

EESTI VANADE SÕNAKUJUDE TUVASTAMISEST SUURTE KEELEMUDELITEGA

Madis Jürviste, Tiina Paet, Sven-Erik Soosaar

Ülevaade. Artiklis kirjeldatakse suurte keelemudelite võimekust vanade (17. ja 18. saj) eesti keelt sisaldavate sõnastike sisu analüüsimisel. Autorid korraldasid kolme suure keelemudeli (GPT-4o, Gemini 1.5 Pro ja Claude 3 Opus) kokku kolm katset. Esimese katse valimis olid vanad ametinimetused ja sotsiaalsed rollid, teises valimis lõunaeesti sõnad, kolmandas vanad laensõnad. Katsete tulemused näitavad keelemudelite suurt potentsiaali vanade sõnakujude ühendamisel nüüdiskujudega: tulevikus saaks seda rakendada sõnastikes märksõnade diakroonilisel kirjeldamisel koos viidetega varasematele esinemisaegadele ja -kohtadele.*

Võtmesõnad: ajalooline leksikograafia, eesti kirjakeele ajalugu, suured keelemudelid, eesti keel

1. Sissejuhatus

Vanade sõnaraamatute uurimine ja analüüs on äärmiselt aja- ja töömahukas, kuna suur osa neist sõnastikest ei ole masinloetavad ja neis sisalduvat keeleainest (märksõnu ja näitelauseid) on sageli keeruline seostada tänapäevase keelega. Seoses suurte keelemudelite (SKM) kiire arenguga on palju uuritud nende potentsiaalset kasutust ka leksikograafias. Siinse uurimuse laiem eesmärk on uurida SKM-ide rakendamise võimalusi sõnakujude (sh piirkondlike) tuvastamisel eesti vanema sõnavara kihistuses. Kitsamalt seadsime eesmärgiks hinnata kolme SKM-i (GPT-4o, Gemini 1.5 Pro ja Claude 3 Opus) tõhusust neile esitatud sõnakujude tuvastamise ülesannete lahendamisel. SKM-ide olulisus leksikograafia jaoks sai ilmseks õige kiiresti pärast seda, kui OpenAI tegi 30.11.2022 kõigile kättesaadavaks ChatGPT esimese avaliku versiooni: vähem kui kolm kuud hiljem, 2023. aasta alguses tutvustasid Gilles-Maurice de Schryver ja David Joffe lahendust, mis näitas, et OpenAI keelemudeli abil on võimalik hõlpsasti genereerida *autentsena näivat*

* Artikkel valmis riikliku programmi EKKD-III1 "Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed" toel. Autorid tänavad Eleri Aedmaad tehniliste nõuannete ja abi eest API-päringute koostamisel.

sõnastikusisu (de Schryver, Joffe 2023). Edasised arengud ja mudelite üha täienev võimekus on võimaldanud keelemudelite laialdast rakendamist leksikograafias (Despot jt 2024), sh väga mahukate ajalooliste sõnaraamatute nagu “Oxford English Dictionary” sõnastiku toimetamisel (Wild 2024). Samal ajal on rõhutatud, et oluline on keskenduda mitte üksnes sellele, *mida* saab (leksikograafias ja mujalgi) SKM-idega teha, vaid tingimata tuleks küsida ka seda, *kuidas* keelemudelid selle tulemuseni jõuavad, mille nad kasutajale esitavad (Geeraerts 2024), kusjuures seda tuleks püüda teha põhjalikumalt kui lihtsalt nentides, et tegu on vektoreid kasutavate transformeritega (Vaswani jt 2017). Olemuselt on tegu tehisintellektsete arvutiprogrammidega, mis on treenitud masinõppe ja närvivõrkude abil töötleva ja genereerima loomulikku keelt.

Siinses artiklis kirjeldatud katsete kaugem eesmärk ja rakendusvõimalus on ajaloolise kihistuse loomine – võimalusel keelemudelite abiga – EKI ühendsõnastikus. Ajalooline kihistus sisaldaks andmeid selle kohta, kus, millal ja mis kujul on märksõna eesti varasemates leksikograafilistes allikates esinenud. Uuringuid selle kohta, kuidas automatiseerida sõnaraamatute koostamise tööd, et vähendada selleks kuluvat aega ja muid ressursse, tehakse mujal maailmas praegu väga aktiivselt (vt nt Lew 2023: 2).

Eesti varasemaist sõnastikest on põhjalikumalt analüüsitud Heinrich Stahli “Vocabulat” (Kikas 2002), samuti Johannes Gutsلاffi grammatikat ja sõnastikku (Gutsلاff 1998), Heinrich Gösekeni grammatikat ja sõnastikku (Kingisepp jt 2010, Tafenau 2011), samuti on käsitletud nende sõnastike ülesehitust (Jürviste 2025). Süsteemne ülevaade kõigi eesti keelt sisaldavate 17. ja 18. sajandi sõnastike leksikast seni puudub. Üks seni koostatud eesti ajaloolise leksikograafia rakendusi on ÕS 1918 kommenteeritud digiväljaanne (ÕS 1918), mis võimaldab linkide kaudu võrrelda 1918. aasta “Eesti keele õigekirjutuse-sõnaraamatu” kirjakujusid “Eesti õigekeelsussõnaraamat ÕS 2018” kirjakujudega. Kuna ÕS 1918 algses väljaandes on tähendusi esitatud vähe, tuli algsete märksõnade sidumiseks ÕS 2018 märksõnadega lisada teiste vanade allikate abiga seletusi, tähendusvihjeid või muid kommentaare (Paet 2023: 27–28). Sellega kaasnevat tehnilist laadi tööd, mis seni on eeldanud ajamahukat andmete käsitsi sisestamist, on edaspidi loodetavasti võimalik kiirendada ja hõlbustada SKM-ide abiga.

Teistes keeltes on varem uuritud ChatGPT kasutamist inglise ükskeelsetes sõnaraamatutes definitsioonide ja näitelause loomiseks ning hinnatud tehisintellekti võimekust genereerida täpset ja kvaliteetset märksõnaartikli sisu, samuti on uuritud võimalust tehisintellekti mudelite peenhäälestamiseks ekspertide tagasiside põhjal, parandamaks mudelite sooritust (Lew 2023: 2). Samuti on katsetatud SKM-ide abil horvaadi keele idioomide semantilist klassifitseerimist (vt nt Beliga, Filipović Petrović 2024).

Eesti 17. ja 18. sajandi saksa-eesti ja eesti-saksa sõnastikud on suuremalt jaolt digiteerimata (v.a Vestring 2000). Need on küll skaneeritud ja kättesaadavad PDF-idena, ent neis sisalduvat keeleainest on ainuüksi tekstituvastuse (OCR-i) vahenditega väga keeruline analüüsida, kuna praeguste tekstituvastusprogrammide veaprotsent on varieeruva kvaliteediga pildifailidelt gooti kirja tuvastamisel väga kõrge. Seetõttu tuli nende sõnastike analüüsimisel siinse uurimuse jaoks tugineda käsitsi sisestatud keeleinfole.

1.1. Mudelite valik ja katsed

Valisime katseteks kolm SKM-i: OpenAI GPT-4o (edaspidi GPT), Anthropic Claude 3 Opus (edaspidi Claude) ja Google'i Gemini 1.5 Pro (edaspidi Gemini). Mudelite valikul lähtusime nende arvestataval tasemel eesti keele genereerimise võimekusest, varasematest projekti käigus tehtud katsetuste tulemustest ja põhimõttest rakendada erinevate pakujate mudeleid. Kõiki katseid kordasime kõigi mudelitega, mis said enne prompti (nimetatud ka viibaks) rollikirjelduse (*system prompt*'i) "Oled ekspert eesti vana sõnavara analüüsimisel". Katsetes rakendatud API-päringute puhul ei jätnud keelemudelid endale eelnevaid päringuid mällu, nagu teevad kõik nendesamade SKM-ide brauseripõhised *chat*'i-rakendused.

Katseid tegime ajavahemikus 17.9–23.10.2024. Esimeses katses uurisime SKM-ide võimekust tuvastada nüüdisõnadele vastavaid vanu ametinimetusi ja sotsiaalsete rollide nimetusi. Teises katses oli keelemudeli ülesanne tuvastada vanu lõunaeestilisi sõnu ja nende põhjal nüüdiskirjakeelseid sõnakujusid. Kolmandas katses oli keelemudeli ülesanne tuvastada vanade laensõnade kirjakuju autorite kaupa.

Kõikide katsete päringute skriptid koos promptidega on nähtaval projekti Githubi repositooriumis (Jürviste jt 2025). Kõigi katsete ja kõigi mudelite puhul valisime temperatuuriparameetrik 0,2, et võimaldada keelemudelil läheneda ülesandele pisut loovamalt kui temperatuuri 0,0 puhul (temperatuurivahemikust – Claude'il 0,0–1,0, Geminil ja GPT-l 0,0–2,0 – sõltub, kui palju kasutab keelemudel oma üldteadmust ja arutlusvõimet), mis omakorda tähendab, et SKM-i vastused võivad katse kordamisel mõnevõrra varieeruda. Promptides antud veebiaadresse ei külastanud ükski keelemudel. Muid koodilisandusi API-päringu Pythoni skriptis etteantud URL-ide külastamiseks ei olnud.

1.2. Materjal

Esimese katse materjal pärineb kuuest allikast: H. Stahli "Vocabulast" (1637), J. Gutsلاffi "Nomenclatorist" (1648), H. Gösekeni "Farragost" (1660), S. H. Vestringi sõnastikust (Vestring 17XX [2000]), A. T. Helle "Vocabulariumist" (1732) ja A. W. Hupeli "Wörterbuchist" (1780). Valimis oli viis ametit või sotsiaalset rolli kirjeldavat nimetust igast sõnastikust, kokku 30 sõna. Teise katse materjal (54 sõna, millest pooled tänapäeva mõistes muredetaustaga sõnad) pärineb Gutsلاffi sõnastikust ja grammatikast. Kolmanda katse materjal (20 laensõna) pärineb Stahli, Gutsلاffi ja Gösekeni sõnastikust.

2. Esimene katse: ametid ja sotsiaalsed rollid

Esimese katse eesmärk oli kontrollida SKM-ide võimekust siduda etteantud sõnastikusisul põhjal vanad sõnakujud nüüdiskujudega. Katsesse valisime ametite ja sotsiaalsete rollide nimetusi, mis erinevad võimalikult palju nüüdiskujudest. Paigutasime tabelisse ametinimetusi ja sotsiaalseid rolle tähistavad vanad sõnakujud, kahel juhul ka nende sünonüümid (*oppeja~opja* ja *Kessineck~Käjemehs*), saksa

vasteid, üksikjuhtudel muid täpsustusi (ladinakeelse kommentaari / murdeala).
 Eraldi tulpades esitasime ka autori nime ja allika leheküljenumbri (vt tabel 1)

Tabel 1. API-päringu lähtematerjal ametinimetuste ja sotsiaalsete rollidega

Vana sõnakuju	Sünonüüm	Saksa vaste v seletus	Täpsustus	Autor
jütteleja		Prediger		Stahl
Kassucksep		Kürßner		Stahl
oppeja	opja	Lehrer		Stahl
Rackel		Hencker		Stahl
sundija		Richter		Stahl
Jeck		Hofnarr		Gutslaff
Kessineck	Käjemehs	Bürge		Gutslaff
Portopähline		Hurer		Gutslaff
Portt		Hure		Gutslaff
Rahhajohataja		Münzmeister		Gutslaff
kachtja		Färber		Göseken
kaffer		Ausgeber		Göseken
Kiwwi Tackar		Stein Metzger	lapidicida	Göseken
Kühna-Mees		Dieb		Göseken
Raamato Rutzoja		Buchdrucker		Göseken
Jooks		Ein Curierer		Vestring
Lämmia		Ein Zauberer		Vestring
Laudamees		Ein unteutscher rechtsfinder		Vestring
Mängimees		Ein Spielmann musicant		Vestring
Pilli ajaja		Ein Sackpfeiffer		Vestring
kolimees		der Schüler		Helle
kulaja		der Kundschaster		Helle
lambrine		der Schäfer		Helle
nikker		der Tischler		Helle
siggur		der Schwein-Hirte		Helle
heikaja		Ausrufer, Herold	d.	Hupel
kalla müja		Fischhändler	r. d.	Hupel
kandijanne		ein Tragender	poetl. r. d.	Hupel
laiwa wallitseja		Steuermann	d.	Hupel
müütnik		der Zöllner	d.	Hupel

SKM-i ülesanne oli lugeda tabelist 1 Pythoni skriptina koostatud API-päringus sisalduva prompti alusel välja kõik read ja näidata, millised vanad sõnakujud vastavad nüüdisõnakujule. Tabeli sisu oli võimalikult lähedane sõnastike sisule, st selles on esitatud sõnastikes sisalduv info keelendite kohta sama lakooniliselt nagu sõnastike skaneeritud algtekstides.

Kahe esimese päringu eesmärk oli katsetada päringu skripti ja prompti toimimist, lastes SKM-il tuvastada sõnadele *kohtunik* ja *varas* vastavad vanad sõnakujud koos allikainfoga. Kõik keelemudelid esitasid mõlemal juhul veatu vastuse, leides esimesel juhul vasteks *sundija* ja teisel juhul *Kühna-meis*. Kahe esimese päringu tulemus näitas, et skript toimib ja promptile vastatakse veatult: seetõttu kordasime katset keerulisema keelendiga. Seekord ei andnud me mudelile ette mitte saksa vastet, vaid saksakeelse selgituse – kuna sellele sõnale ei olegi saksa keeles tõlkevastet – sellisena, nagu ta Vestringi sõnastikus esineb: otsitav sõna oli *laudamees*, mille saksakeelne selgitus on *Ein unteutscher rechtsfinder*. Palusime keelemudelil leida vana sõnakuju sõnale *hirsnik* (sõna *laudamees* kohta tänapäeval kasutusel olev sõna, mida üldkeeles tuntakse siiski vähe). Saksakeelse seletuse põhjal *hirsniku* ja *laudamehe* sidumine osutus ületamatult keeruliseks, kuna ükski keelemudelitest ei teadnud sõna *hirsnik* tähendust: GPT väitis, et *hirsnik* on Helle sõna *nikker* real 22 ja leheküljel 145, Claude ja Gemini nentisid, et sõna *hirsnik* sõnastikufailis ei esine.

Täpsustasime prompti, lastes keelemudelil selgitada, kuidas ta selle tulemuseni jõudis. GPT esitas nn hallutsinatsiooni:

Sõna **hirsnik** esineb [Hupel] real [28] kujul **müütnik** (leheküljel [219]). Selgitus: Otsisin sõnastikufailist sõna **hirsnik** esinemisvorme. Kuna sõna **hirsnik** viitab ajalooliselt tolliametnikule või maksukogujale, siis otsisin sõnastikufailist sõnu, mis võiksid tähendada sama ametit. Leidsin rea 28, kus esineb sõna **müütnik**, mis tähendab **der Zöllner** ehk tolliametnikku.

Ka Claude sidus *hirsniku* sõnaga *müütnik*:

Sõna **hirsnik** esineb Hupel real 28 kujul *müütnik* (leheküljel 219). Tulemuseni jõudsin järgmiselt: Otsisin sõnastikufailist sisestatud sõna “*hirsnik*” täpset vastet veerust “ee van”. Leidsin, et real 28 esineb sõna “*müütnik*”, mis on sisestatud sõna vana esinemiskuju. Vaatasin sama rea veerust “autor”, et tegemist on Hupeli sõnastikuga. [---]

Gemini sidus aga *hirsniku* sõnaga *sigur*:

Sõna **hirsnik** esineb Helle real 24 kujul *siggur* (leheküljel 176). Selle tulemuseni jõudmiseks ma analüüsisin sisestatud teksti ja otsisin sealt sõna “*hirsnik*” vastet. Ma leidsin, et see vaste on real 24, kus “*siggur*” on märgitud “*der Schwein-Hirte*” sünonüümina. “*Schwein*” tähendab siga ja “*Hirte*” karjast. Seega on “*siggur*” vana eesti keeles “*hirsnik*”.

Kõik keelemudelid andsid täpse vastuse alles siis, kui lisasime prompti selgituse, et sõna *hirsnik* sünonüüm on *õigusemõistja*. Kõige täpsema vastuse andis GPT:

Sisestatud sõna oli **hirsnik**, mille sünonüüm on **õiguseleidja**. Otsisin sõnastikufailist sõna, mis tähendaks midagi sarnast. Real 17 leidsin sõna **Laudamees**, mille saksa vaste on “*Ein unteutscher rechtsfinder*”, mis

tõlkes tähendab “mitte-saksa õiguseleidja”. Kuna see vastab sisestatud sõna tähendusele, järeldasin, et *Laudamees* on otsitava sõna vana esinemiskuju.

Kuna ülalkirjeldatud esimesed katsed andsid häid tulemusi, jätkasime loendikatega. Andsime keelemudelitele ette loetelu nüüdissõnadest, millele soovisime leida esinemusi vanades sõnastikes: sõnastikusisu esitasime juba eelnevalt kasutatud tabelina (vt tabel 1). Loetelu sõnadeks olid: *heerold; hirsnik, õiguseleidja; hoor; järelevaataja, valvaja; jutlustaja; käemees, käendaja; kalamüüja; kandja; käskjalg, kuller; kiviraidur; kohtunik; köösner; kupeldaja; lambur; mündivermija; narr; nõid; õpetaja; õpilane; pillimees; raamatutrükkal; seakarjus; spioon; timukas; tisler; tolliametnik, tõlner; torupillimängija; tüürimees; varas; värval, värvija*. Loetelus oli seega 30 rida (5 sõna igast allikast Stahlist (1637) kuni Hupelini (1780)), 6 sõnale oli antud ka sünonüüm, et keelemudel sõna paremini mõistaks. Promptis palusime keelemudelil siduda tekstifailis loeteluna esitatud sõnad sõnastikufaili tabeli ridadega. Tulemuse palusime väljastada tekstiformaadis. Keelemudelite vastuseid hindame tabelis 2 (tabelite värvilised versioonid on avaldatud artikli veebiversioonis).

Tabel 2. Keelemudelite vastuste tulemused vanade ametinimetusi ja sotsiaalseid rolle tähistavate sõnade tuvastamisel

Vana sõnakuju	Nüüdskuju	Claude	GPT	Gemini
heikaja	heerold	0	1	1
Jeck	narr	1	1	1
Jooks	käskjalg, kuller	1	0	0
jütteleja	jutlustaja	1	0	1
kachtja	värval, värvija	1	1	1
kaffer	järelevaataja, valvaja	0	0	0
kalla müja	kalamüüja	1	1	1
kandijanne	kandja	1	1	1
Kassucksep	köösner	1	1	0
Kessineck, Käjemehs	käemees, käendaja	1	1	1
Kiwwi Tackar	kiviraidur	1	0	1
kolimees	õpilane	1	1	1
kulaja	spioon	1	1	0
Kühna-Mees	varas	1	1	1
laiwa wallitseja	tüürimees	1	1	0
lambrine	lambur	1	1	1
Laudamees	hirsnik, õiguseleidja	1	1	1
Lämmia	nõid	1	1	0
Mängimees	pillimees	1 + pilli ajaja	1	0
müütnik	tolliametnik, tõlner	1	1	1
nikker	tisler	1	1	1
oppeja, opja	õpetaja	1	1	1

Vana sõnakuju	Nüüdiskuju	Claude	GPT	Gemini
Pilli ajaja	torupillimängija	1	1	1
Portopählne	kupeldaja	0	0	0
Portt	hoor	1	1	1
Raamato Rutzoja	raamatutrükkal	1	1	1
Rackel	timukas	1+ Hencker	0	0
Rahhajohataja	mündivermija	1	1	1
siggur	seakarjus	1	1	0
sundija	kohtunik	1	1	1
VALESID		3	6	10
ÕIGEID		27	24	20
ÕIGETE OSAKAAL		90%	80%	66,67%

Kõige paremini lahendas ülesande Claude, mille veamäär oli 10%, GPT veamäär oli 20% ja Geminil 33,33%. Sõna *kaffer* ‘järelevaataja, valvaja’ tuvastamisega ei tulnud toime mitte ükski keelemudel. Claude arvas end mitte leidvat ka sõnu *heerold* ja *portopählne*, ent on ometi esitanud mõlema sõna puhul õige vastuse eraldi kommentaaris. Ainsad otseselt kaks vale seost lõi Gemini, öeldes, et *rackel* ehk ‘timukas’ esineb kujul *lämmia* ja et *kaffer* esineb kujul *kulaja*. Claude lisas õigetele vastustele ka seosed, et sõna *rackel* sünonüüm on eesti keeles *Hencker* ning et sõna *mängimees* sünonüüm on *pilli ajaja*. Kui arvestada kaks “mitte leitud, kuid siiski leitud” (st eraldi kommentaaris õigesti esitatud) sõna Claude’i puhul õigeteks vastusteks, võiksime tema veamääraks lugeda 3,33% (üksainus vale vastus), ehkki väärseoseid *rackel* – *Hencker* ja *mängimees* – *pilli ajaja* ei saa seejuures arvestamata jätta. GPT ega Gemini ei kommenteerinud neid sõnu, millele nad vasteid ei leidnud, küll aga tegi seda Claude (prompt ei sisaldanud kommentaarinõuet ühelgi juhul). GPT jättis mainimata 6 rida, aga esitatud vastustes ei eksinud mitte ühegagi, Gemini jättis mainimata 8 rida ja vastas valesti 2 sõnapaari.

3. Teine katse: lõunaeesti taustaga sõnad

Teises katses püüdsime SKM-i abiga leida Johannes Gutsloffi lõunaeestikeelse grammatika (Gutsloff 1648) sõnu etteantud tabelist tänapäeva kirjaviisis sõnade järgi ja vastupidi – leida Gutsloffi teoses esinevatele sõnadele vasted põhjaeesti murdel põhinevas tänapäeva kirjakeeles.

Esimese osakatse aluseks oli tabel, kuhu valisime Gutsloffi sõnade tänapäeva vasteid koos tema antud saksa- ja mõnel juhul ka ladinakeelse tõlkega (tabel 3). 54 sõna ja nende 63 tänapäeva vaste hulgas oli nii kokkulangevaid sõnu (nt *kirp*, *puhas*, *raha*), vähese kirjaviisist tuleneva erinevusega (nt *haud*, *meri*, *panema*) kui täiesti erinevaid (nt *ramato nachk* ‘paber’, *teww* ‘kops’, *rachk* ‘hallitus’). Kui tänapäeva kirjakeeles oli kaks sünonüümset (sageli lõunaeesti päritolu sünonüüm) vastet, anti mõlemad, nt *sissas* ‘ööbik, sisask’, *penni* ‘koer, peni’, *sössar* ‘sõsar, õde’, *siuvck* ‘siug, uss’. Sünonüümi ei antud, kui see ÜS 2024-s puudus või selle andmetel ei olnud kirjakeelne (nt *kaema*, *väits*) või oli vananenud (*häbedus*) – need sünonüümid on tabelis sulgudes (sõna *rääbis* kohta vaata allpool).

Tabel 3. Gutsblaffi sõnadega tehtud katse lähtematerjal

Gutsblaffi kuju	Gutsblaffi saksa v ladina vaste	Kirjakeelne nüüdiskuju
haggo	Morgenröthe	agu, koit
Haugk	Hecht	haug
Hawd	Grab	haud
hebbedus	Scham	suguelundid (häbedus)
helle	Klang, Laute	heli
hemme	Nass	märg
hiuws	Haar (crinis)	juus, juuksekarv
hailmo	florimentum	õisik
iggas	semper	alati
issack	Stieff-Vater	kasuisa
kajema	schawen, versuchen	vaatama (kaema)
kattessas	–	kaheksas
Kattus	Dach	katus
Kirp	Floh	kirp
kôhes	ubi	kus
Kübbar	Hut	kübar
laisk	faul	laisk
lobbisema	schnattern	lobisema
loti	Kugel	kuul
lubbiraya	Aussatz	pidalitõbi; lubiraig
mas	forte, forsan	ehk, võibolla
Merri	Meer	meri
nagl	Nagel	nael
nanne	Nase	nina
neutzick	Magd	näitsik
ombelma	neen	õmblema
pannema	legen	panema
penni	Hund	peni, koer
perrast	nach	pärast
piddama	müssen	pidama
pitk	lang	pikk
puhas	rein, sauber	puhas
pärn	Miltze	põrn
pörss	Arsch	perse
rachk	Kahm	hallitus
Raha	Geld	raha
ramato nachk	Papier	paber
Rasw	Fett	rasv
rawdreuwas	Harnisch	raudrüü (raudrõivas)

Gutslaffi kuju	Gutslaffi saksa v ladina vaste	Kirjakeelne nüüdiskuju
rebes	Murene piscis	mureen (räabis)
saddama	schneyen	sadama
seng	Bett	säng, voodi
Sigga	Sau, Schwein	sig
sissas	Nachtigall	sisask, ööbik
siwck	Schlange	siug, uss
Söma	Essen	sööma
sössar	Schwester	õde, sõsar
Tarck	Klug	tark
teww	Lunge	kops
wagga	fromm	vaga
wannuck	Eid, Schwur	vanne
weitz	Messer	nuga (väits)
wennick	Russa	venelanna
Öh	Nacht	öö

Tabel 4. Teise katse esimese osa tulemused, kus tänapäeva sõnadele tuli leida vasted Gutslaffi sõnastiku sõnade seast. Paksus kirjas on esitatud sünonüümsed sõnad

Sõna	Claude	GPT	Gemini
agu	0	0	0
koit	0	0	1
alati	0	0	1
ehk	0	0	1
võib-olla	1	0	1
hallitus	0	0	0
haud	1	1	1
haug	1	1	1
heli	1	0	0
juuksekarv	1	0	0
juus	1	1	1
kaheksas	1	0	0
kasuisa	1	1	1
katus	1	1	1
kirp	1	1	1
koer	1	0	1
peni	1	0	1
kops	0	0	0
kus	1	0	0

Sõna	Claude	GPT	Gemini
kuul	0	0	1
kübar	1	1	1
laisk	1	1	1
lobisema	1	1	1
lubiraig	1	1	0
pidalitõbi	1	1	0
meri	1	1	1
mureen	0	0	1
märg	0	0	1
nael	1	0	1
nina	1	1	1
nuga	0	0	1
näitsik	1	1	0
paber	1	1	1
panema	1	1	1
perse	1	1	1
pidama	1	1	1
pikk	1	1	1
puhas	1	1	1
pärast	1	1	1
põrn	1	0	1
raha	1	1	1
rasv	1	1	1
raudrüü	1	1	1
sadama	1	0	1
sig	1	1	1
sisask	1	0	0
õõbik	1	1	1
siug	0	1	0
uss	0	0	1
suguelundid		0	0
säng	0	0	1
voodi	0	0	1
sõsar	1	1	1
õde	0	1	1
sööma	1	1	1
tark	1	0	1

Sõna	Claude	GPT	Gemini
vaatama	1	0	0
vaga	1	1	1
vanne	1	1	1
venelanna	0	1	1
õisik	1	0	0
õmblema	1	0	1
öö	1	0	1
VALESID	16	30	15
ÕIGEID	47	33	48
ÕIGETE OSAKAAL	74,6%	52,4%	76,2%

Kolme keelemudeli antud vastuseid Gutsloffi sõnastiku sõnadele hindame tabelis 4. Kõik kolm mudelit leidsid õige vaste 27 sõnale 63-st (42,2%). Need olid valdavalt sõnad, mille tänapäeva vaste ja kirja pilt on sama või lähedane, nt *sõma* 'sööma', *piddama* 'pidama', *perrast* 'päras' ja *pitk* 'pikk'. Kirjakujult erinevamate vastetest leidsid kõik õige vaste sõnadele *paber*, *raudrüü* ja *kasuisa* (vastavalt *ramato nachk*, *rawdreuwas*, *issack*). Sõnu, mille vastega eksisid kaks mudelit, oli 19 (29,7%) ja sõnu, millega eksis üks mudel, oli 14. Kõik mudelid eksisid vaid kolme 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, kaheksal juhul Claude ja vaid ühel juhul GPT. Mudelitest edukaim oli Gemini 48 õige vastusega (veamäär 23,8%), kuid Claude jäi minimaalse vahega teiseks (47 õiget vastust, veamäär 25,4%), selgelt nõrgema tulemuse andis GPT 33 õige vastusega (veamäär 47,6%). Sõna *kattessas*, millele polnud üldse vastet antud, andis õige vastena sõnale 'kaheksas' ainsana Claude. Kui võrrelda kirjakeele tavasõnadele ja nende lõunaestilistele sünonüümidele õige vaste leidmise edukust, siis õigeid tulemusi oli järgnevalt:

- sisask 1, ööbik 3
- säng 1, voodi 1
- peni 2, koer 2
- agu 0, koit 1
- sõsar 3, õde 2

Seega anti muredetaustaga sõnale õige vaste seitsmel juhul ja tavalisele sõnale 9 juhul. Muude sünonüümide puhul oli õigeid vastuseid järgnevalt: *juus* 3, *juuksekarv* 1; *ehk* 1, *võib-olla* 2.

Teise osakatse aluseks oli tabel, mis sisaldas Gutsloffi sõnastiku sõnu nende algsel kujul koos seal antud saksa- ja mõnel juhul ka ladinakeelse tõlkega. Ülesandeks oli leida Gutsloffi sõnadele vasted tänapäeva eesti kirjakeeles (vt tabel 5). Lisaallikaks andsime "Väikese murdesõnaraamatu" veebivariandi (VMS). Gutsloffi sõnu oli 54, nende seas peamiselt grammatika tekstiosast võetud kuus ladina vastega ja üks ilma vasteta (*kattessas* 'kaheksas') sõna. Mudel võis anda ka rohkem kui ühe vaste. Seda võimalust kasutas GPT 9 korral, Claude 13 korral (veel kahele sõnale lisas pikema kommentaari) ning Gemini koguni 42 korral. Kuigi enamasti

¹ Sõna tänapäeva vaste saksa sõna *Kahm* tõlke alusel (Kibbermann jt 1975). Täpsemalt tähendab see sõna ilmselt käärimisel tekkivat vahu- või hallituskihti vedelikul.

oli mitmesõnalise vaste puhul tegu sünonüümidega (nt *märg*, *niiske*, *võõrasisa*, *kasuisa*, *rasv*, *searasv*, *pekk*), siis esines ka ekslikke sõnavorme (nt *vannete*, *töötus*, *laisav*, *laiseldav*), teatud tunnuse alusel sarnase tähendusega sõnu (nt *nael*, *naaskel*, *knopka*; *sig*, *põrsas*, *orikas*; *luts*, *latik*, *särg*) või täiesti erineva tähendusega sõnu (nt *peni*, *penn*; *kübar*, *müts*, *onn*). Sõna *pörss* (sks *Arsch*) puhul tundus Gemini vältivat ebatsensuurse sõna esitamist, andes vastuseks:

Kuigi saksa/ladina vaste viitab teatud kehaosale, ei ole VMSi andmetes ega minu teadmistes eesti keele ajaloost mingeid tõendeid selle kohta, et sõnal “pörss” oleks kunagi selline tähendus olnud. Ilma lisateabeta või kontekstita on võimatu pakkuda usaldusväärseid vasteid.

Tulemuste tõlgendamisel lugesime õigeks need vastused, kus esimene vaste oli tähenduselt õige ja ei oldud pakutud ühtegi täiesti vale vastet, nt sõnale *loti* (sks *Kugel*) antud vaste *lodi*, *loot*, *kuul* lugesime valeks, sest esimesed kaks sõna on valed² ja vaid *kuul* õige, aga *siuck* (sks *Schlange*) vaste *madu*, *uss*, *rästik* õigeks, sest kuigi *rästik* on *mao* hüponüüm, ei ole eksimus suur. Samuti lugesime valeks vasted, kus oli eksitud sõnaliigiga (*söök*, *toit* pro *sööma*) või sooga (*venelane*, *Vene*, *Venemaa* pro *venelanna*, *vene naine*) või oli vaste vormiliselt ebatäpne, nt *neitsi* pro *näitsik*. Valeks lugesime ka kirjakeele sõnaraamatutes (ÜS 2024, EKSS 2009) puuduvad vasted, nt *rahk* või teises tähenduses sõnad, nt *raamatunahk* on tänapäeva erialakeeles tähenduses ‘pärgament’ (RaS). Teises osakatses leidsid 54 Gutsblaffi sõnale õige tänapäevakeele vaste(d) kõik kolm mudelit 26 juhul (48,15%). Kaks mudelit eksisid 5 sõna puhul, neist ainsa õige vaste andis kolmel juhul GPT ja kahel juhul Claude. Üks mudel eksis 16 sõna puhul, neist 10 puhul Gemini ja kuue puhul GPT. Kõik mudelid eksisid 7 sõna puhul, mida vaatleme lähemalt. Sõna *hebbedus* (sks *Scham*) puhul pakuti vasteks *häbi*, mis on ka saksa sõna põhitähenduseks, kuid ootasime vastet *häbedus* või *suguelundid*. Sõna *häbedus* leidub ka ÜS 2024-s, kuigi märgendiga VANANENUD. Sõna *hailmo*, mille vasteks andsime ladina *florimentum* (sõnastikuosas on vaste *Blüete*) vasteteks pakuti *ilu* / *õilmik*, *õis*, *õieke* / ‘kaunistus, ehtimine, iluasi’. Kuna ladina sõna on sõnast *flos* ‘õis’ tuletis kogunimisõnu moodustava suffiksiga *-mentum*, võiks seda tõlkida kui *õisik*. Sel juhul võiks sõna *õilmik* (mis on kasutusel lehmanimena) isegi õigeks lugeda, aga sellist sõna ei leidu kirjakeele sõnaraamatutes. Sõnal *neutzick* (sks *Magd*) on tänapäevakeeles täpne vaste *näitsik*, mida ükski mudel ei pakkunud, selle asemel pakuti *neitsi*; *teenijatüdruk*, *teenijanna*; *kasulik*, *vajalik*, *tarvilik*. Õiget vastet ei antud ka sõnale *rachk* (sks *Kahm*), kuigi Claude pakkus üsna õige tähenduse (*rahk*)³. Samuti jäid mudelid hätta sõnale *rebes* (ld *Murene piscis*) vaste leidmisel. Kui GPT pakkus *räim* ja Gemini pakkus *luts*, *latik*, *särg*, siis Claude andis oma vastetele *angerjas* veel järgmise põhjenduse: “Seda võib järeldada Gutsblaffi sõnale antud ladina vastest “Murene piscis”, mis tähendab angerjat.” Angerjat see sõna ei tähista, aga nutikas on see vastus küll, sest mureenid Eesti vetes ei ela ja angerjas on kujult mureeni sarnane. Õige vaste olnuks *rääbis*, kuid katses andsime vasteks ladina sõna otsetõlke. Tõenäoliselt on Gutsblaffi *rebes* ka *rääbise* esmaesinemus kirjakeeles. Gutsblaff pidas vastega *Murene piscis* ilmselt silmas kala, mis Grimmide sõnaraamatus leidub kujul *maräne* ‘fischname: die grosze maräne, salmo maraena, die kleine maräne, salmo maraenula’ (DWDS: s. v. *maräne*). *Salmo maraena* on

² Sõna *lood*, *luut* esineb lõunaeesti murrakutes (*Rõu Plv Vas Se Lut*) tähenduses ‘kuul’ (EMS).

³ Claude'i selgitus: “Sõna ‘rahk’ tähendab tänapäeval hallitust või kodu, mis tekib niiskes keskkonnas toiduainete või esemete pinnale. Saksa vaste ‘Kahm’ viitab samuti hallitusele.”

tänapäeva ladina nimega *Coregonus maraena* – siirdesiig, kes *pro* mis on samast perekonnast rääbisega (*Coregonus albula*). Ükski mudel ei suutnud täpset vastet anda ka sõnale *Wennick* ‘Russa’ (‘venelanna v vene naine’), mis tänapäeva keeles sellisel kujul (**venik*) ei esine. Seda on varem peetud Gutsloffi omaloominguliseks tuletiseks (Keem 1998: 325), kuid tõenäoliselt oli see siiski tolleaegses rahvakeeles levinud (Mänd, Tóth 2025). Kui valida oli põhja- ja lõunaeeesti päritolu vaste vahel, eelistasid mudelid põhjaeestilist (nt *agu* asemel *koit*, *sõsar* asemel *õde*).

Kuuest sõnast, millele oli antud vaid ladina vaste, leidsid kõik mudelid kolmele õige vaste ja kolmele ei leidnud ükski mudel õiget vastet. Katses oli selgelt edukaim Claude 44 õige vastega (veamäär 18,5%), teiseks tuli GPT 39 õige vastega (veamäär 27,8%) ja kolmandaks Gemini 33 õige vastega (veamäär 38,9%).

Tabel 5. Teise katse teise osakatse tulemused, kus Gutsloffi sõnastiku sõnadele tuli tänapäeva kirjakeele vasted leida. Paksus kirjas on esitatud sõnad, millele oli antud vaid ladinakeelne vaste

Gutsloffi kuju	Saksa või ladina vaste	Claude	GPT	Gemini
haggo	Morgenröthe	1	1	1
Haugk	Hecht	1	1	1
Hawd	Grab	1	1	0
hebbedus	Scham	0	0	0
helle	Klang, Laute	1	1	0
hemme	Nass	1	1	0
hiuws	Haar (crinis)	1	1	1
hällmo	florimentum	0	0	0
iggas	semper	1	1	1
issack	Stieff-Vater	1	0	1
kajema	schawen, versuchen	1	0	1
kattessas	–	0	0	0
Kattus	Dach	1	1	1
Kirp	Floh	1	1	1
kôhes	ubi	1	1	1
Kübbbar	Hut	1	1	0
laisk	faul	1	0	1
lobbisema	schnattern	1	0	1
loti	Kugel	0	1	0
lubbiraya	Aussatz	1	1	1
mas	forte, forsan	1	1	1
Merri	Meer	1	1	1
nagl	Nagel	1	1	1
nanne	Nase	1	1	1
neutzick	Magd	0	0	0
ombelma	neen	0	1	0
pannema	legen	1	1	0
penni	Hund	1	1	0

Gutslaffi kuju	Saksa või ladina vaste	Claude	GPT	Gemini
perrast	nach	1	1	1
piddama	müssen	1	1	1
pitk	lang	1	0	1
puhas	rein, sauber	1	1	1
pärn	Miltze	1	0	0
pörss	Arsch	1	1	0
rachk	Kahm	0	0	0
Raha	Geld	1	1	1
ramato nachk	Papier	0	1	0
Rasw	Fett	1	1	1
rawdreuwas	Harnisch	1	0	1
rebes	Murene piscis	0	0	0
saddama	schneyen	1	1	1
seng	Bett	1	1	1
Sigga	Sau, Schwein	1	1	1
sissas	Nachtigall	1	1	1
siwck	Schlange	1	1	1
Söma	Essen	1	1	0
sössar	Schwester	1	1	1
Tarck	Klug	1	1	1
teww	Lunge	1	1	0
wagga	fromm	1	1	1
wannuck	Eid, Schwur	1	0	0
weitz	Messer	1	1	1
wennick	Russa	0	0	0
Öh	Nacht	1	1	1
VALESID		10	15	21
ÕIGEID		44	39	33
ÕIGETE OSAKAAL		81,48%	72,22%	61,11%

4. Kolmas katse: laensõnad

Kolmanda katse eesmärk oli selgitada, kas ja kuivõrd on SKM-id võimelised siduma kolmes sõnastikus (Stahl 1637, Gutslaff 1648, Göseken 1660) esinevaid vanu kirjakujuksid omavahel ja neile vastavate nüüdiskujudega, nt *mijrrit* (Göseken), *mirra* (Gutslaff) ja nüüdiskuju *mürr* (Stahlil see sõna puudub). Katseid tegime laensõnadega, millest enamus esines juba piiblitõlgetes. Saksa kuju me ei andnud, kuna laenude kujud on üldjuhul suhteliselt sarnased tänapäeva kirjakujuks, st sõnakujudes on üldjuhul palju kokkulangevaid tähti. Ka oli saksa vastete kaudu tuvastamist katsetatud juba esimeses katses.

Sisendmaterjali esitasime neljaveerulise Exceli tabeli kujul allikate kaupa: 1. tänapäeva eesti keele kirjakujud; 2. Stahli sõnaraamatu kirjakujud; 3. Gutsloff'i sõnaraamatu kirjakujud; 4. Gösekeni sõnaraamatu kirjakujud (vt tabel 6). Materjal oli paigutatud veergudesse randomiseeritult, et mudelid ei seoks sõnakujusid lihtsalt ridade kaupa. Uuritavad sõnad olid 20 tänapäevast kirjakuju, st igal real esines üks kirjakuju, samatähenduslikud olid kaks rida – *ahv* ja *pärdik*. Ajaloolist sõnakujude veerud olid järgnevad: Stahli sõnastikus esines 20-st nimetatud tänapäevasest sõnakujust vasteid 14-le, neist kolmel olid esitatud ka sünonüümid. Gösekeni sõnastikust võtsime 20 sõna, millest seitsmel olid ka sünonüümid esitatud. Gutsloff'i sõnastikust võtsime 18 sõna, ühel sõnal ka sünonüümid. Sõnade hulk erineb mõnevõrra, kuna osas sõnastikes puudus osa vasteid, nt pole Stahli vasteid sõnadele *ahv*, *mürr*, *penning*, *viiruk*, *väävel*, Gutsloff'il puuduvad vasted sõnadele *loorber* ja *palsam*; Gösekenil on vasted kõigile sõnadele. Ühel juhul on võetud vaste väljastpoolt sõnastikku: Stahli on antud *palsami* vasteks *üx kaunis heh haisija rocht* (Stahl 1649: 601). Sõnad on valitud nii, et valimis leiduks minimaalse erinevusega kirjakujusid, nt *pipar*, *pippar* ja *pipper*, mõnevõrra erinevaid, kuid siiski kokkulangevate tähtedega, nt *äädikas* ja *Ettickas*, ent ka täiesti erinevaid, nt *väävel* ja *koirasitt*, *pennisitt*.

Tabel 6. Sisendandmed keelemudelitele randomiseeritult

Nüüdiskuju	Stahl	Gutsloff	Göseken
kuningas	Lohrbehr	Kurrat	Picker
paasapüha	Valgkemees, Lendwamees, Püssohand	Pipper	Pertick
trööst	Pippar	Engel	Püssihand, Valgkemees
väävel	Kuningas		Rohst, Rööm
mürr	Ettickas	Juddalinne	Kurrat, Pahharet
loorber	Engel	Pertike	Killing
häärber	Kurrat	Kunningas	Kallis hais, Rocht
pipar	Üx kaunis heh haisija rocht	Maja	Pertick
viiruk		Tröst, Röhm	Ingel
raad		Pennisitt	Loorbeer
ingel			KoiraSitt, Wewel
peeker		Ettikas	Wyrohki, Wijroht
palsam	Karrick	Püssohand	Rahdi
penning		Sauwrocht	Ohstrit, Pahscha põha, Leehawotmepõha
juut	Juddalinne	Pertike	Ettickas
kurat	Rahd	Raht	Pippar
äädikas		Karritz	Kunningas
ahv	Leehawotme Põha, Ohstrit	Mirra	Vöhras kodda
draakon	Öhmaja	Lehawötta põha	Mijrrit
pärdik	Röhm, Trohst	Teng	Judalinne

Kuna eelkatse GPT-ga näitas, et see püüab iga hinna eest sõnakujudele kirjakuult kõige sarnasemat vastet leida, korrigeerisime prompti, lisades, et sünonüümid on komadega eraldatud ja et vastuses tuleb esitada kõik sünonüümid ning kui vastust ei leidu, tuleb öelda, et vastust ei leidu, lisades selle kohta selgituse.

Päringu skript oli kirjutatud nii, et mudel sai ühekaupa sisendiks sõnu nüüdis-sõnade seast, ja siis otsis igast sõnastikufailist ainult sellele sõnale vaste ning jätkas seejärel järgmise sõnaga. Kuna keelemudelid API-päringute puhul eelnevat suhtlust mälu ei talleta, ei võtnud SKM arvesse juba tehtud päringutele antud vastuseid. Katse tulemusi hindame tabelis 7.

Tabel 7. Kolmanda katse tulemused, kus SKM-i ülesanne oli seostada tänapäeva kirjakeele vasted vanade sõnakujudega. Tabelis tähistab x märksõna puudumist: x valgel taustal – keelemudel ei pakkunud vastust, x punasel taustal – keelemudel pakkus sõna, mida antud sõnastikus ei olnud; 0 punasel taustal – keelemudel ei paku vastust või pakub (vähemalt osaliselt) vale vastuse

Nüüdiskuju	Claude			GPT			Gemini		
	Stahl	Gutslaff	Göseken	Stahl	Gutslaff	Göseken	Stahl	Gutslaff	Göseken
ahv	x	0	0	x	0	0	x	0	0
draakon	1	1	1	0	0	0	0	0	0
häärber	0	0	1	0	1	1	0	1	1
ingel	1	1	1	1	1	1	1	1	1
juut	1	1	1	1	1	1	1	1	1
kuningas	1	1	1	1	1	1	1	1	1
kurat	1	1	1	1	1	1	1	1	1
loorber	1	x	1	0	x	1	0	x	1
mürr	x	1	1	x	1	0	x	1	1
paasapüha	1	1	1	1	1	1	1	1	1
palsam	1	x	1	1	x	1	0	x	0
peeker	0	0	1	0	0	1	0	0	0
penning	x	0	0	x	0	0	x	0	0
pipar	1	1	1	1	1	1	1	1	1
pärdik	x	0	1	x	1	1	x	1	1
raad	1	1	1	1	1	1	1	1	1
trööst	1	1	1	1	1	1	1	1	1
viiruk	x	1	1	x	1	0	x	0	0
väävel	x	1	1	x	0	1	x	0	1
äädikas	1	1	1	1	1	1	1	1	1
ÕIGE	12	13	18	10	13	15	9	12	14
VALE	4	6	2	8	6	5	7	7	6
VASTE PUUDUB	4	1	0	2	1	0	4	1	0
Õigeid	60%	65%	90%	50%	65%	75%	45%	60%	70%
Valesid	20%	30%	10%	40%	30%	25%	35%	35%	30%
Puudu	20%	5%	0	10%	5%	0	20%	5%	0

Sõnadest, mille tänapäeva kirjakuju on sama või lähedane vanades sõnastikes esinenud kirjakujudega, viisid veatult kokku kõik mudelid: *kuningas* (*kuningas*, *kunningas*), *pipar* (*pippar*, *pipper*), *ingel* (*engel*, *ingel*), *kurat* (*kurrat*, *pahharet*). Tänapäeva kirjakujudest mõnevõrra erinevad vanad kujud, nagu *juut* (*juddalinne*, *judalinne*), *äädikas* (*ettickas*, *ettikas*) viisid samuti ilma ühegi veata kokku kõik kolm mudelit. Ka sõnale *raad* vastete (*rahd*, *rahdi*, *raht*) ega sõnale *trööst* vastete (*röhm*, *trohst*, *rohst*, *rööm*, *trôst*) leidmisel ei eksinud ükski mudel.

Teistest sõnadest, mille tänapäeva kirjakujudes on väiksemaid erinevusi võrreldes vanade kujudega, olid eri mudelite võimekused erinevad: nt ei eksinud Claude *loorberi* (*loorbeer*, *lohrbehr*) puhul, seevastu eksisid GPT ja Gemini. GPT pakkus vasteks *üx kaunis heh haisija rocht*. Selgituste võrdlusest võib järeldada, et GPT püüab tähendusi analüüsida ja nende põhjal järeldusi teha, aga ka seda, et Gemini “kuulas paremini sõna” ega pakkunud vastust, mida põhjendada ei suutnud.

Veel nüüdiskujudele suhteliselt sarnastest ajaloolistest kirjakujudest on sõnale *pärdik* andnud Claude põhimõtteliselt küll õiged vasted *pertick* ja *pertike*, eksinud aga allikate tuvastamises. Seetõttu oleme hinnanud Claude'i kolmest vastustest kaks valeks. Seega suudab Claude üldjuhul (60–90% juhtudest, vt tabel 7) viia küll sarnased sõnakujud kokku, ent pole mõnel juhul usaldusväärne allika esitamisel. Gemini ja GPT vasted *pärdikule* on õiged.

Osaliselt erinevatest kirjakujudest on kõik mudelid mõnevõrra üllatuslikult suutnud vasted leida sõnale *paasapüha* (*leehawotme põha*, *ohstrit*, *pahscha põha*, *leehawotmepõha*, *lehawötta põha*). Claude'i selgitus *paasapüha* vastete tuvastamise kohta ütleb peamiselt seda, et sõnakujud on kirja pildilt sarnased.

Kirjakujult täiesti erinevatest sõnadest andis Claude *draakonile* kõigi sõnastike puhul õiged vastused, selgitades, et “*püssihand* ja selle variandid viitavad lendamisele nagu ka draakon” ja et *valgemees* viitab draakoni “välgutaolisele välimusele”.

Ka *väävli* ja selle mõlemad vasted (Gutslaffi ja Gösekeni vastavalt *koirasitt*, *wewel* ja *pennisitt*⁴) suutis kokku viia ainult Claude. Lähteandmetes Stahlil vaste puudus ja sellest said aru kõik mudelid, sellised vastused ei kajastu tulemustes ei õigete ega valedena. Gemini annab ühe õige vaste *koirasitt*, *wewel* (Göseken), Gutslaffi *pennisitt* on jäetud esitamata. Claude esitas selgituse, mis lubab oletada, et ta on esmalt *väävli* vastena tuvastanud *weweli* ja seejärel *weweli* sünonüümina esitatud sõna *koirasitt* kaudu leidnud ka vaste *pennisitt*. Seega on Claude võimeline rohkem seostama ja järeldama kui teised mudelid.

Claude'i kolmest vastest *palsami* kohta (*üx kaunis heh haisija rocht*; *kallis hais*; *rocht*) oli kaks õiget (ekslikult lisas vasteks ka *sauwrocht*). Lugesime õigeks lähteandmetele vastavad vastused, sisulist poolt arvestamata. Pooleldi valed vastused lugesime valeks. Selle katse tulemuste tõlgendamine sõltub palju sellest, mida lugeda õigeks vastuseks. Samad küsimused kerkivad ka *viiruki* ja *mürriga*. Osa Claude'i pakutud *viiruki* vasteid (*wyrohki*, *wijroht* ja *sauwrocht*) olid õiged, osa lugesime valeks, ehkki sisuliselt on nad võimalikud ja loogilised. *Mürrile* annavad Claude ja Gemini õiged vasted (*mijrrit*; *mirra*), GPT antud vastetest on esimene vale (*üx kaunis heh haisija rocht*), teine pooleldi vale (*kallis hais*; *rocht*; *mijrrit*) ja viimane õige (*mirra*).

*Häärberile*⁵ esitas Claude Gösekeni vaste (*vöhras kodda*) ja jättis esitamata vasted *öhmaja* (Stahl) ja *maja* (Gutslaff). Gemini esitatud *vöhras kodda* ja *maja*

⁴ Originaalis oli *n* pikenduskrüipsuga, lähtematerjalis *nn*.

⁵ *Häärberi* tähendus on eesti keeles muutunud aja jooksul. Göseken annab *Herberge* kahes eri tähenduses. Esimesel on täpsustus *Losament* – see tähendab ilmselt eluaset üldiselt. Näidetes kasutab Göseken oma sõnastikus aga vastet *Öhmaja*. Vasteks pidanuks võtma kuju *Herpergi*, mitte kuju *Wöhras Kodda*.

on õiged, Claude, nagu ka teised kaks mudelit, ei esita *öhmaja* (Stahl). GPT esitab vasteks *juddalinne*, mida võib pidada hallutsineerimiseks.

Loorberile Claude'i antud vasted *lohrbehr* (Stahlil), *loorbeer* (Gösekenil) olid õiged, nagu ka see, et Gutsblaffil see sõna puudus. Gemini tundis ära vaste *loorbeer*, mis on lähedasem tänapäeva kujule, jättes esitamata kaugema kuju *lohrbehr*. GPT pakub *loorberi* vaelevasteks *üx kaunis heh haisija rocht*, tuues selgituses esile *loorberi* iseloomuliku tunnuse – lõhna.

Peekrile andis Claude vaelevaste *pertike*, selgitus näitab, et selle alus on häälikuline sarnasus. Stahli vaste *karrick* ja Gutsblaffi vaste *karritz* jättis Claude esitamata (nagu ka teised mudelid), Gösekeni vaste *picker* esitas õigesti. Arvestades, et *peeker* ja *karikas* on eri sõnad, ei ole eksimine eriti üllatav. Samas ilmnes siit, et osa sõnade puhul mudelil andmed puuduvad (näiteks see, et *karikas* ja *peeker* on tänapäeval (osa)sünonüümid (vt ÜS 2024)). Gemini andis ainult vaelevasted, GPT ja Claude andsid ühe õige vaste *picker*, mis on ka ainus häälikuliselt sarnane kuju *peekrile*.

Penningu puhul pakkusid Claude ja Gemini vastet *pennisitt* (mis on märk sellest, et mudel on ühestanud sõnakujusid häälikulise kokkulangevuse alusel), jättes esitamata Gösekeni *killing* ja Gutsblaffi *teng*. GPT pakkus lisaks ka *rahd* ja *rahdi*. Selgitusest ilmneb, et *penning* ja *pennisitt* on kokku viidud häälikulise sarnasuse tõttu, *rahd* ja *rahdi* aga tähendusseose tõttu *penningu*⁶ ja *raha* vahel.

Seevastu *ahv* on hea näide sellest, et keerulisi ülesandeid, nagu seda on *draakon* ja *pisuhänna* ning *väävli* ja *penisita* kokkuviimine, lahendav mudel ei näe seost tänapäeva keele sünonüümide vahel (vt ÜS 2024 *ahv* ja *pärdik*). *Ahvi* vasteid *pertick* ja *pertike* mudel ei esitanud, pakkudes vasteteks *püssihand*, *valgkemees* ja *püssohand*.

Kõik kolm mudelit leidsid õige vaste 9 sõnale 20st (45%). Sõnu, mille kõigi vastetega eksisid kõik mudelid, oli kaks: *ahv* ja *penning* (5%).

Mudelite käitumises avaldub suur ebaühtlus, kuivõrd nad suudavad omavahel kokku viia sõnad *draakon* ja *pisuhänd* ning *valgkemees*, samas aga ei suuda siduda tänapäeva eesti keele sünonüüme *ahv* ja *pärdik* ega *peeker* ja *karikas*. On näha, et mudelid seovad sõnu peamiselt kirjakujude sarnasuse põhjal, aga ka tähendusseoste najal.

5. Kokkuvõte

Uurimuses käsitletud keelemudelite analüüs sõnakujude tuvastamisel osutas, et SKM-idel on võimekus kokku viia tänapäevaseid sõnakujusid 17. ja 18. sajandi sõnastikes esinevate sõnakujudega. Kolmest analüüsitud keelemudelist – GPT-4o, Gemini 1.5 Pro ja Claude 3 Opus – osutus kahes katses täpseimaks Anthropicu Claude 3 Opus, esitades seejuures põhjalikke selgitusi vastuseni jõudmise kohta: sellel keelemudelil oli õigeid vastuseid esimeses katses 90% ja kolmandas katses 60–90%. Ka teise katse teises osakatses oli täpseim Claude (81% õigeid vastuseid), teise katse esimeses osakatses oli edukaim Gemini (76% õigeid vastuseid).

Mudelite edukus ja katsete tulemused sõltuvad mõistagi katsematerjali valikust: kas vaadelda n-ö lihtsaid küsimusi (*kohtunik* : *sundija* saksa vaste *Richter* kaudu) või raskemaid (*palsam* : *üx kaunis heh haisija rocht*). Nagu võis eeldada, leidsid

⁶ *Penning* on halvasti valitud sihtmärksõna, kuivõrd tänapäeva keele sõna on *penn*. Samas osutus katse tulemus huvitavaks.

modelid kergemini õiged vasted sõnadele, mille kirja pilt oli sarnane tänapäeva sõnadele. Erineva tähenduse ja kirja pildi puhul lähtusid SKM-id mõnikord vaste pakkumisel ka ainult kirja pildist, arvestamata tähendusi, nt 2. katses *loti* – *lodi* ja *penni* – *penn* ning 3. katses *penning* – *pennisitt*.

SKM-ide potentsiaalset rolli leksikograafias on käsitletud arvukates uurimustes, ent seni nähakse neid leksikograafide abistajana, mitte asendajana. Ennekoike püütakse SKM-e rakendada nende tööprotsesside juures, mis on automatiseeritavad või mille puhul saab vähendada mehaanilist ja ajamahukat käsitööd. Edaspidi jätkame katseid Gutsloffi ja Vestringi sõnastike uusväljaannete põhjal loodud nn digikaksikutega, et uurida keelemudelite võimekust varustada nende sõnastike eesti tõlkevasteid ja märksõnu nüüdisvastetega.

Viidatud kirjandus

- Beliga, Slobodan; Filipović Petrović, Ivana 2024. Large language models supporting lexicography: Conceptual organization of Croatian idioms. – Špela Arhar Holdt, Tomaž Erjavec (Eds.), Proceedings of the Conference on Language Technologies and Digital Humanities. Ljubljana: University of Ljubljana, 26–49. <https://doi.org/10.5281/zenodo.13912514>
- 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. – Konverentsietekanne: 20th CODH Seminar, ROIS-DS Center for Open Data in the Humanities, Tokyo, Japan, 27 February 2023. <http://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227/>
- Despot, Kristina Štrkalj; Ostroški Anić, Ana; Brač, Ivana (Eds.) 2024. Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. 8–12 October. Cavtat, Croatia.
- Geeraerts, Dirk 2024. Lexical Semantics, Lexicography and LLMs. – Konverentsietekanne: XXI EURALEX International Congress. 8–12 October. Cavtat, Croatia https://videolec-tures.net/videos/euralex2024_cavtat_geeraerts_lexical_semantics (10.1.2025).
- Gutsloff, Johannes 1648. Observationes Grammaticae circa linguam Esthonicam. Dorpat: Johannes Vogel. <http://www.digar.ee/id/nlib-digar:100419> (10.9.2024).
- Gutsloff, Johannes 1998. Observationes grammaticae circa linguam Esthonicam. Grammatilisi vaatlusi eesti keelest. Tartu Ülikooli eesti keele õppetooli toimetised, 10. Marju Lepajõe (Tlk ja koost.), Jaak Peebo (Toim.). Tartu: Tartu Ülikool. Tartu: Tartu Ülikool, lk 285–316.
- Göseken, Heinrich 1660. Manuductio ad Linguam Oesthonicam. Anführung zur Öhstni-schen Sprache. Reval: Adolph Simon. <https://kivike.kirmus.ee/meta/AR-11170-72005-62344> (10.9.2024)
- Helle, Anton Thor 1732. Kurtzgefaszte Anweisung Zur Ehstnischen Sprache. Halle: Stephan Orban. <http://www.digar.ee/id/nlib-digar:100071> (10.9.2024)
- Hupel, August Wilhelm 1780. Ehstnische Sprachlehre für beide Hauptdialekte den revalschen und dörptschen; nebst einem vollständigen Wörterbuch. Riga–Leipzig: Johann Friedrich Hartknoch. <http://www.digar.ee/id/nlib-digar:100926> (10.9.2024)
- 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. EKKD-III1 ajaloolise leksikograafia tööühma repositoorium. Failid EKKD-III1 publikatsioonile “Eesti Rakenduslingvistika Ühingu aastaraamatu” 2025. a väljaandes. <https://github.com/keeleinstituut/EKKD-III1/tree/main/ajalooline-leksikograafia/ERY-2025> (20.1.2025).

- Keem, Hella 1998. Johannes Gutsclaffi eesti keel ja Urvaste murrak. – Johannes Gutsclaffi, *Observationes grammaticae circa linguam Esthonicam*. Grammatilisi vaatlusi eesti keelest. Tartu Ülikooli eesti keele õppetooli toimetised, 10. Marju Lepajõe (Tlk. ja koost.), Jaak Peebo (Toim.). Tartu: Tartu Ülikool, lk 317–332.
- Kibbermann, Elisabeth; Kirotar, Salme; Koppel, Paula 1975. Saksa-eesti sõnaraamat = Deutsch-estnisches Wörterbuch. Tallinn: Valgus.
- Kikas, Kristel 2002. Mida sisaldab Heinrich Stahli Vocabula? Tartu Ülikooli eesti keele õppetooli toimetised, 11. Valve-Liivi Kingisepp (Toim.). 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. – *Humanities and Social Sciences Communications*, 10, a704. <https://doi.org/10.1057/s41599-023-02119-6>
- Mänd, Loviisa; Tóth, Szilárd Tibor 2025. Vana kirjakeele feminiinest *ik*-sufiksist murde-sõnastike peeglis. – *Keel ja Kirjandus*, 1–2, 129–138. <https://doi.org/10.54013/kk806a9>
- Paet, Tiina 2023. Võõraainese kinnistumine eesti keeles: keelekorralduslik ja leksikograafiline vaade. *Dissertationes philologiae estonicae Universitatis Tartuensis*, 51. Tartu: Tartu Ülikooli Kirjastus.
- Stahl, Heinrich 1637. *Anführung zu der Esthnischen Sprach*. Reval: Chr. Reusner der älter. <http://www.digar.ee/id/nlib-digar:101060> (10.9.2024).
- Stahl, Heinrich 1649. *Leyen-Spiegels*. II osa. Reval: Heinrich Westphal.
- Tafenau, Kai 2011. Heinrich Gösekeni sõnaraamatu seni märkamata eeskuj. – *Keel ja Kirjandus*, 6, 425–439.
- Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia 2017. Attention is All you Need. – I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (NIPS 2017). Curran Associates, Inc, 5998–6008.
- Wild, Kate 2024. Using large language models for a large historical dictionary: Challenges and opportunities for the Oxford English Dictionary. – *Konverentsietekanne: XXI EURALEX International Congress*. 8–12 October. Cavtat, Croatia https://videolec- tures.net/videos/euralex2024_cavtat_wild_using_language_models (10.1.2025).

Võrguviited

- Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. <https://api.semanticscholar.org/CorpusID:268232499> (10.1.2025).
- DWDS = Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. Berlin-Brandenburgischen Akademie der Wissenschaften. <https://www.dwds.de/> (17.9.2024).
- EKSS 2009 = Eesti keele seletav sõnaraamat. 2., täiendatud ja parandatud tr. Margit Lange-mets, Mai Tiits, Tiia Valdre, Leidi Veski, Ülle Viks, Piret Voll (Toim.). Tallinn: Eesti Keele Sihtasutus. <https://arhiiv.eki.ee/dict/ekss/index.cgi> (17.9.2024).
- EMS = Eesti murrete sõnaraamat. 1994–... (a–roietu). Mari Kendla, Mari-Liis Kalvik, Anu Haak, EviJuhkam, Tiina Laansalu, Varje Lonn, Helmi Neetar, Ellen Niit, Piret Norvik, Vilja Oja, Valdek Pall, Eevi Ross, Aldi Sepp, Mari-Epp Tirkkonen, Jüri Viikberg (Toim.). Tallinn: Eesti Keele Instituut. <https://arhiiv.eki.ee/dict/ems/> (17.9.2024).
- Gemini Team Google; Georgiev, Petko; Lei, Ving Ian; Burnell, Ryan; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (Version 4). arXiv. <https://doi.org/10.48550/ARXIV.2403.05530> (10.1.2025).
- OpenAI; Hurst, Aaron; Lerer, Adam; Goucher, Adam P.; et al. 2024. GPT-4o System Card (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2410.21276> (10.1.2024).

- RaS = Raamatukogusõnastik. Eesti Keele Instituut, Sõnaveeb. <https://sonaveeb.ee/ds/rara> (17.9.2024).
- Vestring 2000 = Vestring, Salomo Heinrich 2000 [17XX]. *Lexicon Esthónico Germanicum*. Võrguteavik. Ellen Kaldjärv, Krista Aru, Arvo Krikmann (Toim.). Tartu: Eesti Kirjandusmuuseum. <https://www.folklore.ee/~kriku/VESTRING/index.htm> (17.9.2024).
- VMS = Väike murdesõnastik. Valdek Pall (Toim.). Anu Haak, Evi Juhkam, Marja Kallasmaa, Arnold Kask, Varje Lonn, Ellen Niit, Piret Norvik (Raud), Vilja Oja, Eevi Ross, Aldi Sepp, Jaak Simm, Jüri Viikberg (Koost.). Indrek Hein (Veebiliides). Tallinn: Eesti Keele Instituut, 1996. <https://arhiiv.eki.ee/dict/vms/> (18.10.2024).
- ÕS 1918 = "Eesti keele õigekirjutuse-sõnaraamatu" kommenteeritud väljaanne. Tiina Paet, Tuuli Rehema (Koost., Toim.). Tallinn: Eesti Keele Instituut, 2018. <http://www.eki.ee/dict/qs1918> (18.10.2024).
- ÜS 2024 = Eesti Keele Instituudi ühendsõnastik 2024. Eesti Keele Instituut, Sõnaveeb. <https://sonaveeb.ee> (18.10.2024).

IDENTIFYING OLD ESTONIAN WORD FORMS USING LARGE LANGUAGE MODELS

Madis Jürviste^{1,2}, Tiina Paet¹, Sven-Erik Soosaar¹

Institute of the Estonian Language¹, University of Tartu²

As large language models (LLMs) have gained more and more visibility and momentum in society since 2022, numerous researchers have studied the possibilities of applying these new technologies for research in lexicography. This article deals with historical sources: how useful are LLMs in identifying old word forms in 17th and 18th-century German-Estonian and Estonian-German dictionaries? More precisely, can these technologies reduce the time burden on human researchers to identify old word forms and connect them with the same words' modern written forms (even if the original word itself has been substituted by a completely new one over the centuries)? To answer these questions, the authors conducted an empirical qualitative study with three major LLMs: GPT-4o, Gemini 1.5 Pro and Claude 3 Opus. The study consisted in analysing the LLMs capacities and success rates using API-request-based prompts in three main tests, each with different samples: 30 old professional titles and societal roles' denominations (6 sources ranging from Stahl 1637 up to Hupel 1780); 54 dialectal words (in Gutsclaff 1648) and 20 borrowed words (in 3 sources: Stahl 1637, Gutsclaff 1648, and Göseken 1660). In these tests, Claude generally outperformed all the others. However, the results show variations due to the sample words' characteristics (words with a similar orthography are more easily recognised). The high success rate, ranging from 74% to 90%, incites the authors to consider the possibility of carrying out tests with a larger sample, possibly encompassing whole dictionaries. This would significantly help lexicographers to create a diachronic historical development path for different words in the entries of large Estonian monolingual explanatory dictionaries.

Keywords: historical lexicography, history of Estonian written language, large language models, Estonian

Madis Jürviste (Eesti Keele Instituut, Tartu Ülikool) peamine uurimisvaldkond on eesti 17. ja 18. sajandi leksikograafia, lisaks tegeleb nüüdisaegsete üks- ja kakskeelsete sõnaraamatute koostamisega ning kaardistab suurte keelemudelite rakendusvõimalusi sõnastike analüüsimisel ja koostamisel.
Roosikrantsi 6, 10119 Tallinn, Estonia
madis.jyrviste@eki.ee

Tiina Paet (Eesti Keele Instituut) tegeleb keelekorralduse aktuaalsete küsimustega (võõrsõnade kohanemine ja keelemuutustega seotud aspektid), keelekorralduse ajaloo ning ajaloolise leksikograafiaga.
Roosikrantsi 6, 10119 Tallinn, Estonia
tiina.paet@eki.ee

Sven-Erik Soosaar (Eesti Keele Instituut) tegeleb soome-ugri keelte leksikograafia ja eesti keele ajaloolise leksikograafiaga ning põliskeelte keeletehnoloogilise arendamisega.
Roosikrantsi 6, 10119 Tallinn, Estonia
svenerik@eki.ee