

Digikratiid avavad vana kirjakeelt: suurtest keelemudelitest 17.–18. sajandi sõnaraamatute uurimisel

Madis Jürviste (EKI nooremteadur-leksikograaf, TÜ doktorant)

Tiina Paet (EKI teadur-vanemkeelekorraldaja)

Sven-Erik Soosaar (EKI vanemteadur)

Taust I

- SKMid ja leksikograafia: arvukalt uurimusi
- Meie: 17.–18. saj sõnastikud

Taust I

- SKMid ja leksikograafia: arvukalt uurimusi
- Meie: 17.–18. saj sõnastikud
- Eesmärk
 - vana sõnavara kaardistamine (esinemused mis kujul ja kus?)
 - ajalooline kihistus ÜSis
 - vanade sõnastike varustamine nüüdisvastetega

Taust II

- Mudelid: GPT-4o, Claude 3 Opus, Gemini 1.5 Pro
- Materjal: Stahl (1637), Gutsclaff (1648), Göseken (1660), Vestring (17XX), Helle (1732), Hupel (1780)

Taust II

- Mudelid: GPT-4o, Claude 3 Opus, Gemini 1.5 Pro
- Materjal: Stahl (1637), Gutsloff (1648), Göseken (1660), Vestring (17XX), Helle (1732), Hupel (1780)
- Katsed:
 - K1: nüüdiskujud > vanad kujud (ametinimetused)
 - K2: vanad lõuna-eesti sõnad (Gutsloff) <> nüüdiskujud
 - K3: vanad laensõnad (Stahl, Gutsloff, Göseken) <> nüüdiskujud

K1: ametid ja rollid

- Eesmärk: leida nüüdisõnadele vastavad vanad ametinimetused / sotsiaalsed rollid
 - Allikad: Stahl 1637 ... Hupel 1780 (6 sõnastikku)
 - Sisend: 30 (5 x 6, raskemini äratuntavad)

Vana sõnakuju

Saksa vaste

Rackel	Hencker
Kassucksep	Kürßner
jütleja	Prediger
oppeja	Lehrer
sundija	Richter
Portt	Hure
Portopähline	Hurer
Kessineck	Bürge
Jeck	Hofnarr
Rahhajohataja	Münzmeister
kaffer	Ausgeber
kachtja	Färber
Kühna-Mees	Dieb
Kiwwi Tackar	Stein Metzger
Raamato Rutzoja	Buchdrucker
Pilli ajaja	Ein Sackpfeiffer

Vana sõnakuju

Saksa vaste

Laudamees	Ein unteutscher rechtsfinder
Mängimees	Ein Spielmann musicant
Lämmia	Ein Zauberer
kulaja	der Kundschaster
lambrine	der Schäfer
nikker	der Tischler
kolimees	der Schüler
siggur	der Schwein-Hirte
heikaja	Ausrufer, Herold
kandijanne	ein Tragender
kalla müja	Fischhändler
müütnik	der Zöllner
laiwa wallitseja	Steuermann
Jooks	Ein Curierer

K1, ametid ja rollid

- Eelkatsed: "Leia vasted" sõnadele (a) *kohtunik*, (b) *varas*,
(c) *hirsnik*
- Põhikatse: *kohtunik* on vanades *otsitav* (f1 loend ja f2 tabel)

K1, ametid ja rollid

- Eelkatsed: "Leia vasted" sõnadele (a) *kohtunik*, (b) *varas*, (c) *hirsnik*
- Põhikatse: *kohtunik* on vanades *otsitav* (f1 loend ja f2 tabel)
- Tulemused:
 - GPT-4o 80%
 - Claude 3 Opus 90%
 - Gemini 1.5 Pro 67%

K2: Vanad lõunaeesti sõnad Gutsclaffil

- Katse aluseks 54 sõna, sh:
 - kokkulangevaid sõnu (*kirp, laisk, puhas, raha*)
 - kirjaviisist tuleneva vähese erinevusega (*Hawd, Merri, pannema*)
 - täiesti erinevaid (*håilmo* 'õisik', *teww* 'kops')

K2/1: tänapäeva sõnad > Gutsclaffi sõnad

- Lähteandmed: Gu sõnad + tnp-kujud + sks / Id tõlkevaste
- Ülesanne: leida tnp-sõnadele vasted Gu sõnastikust
- Sh eL sünod (*ööbik* – *sisask*, *õde* – *sõsar*)

K2/1: tänapäeva sõnad > Gutslaffi sõnad

- Lähteandmed: Gu sõnad + tnp-kujud + sks / Id tõlkevaste
- Ülesanne: leida tnp-sõnadele vasted Gu sõnastikust
- Sh eL sünod (*ööbik* – *sisask*, *õde* – *sõsar*)

Tulemused:

- Gemini: 48 õiget (veamäär 24%)
- Claude: 47 õiget (vm 25%)
- GPT-4o 33 õiget (vm 47,6%)

K2/2: Gutslaffi sõnad > tänapäeva sõnad

- Gu sõnadele vastavad tnp-sõnad
- Lubasime anda mitu vastet + komm. Enim komme: Gemini

1.5 Pro.

K2/2: Gutslaffi sõnad > tänapäeva sõnad

- Gu sõnadele vastavad tnp-sõnad
- Lubasime anda mitu vastet + komm. Enim komme: Gemini 1.5 Pro.
- Tulemused:
 - Kõik mudelid õigesti: 48,15% sõnadest.
 - Kõik mudelid eksisid 6 sõnaga: *håilmo, kattessas, neutzick, rachk, rebes, Wennick*.
 - Claude 3 opus: 44 õiget (vm 19%), GPT-4o 39 õiget (vm 28%, Gemini 1.5 Pro 33 õiget (vm 39%).

K2/2: Sõnad

haggo	Morgenröthe
Haugk	Hecht
Hawd	Grab
Hebbedus	Scham
helle	Klang, Laute
hemme	Nass
hiuws	Haar (crinis)
hailmo	florimentum
iggas	semper
issack	Stieff-Vater
kajema	schawen, versuchen
kattessas	
Kattus	Dach
Kirp	Floh
Kôhes	ubi
Kübbar	Hut
laisk	faul
lobbisema	schnattern

loti	Kugel
lubbiraya	Aussatz
mas	forte,
	forsan
Merri	Meer
nagl	Nagel
nanne	Nase
neutzick	Magd
ombelma	neen
pannema	legen
penni	Hund
perrast	nach
piddama	müssen
pitk	lang
puhas	rein,
	sauber
pärn	Miltze
pörss	Arsch
rachk	Kahm
Raha	Geld

ramato nachk	Papier
Rasw	Fett
rawdreuwas	Harnisch
rebes	Murene
	piscis
saddama	schneyen
seng	Bett
Sigga	Sau,
	Schwein
sissas	Nachtigall
siwck	Schlange
Söma	Essen
sössar	Schwester
Tarck	Klug
teww	Lunge
wagga	fromm
wannuck	Eid,
	Schwur
weitz	Messer
wennick	Russa
Öh	Nacht

K3: Laensõnad

- Eesmärk:
 - vanade kirjakujuude sidumine omavahel
 - vanadele kirjakujuude sidumine nüüdiskujuudega
 - *mijrrit* (Göseken), *Mirra* (Gutslaff) ja *mürr* (tnp)
 - sks kuju ei andnud, promptis korraldus anda selgitus

K3: Laensõnad

- **Eesmärk:**
 - vanade kirjakujuude sidumine omavahel ☺
 - vanadele kirjakujuude sidumine nüüdiskujuudega
 - *mijrrit* (Göseken), *Mirra* (Gutslaff) ja *mürr* (tnp)
 - sks kuju ei andnud, promptis korraldus anda selgitus
- **Materjal:** vasted vähemalt 2 sõnastikus; neist (a) sarnased: *pipar*, *pippar* ja *pipper*; (b) mõnevõrra erinevad: *äädikas* ja *Ettickas*; (c) väga erinevad: *väävel* ja *koirasitt* / *pennisitt*

Nüüdiskuju

kuningas

paasapüha

trööst

väävel

mürr

loorber

häärber

pipar

viiruk

raad

ingel

peeker

palsam

penning

juut

kurat

äädikas

ahv

draakon

pärdik

Stahl

kuningas

Leehawotme Pöha, ohstrit

Trohst, Röhm

lohrbehr

Öhmaja

Pippar

Rahd

Engel

Karrick

üx kaunis heh haisija rocht

Juddalinne

Kurrat

Ettickas

valgkemees, lendwamees,
püssohand**Göseken**

Kunningas

ohstrit, pahscha pöha, leehawotmepöha

Rohst, Rööma

KoiraSitt, wewel

mijrrit

Loorbeer

vöhras kodda

Pippar

Wyrohki, wijroht

rahdi

ingel

picker

kallis hais, rocht

killing

Judalinne

Kurrat, pahharet

Ettickas

pertick

püssihand, valgkemees

pertick

Gutslaff

kunningas

Lehawötta pöha

Trôst, Röhm

Pennisitt

Mirra

Maja

Pipper

Sauwrocht

Raht

Engel

Karritz

Teng

Juddalinne

Kurrat

Ettikas

Pertike

Püssohand

Pertike

K3: Tulemused

- **Kõik õigesti** (ükski mudel ei eksinud): *kuningas (kuningas, Kunningas), pipar (Pippar, Pipper), ingel (Engel, ingel), kurat (Kurrat, pahharet); juut (Juddalinne, Judalinne, Juddalinne), äädikas (Ettickas, Ettikas), raad (Rahd, Rahdi, Raht), trööst (Röhm, Trohst, Rohst, Rööm, Trôst), Paasapüha (Leehawotme Pöha, Ohstrit, pahscha pöha, Lehawötta pöha).*
- **Kaks valesti** (kaks mudelit eksisid): *draakon (valgkemees, lendwamees, püssohand; püssihand, valgkemees; Püssohand).*

K3: Draakon ja väävel

Gutslaff (1648):

Drache Püßtohand/a/

Schwefel Weñißitt/a

Göseken (1660):

Drach / [auffm Rohr] püßihand.
Drach / [in der Luft] walgte mees.

schwefel / (sulphur] Koira Sitt/wewel; wewolit.

Kokkuvõte

- Täpseim mudel: Claude 3 Opus
 - õigeid K1 90%, K2 74–81% ja K3 60–90%
- Edukus sõltub sisendsõnade valikust
 - mida vähem erinevusi tänapäeva kujust, seda täpsem
- Mudelid veebilehti ei külastanud: kasutada nn eellugemist
- Edasised katsed: sõnastikeülesed nüüdisvasted

- Beliga, S., & Filipović Petrović, I. 2024.** Large language models supporting lexicography: Conceptual organization of Croatian idioms. Conference on Language Technologies and Digital Humanities. Ljubljana, Slovenia.
- De Schryver, Gilles-Maurice; Joffe, David 2023.** The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks. 20th CODH Seminar, Center for Open Data in the Humanities. Research Organization of Information and Systems, National Institute of Informatics.
- Jürviste, Madis 2025.** Võrdlevalt eesti vanemast leksikograafiast. – Keel ja Kirjandus, 1–2, 36–53. <https://doi.org/10.54013/kk806a4>
- Jürviste, Madis; Paet, Tiina; Soosaar, Sven-Erik 2025.** Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil. Eesti Rakenduslingvistika Ühingu aastaraamat, 21.
- Keem, Hella 1998.** Johannes Gutsloff eesti keel ja Urvaste murrak, teoses Gutsloff 1998, lk 317–332.
- Kibbermann, E.; Kirotar, S.; Koppel, P. 1975.** Saksa-eesti sõnaraamat = Deutsch-estnisches Wörterbuch. Tallinn: Valgus.
- Kikas, Kristel 2002.** Mida sisaldab Heinrich Stahli *Vocabula*? Toim. Valve-Liivi Kingisepp. Tartu Ülikooli eesti keele õppetooli toimetised 11. Tartu: Tartu Ülikool.
- Kingisepp, Valve-Liivi; Ress, Kristel; Tafenau, Kai 2010.** Heinrich Gösekeni grammatika ja sõnastik 350. Tartu: Tartu Ülikool.
- Lew, Robert 2023.** ChatGPT as a COBUILD lexicographer. Humanit Soc Sci Commun 10, 704 (2023). <https://doi.org/10.1057/s41599-023-02119-6> . ChatGPT as a COBUILD lexicographer. Humanities and Social Sciences Communications - ChatGPT as a COBUILD lexicographer.
- Tafenau, Kai 2011.** *Heinrich Gösekeni sõnaraamatu seni märkamata eeskujud*. – Keel ja Kirjandus, nr 6, lk 425–439.
- Helle, Anton Thor 1732.** Kurtzgefaszte Anweisung Zur Ehstnischen Sprache. Halle: Stephan Orban. <http://www.digar.ee/id/nlib-digar:100071> (10.09.2024)
- Vestring 2000** = Vestring, Salomo Heinrich 2000 [17XX]. Lexicon Esthonio Germanicum. Võrguteavik. Toim. Ellen Kaldjärv, Krista Aru, Arvo Krikmann. Tartu: Eesti Kirjandusmuuseum. <https://www.folklore.ee/~kriku/VESTRING/index.htm> (17.09.2024)
- Wild, Kate 2024.** Using Large Language Models for a Large Historical Dictionary: Challenges and Opportunities for the Oxford English Dictionary.

K1, ametid ja rollid: prompt

```
data = {  
  "model": "claude-3-opus-20240229",  
  "max_tokens": 4096,  
  "temperature": 0.2,  
  "messages": [  
    {  
      "role": "user",  
      "content": f"You are an expert in analyzing old Estonian vocabulary. Sinu ülesanne on leida tekstifailis ('loend_katse2.txt') esitatud sõnade esinemiskujud sõnastikufailist ('Fail_katse2.xlsx'). Allpool on tekstifaili sisu: {words_to_search_content} Allpool on sõnastikufaili sisu: {file_content} Vastuses loetle ridade kaupa, millised tabelis esitatud vanad sõnad vastavad tekstifailis nimetatud ametitele, näiteks nii: * [vastuse rea nr] kohtunik on vanades sõnastikes kujul /sundija/. Allikas: Stahl, lk 101, real nr [siia kirjuta tabeli rea number]. Kui sa tekstifaili sõnade tähendust ei tea, otsi Sõnaveebist: https://sonaveeb.ee/search/unif/dlall/dsall/\[siia kirjuta otsitav sõna ilma kantsulgudeta\]. Tulemus väljasta txt-formaadis failina."  
    }  
  ]  
}
```