

(De)standardiseerides kõnekeelt sõnaraamatus. Kuidas rakendada keelekorpus ja suuri keelemudeleid?

**Lydia Risberg | Maria Tuulik | Margit Langemets | Ene Vainik | Eleri Aedmaa | Esta Prangel |
Kristina Koppel | Hanna Pook | Sirli Zupping**
EKKD-III1 registrite töörühm
Eesti Keele Instituut



Sissejuhatus

EKKD-III1 „Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed“ (1.1.2024–31.12.2027)

- Mida ja miks me registrite töörühmas seni teinud oleme?
- Keskendun suurtele keelemudelitele (SKMidele).
- **Lahtiütlus/diskleimer/disclaimer**: uurimus sõnaraamatutöö hõlbustamiseks, ent praegu alles katsetame.

et **diskleimer** nimisõna

EKI ühendsõnastik 2025 Viimati muudetud 10.03.2025

Tähendused

et **KÕNEKEELNE** tekst, mis lisatakse artiklile vm materjalile, et vältida võimalikke sekeldusi, väärtõlgendusi või -järeldusi, vastutust piirav klausel (nt *Selle veebilehe kasutamine ja sirvimine toimub külastaja enda riisikol*)

Sünonüümid **lahtiütlus, disclaimer** TSITAATSÕNA

“

**Parim viis hinnata,
kas ChatGPT-st on sinu valdkonnas kasu,
on sellele päriselt võimalus anda.
(Gilles-Maurice de Schryver 2023: 2)**

”



Miks valisime sõnaraamatumärgendid?

- ÕS 2025 jaoks vaadatakse mh märgenditega sõnad üle.
- Leksikograafidel on märgendite lisamise või eemaldamise üle olnud tihtilugu keerukas otsustada.
- Sõnaraamatust märgendeid ära jätta ei saa,
 - kogu sõnavara pole ühtviisi neutraalne.
- Märgendite järele on ka nõudlust:
 1. keeleportaali Sõnaveeb kasutajauuringust: võiks olla rohkem märgendeid, et paremini aru saada, millisesse registrisse sõna kuulub (Langemets, Risberg, Algvere 2024: 751);
 2. ka keeleõppijad otsivad sõnastikust infot sõnade kasutusregistrite kohta (Pool, Teral, Kallas 2025).



Märgendid ja fookus

- Sõnaraamatutes on erisuguseid märgendeid:
 - register (nt *kõnekeelne*)
 - eriala (nt *bioloogia*, *õigusteadus*)
 - kasutussagedus (nt *harv*)
 - aeg (nt *uus*, *vanamoodne*)
 - stiil (*stiilitundlik*, *luulekeelne*)
 - suhtumine (nt *halvustav*)



Märgendid ja fookus

- Sõnaraamatutes on erisuguseid märgendeid:
 - register (nt ***kõnekeelne***)
 - eriala (nt *bioloogia*, *õigusteadus*)
 - kasutussagedus (nt *harv*)
 - aeg (nt *uus*, *vanamoodne*)
 - stiil (*stiilitundlik*, *luulekeelne*)
 - suhtumine (nt *halvustav*)
- Näiteks:
 - *mill* vs. *miljon*
 - *?aastanumber* vs. *aastaarv*
- Leksikograaf peab langetama konkreetse otsuse, kas sõna(tähenduse)le kõnekeelsuse märgend lisada või mitte.

“

Iga leksikograaf teab neid piinarikkaid valikuid,
kui otsustamiseks pole aega
ja ainus mõistlik lahendus – aastaid uurida –
pole võimalik.

”

(Ladislav Zgusta 1992: 91)



Miks kõnekeel?

- Sõnaraamatu praktilise töö kaalutlusel.
- Vulgaarsed või lastekeelsed sõnad vs. kõnekeelsed sõnad.
 - Nt *pasarahe* või *notsu* vs. *mill*, *?aastanumber*, *?kirvereeegel*.
- **Kõnekeel on registrina piiripealsem ja subjektiivsem.**
 - (Ent märgendiotsuseid tuleb teha.)
- Ainult SKMide hindamiseks oleks äärmuslikumad sõnad olnud lihtsam valik,
- ent tahtsime võimaluse korral saada kohe ka praktilist kasu kõnekeelseks märgendatud sõnade ülevaatamisel.



Kuidas teised?

- **Miloš Jakubíček ja Michael Rundell (2023: 527–528)**
 - ChatGPT katsetamine sõnaraamatutöö jaoks
 - märgendid vaid üks napp tööloik
 - ChatGPT tuvastas ühe halvustava ja ühe vanamoodsa inglise sõna
 - ära ei tundnud harva sõna
- **Erik Henriksson jt (2024)**
 - liigitasid registrite aspektist veebitekste
 - võrdlesid, kuidas kategoriseerisid neid SKMid ja (korpuse jaoks) inimesed

”

o

^

■ ■

<

2024

x

~

■

/

*



Mida ja miks tegime?

- **Kas SKMidest on eesti keelega üldse kasu?**
 - Pole piiratud ressurssidega keel (ingl *low-resourced language*), nt suur keelekorpus.
 - Keelemudelites alaesindatud võrreldes inglise keelega.
- Kas SKMid suudavad adekvaatselt öelda, missugustes tekstides eesti sõnu kasutatakse?
- ÕS 2018 ja EKI ühendsõnastiku 2024 kõnekeelse märgendiga 100 sõna + võrdluseks 100 märgendita sõna.
- Samadele sõnadele ka korpusanalüüs (SKMide võimekust ei teadnud, algandmetele ligi ei pääsenud).
 - Eesti keele ühendkorpus 2021: **žanrid** ja **teemad** (Koppel, Kallas 2022; alates 2023 alateemad).
 - Žanre ja teemasid polnud selliseks ülesandeks süsteemselt kasutatud.
 - Proovisime orientiiri välja töötada.



Suured keelemudelid

- **GPT-4** ja **Gemini 1.0 Ultra** (märts 2024), **GPT-4o** ja **Claude 3 Sonnet** (mai 2024)
- **Roll:** „Sa oled eesti keele sõnaraamatu koostaja.“
- **Prompt:** „Millistes tekstides kasutatakse sisestatud eesti(keelset) sõna? Kui sul ei ole selle kohta informatsiooni, siis ütle, et sa ei oska öelda.“
- GPT-4o ja GPT-4 koondinformatiivsus: **81%** (oli võimalik määrata, milline märgend sõnale lisada)
 - GPT-4 hallutsineeris, GPT-4o otsekohelem
- Claude ja Gemini jäid lõppkatsest kõrvale:
 - *mutt* – „Sõna „mutt“ viitab üldiselt segaverelisele koerale.“
 - *manu* – „See on ladina päritoluga sõna, mida kasutatakse mitmetes keeltes, sh eesti keeles.“
- Järeldasime: **SKMe tasub eesti keele ülesannetega edasi katsetada.**



Ühendkorpus

- **Korpuse** informatiivsus: **82,5%**
 - *fanatt* vs. *oskar*
- Korpusorientiir:
 - kui sõna esineb **36% või rohkem** blogides ja foorumites (pigem informaalised žanrid vs. seadused, kasutusjuhendid, akadeemilised tekstid),
 - siis võib *kahtlustada*, et sõna kiputakse kasutama informaalsetes kontekstides.
 - Ei saa rakendada midagi muud arvestamata – see on orientiir, vihje.

”

o

^

■ ■

<

2025

x

~

■

/

*



Mida ja kuidas tegime? (I)

- **Jaanuar:** testisime SKMide eri versioone välja valitud 30 sõnaga ja täiustasime prompti.
- Lõppseis: **1) GPT-4o, 2) Gemini 1.5 Pro, 3) Claude 3.5 Sonnet** (Llama ja DeepSeek jäid kõrvale)
- **Veebruar:** 1330 ÕSi sõna, millel kõnekeelne tähendus (veebruari lõpu seisuga),
 - et loodetavasti kaasa aidata ÕSi koostajate tööle.

Roll ja tööülesanne: „Oled eesti keele sõnaraamatu koostaja ja pead otsustama, kas sõnale või väljendile on vaja lisada registrimärgend.“

Prompt: „Kas eesti(keelset) sõna „X“ tähenduses ‘YY’ kasutatakse pigem [informaalsetes, neutraalsetes/formaalsetes] tekstides? Kui sa ei oska eristust teha või see ei tule selgelt esile, siis ütle, et „ei kohaldu“.

Informaalsed tekstid on teiste seas näiteks blogid, foorumid, kommentaariumid, chativestlused, sotsiaalmeedia tekstid, trükivigased täis tekstid, vahel ka raamatutegelaste otsekõne.

Palun põhjenda oma valikut.

Lähtu vastates ainult oma treeningandmetest, mitte välisotsingutest ja välistest andmebaasidest (sh EKI sõnastikest).“



Mida ja kuidas tegime? (II)

- 957 sõna (72% 1330-st) esineb tähenduses ‘YY’ pigem informaalsetes tekstides.
- 373 sõna (28%) puhul vastused erinesid.
- **Märts:** 373 sõna SKMidele võrdlusandmestiku väljatöötamiseks + piiripealsete sõnade endi uurimiseks
 - 0/3: *võpsik, vaevuma* ‘viitsima midagi teha’
 - 1/3: *disklahv, flaier*
 - 2/3: *kainer, diil*
 - (vrd 3/3: *kits* ‘pealekaebaja’, *parkima* ‘rohkesti ja aplalt sööma’)
- Igat sõna hindas 3 tööühma liiget (9 liiget, sõnad jagati võrdselt)
- Tuli valida, kas *kõnekeelne* märgend või mitte (jah/ei).
- SKMide vastuseid kõrval ei olnud, korpust sai kasutada.



Töörühma liikmete hinnangud

- 230 sõnale ehk **61,7% ei paneks** töörühm kõnekeelsuse märgendit.
 - Nii mõnelegi paneks mõne muu märgendi.
- 143 sõnale ehk **38,3% paneks** töörühm just kõnekeelsuse märgendi.
- Parim kokkulangevus töörühma hinnangutega: GPT-4o (61,4%)**
 - Jaauari testides GPT-4o kokkulangevus ka sarnases suurusjärgus.
 - Selliseks sõnaraamatutööks etim?
- Claude ainsana „ei kohaldu“, teised on konkreetse otsuse teinud.
- Kas SKMe on üldse mõtet ühe allikana sõnaraamatutöös kasutada?
 - SKMide vastused on (statistiliselt oluliselt) paremad, kui suvaliselt lahmides saadud vastused.**
 - Jah, on mõtet.**
 - Meie koostatud võrdlusandmestikku on võimalik kasutada ka uute mudeliversioonide hindamiseks.

SKM	Claude	Gemini	GPT
+	212 (56,8%)	205 (55,0%)	229 (61,4%)
–	117 (31,4%)	168 (45,0%)	114 (38,6%)
?	44 (11,8%)	0	0
Kokku	373	373	373



Kõhklusi tööruhma kommentaaridest

- **porken** – „pigem mingi stiilitundliku või vanamoodsa märgendi“
- **noorsand** – 1) „stilistiliselt markeeritud“, 2) „mingi muu, stiilitundlik? vanamoodne?“, 3) „vanamoeline?“
- **lutikas** ‘pealtkuulamiseade’ – „polüseemia tõttu raske korpusest abi saada“
- **küljeluu** ‘ribi’ – „raske korpusest abi saada, põhiliselt küljeluu kui naine“
- **kriimsilm** – „värvikas väljendus (mitte neutr keel), aga ei tea, kas seepärast kõnek tingimata“
- **pritsimaja** – „paneks NALJATLEV, kui saaks“
- **mäluauk** – „,,mäller“ oleks kõnek, pigem PILTLIK?“
- **soendama** ja **tahaplaanile** – „segab teadmine sellest, mis on „õige““



Põhjuseid kõnekeelsuse märgendi taga

- **kas-liide:** *süümekas, huulekas* ‘huulepulk’ (vrd ‘lopsakate huultega’)
- **lühem variant:** *baka, mill* ‘miljon’
- **olemuselt paha:** *jeekim, porduelu*
- **mugandatud võõrsõna:** *diil, (?)volüüm* ‘heli tugevus’

? üle kantud tähendus: *sihverplaat* ‘nägu’, *kullaauk*

? **põhi- või täiendsõna-ellips** (vt Langemets 2009: 176–177): *paralleel* ‘paralleelklass’, *plekk* ‘kontsaplekk’

? **pole termin / rahvakeelne:** *aastanumber* ‘kalendriaastat märkiv arv’, *vereproov* ‘vereproovi uuring’

? **murdest:** *koolma, manu*



SKMide kontroll (märtsi lõpp)

- SKMide vastused võivad muutuda
 - (Ka inimese hinnang võib samu sõnu vaadates erineda.)
- Katse kordamine: **samad kolm SKMi, samad 1330 ÕSi sõna, sama prompt**
 - Kas ja kui palju vastused erinevad?
 - Kas erinevus tuleb statistiliselt oluline?
- **Tulemus 1330 sõnaga:** kõigi kolme SKMi vastused erinevad esimesest korrast **2%** (pole statistiliselt olulised vahed)
- **Tulemus 373 piiripealse sõnaga:** erinevus **4–6%** (pole statistiliselt olulised vahed)
- **Järeldus:** kaks leksikograafi saavad samadel sõnadel sama prompti kasutades **vähemalt 94%** tõenäosusega sama tulemuse.
 - Erinevus korpuse kasutamisest, kus keeleandmed ei muutu, Sketch Engine'is juhuvalimid kõigile samas järjekorras.
 - Ent: tõlgendaja tõlgendab andmeid enda teadmiste ja keeletunnetuse pinnalt,
 - **pisikesed erinevused ja mitmekesisus on sellesse töösse sisse kirjutatud.**



Näide SKMi vastuse muutusest

Claude 3.5 Sonnet: *kroku*

- 1. ei kohaldu:** „Minu treeningandmetes ei ole piisavalt näiteid sõna „kroku“ kasutuse kohta, et teha kindlat järeldust selle registrilise kuuluvuse kohta. Kuigi see on ilmselt kõnekeelne lühend sõnast „krokodill“, ei saa ma ilma piisavate kasutusandmeteta kindlalt väita, kas seda kasutatakse pigem informaalsetes või neutraalsetes kontekstides. Sõna registrilise kuuluvuse määramiseks oleks vaja näha selle tegelikku kasutust erinevates tekstides.“
- 2. informaalsetes:** „„Kroku“ on krokodilli lühendatud, hellitav või lapsekeelne vorm, mida kasutatakse pigem mitteametlikes olukordades ja informaaelses kõnes. Neutraalsetes ja formaalsetes tekstides (nt teadusartiklid, entsüklopeediad) kasutatakse täispikka vormi „krokodill“. „Kroku“ esineb tõenäolisemalt lastejuttudes, igapäevases kõnes, sotsiaalmeedias ja muudes vabamas vormis tekstides.“

”

o

^

■ ■

<

Lõpp sõnaks

/

~

x

■

*

**SKMide kasutamisel on korpuse ees küll puudusi
(algandmeid ei saa kontrollida, vastused pisut muutuvad jm).**

**Ent katsetused on näidanud, et
SKMe on mõtet proovida leksikograafias kasutada.**

On palju sõnu (tähendusi), mis ei kuulu mustvalgelt kõnekeelde.

**Loome leksikograafide tuge märgendi üle otsustamiseks,
proovides korpuse ja SKMide abil otsustusprotsessi
kiirendada ja süsteemsemaks muuta.**

**Lõplik otsustaja on aga endiselt leksikograaf, keeletekspert,
mis on ühtlasi nii eelis (ta saab andmeid vajadusel juurde vaadata)
kui ka puudus (mõjutama hakkavad eelteadmised, isiklik keeletunnetus jm).**

Sõnaraamatumärgend markeerib ja
standardiseerib sõna(tähenduse) ära, eriti ÕSis.
(Märgendeid on sageli tõlgendatud normatiivsena,
Langemets, Risberg, Algvere 2024; Trap-Jensen 2002.)

Sõna(tähenduse) saab (de)standardiseerida,
jättes märgendi ära.

Kirjandus

- Davies, Mark 2025.** Corpora and AI / LLMs: Overview. English-Corpora.org. <https://www.english-corpora.org/ai-llms/>
- de Schryver, Gilles-Maurice 2023.** Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4), 355–387.
- Henriksson, Erik, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo, Veronika Laippala 2024.** From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations. *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, 308–318.
- Jakubíček, Miloš, Michael Rundell 2023.** The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? *Electronic Lexicography in the 21st Century: Invisible Lexicography, Brno, Czechia. eLex 2023*, 518–533.
- Koppel, Kristina, Jelena Kallas 2022.** Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat* 18, 207–228.
- Langemets, Margit 2009.** Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras. Doktoritöö. Tallinna Ülikool.
- Langemets, Margit, Lydia Risberg, Kristel Algvere 2024.** To Dream or Not to Dream About ‘Correct’ Meanings? Insights into the User Experience Survey. *XXI EURALEX International Congress. Cavtat, Croatia*, 700–719.
- Pool, Raili, Marika Teral, Jelena Kallas 2025.** Keeleminutid. Eesti keele õppijad sõnaraamatute maailmas. *ERR Kultuur*, 24.03. <https://kultuur.err.ee/1609642106/keeleminutid-eesti-keele-oppijad-sonaraamatute-maailmas>
- Trap-Jensen, Lars 2002.** Descriptive and Normative Aspects of Lexicographic Decision-Making: The Borderline Cases. *Proceedings of the Tenth EURALEX International Congress. Copenhagen*, 503–509.
- Zgusta, Ladislav 1992.** The Czech-Chinese Dictionary and the theory of lexicography. *International Journal of Lexicography* 5(2), 85–128.

Suured keelemudelid

Claude 3 Sonnet = Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku.

<https://api.semanticscholar.org/CorpusID:268232499>

Claude 3.5 Sonnet = Anthropic. Claude 3.5 Sonnet Model Card Addendum. <https://paperswithcode.com/paper/claude-3-5-sonnet-model-card-addendum>

Gemini 1.0 Ultra = Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023).

Gemini 1.5 Pro = Team, Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530

GPT-4 = Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774

GPT-4o = Hurst, Aaron, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow et al. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276

”

o

^

■ ■

<

...töö jätkub

x

~

■

/

*