EESTI
KEELE
INSTITUUT

# Towards the Estonian constructicon: the first vision

Geda Paulsen, Heete Sahkai, Ene Vainik, Jelena Kallas, Kertu Saul, Arvi Tavast, Kristina Koppel, Katrin Tsepelina

21st Annual Conference of Applied Linguistics
18.–19.4.2024

# Topics

To envision the ways towards an **Estonian constructicon** – a representation of constructions

1. Why an Estonian constructicon?
2. Which constructions should be included?
3. What should the constructicon entry contain?
4. Lexical resources to integrate the constructicon to?
5. **How to relate constructional information to lexicographic information?**

# Building a constructicon

## Prerequisites:

1. a theoretical conception of language as a network of constructions,
2. a constructicographic collection of defined constructional patterns,
3. a technical corpus linguistic usage-based approach to compilation of the constructional network.

# Why an Estonian constructicon?

- integrating the **existing knowledge** of Estonian constructions into a single resource
- **expanding** the research of constructions
- **systematising** constructional description, establishing networks
- making constructions an explicit **part of the lexicographic description**
- for different user groups (learners, experts…)

# Lexical resources

## to integrate the construction with

- Constructicons typically linked to an existing lexical resource
  - FrameNet of respective language
  - databases of grammatical patterns
  - general dictionaries
  - corpora
- EKI Combined Dictionary (Langemets et al 2023), the underlying database in the dictionary writing system Ekilex (Tavast et al 2018, Langemets et al 2021), accessible to users via the language portal Sõnaveeb

# Why EKI Combined Dictionary?

- The most complete and up-to-date lexical data collection over Estonian language,
- contains a variety of syntactic patterns (government patterns, collocations, word classes), morphological paradigms, and semantic types; proficiency level markers (A1–B1).
- practical considerations
- theoretical considerations
- pedagogical considerations
- bringing together language information
- increasing the language technology potential of the Combined Dictionary

# Current state

## The verb *armastama* ’to love’ in EKI Combined Dictionary

*love* (verb)

EKI Combined Dictionary

meaning

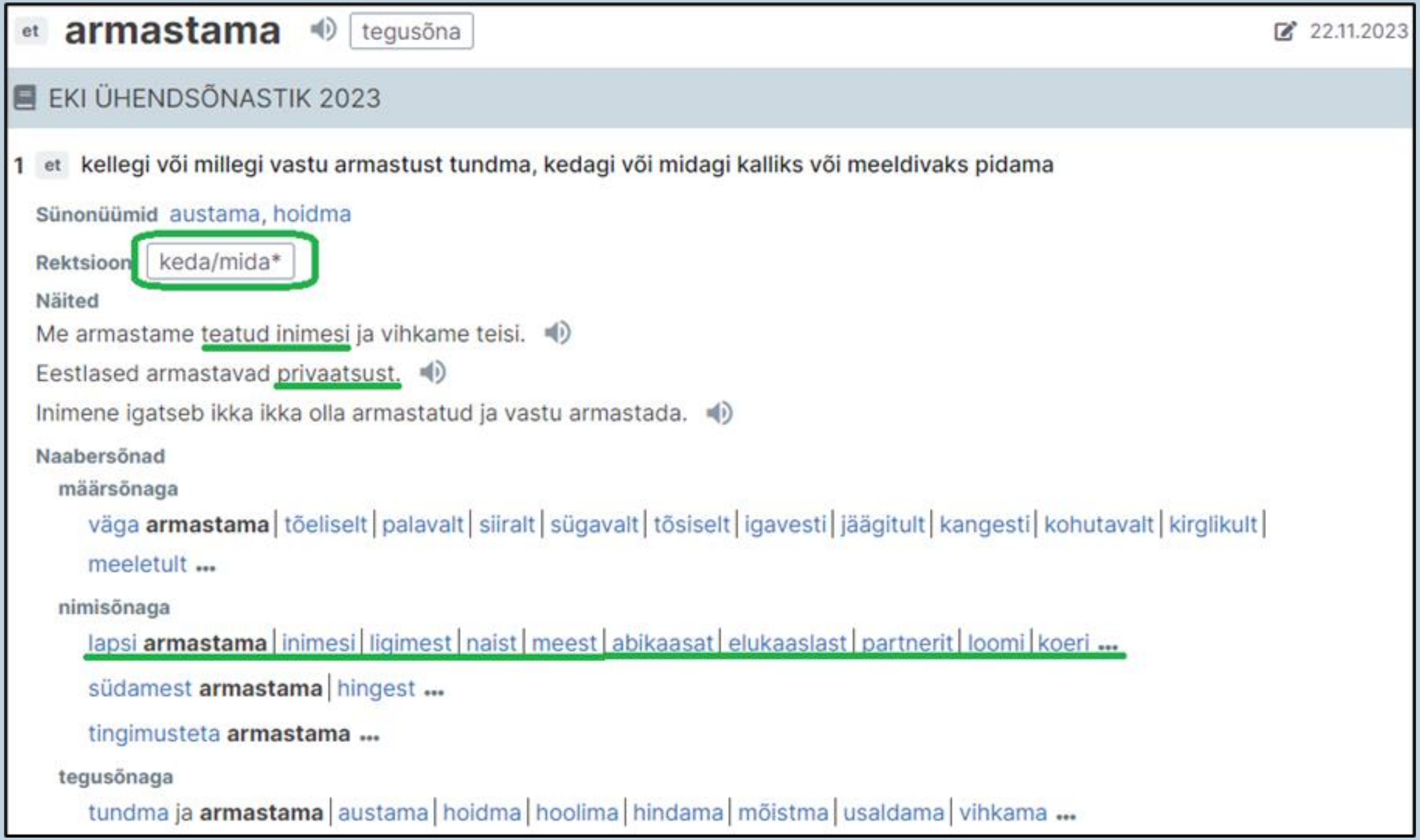
synonyms

government pattern: **OBJ**$_{\textbf{partitive}}$

examples with **OBJ**$_{\textbf{partitive}}$

collocations

collocate NPs: **OBJ**$_{\textbf{partitive}}$



et **armastama** tegusõna 22.11.2023

EKI ÜHENDSÕNASTIK 2023

1 et kellegi või millegi vastu armastust tundma, kedagi või midagi kalliks või meeldivaks pidama

**Sünonüümid** austama, hoidma

**Rektsioon** keda/mida*

**Näited**

Me armastame teatud inimesi ja vihkame teisi.

Eestlased armastavad privaatsust.

Inimene igatseb ikka ikka olla armastatud ja vastu armastada.

**Naabersõnad**

**määrsõnaga**

väga **armastama** | tõeliselt | palavalt | siiralt | sügavalt | tõsiselt | igavesti | jäägitult | kangesti | kohutavalt | kirglikult | meeletult ...

**nimisõnaga**

lapsi **armastama** | inimesi | ligimest | naist | meest | abikaasat | elukaaslast | partnerit | loomi | koeri ...

südamest **armastama** | hingest ...

tingimusteta **armastama** ...

**tegusõnaga**

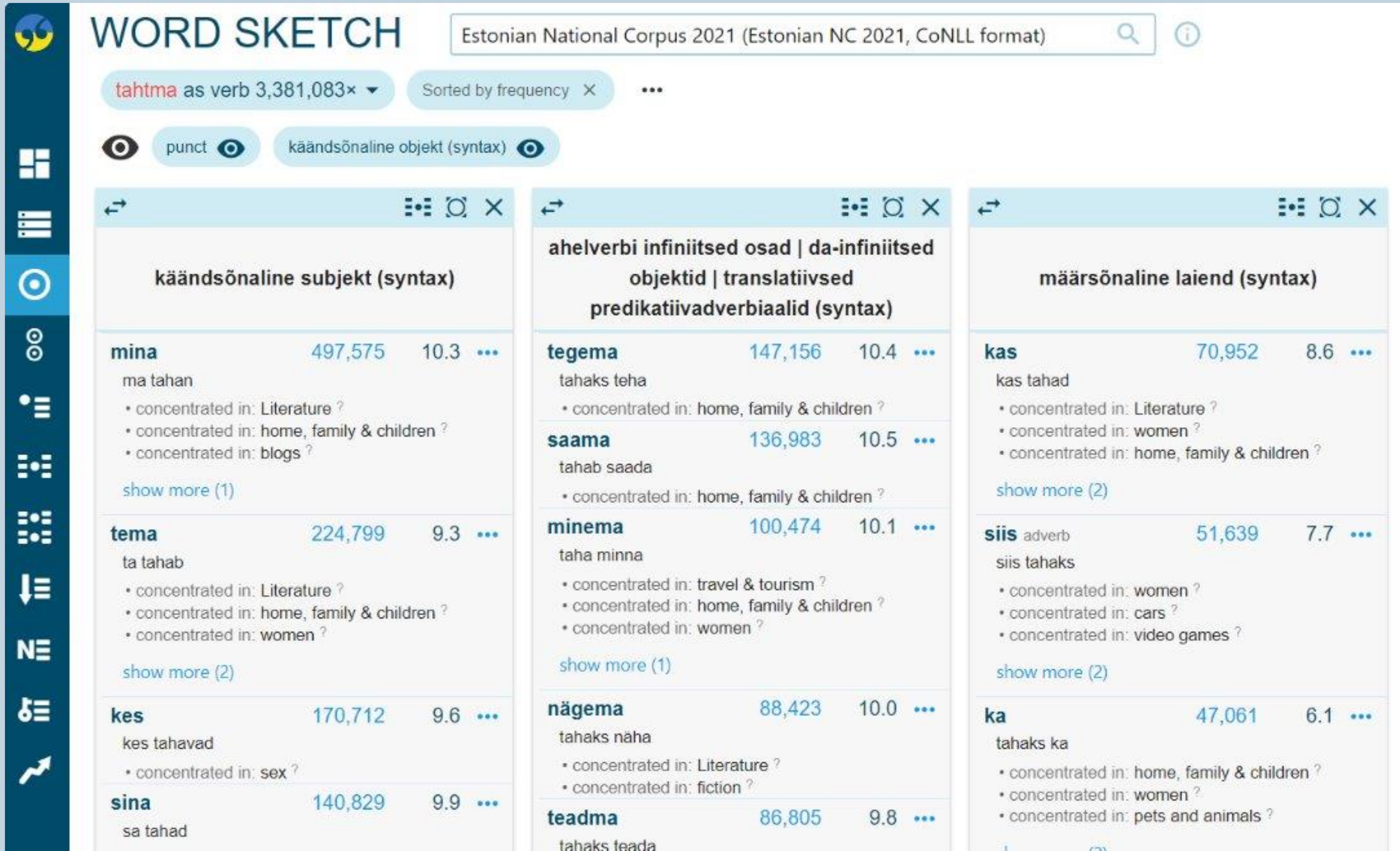
tundma ja **armastama** | austama | hoidma | hoolima | hindama | mõistma | usaldama | vihkama ...

IMPLICIT STRUCTURE

USER VIEW

# Corpus-lexicographic methods

- Estonian National Corpora (2013, 2027, 2019, 2021, 2023), multi-layer annotation
- Annotation: estNLTK, EstNLTK's StanzaSyntaxTagger
- Starting from 2024 – working with dependency relations not just with word sequences and morphological annotation
- Sketch Engine (Kilgarrif jt 2004)
- Corpus-based identification and extraction of
  - government patterns
  - collocations
  - example sentences



WORD SKETCH

Estonian National Corpus 2021 (Estonian NC 2021, CoNLL format)

tahtma as verb 3,381,083× · Sorted by frequency ✕

punct · käändsõnaline objekt (syntax)

**käändsõnaline subjekt (syntax)**

**mina** 497,575 10.3
ma tahan
- concentrated in: Literature ?
- concentrated in: home, family & children ?
- concentrated in: blogs ?

show more (1)

**tema** 224,799 9.3
ta tahab
- concentrated in: Literature ?
- concentrated in: home, family & children ?
- concentrated in: women ?

show more (2)

**kes** 170,712 9.6
kes tahavad
- concentrated in: sex ?

**sina** 140,829 9.9
sa tahad

**ahelverbi infiniitsed osad | da-infiniitsed objektid | translatiivsed predikatiivadverbiaalid (syntax)**

**tegema** 147,156 10.4
tahaks teha
- concentrated in: home, family & children ?

**saama** 136,983 10.5
tahab saada
- concentrated in: home, family & children ?

**minema** 100,474 10.1
taha minna
- concentrated in: travel & tourism ?
- concentrated in: home, family & children ?
- concentrated in: women ?

show more (1)

**nägema** 88,423 10.0
tahaks näha
- concentrated in: Literature ?
- concentrated in: fiction ?

**teadma** 86,805 9.8
tahaks teada

**määrsõnaline laiend (syntax)**

**kas** 70,952 8.6
kas tahad
- concentrated in: Literature ?
- concentrated in: women ?
- concentrated in: home, family & children ?

show more (2)

**siis** adverb 51,639 7.7
siis tahaks
- concentrated in: women ?
- concentrated in: cars ?
- concentrated in: video games ?

show more (2)

**ka** 47,061 6.1
tahaks ka
- concentrated in: home, family & children ?
- concentrated in: women ?
- concentrated in: pets and animals ?

# Problems of the current state

## From the general dictionary user perspective:

- information about government patterns is incomplete
- government patterns, collocations, and example sentences are presented as **unrelated information layers**
- dictionary does not allow for **systematic search for patterns** or to find other verbs behaving in a similar way.

## From the perspective of language learners and teachers:

- creators of teaching materials and learners cannot get an overview of verbs with shared constructional patterns and proficiency level.

# Which constructions should be included?

**1. Fully schematic** constructions

subject_predicate construction ~ [NP V]

*Mia loeb*

‘Mia reads’

**2. Partially schematic** constructions

$X_{\text{elative}}$ *pole* $Y_{\text{partitive}}$ ~ [NP *doesn’t* NP]

*Mia-st* *pole* *koristaja-t*

Mia-ELA be-NEG cleaner-PART

‘Mia will not manage to clean the mess’

**3. Lexically fully specific** constructions

*Kuidas külvad, nõnda lõikad*

’You reap as you sow’

# Where are the construction types now?

**Collocations** -> constructs, i.e. realizations of different types of (e.g. noun, adjective, adverb and verb phrase) constructions

*lapsi* / *inimesi* / *ligimest armastama*

to love children / people / fellow beings

**Government patterns** -> semi-schematic constructions (lexeme + dependent's morphological form – noun/adjective in certain case form, verb in certain infinitival form)

*armastama* keda / mida
to love who / what

**Example sentences** -> realizations of fully schematic constructions?

*Eestlased armastavad privaatsust.*

Estonians love privacy.

# Argument structure constructions

- Abstract meaningful constructions that in right conditions can fuse with lexical entries of verbs in order to provide them with **additional constructional roles** that then in turn are realized syntactically (Goldberg 1995).

Caused Motion Construction:

*Mia peeled a bunch of oranges into the bowl.*

X CAUSES Y TO MOVE Z

$NP_{subj}$ VP $NP_{obj}$ $NP_{obl}$

- Learned form-function pairings that exist independently of the specific verbs.
- Applying the idea of ASCs to a lexicographic resource enables to **generalise the patterns certain verbs share**.

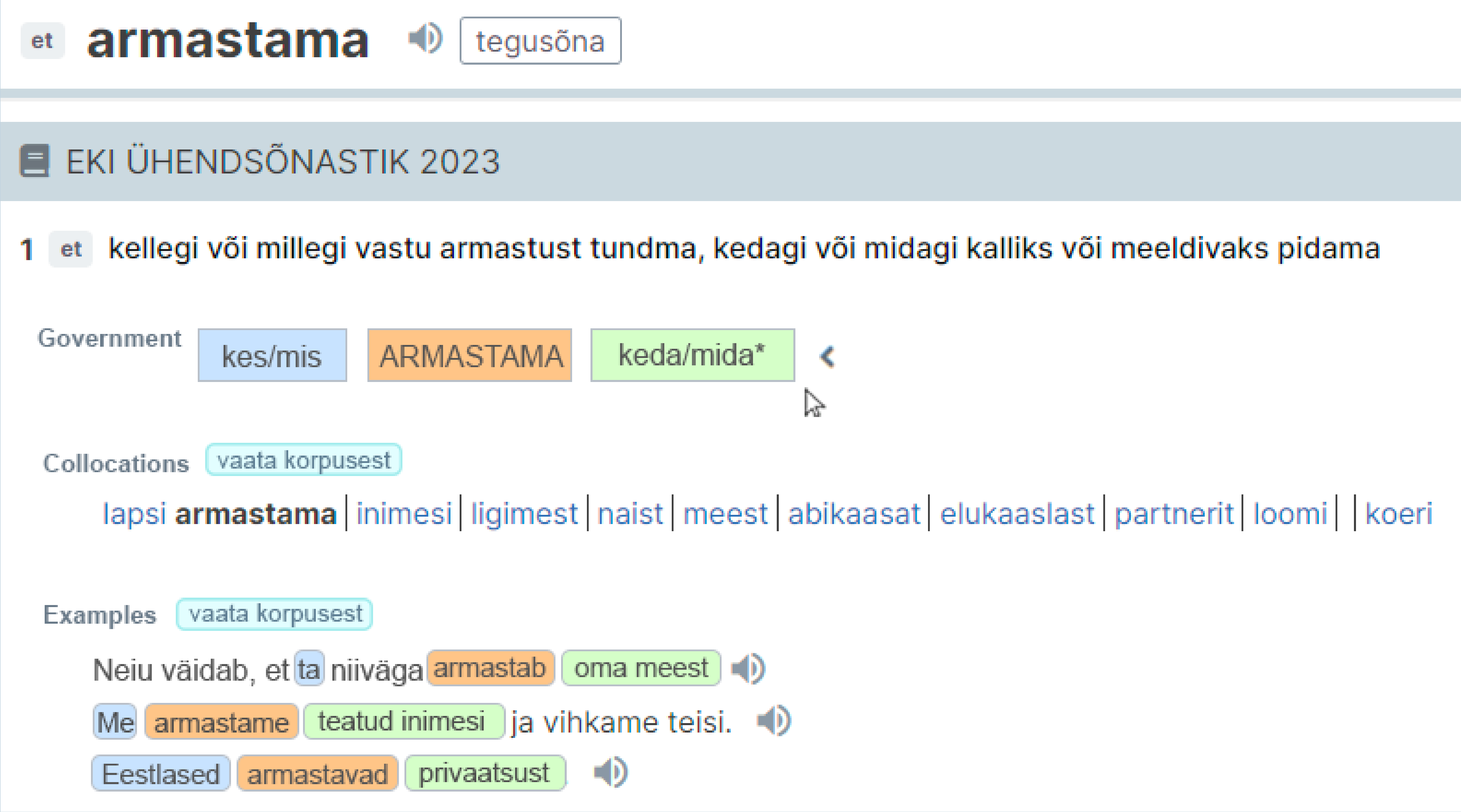
# Draft 1

## Highlighting parts of the construction

Transitive construction

**P**artitive **O**bject **C**onstruction

$\{[NP_{nominative}]\ [VP]\ [OBJ_{partitive}]\}$



et **armastama** tegusõna

EKI ÜHENDSÕNASTIK 2023

1 et kellegi või millegi vastu armastust tundma, kedagi või midagi kalliks või meeldivaks pidama

Government kes/mis ARMASTAMA keda/mida*

Collocations vaata korpusest

lapsi **armastama** | inimesi | ligimest | naist | meest | abikaasat | elukaaslast | partnerit | loomi | | koeri

Examples vaata korpusest

Neiu väidab, et ta niiväga armastab oma meest

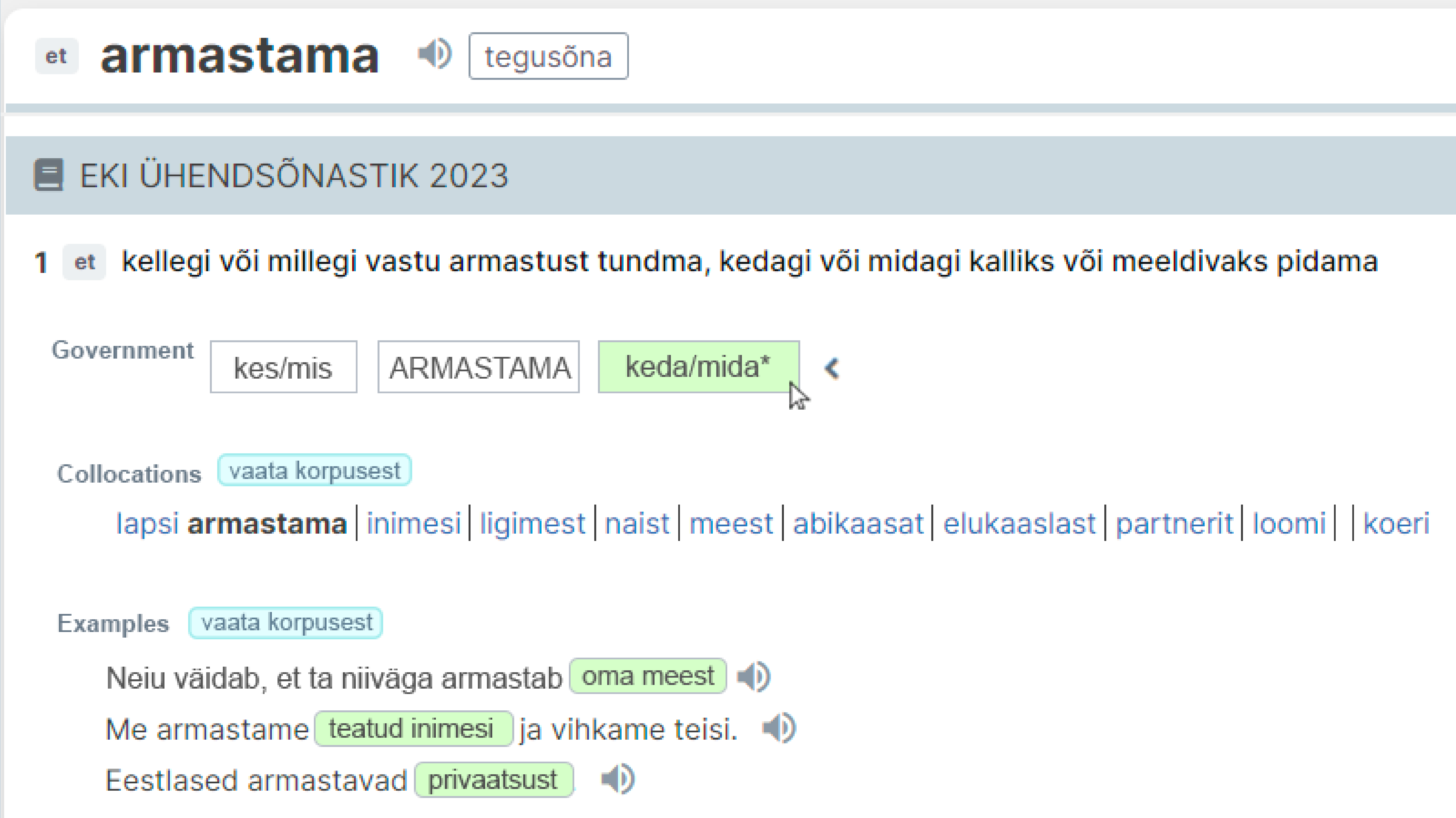
Me armastame teatud inimesi ja vihkame teisi.

Eestlased armastavad privaatsust

parts of POC highlighted

# Draft 2

## highlighting only certain parts of the construction



et **armastama** tegusõna

EKI ÜHENDSÕNASTIK 2023

1 et kellegi või millegi vastu armastust tundma, kedagi või midagi kalliks või meeldivaks pidama

Government kes/mis ARMASTAMA keda/mida*

Collocations vaata korpusest

lapsi **armastama** | inimesi | ligimest | naist | meest | abikaasat | elukaaslast | partnerit | loomi | | koeri

Examples vaata korpusest

Neiu väidab, et ta niiväga armastab oma meest

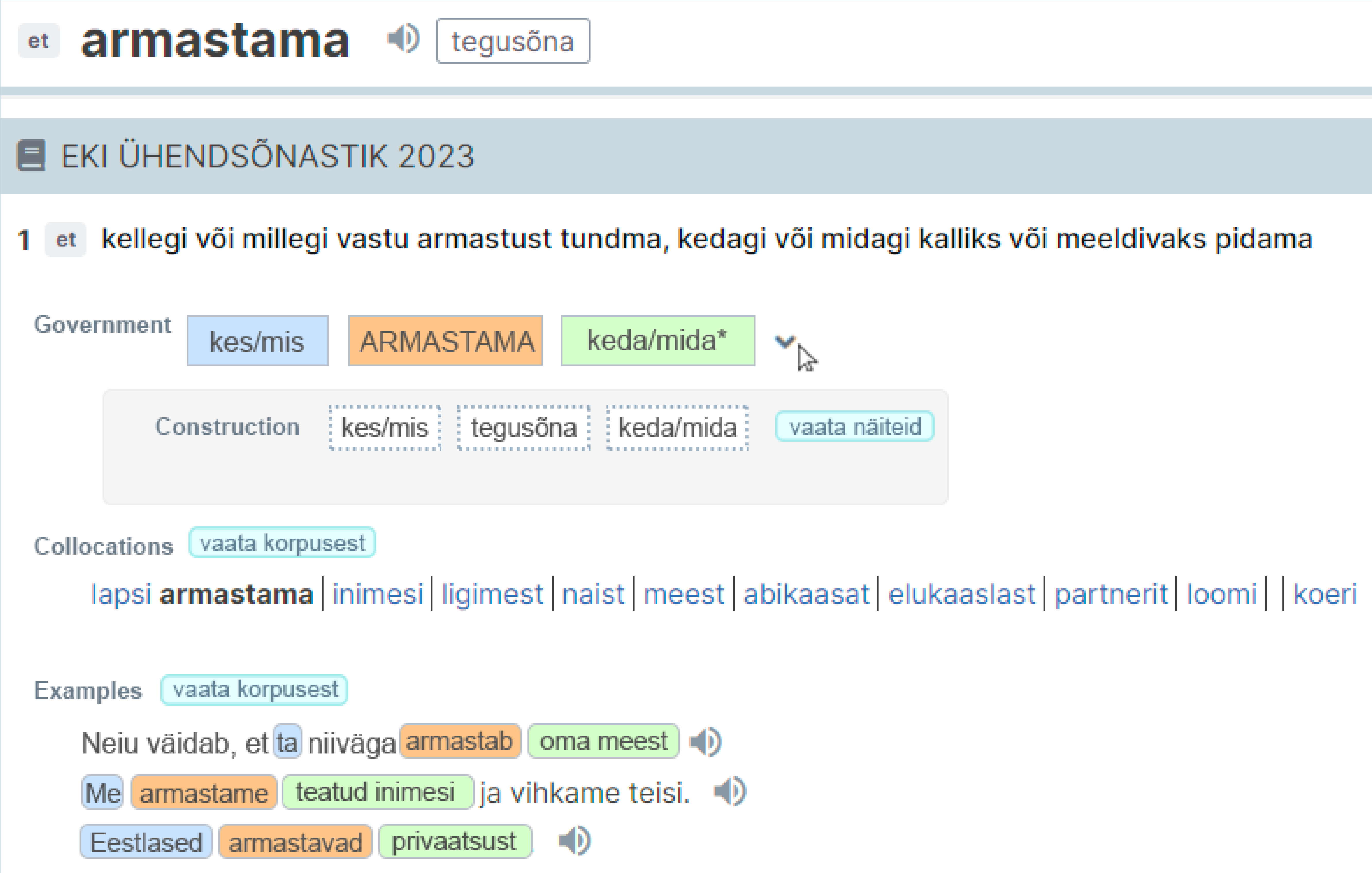
Me armastame teatud inimesi ja vihkame teisi.

Eestlased armastavad privaatsust

only certain parts of POC highlighted

# Draft 3

## Highlighting the construction



et **armastama** tegusõna

EKI ÜHENDSÕNASTIK 2023

1 et kellegi või millegi vastu armastust tundma, kedagi või midagi kalliks või meeldivaks pidama

Government kes/mis ARMASTAMA keda/mida*

Construction kes/mis tegusõna keda/mida vaata näiteid

Collocations vaata korpusest

lapsi **armastama** | inimesi | ligimest | naist | meest | abikaasat | elukaaslast | partnerit | loomi | | koeri

Examples vaata korpusest

Neiu väidab, et ta niiväga armastab oma meest

Me armastame teatud inimesi ja vihkame teisi.

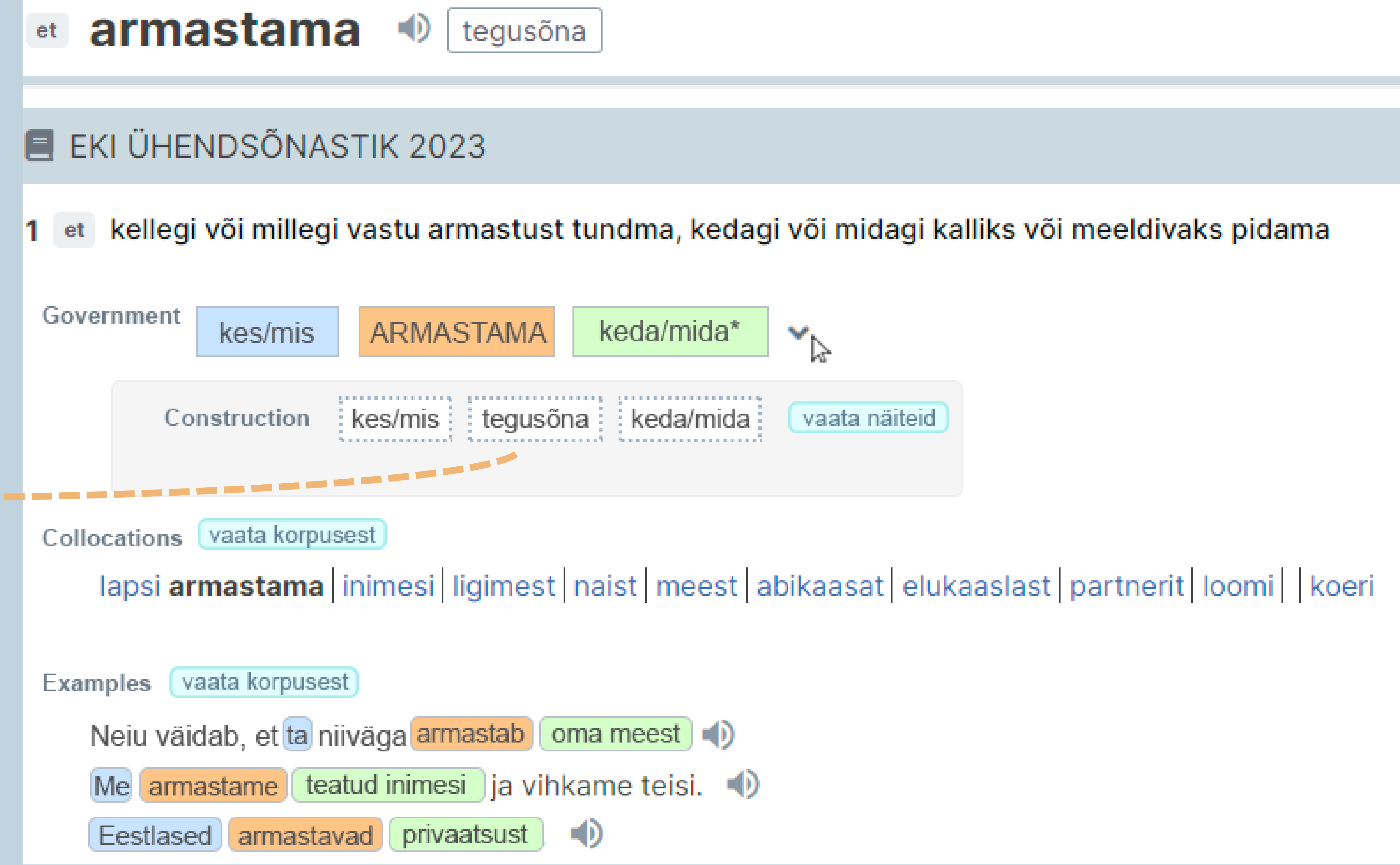
Eestlased armastavad privaatsust

abstract description of the POC

parts of POC highlighted

# List of POC-verbs

*austama* ’respect’
*ette heitma* ’reproach’
*häbenema* ‘be ashamed’
*igatsema* ’miss’
*ihkama* ’desire’
*kartma* ’fear’
*pidama* ’regard’
*põlgama* ’despise’
*kavatsema* ’itnend’
*lootma* ’hope’
*nõudma* ’demand’
*paluma* ’ask for’
*soovima* ‘wish’
*kuulama* ’listen’
*kuulma* ’hear’
*maitsma* ’taste’



et **armastama** tegusõna

EKI ÜHENDSÕNASTIK 2023

1 et kellegi või millegi vastu armastust tundma, kedagi või midagi kalliks või meeldivaks pidama

Government kes/mis ARMASTAMA keda/mida*

Construction kes/mis tegusõna keda/mida vaata näiteid

Collocations vaata korpusest

lapsi **armastama** | inimesi | ligimest | naist | meest | abikaasat | elukaaslast | partnerit | loomi | | koeri

Examples vaata korpusest

Neiu väidab, et ta niiväga armastab oma meest

Me armastame teatud inimesi ja vihkame teisi.

Eestlased armastavad privaatsust

# Constructicon entry

- type (schematic, partially schematic, lexical cxn)
- name
- form
  - constituent and dependency structure
  - morphological form of the components
  - syntactic function
  - phrasal type
- meaning
- productivity description
- language proficiency level
- frequency information
- regular expression for corpus query

# Steps towards the Estonian constructicon

1. Organising and unifying government patterns, collocational information and example sentences already provided in the Combined Dictionary.
2. Adding **fully schematic constructions** in the constructicon (e.g. represented by collocations or argument structure cxns)
3. Adding **semischematic constructions**
4. Adding **lexically specified constructions** (various types of idiomatic expressions)
5. **Theoretical** work with gathering existing constructional description of Estonian and developing it further.
6. Methods to identify constructions in the corpus.
7. **Technical** work – enriching the Combined Dictionary with constructions.
8. Methods to define and display **L2 proficieny levels**

# Questions

1. How to realise the network idea of the constructicon?
2. …inheritance relations?
3. Flexibility-adaptability vs. limits of the data model and the database?
4. What kind of constructicon regarding different user groups and the “backbone” structure?
   - display view – the “approachable” constructicon
   - internal view – the underlying constructicon
5. L2 proficiency levels – how to define and how to display?
6. How to connect the constructicon to the corpus?
7. How to identify novel/undescribed constructions?

# CONCLUSION OF THE FIRST VISION:

THERE IS LIGHT AT THE END OF THE TUNNEL

# References

The EKI Combined Dictionary. (2024). Hein, I., …, & Voll, P. Institute of the Estonian Language. https://sonaveeb.ee

Fillmore, C. J.; Kay, P.; O'Connor, M. C. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language* 64: 501–538.

Goldberg, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Langemets, M., Koppel, K., Kallas, J., Tavast, A. 2021. Sõnastikukogust keeleportaaliks. *Keel ja Kirjandus*, 8-9, 755−770.

Langemets, M. 2010. *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras*. Tallinn: Eesti Keele Sihtasutus.

Lyngfelt, B. 2018. Introduction: Constructicons and constructicography. B. Lyngfelt et al., Constructicography: Constructicon development across languages. *Constructional Approaches to Language 22*. Amsterdam: John Benjamins, 1–18.

Tavast, A., Langemets, M., Kallas, J., & Koppel, K. 2018. Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts*, Ljubljana, 17-21 July 2018. Ed. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana University Press, Faculty of Arts, 749−761.

Tuulik, M. 2022. Adjektiivide süstemaatiline polüseemia eesti keeles tajuadjektiivide näitel. Tartu: Tartu Ülikooli toimetised.

Vainik, E., Paulsen, G., Kallas, J. (submitted). Sõnastikust konstruktikoniks? Taustu, eeskujusid ja väljakutseid.