

Keelemudelid kui leksikograafi pisuhännad vanade sõnakujude tuvastamisel?

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: SKMid + 17.–18. saj sõnastikud
- Eesmärk
 - ajalooline kihistus (esinemused mis kujul ja kus?)
 - 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)
- Katsed:
 - K1: nüüdiskujud > vanad kujud (ametinimetused)
 - K2: vanad lõuna-eesti sõnad (Gutsclaff) <> nüüdiskujud
 - K3: vanad laensõnad (Stahl, Gutsclaff, 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)
- Tulemused:
 - Kõik õigesti (ükski mudel ei eksinud): 17/30
 - Kaks valesti (kaks mudelit eksisid): *rackel* 'timukas' ja *jooks* 'käskjalg'
 - Kõik valesti (kõik mudelid eksisid): *portopähline* 'kupeldaja' ja *kaffer* 'järelevaataja'

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."  
    }  
  ]  
}
```

K1, ametid ja rollid: Tulemused

- Õigeid vastuseid:
 - GPT-4o 80%
 - Claude 3 Opus 90%
 - Gemini 1.5 Pro 67%

K2: Vanad lõunaeesti sõnad Gutsclaffil

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

K2/1: tänapäeva sõnad > Gutsclaffi sõnad

- Lähteandmed tabelis Gutsclaffi sõnade tänapäeva vastetega koos saksakeelse (v ladinakeelse) tõlkevastega
- 3 SKM-le anti ülesanne leida tänapäeva eesti kirjakeele sõnadele etteantud tabeli põhjal Gutsclaffi vaste
- Lähteandmete seas oli ka kirjakeele sõnade lõunaeestilisi sünonüüme (nt *ööbik* – *sisask*, *õde* – *sõsar*)

K2/1: Tulemused

- Kõik kolm mudelit leidsid õige vaste 27 sõnale 63-st (42,2%)
- Kõik mudelid eksisid vaid 3 sõna puhul (*kops* : *teww* 'Lunge',
agu : *haggo* 'Morgenröthe' ja *hallitus* : *rachk* 'Kahm')
- 19 sõnast, millele leidis õige vaste vaid üks mudel kolmest, oli
10 juhul selleks Gemini, 8 juhul Claude 3 opus ja vaid 1 juhul
GPT-4o

K2/1: Tulemused

- Edukaim oli Gemini 48 õige vastusega (veamäär 23,8%), Claude jäi minimaalse vahega teiseks – 47 õiget vastust (veamäär 25,4%), selgelt nõrgema tulemuse andis GPT-4o 33 õige vastusega (veamäär 47,6%)
- Sünonüümide puhul anti murdesõnale või murdetaustaga sõnale õige vaste 7 juhul ja kirjakeele sõnale 9 juhul

K2/2: Gutslaffi sõnad > tänapäeva sõnad

- Leida Gutslaffi sõnadele etteantud tabeli põhjal vaste tänapäeva eesti kirjakeeles
- Lubasime anda igale sõnale mitu vastet ja palusime selgitada, kui vastet ei leitud, miks ei leitud. Seda võimalust kasutas kõige rohkem Gemini 1.5 Pro.

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
lubbiraya
mas

Merri
nagl
nanne
neutzick
ombelma
pannema
penni
perrast
piddama
pitk
puhas

pärn
pörss
rachk
Raha

Kugel
Aussatz
forte,
forsan
Meer
Nagel
Nase
Magd
neen
legen
Hund
nach
müssen
lang
rein,
sauber
Miltze
Arsch
Kahm
Geld

ramato nachk
Rasw
rawdreuwas
rebes

saddama
seng
Sigga

sissas
siwck
Söma
sössar
Tarck
teww
wagga f
wannuck

weitz
wennick
Öh

Papier
Fett
Harnisch
Murene
piscis
schneyen
Bett
Sau,
Schwein
Nachtigall
Schlange
Essen
Schwester
Klug
Lunge
romm
Eid,
Schwur
Messer
Russa
Nacht

Gutsl.	Sks / ld	GPT-4o	Claude 3 Opus	Gemini 1.5 Pro
hemme	Nass	niiske, märg	märg, niiske	-
hailmo	florimentum	ilu	õilmik, õis, õieke	kaunistus, ehtimine, iluasi
issack	Stieff-Vater	isasukk	kasuisa, võõrasisa	võõrasisa, kasuisa
kajema	schawen, versuchen	kahetsema, proovima	kaema, vaatama, uurima, proovima	vaadata, proovida, püüda
kattessas		katus	-	-
Kübbär	Hut	kübar	kübar	kübar, müts, onn
laisk	faul	laisav, laiseldav	laisk	laisk, loiud, lodev
lobbisema	schnattern	prääksuma	lobisema	lobiseda, lobisema, jutustama
loti	Kugel	kuul	lodi	lodi, loot, kuul
neutzick	Magd	neitsi	teenijatüdruk, teenijanna	kasulik, vajalik, tarvilik
penni	Hund	koer	peni, koer	peni, penn
pitk	lang	pidu	pikk	pikk, pikaajaline, kaua kestev
pörss	Arsch	perse	perse, tagumik, tuhar	Kuigi saksa/ladina vaste ...
rachk	Kahm	rahk, kärn	rahk (Sõna 'rahk' ...	rahakott, kukkur, rahapaun
rebes	Murene piscis	räim	angerjas (Seda ...	luts, latik, särg
teww	Lunge	kops	kops	hoog, sööst, rünnak
wennick	Russa	vene	-	venelane, Vene, Venemaa

K2/2:
tulemused

K2/2: Tulemus

- 54 Gutsclaffi sõnale andsid õige tänapäevakeele vaste(d) kõik kolm mudelit 26 juhul (48,15%).
- Kõik mudelid eksisid 6 sõna puhul: *hailmo, kattessas, neutzick, rachk, rebes, Wennick*.
- Katses oli selgelt edukaim Claude 3 opus 44 õige vastega (veamäär 18,5%), teiseks tuli GPT-4o 39 õige vastega (vm 27,8%) ja 3. Gemini 1.5 Pro 33 õige vastega (vm 38,9%).

K3: Laensõnad

- **Eesmärk:** selgitada SKMide võimekust siduda vanu kirjakujuisid omavahel ja neile vastavate nüüdiskujudega, nt *mijrrit* (Göseken), *Mirra* (Gutslaff) ja *mürr* (tnp)
 - Sks kuju ei andnud, promptis selgituste andmise käsk, sõnade hulk erinev
- **Materjal:** vasteid vähemalt 2 sõnastikus; neist 1) sarnased: *pipar*, *pippar* ja *pipper*; 2) mõnevõrra erinevad: *äädikas* ja *Ettickas*; 3) 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 1

- **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: Tulemused 2

	Claude 3 Opus			Gemini 1.5 Pro			GPT-4o		
Nüüdiskuju	Stahl	Göseken	Gutslaff	Stahl	Göseken	Gutslaff	Stahl	Göseken	Gutslaff
väävel	x	1	1	x	1		x	1	
mürr	x	1	1	x	1	1	x		1
loorber	1	1	x		1	x		1	x
häärber		1			1	1		1	1
viiruk	x	1	1	x			x		1
peeker		1						1	
palsam	1	1	x			x	1	1	x
ahv	x			x			x		
draakon	1	1	1						
pärdik	x	1		x	1	1	x	1	1
Õigeid	60%	90%	65%	45%	70%	60%	50%	75%	65%
Valesid	20%	10%	30%	35%	30%	35%	40%	25%	30%
Puudu	20%	0	5%	20%	0	5%	10%	0	5%

Kõigil mudelitel õigesti:

kuningas
pipar
ingel
kurat
kuut
äädikas
paasapüha
praad
trööst

K3: Draakon ja väävel

Gutslaff (1648):

Drache Püßtohand/a/

Schwefel Weißitt/a

Göseken (1660):

Drach / [auffm Rohr] püßihand.
Drach / [in der Luft] walgtē mess.

schweffel / (sulphur) Kotra 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.
- Despot, Kristina Štrkalj; Ostroški Anić, Ana; Brač, Ivana (Toim).** Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. 8–12 October 2024. Cavtat, Croatia.
- Geeraerts, Dirk 2024.** [Viide Euralex 2024 *keynote*'ile.]
- Keem, Hella 1998.** Johannes Gutsclaffi eesti keel ja Urvaste murrak, teoses Gutsclaff 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 Esthonicum 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.