paulsen@eki.ee, ene.vainik@eki.ee, jelena.kallas@eki.ee, ahto.kiil@gmail.com, katrin,tsepelina@eki.ee, kertu.saul@eki.ee, arvi.tavast@eki.ee

Abstract

Constructicography, or the description of grammatical constructions in a lexicographic format, is an emerging field currently in the stage of developing and automating methods for treating large numbers of (semi-)schematic constructions. This study explores how existing lexicographic data and language models can be used to facilitate the constructicographic workflow. Our results suggest that (1) collocations and semantic relations represented in a lexicographic database can be used to identify the collexemes of constructions, that is, the lexemes occurring in the open slot(s) of schematic constructions, (2) BERT-based language models can be trained to identify instances of constructions in corpora, using collocations as the starting point to create appropriate training data, and (3) commercial large language models can be prompted to identify constructional instances, using a small number of examples. The identification of the collexemes and corpus instances of constructions provide several pieces of information that can be represented in constructicon entries: the meaning, form, frequency and productivity of constructions, the frequency and association strength of particular collexemes, the CEFR-level of the construction, etc.

Keywords: Constructicography; BERT-based models; Large Language Models; Lexicography; Collostructional Analysis; Estonian

1. Introduction

Construction-based view of linguistic knowledge (e.g. Fillmore et al., 1988; Goldberg, 1995; Kay & Fillmore, 1999) has led to Charles Fillmore's idea of a constructicon (Fillmore, 2006; 2008) – an assemblage of complex linguistic units with variable components associated with a particular meaning in a resource like an electronic dictionary. This kind of approach has resulted in several constructicographic projects, and the formation of a field known as constructicography (see Lyngfelt et al., 2018; Borin & Lyngfelt, 2025 for an overview).

In addition to determining what constructions exist, what their typical properties are, and how they can be presented to users in a lexico-grammatical resource in an accessible manner, it is essential to develop an efficient and consistent workflow (cf. Ziem et al., 2019) that establishes systematic and replicable procedures. These procedures should enable the gathering of information about the wide range of (semi-)schematic constructional candidates, as well as the identification of information necessary for describing constructional properties, including the slot-fillers of (semi-)schematic constructions.

For gathering constructional candidates different approaches exist (cf. Barteld & Ziem, 2020): (1) the constructicon is built on existing lexicographical and/or lexical-semantic resources (e.g. Herbst & Hoffmann, 2024; Perek & Patten, 2019; Sass, 2023); (2) construction candidates are identified using a "full-text" annotation approach, allowing researchers to detect and define construction candidates on the fly by manually annotating text (e.g. Lee-Goldman & Petruck, 2018); (3) manual collection of the construction candidates from different types of L1 and L2 language teaching and learning resources (e.g. Janda et al., 2020; Lyngfelt et al., 2018); (4) construction mining, including techniques based on n-grams (Shibuya & Jensen, 2015; Wible & Tsao, 2010), hybrid n-grams and skip-grams (Bäckström et al., 2013; Dunn, 2017; 2023), and syntactic n-grams (Sidorov, 2019). Barteld and Ziem (2020) have also developed a dedicated tool that allows users to define both the pattern type and parameters to filter extracted patterns, resulting in a list of constructional candidates generated from corpus data.

For describing constructions, constructicographers typically assess their frequency, productivity, and semantic properties. One of the most widely used methods is collostructional analysis (Stefanowitsch & Gries, 2003). Novel approaches involve applying language models to identify constructional collo-profiles, which help in describing the meaning potential of constructions (Ziem & Feldmüller, 2023). However, some studies suggest that LLMs do not acquire representations of fully schematic constructions, as they have access only to the lower levels of the constructional hierarchy (Bonial & Tayyar Madabushi, 2024).

This study continues the latter line of research, exploring the possibility of using existing lexicographic data and language models to facilitate the description of constructions. In particular, we will examine how existing lexicographic data can be used to identify the lexical slot-fillers, or collexemes (Stefanowitsch & Gries, 2003), of (semi-)schematic constructions, how to train BERT-based language models to identify the instances of constructions in corpora, and whether LLMs can be used to identify constructional instances in corpus data.

These tasks – and the retrieval of the complete corpus concordance of a construction, and the identification of constructional collexemes – can serve several purposes in a constructicographic workflow. The corpus retrieval of the instances of constructions is

necessary for the identification of several pieces of information that could be represented in a constructicon entry: the formal details of its components, its frequency and productivity, its collexemes and their association strength with the construction (Stefanowitsch & Gries, 2003). The semantic analysis and categorization of the construction's (strongly associated) collexemes, in turn, allows to identify any semantic preferences regarding the collexemes of the construction, as well as the meaning of the construction (Stefanowitsch & Gries, 2003; Ziem & Feldmüller, 2023). Collexeme identification is particularly important in case of constructicons that list the collexemes of a construction in the constructicon entry (e.g. Patel et al., 2023; Ziem & Feldmüller, 2023; Perek & Patten, 2019). One step further would be to link construction entries with the entries of their collexemes in a lexicographic database. This is the goal of the recent initiative by the Institute of the Estonian Language (Vainik et al., 2024), which aims to develop the constructicon as an extension of a lexicographic database, and to describe, on the one hand, (semi-)schematic constructions, and, on the other hand, the grammatical behavior of lexemes by linking them to constructicon entries that represent their typical usage patterns.

Additional pieces of information can be identified via the extraction of constructional instances from specialized corpora. In particular, the distribution of constructional instances in L2 corpora can reveal information that is relevant to L2 learners, teachers, assessors, and teaching materials' compilers. This kind of information includes, for instance, the CEFR-level at which a particular construction is first acquired or should be taught, as well as the distribution of particular construction-collexeme pairings across CEFR-levels. Similarly, the distribution of a construction's instances across genre-specific corpora can yield information about the construction's register (Risberg et al., 2025; Pilvik et al., 2025).

The corpus retrieval of the instances of a schematic construction is, however, challenging: constructions are pairings of form and meaning and are thus rarely identifiable by a unique combination of morphosyntactic labels. This paper will therefore explore solutions that facilitate the identification of constructional collexemes and the corpus retrieval of constructional instances, making use of existing lexicographic data and language models.

The solutions will be put to test, using the Estonian Nominal Quantifier Construction (NQC) as a case study, see e.g. Metslang 2017; Pilvik et al., 2025. NQC is a (pseudo-)partitive construction consisting of two nouns (see ex. 1a, b[1]). The first noun functions as a quantifier, and the second noun denotes the referent (in partitive use) or the kind of entity (in pseudo-partitive use) quantified over (Koptjevskaja-Tamm, 2001).

[1] Examples are taken from the Balanced Corpus of Estonian.

(1) a. Partitive use

Enamik ***selle*** ***rühma*** ***keeli*** *on* *kasutusel* *kui* *lingua*
majority DEM.GEN group.GEN language.PL.PAR be.3SG use.ADE as lingua

franca erinevates *maailma* *regioonides.*
franca different.PL.INE world.GEN region.PL.INE

'The majority of the languages in this group are spoken as a lingua franca in different regions of the world.'

b. Pseudo-partitive use

Ettevõte *kulutab* ***hulga*** ***raha*** *paludes* *mõnel*
company spend.3SG amount.GEN money.PAR ask.CONV some.ADE

reklaamifirmal *oma* *tootele /* *teenusele* *kaubamärk* *välja* *töödata.*
advertising.company.ADE REFL product.ADE service.ADE brand out work.INF

'The company spends a large amount of money asking an advertising company to develop a brand for its product/service.'

The quantifier usually functions as the head of the construction. The complement noun shows complex inflectional behavior: its number depends on its countability and its case varies depending on the syntactic function of the construction in the sentence. When the construction functions as a subject or object, the complement is in partitive case; otherwise the quantifier and the complement agree in case (four cases, terminative, essive, abessive, and comitative, show suspended affixation where the case marker appears only in the last case-marked constituent of the NP).

Nouns vary as to their ability to be used as quantifiers. Some nouns can be used as quantifiers directly; others can function as quantifiers via a productive word formation process whereby they are compounded/affixed with the element *-täis* '-ful' (*lusikas* 'spoon' > *lusikatäis* 'spoonful').

From a constructicographic perspective, especially with the view of listing the potential lexical fillers of the nominal quantifier (NQ) slot and serving a pedagogical purpose, some relevant questions regarding the construction are: which nouns can function as NQs, is the construction productive and if yes, which semantic classes of nouns are attracted to it, at which CEFR-level is the construction acquired and should be taught. To answer these questions, we need an overview of the NQ collexemes of the construction as well as corpus data reflecting the productivity of the construction, the association strength between the construction and the collexemes, and the distribution of the constructional instances in L2 corpora. Due to the complex and, at the same

time, non-unique formal properties of the construction, its corpus extraction is, however, laborious. We will therefore explore ways to facilitate the data-gathering process.

In particular, we will address the following three research questions (RQ):

> RQ1. To what extent can potential fillers of the NQ slot be identified based on an existing lexicographic database, using (1) the collocations that represent the NQC, and (2) synonymy relations represented in the database?
>
> RQ2. Can BERT-type language models be trained to identify instances of the NQC in corpora, using lexicographic collocation data as a starting point to create training data?
>
> RQ3. Can commercial large language models be prompted to identify instances of the NQC in corpora, using only a small number of examples?

Previous research (Ziem & Feldmüller, 2023) has made use of BERT models in order to predict the collo-profiles of constructions, that is, their typical collexemes, for the purpose of their semantic description. Our goal is to go a step further and to use BERT-type models and LLMs for the corpus retrieval of constructional instances. In addition to the identification of the semantic profile of the construction, the corpus retrieval of its instances would allow to identify its frequency and productivity, and to conduct a full collostructional analysis. Additionally, the extraction of the instances of the construction from L2 corpora would allow for the identification of CEFR-related information.

The paper is structured as follows: the lexicographic data, resources and tools used in the study are described in Section 2; the procedures undertaken are described in Section 3, and the results and discussion in Section 4. The paper ends with conclusions and future work in Section 5.

2. Data, tools and resources

This Section describes the lexicographic data (2.1), corpora (2.2), and language models (2.3) used in the study.

2.1 Lexicographic data

The lexicographic data used in the study come from the EKI Combined Dictionary (CombiDic) (Koppel et al., 2019). The dictionary is compiled within Ekilex (Tavast et al., 2018), a relational database that employs a unified data model across multiple dictionaries. There are 188,251 headwords in the CombiDic, and this number is gradually increasing as new words and meanings are continuously added to the

database. CombiDic's lexicographic information layers include definitions, examples, collocations, semantic relations (synonyms, antonyms, co-hyponyms), government patterns, inflectional paradigms, semantic types, etymological information, translations, grammatical comments (e.g. usually in plural, usually in the 3rd person), language planning notes, CEFR levels for senses, and pronunciation information.

In the present study, we made use of synonyms and collocations. Two types of synonymic relations are distinguished: total synonyms and partial synonyms. Total synonyms have been defined for 25,777 lexemes. They are manually added to the database by lexicographers and share the same definition. Partial synonyms, covering 93,074 lexemes, are selected manually by lexicographers from candidate lists (Tavast et al., 2020) using a dedicated interface in Ekilex, which allows linking a suitable synonym to the appropriate sense of a given headword. As for collocations, CombiDic includes collocations for the 10,000 most frequent Estonian words (nouns, adjectives, adverbs, verbs, numerals, and proper nouns). Altogether, it contains 315,000 collocations tagged with 87 types of morphosyntactic classifiers, using both categorical and functional-relational labels – for example, Adj_modifies/Adj_modifier, subject/subject_of (see Kallas et al., 2015 for details). Collocations were first detected from the Estonian National Corpus 2013 using the Sketch Engine corpus query system (Kilgarriff et al., 2004; 2014), and then post-edited manually. As collocates, nouns – including those functioning as subjects, objects, or adverbials – adjectives, participles, adverbs, pronouns, prepositional phrases, non-finite verbs, and subordinate clauses are registered across different parts of speech. In addition, some collocations are supplemented with example sentences.

2.2 Corpora

The corpora used in the current study are from the Estonian National Corpora series (Koppel & Kallas, 2022): Estonian National Corpus 2023 (henceforth *ENC 2023*), the Estonian National Corpus 2021 (henceforth *ENC 2021*), and the Estonian National Corpus 2017 (henceforth *ENC 2017*). All corpora are accessible via the Sketch Engine interface and are also available as relational databases containing proprietary collection structures, as documented on GitHub[2]. ENC 2023 comprises the following subcorpora: the Estonian Reference Corpus (henceforth *Reference Corpus*), incl. the Balanced Corpus (henceforth *Balanced Corpus*); the Estonian Web Corpora (2013, 2017, 2019, 2021, 2023) (henceforth *Estonian Web*); Wikipedia (2017, 2019, 2023) and Wikipedia Talk 2017; Old Literature (1864–1945), Contemporary Literature (2000–2023); Estonian Feeds (2014–2021, 2020–2023), and the corpus of open access journals DOJA (2020–2023). ENC 2021 includes Reference Corpus, Balanced Corpus; Estonian Web (2013, 2017, 2019, 2021); Wikipedia 2021 and Wikipedia Talk 2017; DOAJ (2020-2021), Estonian Feeds (2014–2021), and Literature. ENC2017 comprises four subcorpora: Reference Corpus, Web Corpora (2013, 2017), and Wikipedia 2017.

[2] https://github.com/estnltk/estnltk-workflows/tree/master/enc_workflows (7 October 2025)

For evaluation purposes, we also used the Estonian as a Second Language Coursebook Sentences Corpus 2021 (henceforth *Coursebook Corpus*) and the Estonian as a Second Language School Coursebook Sentences Corpus 2021 (henceforth *School Coursebook Corpus*). Both corpora are accessible via the corpus query system KORP[3]. These corpora are divided into subcorpora corresponding to different CEFR proficiency levels (A1–C1), as well as to specific stages of the Estonian school system (e.g. 3rd grade, 7th grade, 9th grade, gymnasium), making them suitable for level-specific analysis.

2.3 Language models

The language models tested in this study can be divided into two general groups: BERT-based and commercial large language models. We tested the following three transformer-based Estonian-specific BERT-models: the EstBERT and two EstRoBERTa models. The EstBERT model (Tanvir et al., 2021) was trained using ENC2017 on both 128- and 512-token sequence lengths. The first EstRoBERTa model is a monolingual Estonian BERT-like model, closely related to the French CamemBERT model. The corpora used for training this model have 2.51 billion tokens in total, which is a much larger dataset than used for training the EstBERT (Ulčar & Robnik-Šikonja, 2021). The second EstRoBERTa model has been trained by the TartuNLP work group (henceforth *TartuNLP/EstRoBERTa*).

Among the commercial large language models, we selected three models: Claude-Sonnet-4, OpenAI's GPT-4.1, and o3-mini. Claude-Sonnet-4 is a hybrid reasoning large language model from Anthropic, available on the Claude AI home page. We tested this model directly in the chat window. GPT-4.1 was the most common generative pre-trained transformer (gpt) model at the testing time. o3-mini is an affordable version of OpenAI's "reasoning" model. Both OpenAI models were tested via API.

3. The procedure

In this section, we describe the procedure used for answering the research questions (cf. Section 1). Section 3.1 presents the creation of the evaluation benchmarks – the NQC gold standard dataset and the model based on this data. The process of collecting all the collexemes of the NQC in CombiDic (answering RQ1) is reported in Section 3.2. Sections 3.3 and 3.4 are devoted to the training and prompting processes of the BERT-based language models (answering RQ2) and the large commercial language models (answering RQ3) to identify instances of NQC.

[3] https://korp.keeleressursid.ee/?mode=coursebook2021#?lang=et (7 October 2025)

3.1 Creating the evaluation benchmarks

To evaluate our results, we started by creating the following three benchmarks: a manually verified dataset of the corpus instances of the NQC (henceforth *the Gold Standard Dataset*); list of the collexemes of the NQC along with their frequency and association strength, identified from the Gold Standard Dataset; and a language model trained on the Gold Standard Dataset (henceforth *the Gold Model*).

The Gold Standard Dataset was created by extracting the instances of the NQC from the Balanced Corpus, using morphosyntactic queries (all possible ordered combinations of dependency relation, part-of-speech, case and number tags that instances of NQC can have) and manual verification. The discarded sentences in the query result were later used as the source of negative examples. The Gold Standard Dataset consisted of 9157 sentences with 436 different nominal quantifiers, including 191 hapax legomena. The hapaxes/tokens ratio is 0.02, suggesting that the construction is partially productive (Baayen, 2009).

The Gold Standard Dataset was used to perform a collostructional analysis, using the program Coll.analysis 4.0 (Gries, 2022). The analysis provided an overview of the type frequency of the NQC as well as the frequency and association strength of its collexemes, allowing us to evaluate the results of RQ 1. The statistics used to measure the association strength included among others Log-Likelihood Ratio, Log Odds Ratio, and Pointwise MI.

To evaluate the results of RQs 2 and 3, we aimed to develop the best possible BERT-based model for NQC identification, using the Gold Standard Dataset to fine-tune EstBERT (the 512 tokens sequence length version) and the two EstRoBERTa models. To simplify the task, we only used the NQC instances where the complement noun is in partitive case, which form most of the Gold Standard Dataset (8659 instances, 159 were used as test set during development, 8500 for training). We trained all models in the Google Colab environment using T4 GPUs; the batch size of 8 gave the best results. As the models are relatively small (~ 500 MB to 1.1 GB), this approach is suitable for development and testing purposes.

We experimented with 4 different ratios of positive and negative examples in the training data: 1:1, 1:2, 2:1, and 3:2. The performance metrics showed that for this task, TartuNLP/EstRoBERTa model performed slightly better than the other tested models in precision (0.8788), recall (0.8561), F1-score (0.8609) and accuracy (0.9646), with the ratio of positive-negative examples 1:2. We aimed for the highest recall as it results in the largest number of NQC instances extracted from a given input text. The Gold Model was used as the benchmark model to evaluate the results of research questions 2 and 3.

3.2 Identifying the NQs in the lexicographic data

To answer RQ1 – to what extent can the collexemes of the NQC, that is, the potential fillers of the NQ slot, be identified based on CombiDic data – we proceeded as follows: (1) we extracted all the collocations tagged with the morphosyntactic classifiers 'partitive_modifier' and 'partitive_modifies', i.e., co-occurances of two nouns the second of which is in partitive case, and checked them manually. The CombiDic database contained altogether 1367 unique collocations instantiating the NQC; (2) we retrieved all unique NQs from the collocations instantiating the NQC; (3) we retrieved all the synonyms of these NQs from the CombiDic database; (4) we verified whether the retrieved synonyms occur as nominal quantifiers. To do this, we first examined whether they occur as quantifiers in their dictionary example sentences; if not, we also examined whether they occur as quantifiers in ENC 2021, using morphosyntactic corpus queries. Some words could be eliminated without further verification; for example, the word *päts* 'loaf' has a homonym that is a folk term for 'bear', meaning that the list of words retrieved via semantic relations included the word *karu* 'bear', which we excluded as a potential quantifier without further checking; (5) we compared the result with the Gold Standard Dataset and the list of collexemes based on the Gold Standard Dataset.

The training procedure of both models described in Section 3.3 below was based on the CombiDic NQC collocations.

3.3 Training process of the Tartu NLP/EstRoberta model

To answer RQ2, we chose the BERT-based model for Estonian that performed best in the NQC extraction task (cf. Section 3.1) – the Tartu NLP/EstRoBERTa model. In the training process of the Tartu NLP/EstRoBERTa model, we used the CombiDic collocations (cf. Section 3.2) to create a set of training data. This procedure resulted in two submodels of Tartu NLP/EstRoBERTa – M1 and M2. The difference between the two models lies in the degree of automaticity: the M1 can be implemented as a fully automated workflow, M2 involves human input in the form of validating the training data.

For best comparison, both M1 and M2 models were trained on training sets of approximately the same size (Gold Model – 8500 positive cases, M1 – 8807 positive cases, M2 – 8523 positive cases), with the same 1:2 (positive : negative) ratio. The fluctuation in training data size is caused by the number of unique phrases – we aimed for a balanced distribution of training examples in the M1 (1329 unique phrases) and M2 (5394 unique phrases) models. We finetuned the TartuNLP/estRoBERTa pretrained model in Google Colaboratory environment for 5 epochs, batch_size=8.

The procedure for creating the M1 was the following:

(1) We used the 1367 unique NQC collocations identified from CombiDic (see previous Section) to extract sentences with these exact collocations from different subcorpora of ENC 2023: Wikipedia (2017, 2019, 2023), Estonian Web (2019, 2021) and Contemporary Literature (2021) corpora. The final training set contains 8807 positive examples with 1329 unique phrases (38 phrases from the collocations list had no exact matches in corpora) – 97.2% of the initial list.

(2) We added 17614 negative examples (sentences without quantifier construction) that we extracted from the ENC 2023 Wikipedia 2023 subcorpus. The selection criteria were: no words from the quantifier set in the sentence; sentence must contain two nouns – first noun in [sg/pl nom, gen, part], second noun in [sg/pl part]; 'parent-child' condition ignored.

To finetune the M2 model, we performed the following steps:

(1) Using fine-tuned M1, we predicted quantifier expressions from Wikipedia 2023 and Literature corpora.

(2) 8920 unique phrases were found, and we checked their validity manually (~ 7 hours manual labor). In result, 4192 new phrases were selected.

(3) We extracted additional training sentences with these new phrases from ENC 2023, Wikipedia (2017, 2019, 2023), Estonian Web (2019, 2021), and Contemporary Literature (2021) subcorpora. The final training set contains 8523 positive examples with 5394 unique phrases.

(4) We added 17046 negative examples (1:2 ratio between positive/negative cases has shown the best results so far); 25569 examples in total. The selection criteria were the same as for model M1.

To evaluate the performance of M1 and M2 in comparison with the Gold Model, we used the C1-level subcorpus of the Coursebook Corpus as a test set. The subcorpus contains 745 sentences and 13 instances of NQC. To increase the proportion of positive instances in the test set, we added 71 positive instances from the gymnasium-level subcorpus of the School Coursebook Corpus. In total, the test set contained 816 sentences, of which 84 with an instance of the NQC.

3.4 Testing process of the selected commercial LLMs

To answer RQ3, we tested three commercial LLMs: Claude-Sonnet-4, gpt-4.1, and o3-mini, using the test set described in Section 3.3. Instructions to all three models were given in the form of this prompt:

```
TASK: In each input sentence find the first quantifier construction
that satisfies ALL rules 1-8 below and output it in the format
```

<quantifier-phrase>;<space><original sentence> If the sentence contains **no** valid quantifier construction, output NO_QUANT;<space><original sentence>
Definitions & rules:
1. A quantifier construction always consists of **exactly two words**.
2. **Word 1** is a noun in nominative, genitive, or partitive (sg./pl.) that expresses a measurable quantity (time, volume, amount, etc.).
3. **Word 2** is a noun in **partitive** (sg./pl.) that denotes the thing being quantified.
4. The two words may be adjacent or separated by other words.
5. Morphosyntactically, Word 2 is the **parent** (head) of Word 1.
6. Word 1 always **precedes** Word 2 in the sentence.
7. Exclude numeral-like words that are not general quantifiers (miljon, tuhat, sadakond, mustmiljon, …). The following units ARE valid quantifiers: gramm, kilo, meeter, liiter, tonn etc. (and their inflected forms).
8. Pure numerals (1, 12, 110) and their written-out forms (üks, kaksteist, sadakümme, …) are **not** quantifiers.
9. Adjectives (pikk, suur, lai etc.) and adverbs (mitu, palju, natuke) are **not** quantifiers.

After the instructions, we provided for Claude-Sonnet-4 many-shot (50 of each) examples with consecutive ("*aasta aega*; *Juba peaaegu **aasta aega** olen pidanud kirjavahetust kloostri abti isa Jeaniga.*"), non-consecutive ("*hulka rühmitusi; Algkristlus hõlmas suurt **hulka** erinevaid **rühmitusi** ja **arusaamu**.*") and negative cases ("NO_QUANT; *Loodan väga, et keegi neist ei kavatse elustada veritasu traditsioone.*").

For the OpenAI models, we provided few-shot (7 of each) examples with consecutive, non-consecutive and negative cases for every query. To avoid request time-out errors, we only queried one sentence per two seconds; this time interval can be reduced.

For Claude-Sonnet-4, we first provided the entire test set as a text file. This attempt was a failure: the model included training sentences and did not follow the instructions. We then provided sentences in smaller batches (starting from 100) and with 25 sentences per query the results were satisfactory. We had no access to test the model via API (this should be considered as a possibility in the future). The other commercial LLMs, the OpenAI models gpt-4.1 and o3-mini, were tested via API.

4. Results and discussion

Section 4.1 will discuss RQ1 and Section 4.2 will discuss RQs 2 and 3.

4.1 Identification of NQs from lexicographic data

Our first research question was to what extent potential fillers of the nominal quantifier slot of the NQC can be identified based on the CombiDic database, making use of (1) the collocations that represent the NQC, and (2) synonymy relations represented in the database.

The CombiDic contains altogether 1688 collocations with the labels 'partitive_modifes/modifer', 1634 of which instantiate the NQC. The number of unique NQs included in the collocations is 89 and seven of them are absent from the Gold Standard Dataset. The 20 most frequently occurring NQs in the CombiDic collocations largely coincide with the 20 most frequent and the 20 most strongly associated NQs in the Gold Standard Dataset (see Table 1).

20 most strongly associated NQs in the Gold Standard (statistic LLR)	**20 most frequent NQs in the Gold Standard**	**20 most frequent NQs in the CombiDic collocations**
hulk 'amount'	hulk 'amount'	**hulk** 'amount'
enamik 'majority'	enamik 'majority'	**rida** 'row, range'
tükk 'piece'	tükk 'piece'	**osa** 'part'
osa 'part'	osa 'part'	**enamik** 'majority'
rida 'row, range'	rida 'row, range'	**tonn** 'tonne'
kuu 'month'	*kuu* 'month'	**kilo** 'kilo'
tonn 'tonne'	*tund* 'hour'	**pakk** 'package'
tund 'hour'	*kroon* 'crown'	**liiter** 'litre'
liiter 'litre'	tonn 'tonne'	**hunnik** 'heap'
enamus 'majority'	*aasta* 'year'	**tükk** 'piece'
pudel 'bottle'	nädal 'week'	valik 'selection'
kilo 'kilo'	*arv* 'number'	**klaas** 'glass'
jagu 'size'	*enamus* 'majority'	**pudel** 'bottle'
kroon 'crown'	pudel 'bottle'	**rühm** 'group'
pakk 'package'	liiter 'litre'	peotäis 'handful'
kogus 'amount'	*kogus* 'amount'	kimp 'bunch, bouquet'
nädal 'week'	*jagu* 'size'	minut 'minute'
klaas 'glass'	kilo 'kilo'	**nädal** 'week'
hunnik 'heap'	rühm 'group'	ports 'portion'
hektar 'hectare'	klaas 'glass'	kiht 'layer'

Table 1: Top 20 nominal quantifiers in the Gold Standard Dataset by collocation strength (according to the Log-Likelihood Ratio statistic) and raw frequency, and the 20 most frequent nominal quantifiers in the CombiDic collocations. The CombiDic NQs that coincide

with the Gold Standard NQs are in bold; the Gold Standard NQs not in the CombiDic top 20 are in italics

The extraction of the synonyms of the 89 NQs identified from the collocations resulted in 906 NQ candidates, 318 of which were identified as NQs after checking their example sentences or the ENC 2021. Thus, in total 412 NQs were identified from the CombiDic collocations and their synonyms. This number is comparable to the number of NQs in the Gold Standard Dataset, which is 435. Only 193 NQs were contained in both lists, confirming the productivity of the construction.

These results suggest that CombiDic's collocations and synonyms may provide representative information on the collexemes of a schematic partially productive construction, and even their association strength with the construction, as the 20 most frequent NQs in the collocation data largely overlap with the top 20 most frequent and mostly strongly associated NQs in the Gold Standard. More generally, it can be concluded from the results that a lexicographic resource that has been compiled on the basis of corpus data can closely reflect the collocational corpus profile of a construction, which is probably not surprising. Corpus-based lexicographic data can thus provide a less laborious alternative to corpus data for the identification of constructional collexemes. This is particularly useful when the lexicographic database that is used for this purpose is linked to the constructicon, and the identified collexemes can thus be linked to the relevant construction entries. Still, a step that requires further automation is the checking of the collexeme candidates retrieved via semantic relations from a lexicographic database.

However, NQC is a construction that is specifically targeted by a separate rule of the Estonian Sketch Grammar (Kallas, 2013), which was used to identify the collocations represented in CombiDic. The usefulness of collocation data may thus be limited to the constructions that were targeted by the corpus queries used to identify the collocations. For instance, in addition to the NQC, the classifiers 'partitive_modifies/modifier' cover another construction, which was not specifically targeted by the Sketch Grammar, the so-called Partitive Parameter Word Construction. The construction is headed by a noun in partitive case form, modified by an agreeing adjective or a noun in genitive, and it expresses a parameter of a referent, e.g. *kindlat tüüpi* [certain.par type.par] 'of a certain type', *ühiselamu tüüpi* [hall.of.residence.gen type.par] 'hall of residence-type'. In the collocations, the construction is represented by 43 instances and 9 types. While the construction is considerably less frequent and less productive than the NQC, it seems to be underrepresented in the collocations and, consequently, other methods are needed to obtain sufficient data for a constructicographic description of the construction.

4.2 Language models as tools for identifying instances of NQC in corpus data

Our second research question was whether BERT-type language models can be trained to identify instances of the NQC in corpus data, using CombiDic's collocations as a

starting point to create training data. The third research question was whether commercial large language models can be prompted to identify instances of the NQC in corpora, using only a small number of examples.

To answer RQ2, two EstRoBERTa models were trained, M1 and M2. To answer RQ3, three models were tested: Claude-Sonnet-4, o3-mini and gpt-4.1. The models were compared to the benchmark Gold Model, an EstRoBERTa model trained on the Gold Standard Dataset. To test the models' performance, all models were used to identify instances of the NQC in the test set, which was based on the Coursebook Corpus. The results in terms of four metrics (precision, recall, F-score and accuracy) are presented in Table 2.

In comparison with the Gold Model, the tested language models performed slightly worse on all statistics, but not significantly. Regarding the EstRoBERTa models, the manually improved M2 (with three times more unique NQC instances than M1) performed slightly better than M1 in three metrics but received the same result in recall (0.8095). Hence, these models do not differ significantly, and the manually assisted training process (M2) did not bring notable benefits. The results suggest, overall, that collocations can be successfully used to create training data to train language models for the purpose of extracting instances of constructions from corpora.

Model	**Precision**	**Recall**	**F-score**	**Accuracy**	**Dev Time (day)**	**Exec Time (sec)**
Gold model	0.9747	0.9167	0.9448	0.9890	2	21
M1	0.9315	0.8095	0.8660	0.9743	2	22
M2	0.9444	0.8095	0.8700	0.9755	3	21
Claude-Sonnet-4	0.9394	0.7381	0.8255	0.9684	N/A	~ 1200
o3-mini	0.8721	0.8929	0.8814	0.9755	N/A	816
GPT-4.1	0.7293	0.7976	0.7613	0.9485	N/A	21

Table 2. Comparison of the model performance (Dev time – development time; Exec time – execution time for extraction after the model has been loaded)

Regarding the commercial LLMs, o3-mini outperforms the EstRoBERTa-based models in terms of recall. The highest recall is an important performance metric as it results in the largest number of quantifier constructions extracted from a given input text. This means that commercial LLMs can be successfully used to identify instances of partially productive schematic constructions in corpus data, using only a few-shot fine-tuning. Interestingly, these results go against the schematicity hypothesis according to which LLMs do not learn representations of fully schematic constructions (Bonial & Tayyar Madabushi, 2024)[4]. In terms of resources, time, and cost, we can conclude that the effort required is significantly lower compared to manual approaches. In terms of workflow, BERT remains the more reliable and controllable option. However, it requires separate

[4] Bonial & Tayyar Madabushi's (2024) results were based on the GPT-3.5 and GPT-4 models; the o3-mini was not available yet, however, GPT-family models are comparable.

training for each construction, though once trained, it can be applied repeatedly to different types of corpora, such as learner corpora or textbook corpora.

5. Conclusions and next steps

The results suggest that a lexicographic database can provide representative information on the collexemes of a construction: the number of NQs retrieved from the CombiDic database via collocations and semantic relations was nearly equal to the number of NQs identified from corpus, and the 20 most frequent NQs occurring in the CombiDic collocations largely coincided with the top 20 most frequent and most strongly associated NQs in the corpus data. However, the usability of collocation data may be limited to constructions that were specifically targeted during the identification of the collocations. Also, automation of the verification of the collexeme candidates retrieved via semantic relations is needed. The results additionally showed that collocations can be successfully used to create training data to train language models for the purpose of extracting instances of constructions from corpora.

The most promising result of the study is the potential of LLMs to identify instances of constructions in corpus data with few-shot fine-tuning: with only 14 positive and 7 negative example sentences, o3-mini achieved a recall (0.89) that is comparable to the benchmark EstRoBERTa model trained on 8500 positive and 17,000 negative examples based on manually verified corpus data.

Our next steps will include three lines of research. The first line of research continues to explore the potential of large language models to retrieve corpus instances of constructions, focusing on different types of constructions, smaller sets of training data, larger sets of corpus data, prompt design, and developing a user-friendly interface to exploit language models. The second line of research will continue to test the constructicographic potential of further types of lexicographic data, like government patterns, definitions, and semantic types. The third line of research will develop solutions for modelling the constructicon data model, including linking lexical entries with (semi-)schematic constructions, as well as solutions for the user interface, e.g. how to display these links within the entries of lexical items and within the constructicon entries.

6. Acknowledgements

This study was supported by the Estonian Research Council grant PRG1978.

7. References

Baayen, H. (2009). Corpus linguistics in morphology: Morphological productivity. *Corpus Linguistics: An International Handbook*, pp. 899–919. doi.org/10.1515/9783110213881.2.899

Barteld, F., & Ziem, A. (2020). Construction mining: Identifying construction candidates for the German constructicon. *Belgian Journal of Linguistics*, *34*, pp. 5–16. doi.org/10.1075/bjl.00030.bar

Bonial, C., & Tayyar Madabushi, H. (2024). Constructing understanding: On the constructional information encoded in large language models. *Language Resources and Evaluation*. doi.org/10.1007/s10579-024-09799-9

Borin, L., & Lyngfelt, B. (2025). Framenets and constructiCons. *The Cambridge Handbook of Construction Grammar*. Cambridge University Press. https://www.academia.edu/95301779/Framenets_and_constructiCons

Bäckström, L., Borin, L., Forsberg, M., Lyngfelt, B., Prentice, J., & Sköldberg, E. (2013). Automatic identification of construction candidates for a Swedish constructicon. In *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013. NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88*, pp. 2–11.

Dunn, J. (2017). Computational learning of construction grammars. In *Language and Cognition: An Interdisciplinary Journal of Language and Cognitive Science*, *9*(2), pp. 254–292. doi.org/10.1017/langcog.2016.7

Dunn, J. (2023). *Exploring the Constructicon: Linguistic Analysis of a Computational CxG*. arxiv.org/abs/2301.12642

Fillmore, C. J. (2006, September 3). *The articulation of lexicon and constructicon* [Plenary lecture]. Fourth International Conference on Construction Grammar (ICCG4), University of Tokyo, Japan.

Fillmore, Charles J. (2008). Border Conflicts: FrameNet Meets Construction Grammar. In *Proceedings of the XIII EURALEX International Congress.* Barcelona: Universitat Pompeu Fabra, pp. 49–68.

Fillmore, C. J., Kay, P., O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions. In *Language* 64, 501–538.

Forsberg, M., Johansson, R., Bäckström, L., Borin, L., Lyngfelt, B., Olofsson, J., & Prentice, J. (2014). From construction candidates to constructicon entries: An experiment using semi-automatic methods for identifying constructions in corpora. In *Constructions and Frames*, *6*(1), pp. 114–135. doi.org/10.1075/cf.6.1.07for

Gries, S. T. 2022. *Coll.analysis 4.0. A script for R to compute perform collostructional analyses*. https://www.stgries.info/teaching/groningen/index.html

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Herbst, T., & Hoffmann, T. (2024). *A Construction Grammar of the English Language: CASA – a Constructionist Approach to Syntactic Analysis*. John Benjamins Publishing Company. doi.org/10.1075/clip.5

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), pp. 161–173. doi.org/10.1075/bjl.00043.jan

Kallas, J. (2013). *Eesti keele sisusõnade süntagmaatilised suhted korpus- ja*

õppeleksikograafias [Tallinn University. Dissertations on humanities]. Tallinna Ülikool.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: Linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom.* Trojina, Institute for Applied Slovene Studies / Lexical Computing Ltd., pp. 49–68.

Kay, P., Fillmore, C. J. (1999) Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. In *Language* 75, 1–33.

Kilgarriff, A.; Rychly, P.; Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the XI Euralex International Congress.* Lorient: Université de Bretagne Sud, pp. 105–116.

Kilgarriff, A., Rychlý, P., Jakubicek, M., Kovář, V., Baisa, V. & Kocincová, L. (2014). Extrinsic Corpus Evaluation with a Collocation Dictionary Task. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavik, Iceland, pp. 454–552.

Koppel, K., & Kallas, J. (2022). Eesti keele ühendkorpuste sari 2013–2021: Mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat*, *18*, pp. 207–228. doi.org/10.5128/ERYa18.12

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In *Proceedings of the eLex 2019 conference.* Lexical Computing CZ, s.r.o., pp. 434−452.

Koptjevskaja-Tamm, M. (2001). Partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages: "A piece of the cake" and "a cup of tea". In Ö. Dahl & M. Koptjevskaja-Tamm (eds.) *Circum-Baltic Languages: Volume 2: Grammar and Typology.* John Benjamins Publishing Company, pp. 523–568. doi.org/10.1075/slcs.55.11kop

Lee-Goldman, R., & Petruck, M. R. L. (2018). The FrameNet constructicon in action. *Constructicography: Constructicon development across languages.* John Benjamins Publishing Company, pp. 19–40. doi.org/10.1075/cal.22.02lee

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

Metslang, H. 2017. Kvantorifraas. – M. Erelt & H. Metslang (eds.) *Eesti keele süntaks. Eesti keele varamu, 3.* Tartu: Tartu Ülikooli Kirjastus, pp. 463–478.

Patel, M., Garibyan, A., Winckel, E., & Evert, S. (2023). A reference constructicon as a database. *Yearbook of the German Cognitive Linguistics Association*, *11*(1), pp. 175–202. doi.org/10.1515/gcla-2023-0009

Perek, F., & Patten, A. L. (2019). Towards an English Constructicon using patterns and frames. *International Journal of Corpus Linguistics*, *24*(3), pp. 354–384.

doi.org/10.1075/ijcl.00016.per

Pilvik, M.-L., Lindström, L., Plado, H., & Simmul, C. E. (2025). Nimisõnafraasi ja hulgafraasi piirimail: „osa“, „enamik“ ja „enamus“ hulgasõnadena. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, *21*, pp. 237-261. doi.org/10.5128/ERYa21.13

Risberg, L., Tuulik, M., Langemets, M., Koppel, K., Vainik, E., Prangel, E. & Aedmaa, E. (in press). Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel [Using corpus data to support lexicographers in identifying informal language]. Keel ja Kirjandus, 7.

Sass, B. (2023). From a dictionary towards the Hungarian Constructicon. In *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.*, pp. 534–544.

Shibuya, Y., & Jensen, K. E. (2015). Mining for constructions in texts using N-gram and network analysis. *Globe: A Journal of Language, Culture and Communication, 2*, pp. 23–54.

Sidorov, G. (2019). *Syntactic n-grams in Computational Linguistics*. Springer International Publishing. doi.org/10.1007/978-3-030-14771-6

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), pp. 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), pp. 203–226. doi.org/10.1515/gcla-2023-0010

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

Tanvir, H., Kittask, C., Eiche, S., & Sirts, K. (2021). EstBERT: A Pretrained Language-Specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 11–19.

Tavast, A., Koppel, K., Langemets, M., & Kallas, J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Alexandroupolis, Greece: Democritus University of Thrace, pp. 215—223.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018*. Ljubljana University Press, Faculty of Arts, pp. 749—761.

Ulčar, M., & Robnik-Šikonja, M. (2021). *Training dataset and dictionary sizes matter in BERT models: The case of Baltic languages* arXiv:2112.10553; Version 1

Vainik, E., Paulsen, G., Sahkai, H., Kallas, J., Tavast, A., & Koppel, K. (2024). From a

Dictionary to a Constructicon – Putting the Basics on the Map. In *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language, pp. 209-216

Wible, D., & Tsao, N.-L. (2010). StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics.* Los Angeles, California, Association for Computational Linguistics, pp. 25–31.

Corpora and datasets

Balanced Corpus of Estonian https://www.cl.ut.ee/korpused/grammatikakorpus/ (1 October 2025)

Estonian as a Second Language Coursebook Sentences Corpus 2021 doi.org/10.15155/3-00-0000-0000-0000-0885AL (1 October 2025)

Estonian as a Second Language School Coursebook Sentences Corpus 2021 doi.org/10.15155/3-00-0000-0000-0000-0888DL (1 October 2025)

Estonian National Corpus 2017. doi.org/10.15155/3-00-0000-0000-0000-071E7L (1 October 2025)

Estonian National Corpus 2021. doi.org/10.15155/3-00-0000-0000-0000-08D17L (1 October 2025).

Estonian National Corpus 2023 doi.org/10.15155/3-00-0000-0000-0000-08C04M (1 October 2025)

Estonian Reference Corpus https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=en (1 October 2025)

Nominal quantifier constructions_ Gold Standard Dataset https://github.com/keeleinstituut/PRG1978/tree/main/constructions/gold_standards

Language models

CamemBERT – https://huggingface.co/docs/transformers/model_doc/camembert)

Claude-Sonnet-4 – https://claude.ai

EstBERT – https://huggingface.co/tartuNLP/EstBERT

Est-RoBERTa – https://huggingface.co/EMBEDDIA/est-roberta

GPT-4.1 – https://openai.com/index/gpt-4-1/

o3-mini – https://openai.com/index/openai-o3-mini/

TartuNLP/EstRoBERTa – https://huggingface.co/tartuNLP/EstRoBERTa