EESTI
KEELE
INSTITUUT

Methods for Pedagogical Constructicography

Identifying the CEFR Level of Constructions, Construction-Collexeme Pairings, and Collocations on the basis of CEFR-Graded L2 Textbook and Learner Corpora

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

Constructionist Approaches to Language Pedagogy
March 4–6 2026

Background

Research project: Expanding the scope of a multi-purpose lexicographic resource to grammar and L2 competence (PRG 1978), 2023–2027

Estonian Constructicon as an extension to the EKI Combined Dictionary
Developed within the Ekilex Dictionary Writing System (PostgreSQL)

Current stage: workflow, inventory, data model, database, user interface

Semi-automatic compilation

Corpus-based

CEFR levels

Resources

- **L1 corpora:** construction frequency, productivity, and collexemes (incl. frequency and association strength)

- **L2 textbook and** & **learner corpora**: CEFR- level information, learner difficulties and error-prone properties of constructions

- **CEFR-based resources**: vocabulary, grammar, and morphological profiles

- **Lexicographic database:** CEFR levels assigned to headwords and senses

CEFR-level assignment

1. **Construction as a whole**

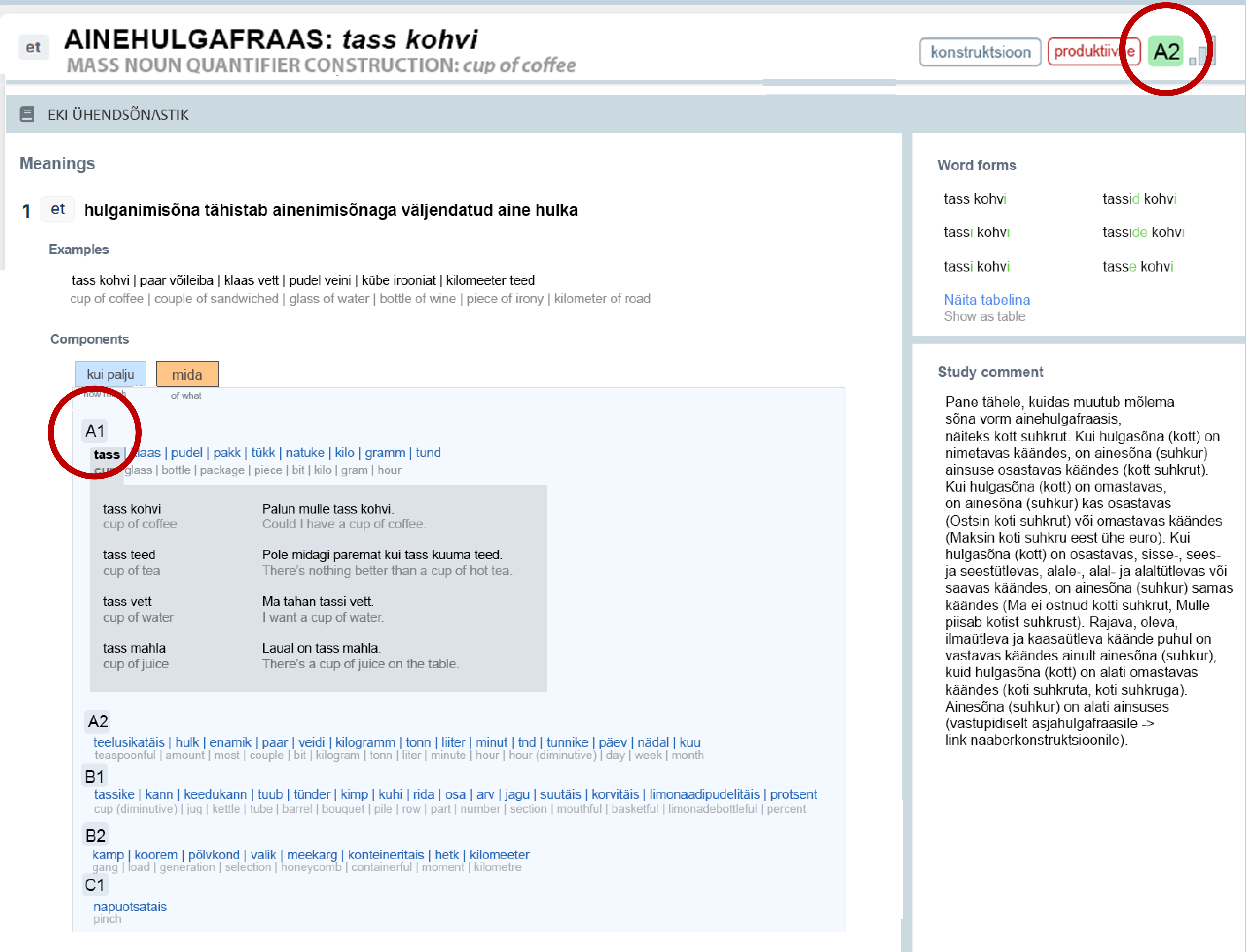
   The lowest level at which a schematic construction appears with different collexemes in L2 data

2. **Construction–collexeme pairings**

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

3. **Collocations instantiating the construction**

   The lowest level at which a collocation that represents the schematic construction appears more than once in L2 data



et **AINEHULGAFRAAS:** ***tass kohvi***

MASS NOUN QUANTIFIER CONSTRUCTION: *cup of coffee*

konstruktsioon produktiivne A2

EKI ÜHENDSÕNASTIK

Meanings

1 et **hulganimisõna tähistab ainenimisõnaga väljendatud aine hulka**

Examples

tass kohvi | paar võileiba | klaas vett | pudel veini | kübe irooniat | kilomeeter teed

cup of coffee | couple of sandwiched | glass of water | bottle of wine | piece of irony | kilometer of road

Components

kui palju — how much

mida — of what

A1

tass | klaas | pudel | pakk | tükk | natuke | kilo | gramm | tund

cup | glass | bottle | package | piece | bit | kilo | gram | hour

tass kohvi cup of coffee	Palun mulle tass kohvi. Could I have a cup of coffee.
tass teed cup of tea	Pole midagi paremat kui tass kuuma teed. There's nothing better than a cup of hot tea.
tass vett cup of water	Ma tahan tassi vett. I want a cup of water.
tass mahla cup of juice	Laual on tass mahla. There's a cup of juice on the table.

A2

teelusikatäis | hulk | enamik | paar | veidi | kilogramm | tonn | liiter | minut | tnd | tunnike | päev | nädal | kuu

teaspoonful | amount | most | couple | bit | kilogram | tonn | liter | minute | hour | hour (diminutive) | day | week | month

B1

tassike | kann | keedukann | tuub | tünder | kimp | kuhi | rida | osa | arv | jagu | suutäis | korvitäis | limonaadipudelitäis | protsent

cup (diminutive) | jug | kettle | tube | barrel | bouquet | pile | row | part | number | section | mouthful | basketful | limonadebottleful | percent

B2

kamp | koorem | põlvkond | valik | meekärg | konteineritäis | hetk | kilomeeter

gang | load | generation | selection | honeycomb | containerful | moment | kilometre

C1

näpuotsatäis

pinch

Word forms

tass kohvi	tassid kohvi
tassi kohvi	tasside kohvi
tassi kohvi	tasse kohvi

Näita tabelina

Show as table

Study comment

Pane tähele, kuidas muutub mõlema sõna vorm ainehulgafraasis, näiteks kott suhkrut. Kui hulgasõna (kott) on nimetavas käändes, on ainesõna (suhkur) ainsuse osastavas käändes (kott suhkrut). Kui hulgasõna (kott) on omastavas, on ainesõna (suhkur) kas osastavas (Ostsin koti suhkrut) või omastavas käändes (Maksin koti suhkru eest ühe euro). Kui hulgasõna (kott) on osastavas, sisse-, sees- ja seestütlevas, alale-, alal- ja alaltütlevas või saavas käändes, on ainesõna (suhkur) samas käändes (Ma ei ostnud kotti suhkrut, Mulle piisab kotist suhkrust). Rajava, oleva, ilmaütleva ja kaasaütleva käände puhul on vastavas käändes ainult ainesõna (suhkur), kuid hulgasõna (kott) on alati omastavas käändes (koti suhkruta, koti suhkruga). Ainesõna (suhkur) on alati ainsuses (vastupidiselt asjahulgafraasile -> link naaberkonstruktsioonile).

CEFR-Level Information in the dictionary portal Sõnaveeb (prototype)

Research questions

Overall goal: to design the workflow for assignment CEFR levels based on L2 textbook/learner corpora

RQ1: Can LLMs be prompted to identify constructional instances in an unannotated **L2 textbook** corpus?

RQ2: Can LLMs be prompted to identify constructional instances in unannotated **L2 learner** data?

RQ3: Do L2 textbook and learner corpora correlate in providing CEFR-level evidence for a construction?

RQ4: Can CEFR levels of collexeme-construction pairings be predicted from the frequency and association strength of construction–collexeme pairings in L1 data?

Constructions

The Nominal Quantifier Construction (NQC)

Two nominal slots:

- quantifier noun
- noun in the partitive case referring to the quantified entity

karp	*komme*
box-NOM	sweet-PART

‘a box of sweets’

DA-infinitive Construction (DA-INF)

Argument structure construction involving

- infinitival object complement with -DA
- subject control
- finite verbs expressing desire, attitude or ability

Ta	*otsusta-s*	*kolida-da.*
3SG	decide-PST.3SG	move-DA

‘S/he decided to move.’

Corpora

L1: The Balanced corpus of Estonian:

- fiction, journalistic and scientific writing
- 15 million words

L2 Textbook corpus:

- Estonian as a Second Language Coursebook Sentences Corpus 2021
- Estonian as a Second Language School Coursebook Sentences Corpus 2021
- 500 000 words
- 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)

Estonian L2 Learner corpus

- EMMA corpus /EKI subsection 2025:
- 12 076 texts from tests, assignments, and exams (134 551 sentences)
- 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)
- Russian as L1

RQ1: Can LLMs be prompted to identify constructional instances in an unannotated L2 textbook corpus?

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

- Estonian Nominal Quantifier Construction
- 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

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.

RQ2: Can LLMs be prompted to identify constructional instances in unannotated L2 learner data? What is the minimal prompt design and fine-tuning data needed for robust results?

- **Task**: identify NQC
- **Test-set**: C1-level learner texts, 1,418 sentences.
- **Models**: 9 commercial LLMs
- **Prompt types**
    - **baseline** (rules, 15 examples)
    - **extended** (detailed rules, stopwords, 15 examples)
    - **reduced** (detailed rules, stopwords, 3 examples)
    - **minimal** (compact rules, 3 examples)
    - **zero-shot** (compact rules, 0 examples)

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: Kallas, J., Kiil, A., Sahkai, H., Paulsen, G., & Saul, K. (2026, in press). *Using LLMs to extract instances of schematic constructions from unannotated L2 learner corpora*. In Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC 2026). European Language Resources Association (ELRA).

RQ3: Do the textbook and learner corpora correlate in providing CEFR-level evidence for a construction?

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

Method:

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

RQ3 results 1: Collexemes of NQC in textbook and learner copora

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*	–

Table 3. 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.

RQ3 results 2: Correspondence of the CEFR levels of collexeme-NQC pairings 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

Table 4. CEFR levels of collexeme-NQC pairings in learner and textbook corpus

- The vast 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

RQ3 results 3: Collexemes of DA-INF in textbook and learner copora

CEFR level	Collexemes in textbook corpus	Collexemes in learner corpus
A1	võima ‘can/may’, saama ‘be able’, tahtma ‘want’, tohtima ‘be allowed’, oskama ‘know how’, jõudma ‘manage’, lubama ‘allow’	tahtma ‘want’
A2	suutma ‘be capable’, püüdma ‘try’, kavatsema ‘intend’, soovima ‘wish’, julgema ‘dare’, proovima ‘try’, jaksama ‘have strength’, plaanima ‘plan’, armastama ‘love’, unistama ‘dream’, unustama ‘forget’	–
B1	üritama ‘attempt’, otsustama ‘decide’, nägema ‘see’, teadma ‘know’, katsuma ‘try’, kartma ‘fear’	saama ‘be able’, julgema ‘dare’, armastama ‘love’
B2	pruukima ‘use’, viitsima ‘bother’, lootma ‘hope’, eelistama ‘prefer’, ihkama ‘desire’, häbenema ‘be ashamed’	katsuma ‘try’

Table 5. DA-INF collexemes in textbook (both school and adult textbooks) and learner corpus

Similar trends as observed for the NQC

- All the collexemes in the learner corpus also appear in the textbook corpus.
- Many more collexemes in the textbook vs. learner corpus: 30 vs. 5.
- The construction appears with more than one collexeme two levels lower in the textbook vs. learner corpus: A1 vs. B1.

RQ3 Conclusions

- There is a systematic correspondence between the CEFR levels identified based on textbook and learner corpora: in the textbook corpus, constructions and construction-collexeme pairings appear one or two levels lower.

  The correspondence may indicate a natural learning order:

  - The level based on textbook corpus indicates the level at which a construction or construction-collexeme pairing is included in the teaching materials – useful for teaching

  - The level based on learner corpus indicates the level at which a construction or construction-collexeme pairing should have been acquired – useful for assessment

- 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, level-inappropriate collexemes in textbooks

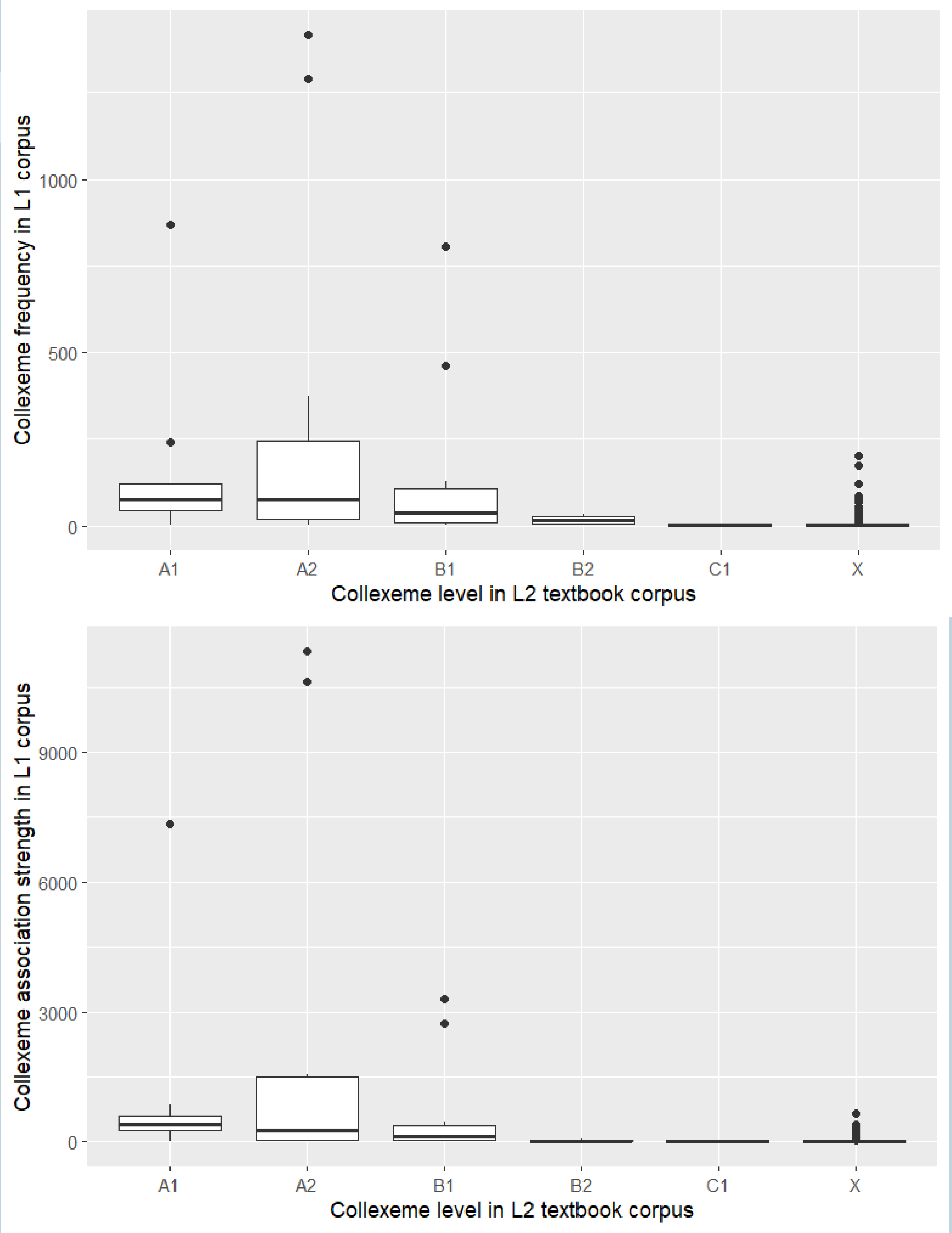
RQ4: Can CEFR levels of construction-collexeme pairings be predicted from the frequency and association strength of construction-collexeme pairings in L1 data?

Goal: avoid the need to rely on textbook/learner corpora for CEFR level identification when L1 frequency and association strength data are available

Method:

- Assign CEFR levels to construction-collexeme pairings based on the L2 textbook corpus (= RQ3)
- Collostructional analysis of the NQC and DA-INF constructions in L1 data
- Examination of the L1 frequency and association strength of the construction-collexeme pairings at each CEFR level

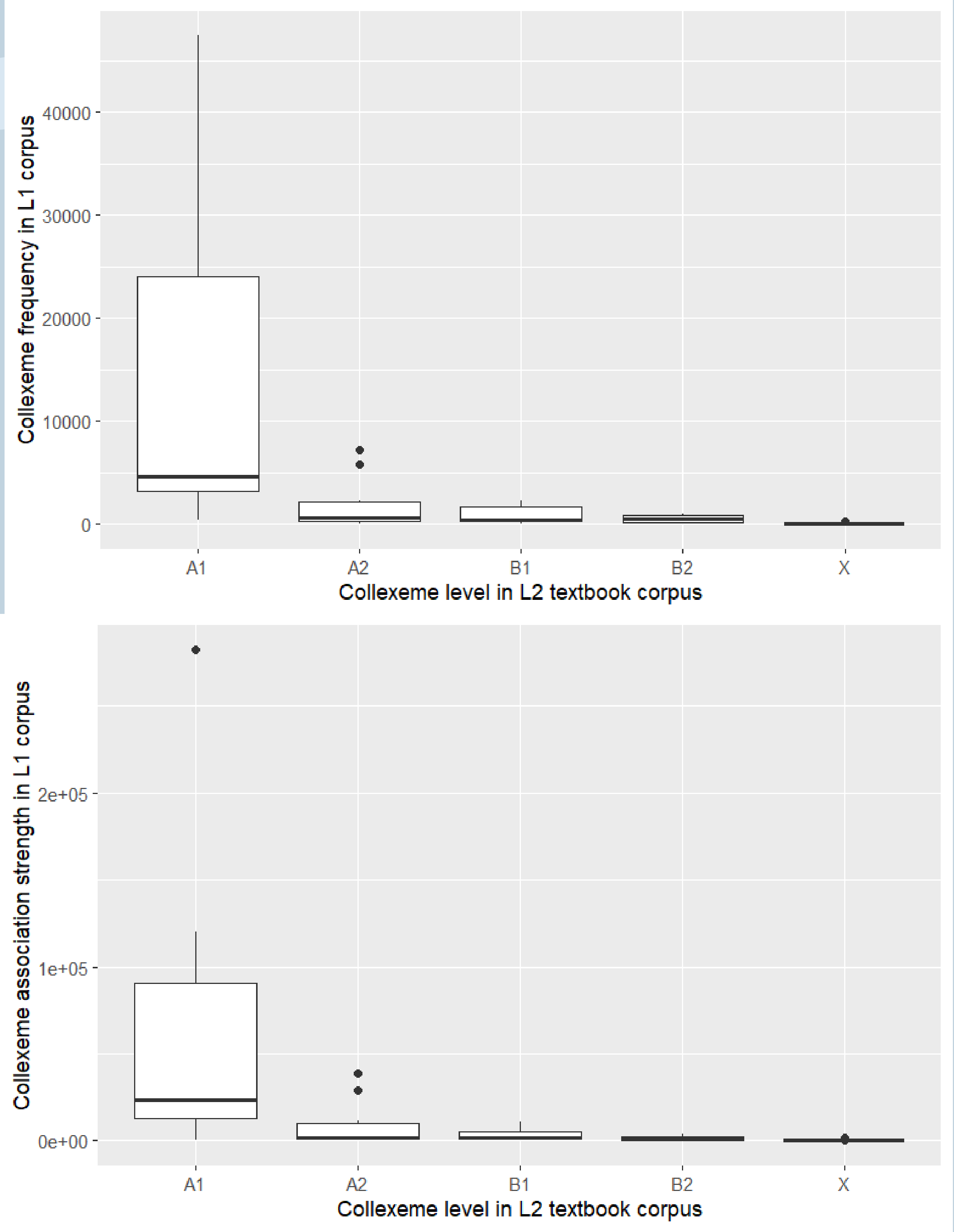
RQ4 results: NQC



- Construction-collexeme pairings at **levels A1-B1 are generally more frequent and more strongly associated** than the pairings at higher levels or absent from the textbook corpus
- **No perfect correspondence** between L1 frequency/association strength and CEFR level

Figure 1. L1 frequency (above) and association strength (below) of NQC-collexeme pairings at the different CEFR levels identified from the textbook corpus; X = collexemes that appear in L1 corpus but are absent from the textbook corpus

RQ4 results: DA-INF



- **Similar trend as for the NQC**:
  - Overall correspondence between CEFR level and L1 frequency/association
  - Insufficient to predict CEFR level from L1 data

Figure 2. L1 frequency (above) and association strength (below) of DA-INF construction-collexeme pairings at the different CEFR levels identified from the textbook corpus; X = collexemes that appear in L1 corpus but are absent from the textbook corpus

RQ4 conclusions

- There is a certain correlation between L1 frequency / association strength and L2 textbook-based CEFR-levels
- However, the CEFR level cannot be predicted from L1 frequency / association strength.
- Therefore, L2 textbook/learner corpora are a more reliable source for the identification of CEFR levels.
- L1 frequency data may still be useful for the correction of textbook corpus data.
    - For example, there are two words at the A2 level in the textbook corpus whose frequency in the L1 corpus is less than 10 (*teelusikatäis, tunnike*).
    - In addition, there are five words in the textbook corpus that are completely absent from the Balanced Corpus.

Overall conclusions

- 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 construction-collexeme pairings
    - The level based on textbook corpus indicates the level at which a construction or construction-collexeme pairing is included in the teaching materials
    - The level based on learner corpus indicates the level at which a construction or construction-collexeme pairing is acquired
- L1 frequency and association strength do not predict CEFR level but could be taken into account when a collexeme appearing in the textbook corpus is rare or absent in L1 corpus.