Estonian National Corpora 2013–2025: methodological foundations and constucticographic applications

Jelena Kallas, PhD jelena.kallas@eki.ee
Kristina Koppel, PhD kristina.koppel@eki.ee

INSTITUTE OF THE ESTONIAN LANGUAGE

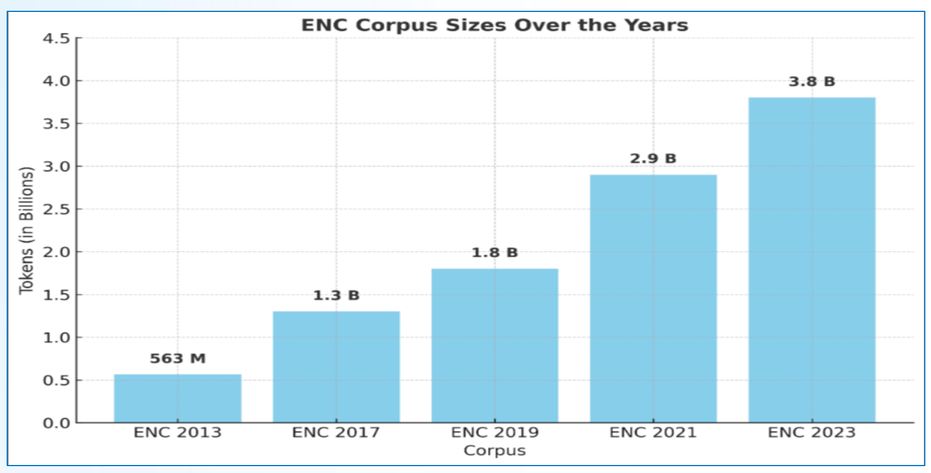
The series of the Estonian National Corpus (ENC) 2013–2025

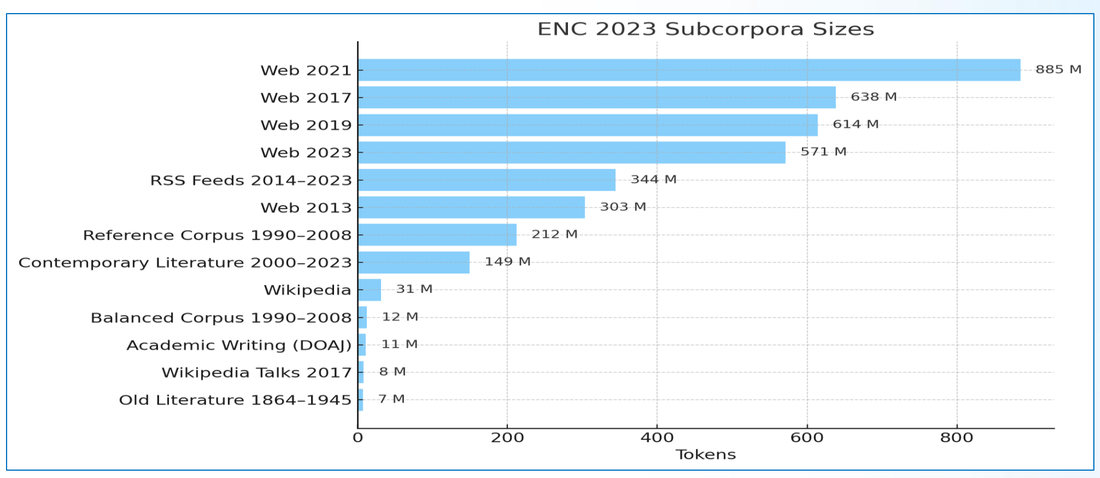
Developed by the Institute of the Estonian Language in collaboration with Lexical Computing Ltd.
Accessible via the Sketch Engine interface (Kilgarriff et al., 2004)
ENC 2021: Fully dependency-annotated (CONLL format)
ENC 2023:

- Accessible also as a PostgreSQL database (workflows in the EstNLTK GitHub repository)
- Semi-automatic classification: 7 genres (e.g., periodicals, blogs, literature), 21 topics (e.g., history, sports, politics & government)

ENC 2025:

- Collaboration with National Library; new types of sub-corpora (e.g., domain-specific books)
- Data from public sector, and from the Web (e.g., cleaned Common Crawl)
- Size: up to 15 billion tokens; to be used also as training data for LLMs

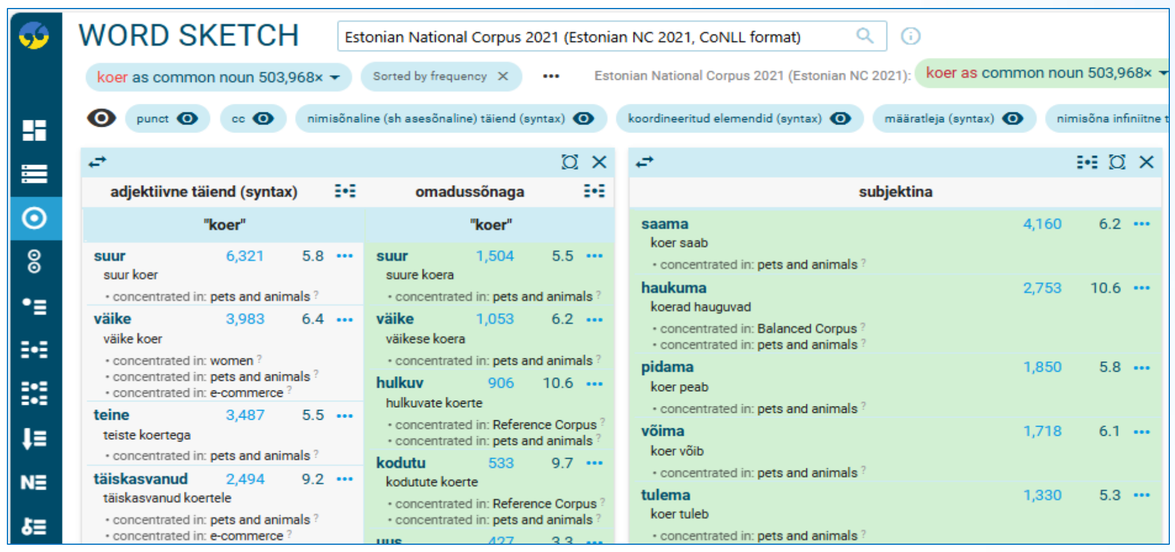




Lexicographic and constructicographic applications

Automatic detection of collocations:

- Part-of-Speech pattern approach: the Estonian Sketch Grammar v2.1 defines 113 grammatical relations.
- Dependency-based approach (cf. Uhrig & Proisl, 2012): 23 syntactic relations (e.g., nsubj, obj, obl, advmod, amod, nummod etc.).



Automatic detection of Good Dictionary Examples (Kosem et al., 2019)

Estonian Constructicon project (2022–2027) (Vainik et al., 2024): Developing a Constructicographic Workflow

- Training language models (EstBERT, Est-RoBERTa, Claude-Sonnet-4, GPT-4.1) to identify construction instances in corpora
- Applying collostructional analysis to identify the lexical slot-fillers (cf. Stefanowitsch & Gries, 2003)
- Automatically detecting argument structure constructions on the basis of morpho-syntactically annotated corpora (UD schema) (Saul et al., 2025)

References

DIGAR. https://www.digar.ee/arhiiv/en
EstNLTK Workflows. https://github.com/estnltk/estnltk-workflows
Kilgarriff, A. et al. (2004). *The Sketch Engine.* In *EURALEX Proc.*, 105–115.
Kosem, I. et al. (2019). *GDEX: Automatic extraction of good dictionary examples. Int. J. Lexicography*, 32(2), 119–137.
Saul, K., Muischnek, K., Kallas, J. (2025). *Lausemallide automaatne tuvastamine. ERYa*, 21, 297–312. https://doi.org/10.5128/ERYa21.16
Stefanowitsch, A. (2013). *Collostructional Analysis.* In Hoffmann & Trousdale (Eds.), *Oxford Handbook of Construction Grammar*, 290–306.
Uhrig, P., Proisl, T. (2012). *Using dependency corpora for collocation extraction. Lexicographica*, 28, 141–180.
Vainik, E. et al. (2024). *From a dictionary to a constructicon.* In *EURALEX XXI Proc.*, 196–203.

This work was partly supported by the Estonian Research Council grant "Expanding the scope of a multi-purpose lexicographic resource to grammar and L2 competence (PRG 1978)" and EKKD-III1 R&D project "Applying LLMs to lexicography: new opportunities and challenges".