doi:10.5128/ERYa22.10

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 22, 163–182

KORPUSE KONTEKSTIGA VÕI ILMA?
SUURTE KEELEMUDELITE VÕIMEKUS TUVASTADA JA MÄRGENDIGA VARUSTADA EESTI KEELE SÕNATÄHENDUSI

Lydia Risberg, Eleri Aedmaa, Hanna Pook, Kristina Koppel, Maria Tuulik, Esta Prangel, Margit Langemets

Ülevaade. Artiklis kirjeldame katset, milles võrdlesime suurte keelemudelite (SKM) võimekust tuvastada sõnatähendusi, määrata neile register ja pakkuda sobivaid märgendeid, ühel juhul eesti keele ühendkorpusest (2023) pärit kontekstide toel ja teisel juhul eeltreenitud SKM-ide keelelistele teadmistele tuginedes. Kõigi SKM-ide väljund hinnati adekvaatsemaks, kui need toetusid korpusandmetele, sealjuures parima tulemuse eri aspektides andis Claude Opus 4.1. Näiteks sõnatähenduste tuvastamisega said SKM-id edukamalt hakkama korpusematerjali toel, eeltreenitud SKM-ide pakutud tähendused jäid oletuslikumaks. Registri määramisega tulid SKM-id mõlemal juhul hästi toime. Pakutud märgendite kattuvus EKI ühendsõnastiku (ÜS 2025) omadega aga varieerus SKM-iti rohkem.*

Võtmesõnad: suured keelemudelid, eesti keele ühendkorpus, leksikograafia, register, sõnastikumärgendid, eesti keel

1. Sissejuhatus

Leksikograafi ehk sõnaraamatukoostaja töölauale satub aeg-ajalt sõnu või õigupoolest sõnatähendusi, mille puhul on keerukas otsustada, kas sellele tuleks sõnaraamatus lisada märgend, osutamaks, et seda kasutatakse mingil viisil eripäraselt (nt KÕNEKEELNE, VULGAARNE, HARV). Deskriptiivses sõnaraamatus tugineb keeleinfo, sh märgendid, tegeliku keelekasutuse andmetele. Samas on teada, et osa kasutajaid kipub ka sel juhul kogu keeleinfot tõlgendama preskriptiivselt (Trap-Jensen 2002). Siiski, nagu on öelnud keeleteadlane Geoffrey K. Pullum (2023: 7): "Piirangute kirjeldamine ei ole sama, mis soovitada nende järgimist."

Nii mujal maailmas (nt Salgado 2025, Klosa-Kückelhaus, Tiberius 2024) kui ka Eestis on aja jooksul muutunud see, millele leksikograaf otsuse langetamisel toetuda

* Artikkel on valminud teadusprojekti EKKD-III1 "Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed" toel. Täname ka Ene Vainikut panuse eest katse kujundamise algfaasis!

on saanud. Kui 20. sajandil (ja varem) toetusid koostajad märgendite lisamisel oma intuitsioonile või väikesemahulistele sedelkartoteekidele (Karelson 1990, EKSS 2009: 5), mis ei pruukinud esindada laiema keelekõnelejaskonna kasutusmustreid, siis 20. sajandi lõpust alates on aina enam saanud toetuda mahukatele korpusandmetele (Risberg jt 2025a, Langemets jt 2018). Korpuse kasutamiseks on leksikograafidel üldised tööpõhimõtted juba välja kujundatud (ehkki märgendite vallas on veel küsimusi õhus), ent suhteliselt uute abivahendite, suurte keelemudelite (SKM-ide) sõnaraamatutöös kasutamise võimaluste uurimises ja ühtsete aluste väljatöötamises on astutud alles esimesi samme (vt nt Risberg jt 2025b, Tuulik jt 2025, Jürviste & Jakobson 2025, mujal nt Trap-Jensen 2024).

Artiklis uurime, milline SKM saab leksikograafile olla potentsiaalselt abiks märgendite üle otsustamisel. Mille poolest erinevad SKM-id, kui need peavad tuvastama eesti sõnade tähendusi ja määrama neile registri ning sõnastikumärgendid, tuginedes kas eesti keele ühendkorpuse 2023 (Koppel jt 2023) materjalile või ainult eeltreenimisel omandatud teadmistele? Näiteks inglise keele puhul on SKM-e korpuseandmete analüüsiks juba edukalt rakendatud, sealjuures ennustasid SKM-id väga hästi sõna kasutuserinevusi ilukirjanduses ja formaalsetes akadeemilistes tekstides (Davies 2025). Eesti keel on seevastu SKM-ide treeningandmetes alaesindatud (on vähem erisugust materjali kui inglise keele kohta), samas pole see piiratud ressurssidega keel, eesti keele kohta oli 2025. a suvel 3,8 mld sõnega ühendkorpus (vt ka Koppel & Kallas 2022).

Kui eelmise katsega (Risberg jt 2025b) saime teada, et parimad eeltreenitud SKM-id saavad sõnatähenduste (in)formaalsuse üle otsustamisega ~76%-l juhtudest adekvaatselt hakkama, siis uue katsega tahame teada, kas ja kuidas parandab konteksti lisamine korpusandmete kujul SKM-ide tööd tähenduste tuvastamise ning neile registri ja sõnastikumärgendi määramise ülesannetes.

2. SKM-ide hindamine

SKM-ide oskuste ja piirangute mõistmiseks tuleb neid hinnata, ent valdkonna uudsuse tõttu ei ole veel tekkinud üldtunnustatud hindamisstandardeid. SKM-ide hindamiseks on erisuguseid viise, näiteks pakuvad standardsete mõõdikutega automaatse hindamise meetodid arvutuslikku tõhusust, samal ajal kui inimhinnangud võivad pakkuda nüansirikast teavet SKM-ide vastuste kvaliteedi ja täpsuse kohta.[1] (McIntosh jt 2024: 1)

SKM-ide hindamisel on sageli keskendutud ainult ChatGPT-le (Laskar jt 2024: 13785). Lisaks on SKM-e tihti hinnatud kindlate "õigete" vastuste põhjal, aga see ei näita SKM-ide päriselus rakendatavust. SKM, mis saab hästi hakkama standardiseeritud ülesannetega, võib aga raskustesse sattuda kultuuri tundmist nõudvates olukordades või otsuse tegemisel eetiliste kaalutluste arvestamisega. Niisiis vajab SKM-ide hindamine meetodit, mis arvestaks ka SKM-ide tegelikku rakendatavust ja samuti ohutust päriselus – SKM võib põhjustada riske nagu ühe keele, kultuuriruumi, maailmavaate vms poole kallutatuse süvendamine, ohtlike otsuste tegemine või manipuleerimisele allumine. (McIntosh jt 2024: 1–2)

Meiegi ülesande puhul on oluline SKM-ide rakendatavus päriselus, sõnaraamatutöös. Inimesiti võib vägagi erineda, kuidas sõna mingis tähenduses

[1] Näiteks eesti keele ja kultuuri tundmise hindamiseks on Tartu Ülikooli, Tallinna Tehnikaülikooli, Tallinna Ülikooli ja Eesti Keele Instituudi koostöös loodud tehisaru baromeeter, vt https://baromeeter.ai/ (5.1.2026). SKM-e on eesti keeles hinnanud ka Lillepalu & Alumäe (2025).

tajutakse. Isegi kui on kokku lepitud ühistes juhistes, võib leksikograafil olla märgendi üle keeruline otsustada, sest keel ei ole alati nii selgete piiridega, kui sõnaraamatu jaoks tarvis oleks (Risberg jt 2025b: 345, Rundell 2002). Seepärast on oluline märgendi üle otsustades tugineda laiema kõnelejaskonna kasutusmustritele, millele jälilesaamiseks ongi saanud uurida keelekorpust. Mahukatest korpusandmetest olulise väljasõelumine on üks ülesanne, millega SKM-id saavad leksikograafile potentsiaalselt abiks olla. Kas see nii on, selleks peab SKM-ide väljundit esmalt hindama. Kuigi üha rohkem rakendatakse kiireks ja odavaks hindamiseks inimhindajate asemel SKM-e endid (Wang jt 2025: 1956), on jätkuvalt oluline inimeste hinnangute põhjal välja töötada pädevad ja objektiivsed võrdlusandmestikud. Selgi juhul tuleb arvestada, et iga hindamisandmestik on oma koostajate nägu.

3. Katse disain

Katsega tahtsime teada, kuidas saavad SKM-id hakkama sõnastikumärgendi määramise ülesandega, milleks on oluline mõista sõnatähenduse kasutusregistrit.[2] Registrit on mõtestatud üsnagi erinevalt – vahel on selle all mõeldud žanreid ehk tekstiliike, teinekord eri formaalsusastmega kategooriaid nagu (formaalne) ametlik keel ja (informaalne) kõnekeel (vt Biber & Egbert 2023, Vaik 2024). Siin artiklis käsitleme registrina formaalsust, lähtudes sellest, kas sõna mingis tähenduses kiputakse keeleandmete põhjal kasutama pigemini mõne formaalsusastme poole kaldu olevas ümbruses.

Märgendi lisamise üle otsustamiseks uurib leksikograaf sõna tähendusi ja nende registrit eesti keele ühendkorpuses (korpuspäringutööriistas Sketch Engine, Kilgarriff jt 2014), kasutades enamasti konkordantsotsingut. Sketch Engine'is saab küll vaadata sõna kasutuskonteksti kohta žanrite, teemade jm kokkuvõtteid, ent iga blogitekst ei pruugi alati olla kirjutatud informaalselt ega iga ajaleheartikkel alati olla neutraalne (vt ka Risberg jt 2025a: 614). Kuna teksti formaalsusaste sõltub paljuski sihtrühmast, siis otsustasime katses SKM-idele kontekstide metaandmeid kaasa mitte anda.

Korpusest tuleb uurijal sõnatähendusi ise tõlgendada, see on aga sagedate sõnade puhul keerukas ülesanne. Ka see, kui sõna kattub isiku-, koha- vm nimega, raskendab muu, nt harva tähenduse leidmist. Seepärast uurisime, kas ja mille poolest erinevad SKM-ide väljundid sõnatähenduste tuvastamises ning neile vastavalt registrile sõnastikumärgendite määramises sõltuvalt sellest, kas SKM-idele antakse lisainfona kaasa autentne kontekst eesti keele ühendkorpusest (2023) või need tuginevad ainult oma eeltreeningus nähtud tekstidele.

3.1. SKM-ide kasutamine ja valik

Meie eesmärk oli võrrelda eri SKM-e. Ülesande sarnasest iseloomust tulenevalt valisime katsesse SKM-id eelmise katse (Risberg jt 2025b) tulemuste põhjal, võttes kolmelt parimalt tootjalt parima ja kättesaadava versiooni. Lõppkatsesse jäid

[2] Enne põhikatset tegime augustis 2025 pilootkatse, mille põhjal lõppkatse prompti pisut muutsime. Näiteks otsustasime lasta SKM-idel tähendusi ise tuvastada, mitte anda sõna konkreetsed tähendused ette. Üldiselt seisnesid promptitäiendused väikestes detailides, nii et artiklis me pilooti ei kirjelda. Kõik katsega seotud materjalid (sh promptid, andmed, juhendid) leiab GitHubist: https://github.com/keeleinstituut/EKKD-III1/tree/main/registrid/katse3 (5.1.2026).

Claude Opus 4.1 (Anthropic 2025),[3] Gemini 2.5 Pro (Gemini Team, Google 2023) ja GPT-4o (OpenAI 2024;[4] edaspidi nimetame SKM-e üldnimega).

SKM-ide kasutamisel tuleb olla teadlik nende puudustest. Näiteks pole kõikide SKM-ide tehniline arhitektuur ning treenimiseks kasutatud andmed ja juhendid avalikud, mistõttu ei ole nende väljundeid alati võimalik lahti seletada ega põhjendada. SKM-id võivad genereerida ebatäpset infot ja nende väljundid ei ole alati reprodutseeritavad. (Lappin 2024) Kommertsmudeleid saab üldjuhul kasutada nii vestlusrakenduse (nt ChatGPT) kui rakendusliidese (API) kaudu. Meie kasutasime neid API-de kaudu, sest nii on SKM-ile võimalik anda automaatselt analüüsimiseks suuri andmehulki, nagu on meie katses kasutatud korpusandmed, ning vestlusrakendusega võrreldes on vastuste struktuuri üle suurem kontroll.

3.2. Promptid

Promptisime SKM-e kahel viisil – täiendava kontekstiga (ingl *context-augmented prompting*) ja ilma, st eeltreenitud SKM-i selle teadmistele tuginedes. Promptid erinesidki selles, kas SKM-ile anti lisaks korpusandmeid või mitte (vt Github). Koostasime erinevad promptid selleks, et lisakontekst ei mõjutaks SKM-ide vastuseid, kui need pidid tuginema vaid eeltreenimisel õpetatud teadmistele. Mõlemad promptid struktureerisime ülesannete kaupa. Kui üldiselt kirjeldasime ülesandeid promptis ilma näideteta (ingl *zero-shot prompting*), siis märgendi määramise ülesandes lisasime ka näiteid selle kohta, missugustele sõnadele on sõnaraamatus vastav märgend lisatud. Lisakontekstiga promptimisel lisasime korpusest saadud kontekstid pikkusega (kontekstiaknaga) 7 lauset. Andsime kõigi sõnadega SKM-ile koos promptiga kaasa maksimaalse võimaliku hulga konteksti. Sagedasemate sõnade puhul tähendas see konteksti eelnevat vektoriseerimist ning vektoriseerimise järel kõige relevantsemate kontekstide valikut.[5] Kontekstid vektoriseerisime Multilingual E5 baasmudeliga (Wang jt 2024). See, milliste sõnade kontekst tuli enne SKM-ile etteandmist vektoriseerida, olenes SKM-i maksimaalsest sisendi suurusest.

Mõlemas promptis kirjeldasime SKM-idele sama rolli: "Oled eesti keele sõnaraamatu koostaja" ning tööülesande vastavalt sellele, millele tuginedes SKM oma vastuseid andma pidi: "Sinu ülesanne on analüüsida sõna "{katsesõna}" kasutust [etteantud tekstimaterjalis / enda treeningandmetes] ja otsustada, kas selle tähendustele tuleks lisada registrimärgend." Instrueerisime SKM-e (1) tuvastama sõna tähendusi, (2) nimetama tähenduste koguarvu, (3) määrama tähenduse sageduse,[6] (4) tooma tähenduse kohta 5 näitelauset, (5) määrama tähendusele registri, (6) põhjendama valikut, (7) ütlema, kui kindel see oma registrivalikus on, (8) pakkuma märgendi(d) ja (9) märgendivalikut lühidalt põhjendama.

SKM-il oli valik märgendit mitte lisada, kui see andmete põhjal seda vajalikuks ei pidanud. Teisalt võis ühele tähendusele lisada mitu märgendit. Otsustasime toetuda EKI ühendsõnastiku 2025 (ÜS)[7] jaoks väljatöötatud märgenditele ja selgitasime igaühe kohta promptis, millisel juhul see tuleks valida. Nii oli SKM-ide

[3] Piloodis katsetasime Claude 3.7 Sonnet'ga. Vahepeal tuli välja uuem versioon, Claude Opus 4.1, mis andis võrdlusandmestikuga parema tulemuse. Seega valisime lõppkatsesse uuema versiooni.

[4] Tol hetkel uusim, GPT-5 andis eelmise katse võrdlusandmestikuga kehvema tulemuse kui GPT-4o. GPT-4.1 tulemused ei erinenud aga oluliselt ning kuna ka pilootuurimuses saadud väljundite võrdlus ei näidanud märkimisväärseid erinevusi, otsustasime lõppkatses jääda end tõestanud GPT-4o juurde.

[5] Vektoriseerimisest SKM-ide kontekstis on kirjutanud nt Eiche jt 2025.

[6] Lasime SKM-il konkreetsete numbrite asemel öelda suurusjärke, sest pilootkatsest oli näha, et ka vektoriseerimata puhkudel ei andnud SKM-id tõeseid numbreid. See polnud üllatav, sest SKM-idele on palju lihtsamadki kokkulugemisülesanded keerukaks osutunud (vt nt Fu jt 2024).

[7] Kasutasime 2025. a augusti keskpaiga seisuga EKI ühendsõnastikku.

väljundit lihtsam hinnata, sest kui oleksime lasknud SKM-idel ise märgendeid luua ja kategooriaid välja mõelda, poleks me saanud neid omavahel sama hästi ja samadel alustel võrrelda. Lisaks jätsime promptis ütlemata, et SKM ei arvestaks Eesti Keele Instituudi (EKI) materjale, sest sellisel juhul SKM lihtsalt ei ütleks seda, kuigi võis ennustuses nendel siiski põhineda (vt ka ptk 4.4.2).

3.3. Katsesõnade valik ja märgendid

Katsesse valisime EKI ühendsõnastiku (ÜS) märgendid HALVUSTAV (oktoobri 2025 seisuga kokku 188 tähendusel), HARV (3478), KÕNEKEELNE (5008), LASTEKEELNE (62), LUULEKEELNE (28), MURDEKEELNE (371), RAHVAPÄRANE (358), STIILITUNDLIK (24), VANANENUD (1617) ja VULGAARNE (61).[8] Lisaks kaasasime katsesse ilma märgendita sõnu (ÜS-is on kokku ligi 180 000 keelendit, neist u 169 000 ilma märgendita). Igaühe kohta võtsime algul ÜS-ist augustis 2025 juhuvalimiga 10 sõna, mille tähenduse küljes vastav märgend oli. Vaatasime töörühmaga katsesõnad üle ja olime nõus, et tähenduse juures olemasolevad märgendid pädevad. Valisime iga märgendi 10 sõna seast välja 5 erisugust juhtumit, et esindatud oleksid sagedad ja harvad sõnad, mitme- ja ühetähenduslikud sõnad, eri sõnaliigid, ühe või mitme märgendiga ja ilma märgendita sõnad. Kokku oli katses 55 sõna (vt tabel 1). Niisiis võis sõna mingi märgendi kaudu valimisse sattuda, aga kuna märgend on tähenduse, mitte sõna kui terviku küljes, võis sõnal olla teisi tähendusi ja alamtähendusi või homonüüme, milledel kõigil võis olla kas sama või muu märgend või mitte ühtegi märgendit.

Kaalusime seekord registri kolmeks (informaalne, neutraalne, formaalne) jagamist, sest esialgu mõtlesime kaasata ka formaalseks peetavaid sõnu. Neid oli aga raske leida, sest kuigi 2018. a õigekeelsussõnaraamatus on kasutatud liigformaalsele osutavat märgendit PABERL (ÕS 2025-s seda ei ole), siis töötasime need ligi 70 vastet läbi ning leidsime, et neile on märgend lisatud pigem keelekorralduslikel kaalutlustel ja nii ei peegelda märgend formaalsust laiemas keelekasutuses eriti hästi. Seepärast loobusime neutraalset ja formaalset eristamast ning grupeerisime need SKM-i jaoks üheks valikuks: informaalne *vs*. neutraalne/formaalne register.

Tabel 1. Katsesõnad märgendi järgi, millega need juhuvalimisse sattusid

Märgend	Katsesõnad
HALVUSTAV	*idioot*, *kobima*, *parasiit*, *pursui*, *sabarakk*
HARV	*ajaldi*, *halvakvaliteediline*, *krampima*, *leivaline*, *õlikas*
KÕNEKEELNE	*burks*, *kesse*, *kõmmima*, *soperdis*, *tele*
LASTEKEELNE	*jänku*, *pepu*, *punnu*, *tuduma*, *vissi*
LUULEKEELNE	*sala*, *askus*, *neelma*, *puhk*, *turm*
MURDEKEELNE	*hüva*, *kargutama*, *pedakas*, *sirk*, *õkva*
RAHVAPÄRANE	*jänesemokk*, *mihklikuu*, *nääl*, *talinelk*, *ute*
STIILITUNDLIK	*aulik*, *austet*, *elik*, *manu*, *seitung*
VANANENUD	*inglistina*, *kihvtitama*, *maakuulamine*, *rüve*, *töötatööline*
VULGAARNE	*emane*, *molu*, *munn*, *pasapea*, *perselakkuja*
märgendita	*süvasadam*, *teraapia*, *algatuslik*, *otseteed*, *palmima*

[8] Välja jäid märgendid UUS, sest need sõnad ei pruukinud 2023. a ühendkorpuses veel (palju) esineda, ning SÕNAETTEPANEK ja UNARSÕNA, sest selliseid sõnu pole kasutatud ja seega polnuks SKM-il neid korpuse põhjal võimalik analüüsida – erinevalt harva, kuid siiski kasutatavatest sõnadest (HARV), või varem kasutatud, aga vananenuks muutunud sõnadest (VANANENUD).

Tabelis 1 esitatud märgendite lisamise juhistes on ÜS-i koostajad omavahel kokku leppinud. Meiegi andsime SKM-idele promptides ette uurimisrühmas läbi arutatud lühijuhised. Näiteks märgend RAHVAPÄRANE "vali siis, kui sõna selles tähenduses on rahva seas levinud, aga pole ametlik termin, tihti näiteks kuude, taimede, loomade, haiguste, sugulaste nimetused", samuti lisasime katsesõnadest erinevaid näitesõnu: *heinakuu*, *jooksva*, *männiseen* (teisi juhiseid vt Githubist). SKM sai märgendit pakkuda ka registriliselt neutraalsele/formaalsele tähendusele, sest kuigi enamus märgendeid osutavad neutraalsest irduvale kasutusele (nt KÕNEKEELNE, STIILITUNDLIK), antakse märgenditega muudki infot, nagu aeg (VANANENUD) ja sagedus (HARV).

3.4. Väljundi hindamine

SKM-ide vastused küsisime augustis 2025. Osa väljundist hindasime käsitsi: hindaja pidi tegema SKM-ide pakutud tähenduste, registri ja märgendite adekvaatsuse kohta kategoorilise jah/ei-valiku. Lisada sai ka kommentaare. Adekvaatsuse hindamine on üsnagi subjektiivne ning kuigi näha olid (autentsed) näitelaused ja SKM-i põhjendused ning hindajal oli võimalus ka ise korpust vm lisamaterjali uurida, siis oleme teadlikud, et inimhinnangute puhul võib teine rühm inimesi jõuda teise tulemuseni.

Siin katses olid kõik hindajad meie uurimisrühma liikmed, kes kõnelevad eesti keelt emakeelena. Igat elementi hindas kaks hindajat. Erinevuse korral küsisime kolmanda hinnangu ja tulemuseks jäi ülekaalu saanud hinnang. Hindamisel üksteise arvamusi ei nähtud, küll aga arutasime koos läbi elementide adekvaatseks määratlemise põhimõtted, et tagada ühine arusaam hindamisjuhistest. Näiteks kui SKM oli tähenduse halvasti sõnastanud, ent näidete põhjal sai hindaja aru, mida SKM oli pakkunud ja et see tähendus on eesti keeles olemas, siis märgiti tähendus adekvaatseks (nt pakkus Claude korpusandmete pealt sõna *hüva* üheks tähenduseks 'omadussõnana', selmet pakkuda 'hea'). Kui SKM-i toodud näitelausetes esines katsesõnast erinevas sõnaliigis sõna (nt *kramp*, mitte katsesõna *krampima*), märgiti tähendus ebaadekvaatseks, sest see ei kirjeldanud otsitava sõna tähendust.

Hoolimata kokkulepetest ilmnes osade tähenduste ja märgendite kohta vastandlikke hinnanguid, sest hindamisjuhiseid oli erinevalt tõlgendatud – neil juhtudel jäi koondhinnanguks kokkulepitud juhiste kohane hinnang. Kokkuvõtlikult kattusid üksikhindaja otsused lõpliku konsensusotsusega 90–99% juhtudest, olenevalt hinnatavast aspektist.

4. Tulemused

Üldiselt vastasid SKM-id mõlema prompti puhul sellele, mida küsisime (kuigi formaadinõuetest alati kinni ei peetud).[9] SKM-ide lähemalt võrdlemiseks analüüsime tähenduste tuvastamist ja arvu, registri ja märgendite määramist nii korpuse konteksti toel kui ka ilma välise kontekstita antud vastuste põhjal.

[9] Eelmises, kõnekeelsete sõnade katses (Risberg jt 2025b) tegime ka järjepidevuse kontrolli, mis näitas, et vähemalt 97% tõenäosusega saavad kaks leksikograafi samade sõnade puhul samale promptile samad vastused. Kuna olemuslikku muutust SKM-ide arhitektuuris kahe katse vahel ei olnud, siis siin me järjepidevust ei kontrollinud.

4.1. Tähenduste tuvastamine ja arv

4.1.1. Tuvastamise adekvaatsus

Väljundit hinnates lähenesime tähendustele kitsalt leksikograafi praktilise pilguga, vastates küsimusele "Kas tähendus on eesti keeles olemas ja adekvaatne, et seda kirjeldada eesti keele sõnaraamatus?". Läheneda saanuks ka laiemalt keeleuurija vaatevinklist, kes tahab teada, kuidas sõna korpusematerjalis üleüldiselt kasutatud on: SKM-id tuvastasid materjalist ka selliseid tähendusi nagu *manu* 'jalgpalliklubi Manchester United' ja *ute* 'Austraalia kastiauto' ning kuigi see on uurijale kasulik info korpusematerjali kohta, ei ole need eestikeelsete sõnade tähendused, mida leksikograaf eesti sõnaraamatus tähendusena esitaks – seepärast me selliseid tähendusi adekvaatseks ei pidanud. Me ei lugenud adekvaatseks isiku-, pere-, hüüd- ega tootenimesid ega ka muid lühendeid (nt *elik* kui Euroopa Liidu Infokeskus) või pealkirju (*idioot* kui Dostojevski romaan), samas lugesime EKI ühendsõnastiku spetsiifikast tulenevalt adekvaatseks kohanimed (nt *kesse* 'kes see' kattub kohanimega Kesse). Tähenduste pakkumise adekvaatsuse võrdlus on näidatud tabelis 2.

Tabel 2. SKM-ide tulemused tähenduste pakkumise adekvaatsuses[10]

Tähenduse adekvaatsus	Claude		Gemini		GPT	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
jah (%)	78 (66%)	43 (45%)	102 (58%)	70 (50%)	102 (64%)	52 (46%)
ei (%)	40 (34%)	52 (55%)	75 (42%)	69 (50%)	58 (36%)	62 (54%)
Kokku	118	95	177	139	160	114

Sõnatähendusi pakkus kõige adekvaatsemalt korpuse kontekste kasutav Claude (66%), kuid statistiliselt oluliselt parem see teistest SKM-idest ei olnud[11] – selle ülesande puhul tuleks parima SKM-i väljaselgitamiseks valimit suurendada. GPT ja Claude'i tulemused erinesid oluliselt olenevalt sellest, kas need said korpuste kontekste kasutada või mitte. Nähtuski, et SKM-id said korpuse konteksti kasutades tähenduste tuvastamisega paremini hakkama. Ilmselt seepärast, et eeltreenitud SKM-id võivad hallutsineerida, sest peavad rohkem oletama kui autentseid kontekste kasutades ning samuti on neid instrueeritud alati vastus andma (Kalai jt 2025). Näiteks eeltreenitud GPT pidas sõna *kihvtitama* üheks tähenduseks 'lahe, äge' ja eeltreenitud Claude pakkus sõna *burks* ainsaks tähenduseks 'purk'. Üksnes eeltreenitud GPT märkis, kui see sõna ei teadnud (20 sõna juures), ent Claude ja Gemini olid enda jaoks tundmatutele sõnadele alati midagi pakkunud (nt *ajaldi* 'aeg-ajalt', *sala* 'salat').

4.1.2. Tähenduste arv ja eristamine

Tähenduste eristamiseks ei olnud SKM-idele teoreetilist alust ette anda, see on ka leksikograafidele keerukas ülesanne: ei ole lihtne otsustada, kas tegemist on eraldiseisva tähendusega või (enam/veel) mitte. SKM-id võisid mõne tähenduse

[10] Leksikograafi ja keeleuurija vaatenurgast erinevalt lahendatavaid juhtumeid oli küll märkimisväärselt, ent SKM-ide paremusjärjestus poleks muutnud. Korpuse kontekste kasutav Claude pakkus keeleuurija silmis adekvaatseid tähendusi 22 (sel juhul oleks adekvaatseid tähendusi kokku 85%) ja eeltreenitud Claude 0 (45%). Korpuse kontekste kasutav GPT 28 (81%) ja eeltreenitud GPT 3 (48%). Korpuse kontekste kasutav Gemini 27 (73%) ja eeltreenitud Gemini 6 (55%).

[11] Siin ja edaspidi oleme gruppidevahelise erinevuse hindamiseks kasutanud hii-ruut testi ning olulisuse nivood 0,05.

liiga peeneks ajada (promptis oli küll suunis seda mitte teha): nt korpuse kontekste kasutav GPT pakkus sõna *burks* eraldi tähendusteks 'burger' ja 'taimne burger'. Gemini tõi igasuguseid trükivigugi sõnatähendustena esile, nt pakkus see sõnale *tuduma* üheks tähenduseks 'tunduma', kuigi põhjendab ka ise, et "Tegu on tõenäoliselt trükivea või foneetilise kirjaviisiga". Kirjeldasime SKM-i rolliks olla sõnastiku koostaja, seega see võinuks teada, et ei pea üksikuid trükivigu tähendustena esile tooma, vaid tähendus tuleb tõlgendada kontekstist ja esinemissagedusest.

EKI ühendsõnastikus oli 55 katsesõnal kokku kirjeldatud 98 tähendust. Tabelis 3 on võrreldud SKM-ide pakutud tähenduste (arvestamata seda, kas need olid adekvaatsed) arvu ÜS-is olevate tähenduste arvuga. Nähtus, et Claude on pea pooltele sõnadele määranud sama palju tähendusi kui ÜS-is, erinedes statistiliselt oluliselt teistest SKM-idest, millel on ÜS-iga võrdsete tähenduste osakaal märksa väiksem. Kõiki SKM-e koos hinnates oli vaid Gemini puhul oluline, kas seda rikastati korpuse kontekstidega või mitte. Järeldasime, et SKM-ide pakutud tähenduste arv sõltub tõenäoliselt pigem SKM-ist, mitte sellest, kas seda lisaandmetega promptitakse või mitte.

Tabel 3. SKM-ide pakutud tähenduste arv võrreldes ÜS-is olevate tähenduste arvuga

Tähenduste arv *vs.* ÜS	**Claude**		**Gemini**		**GPT**	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
sama (%)	25 (46%)	23 (42%)	7 (13%)	19 (35%)	12 (22%)	8 (15%)
erinev (%)	30 (54%)	32 (58%)	48 (87%)	36 (65%)	43 (78%)	47 (85%)
kui erinev, siis kas rohkem (%)	22 (73%)	17 (53%)	44 (92%)	31 (86%)	38 (88%)	23 (49%)
kui erinev, siis kas vähem (%)	8 (27%)	15 (47%)	4 (8%)	5 (14%)	5 (12%)	24 (51%)

Tabelis 3 on näha ka see, et kui tähenduste arv ei klappinud, siis kas SKM pakkus ÜS-ist rohkem või vähem tähendusi. Erinevuse korral leidsid SKM-id üldjuhul sõnadele ÜS-iga võrreldes protsentuaalselt rohkem tähendusi, olenemata sellest, kas need kasutasid korpuse konteksti või mitte. Tähenduste arvus erinesid eeltreenitud SKM-id üksteisest statistiliselt olulisel määral. Kui eeltreenitud Claude ja GPT pakkusid enam-vähem võrdselt rohkem ja vähem tähendusi, kui on ÜS-is, siis eeltreenitud Gemini pakkus neid pigem rohkem. Kõiki SKM-e koos vaadates (ning GPT puhul eraldi) saab üldistada, et kontekste kasutavad SKM-id leidsid rohkem tähendusi kui eeltreenitud SKM-id.

Uurisime lähemalt, mis sõnadega oli väga suur erinevus SKM-i pakutud tähenduste arvu ja ÜS-is kirjeldatud tähenduste arvu vahel. Selgus, et enim tähendusi genereerinud Gemini vastused erinesid ÜS-ist rohkem kui teiste SKM-ide omad. Korpusest pakkus see 8 sõnale 3 või rohkem tähendust enam, kui on ÜS-is, sealjuures sõnale *kõmmima* lausa 9 tähendust (sh kõnekeelse nimisõna *kõmm* tähendused 'käibemaks' ja 'kilomeeter'), kui ÜS-is oli kirjeldatud vaid 1 tähendus 'tumedalt kõlavaid lööke, hoope andma, taguma'. Eeltreenitud Gemini pakkus ühele sõnale 3 tähendust rohkem, kui on ÜS-is (*neelma*).

Kõik SKM-id pakkusid sõnale *vissi* 3 või enam tähendust vähem, kui on kirjeldatud ÜS-is. Teisalt pakkusid kõik korpuse konteksti kasutavad SKM-id sõnale *manu*

3 või enam tähendust rohkem, sest ÜS-is on vaid stiilitundlik tähendus 'juurde', aga SKM-id tuvastasid kasutusest näiteks ka Manchester Unitedi lühendi ja india mütoloogias inimsoo esiisa Manu, mida üldkeele sõnaraamatus ei kirjeldata (küll aga ida mõtteloo leksikonis).

4.1.3. SKM-i ja sõnastikutähenduste kattuvus

Uurisime, kas SKM-i pakutud tähendused kattuvad EKI ühendsõnastikus kirjeldatud tähendustega ka sisuliselt ja milline SKM pakkus enim ÜS-i tähendusi. Kuna see ei olnud meie uurimuse põhiküsimus, siis testisime siin SKM-i potentsiaali olla iseenda (ja teiste omasuguste) hindaja. Kasutasime korpuse konteksti kasutavat Claude'i, kuna see oli kõige adekvaatsem tähenduste eristaja. Instrueerisime Claude'i igale ÜS-i tähendusele leidma enda antud tähenduste seast vaste või ütlema, et vastet pole.

55 katsesõna ÜS-i 98 tähendusest leidis Claude enda vastuste seast vaste 80-le (82%). Käsitsi üle kontrollides selgus, et täpsed vasteid oli siiski 68 (69%; nt tuvastas Claude valesti, et *hüva* 'parempoolne, parem' on sama mis tema pakutud 'hea, meeldiv, kasulik'). Tulemus ei anna küll alust kaugeleulatuvaid järeldusi teha ja lähenemine vajab edasist uurimist, aga viitab SKM-ide potentsiaalile toimida automaatsete hindajatena tähenduste tuvastamisel, et tulevikus vähendada käsitsi kontrollimise mahtu.

4.2. Register

Registri ehk formaalsuse asjus lasime SKM-idel tuvastada, kas sõna tema pakutud tähenduses esineb pigem informaalsetes või neutraalsetes/formaalsetes tekstides.[12] Valikus oli ka võimalus, et SKM ei oska eristust teha. Tabelis 4 on näha, kui paljude sõnatähenduste registrit (formaalsust) on SKM-id pakkunud meie töörühma hinnangul adekvaatselt (siin arvestasime üksnes adekvaatsetele tähendustele märgitud registreid). Registreid pakkusid statistiliselt olulisel määral adekvaatsemalt korpuse kontekste kasutavad Claude ja Gemini (90% ja 89%), märgatavalt kehvema tulemuse andis GPT (70%). Võime oletada, et kui SKM tuvastas sõnale juba adekvaatse tähenduse, suutis see üsna edukalt määrata ka adekvaatse registri, olenemata sellest, kas kasutas korpusekontekste või mitte. Samas ei luba statistiline analüüs nende andmete pealt põhjalikke järeldusi teha.

Tabel 4. SKM-ide tulemused tähendusele registri pakkumise adekvaatsuses

Registri adekvaatsus	**Claude**		**Gemini**		**GPT**	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
jah (%)	70 (90%)	34 (79%)	91 (89%)	62 (89%)	71 (70%)	37 (71%)
ei (%)	8 (10%)	9 (21%)	11 (11%)	8 (11%)	31 (30%)	15 (29%)
Kokku	78	43	102	70	102	52

[12] Andmeid analüüsides selgus, et oleksime võinud lasta eristada ka neutraalset ja formaalset registrit, sest märgendid STIILITUNDLIK ja LUULEKEELNE võisid osutada justnimelt formaalsele või kõrgstiilsemale esinemiskontekstile.

Uurisime ka seda, kui paljudel kõikidest juhtudest olid tähendus ja register mõlemad adekvaatselt määratud (vt tabel 5). Näeme, et Claude'i korpusandmetelt saadud tulemused on parimad (59%), ent seda tulemust ei saa pidada piisavalt kõrgeks. SKM-ide vahel on korpusandmete põhjal napilt statistiliselt oluline erinevus (GPT näitas taas kehvemaid tulemusi). Statistiliselt olulist erinevust täheldasime vaid eeltreenitud ja korpuse kontekste kasutava Claude'i puhul.

Seda kõike arvesse võttes järeldame, et sõnadele adekvaatsete tähenduste leidmine on SKM-ide jaoks keerukam ülesanne (adekvaatseid tähendusi oli 45–65%) kui registri määramine (adekvaatseid registreid 70–90%).

Tabel 5. SKM-ide tulemused selles, kas tuvastatud tähendus ja pakutud register on mõlemad adekvaatsed

Tähenduse + registri adekvaatsus	**Claude**		**Gemini**		**GPT**	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
jah (%)	70 (59%)	34 (36%)	91 (51%)	62 (45%)	71 (44%)	37 (33%)
ei (%)	48 (41%)	61 (64%)	86 (49%)	77 (55%)	89 (56%)	77 (67%)
Kokku	118	95	177	139	160	114

Kuna meie uurimuses oli registri määramine olulisel kohal, huvitas meid seegi, kas SKM on oma vastuses kindel või ebakindel. Seepärast analüüsisime, kas leidub seos SKM-i enesekindluse ja adekvaatse registripakkumise vahel. Tabelist 6 on näha, et kui SKM on olnud oma vastuses kindel ("väga kindel" või "pigem kindel"), hindas see ka registrit protsentuaalselt adekvaatsemalt kui siis, kui see oli ebakindel ("väga ebakindel" või "pigem ebakindel"). Kõiki SKM-e arvesse võttes on see erinevus statistiliselt oluline, SKM-e eraldi vaadates on erinevus kindlate ja ebakindlate vastuste adekvaatsuses statistiliselt olulisel määral erinev vaid GPT (nii korpuse kontekste kasutava kui ka eeltreenitud) puhul.

GPT oli enda pakutud registrihinnangutes pisut ebakindlam kui teised ja tundub, et põhjusega, sest sel juhul märkis see registrit ka vähem adekvaatselt. GPT kvaliteet aga ei küündinud teiste tasemele: Claude ja Gemini olid registreid hinnates enesekindlad, kahtlesid üksikute tähenduste puhul ning ebakindlus hinnangu adekvaatsust suurt ei mõjutanud.

Tabel 6. Seos SKM-ide enesekindluse ja registrimärgendi adekvaatsuse vahel

SKM-i enesekindlus + registri adekvaatsus inimhinnangul	**Claude**		**Gemini**		**GPT**	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
kindel + jah (%)	68 (91%)	34 (81%)	89 (90%)	60 (88%)	66 (75%)	34 (79%)
ebakindel + jah (%)	2 (67%)	0 (0%)	2 (67%)	2 (100%)	5 (36%)	3 (33%)
kindel + ei (%)	7 (9%)	8 (19%)	10 (10%)	8 (12%)	22 (25%)	9 (21%)
ebakindel + ei (%)	1 (33%)	1 (100%)	1 (33%)	0 (0%)	6 (64%)	6 (67%)

4.3. Märgendid

EKI ühendsõnastikus ei panda sõnatähendusele üle kahe märgendi korraga, prompti me aga sellist suunist ei lisanud. Nii olid SKM-id osadele tähendustele määranud ka kolm märgendit, nt eeltreenitud Gemini pakkus sõna *pasapea* tähendusele 'rumal, ebameeldiv või vastik inimene; sõimusõna' kolme märgendit: VULGAARNE, HALVUSTAV ja KÕNEKEELNE. Kuigi need kõik hinnati adekvaatseks, jätaks leksikograaf kõnekeelsuse siin eraldi välja toomata, kuna teised n-ö tugevamad märgendid juba osutavad, et neutraalsesse keelde selline kasutus ei sobi. Ühtlasi määraks leksikograaf tähendusele kahest tugevast märgendist vaid ühe. Hinnates võis märgendeid sobivuse korral märkida adekvaatseks ka juhul, kui need ei olnud samad, mis ÜS-is.

Vaatasime alustuseks üldiselt, kui paljud SKM-ide pakutud märgendid olid inimhinnangute järgi adekvaatsed (vt tabel 7), jättes siingi kõrvale ebaadekvaatsed tähendused. Kõikide SKM-ide adekvaatsete märgendite osakaal oli üsna sarnane. Üksnes korpuse kontekste kasutav Claude pakkus statistiliselt olulisel määral rohkem adekvaatseid märgendeid kui teised SKM-id.

Tabel 7. SKM-ide tulemused tähendusele märgendi pakkumise adekvaatsuses

Märgendi adekvaatsus	**Claude**		**Gemini**		**GPT**	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
jah (%)	86 (83%)	36 (59%)	125 (68%)	81 (67%)	87 (66%)	40 (63%)
ei (%)	18 (17%)	25 (41%)	59 (32%)	40 (33%)	44 (34%)	24 (37%)
Kokku	104	61	184	121	131	64

4.3.1. SKM-i ja sõnastikumärgendite kattuvus

Võrdlesime seda, kuivõrd kattusid SKM-ide pakutud märgendid EKI ühendsõnastikus sõna eri tähendustel olevate märgenditega (siin me ei vaadanud, et märgend oleks õige tähenduse juures, seda vt ptk 4.3.2). Katsesõnu oli 55, neil oli erinev arv tähendusi ja seega erinev arv märgendeid. Tabelis 8 on näidatud, kui mitmel juhul SKM-i pakutud märgendid ÜS-is olevatega 1) kattusid täielikult (100%); 2) kattusid, aga SKM pakkus midagi lisaks; 3) kattusid osaliselt (midagi oli sama, midagi erinevat); 4) ükski märgend polnud sama mis ÜS-is ehk kattuvus puudus.

Tabel 8. SKM-i pakutud märgendite kattuvus ÜS-is katsesõnade tähendustel olevate märgenditega

Kattuvus ÜS-iga	**Claude**		**Gemini**		**GPT**	
	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita	Kontekstiga	Kontekstita
100%	17	9	4	8	4	2
100%, aga SKM pakkus lisaks	6	5	35	18	17	7
osaliselt sama, osaliselt mitte	9	5	4	8	10	7
üldse mitte	23	36	12	21	24	39

Kõige täpsemalt määras ÜS-iga samu märgendeid korpuse kontekste kasutav Claude (17 sõnal täpselt samad märgendid). ÜS-iga võrreldes oli kõige kehvemini märgendeid määranud eeltreenitud GPT (39 sõna puhul ei klappinud mitte ükski märgend). GPT ja Gemini puhul oli lisakonteksti kasutamisel märgendite määramisele statistiliselt oluline mõju.

Nii eeltreenitud kui ka korpuse konteksti kasutav GPT pakkusid kõige enam üle märgendit KÕNEKEELNE. Claude ja Gemini pakkusid mõlemal juhul enim üle märgendit HARV – samas oligi promptis juhis panna see märgend iga kord, kui tähendust on andmetes vähe. Kokkuvõttes oli korpuse kontekste kasutav Gemini ainuke SKM, kes kõiki märgendeid üle pakkus.

Veel huvitas meid see, milliste märgenditega oli SKM-idel rohkem raskusi. Ülepakkujal Geminil ei olnud ühtegi märgendit, millele ta üldse pihta ei saanud, samas kui eeltreenitud GPT ja korpuse kontekste kasutav Claude ei tabanud nt märgendit STIILITUNDLIK. Kokkuvõttes osutusid aga kõikidele SKM-idele keerulisimaks märgendid VANANENUD (tabati 1–3 korda 7-st) ja HARV (1–3/6). Tulemus on osalt üllatav, kuna inimhindajate jaoks oli keerukas vahet teha näiteks kõne- ja rahvakeelsusel ning määrata stiilitundlikkust. Küll aga panime andmeid analüüsides tähele, et sõnade *inglistina* ja *maakuulamine* märgend VANANENUD sõltus pigem terminiotsusest, mitte tingimata korpuses nähtavast üldkeelsest kasutusest. Sõnal *halvakvaliteediline* oli ÜS-is märgend HARV, ent korpuses esineb seda ligi 200 korda.[13]

4.3.2. Tähendusele sõnastikumärgendi määramise täpsus

Uurisime, kas EKI ühendsõnastikus kirjeldatud tähendusel olev märgend on SKM-il pandud õige tähenduse külge. Aluseks võtsime ptk 4.1.3-s kirjeldatud eksperimendi tulemused õigete tähenduste määramisest Claude'iga ning ptk 4.3.1 andmed SKM-i ja ÜS-i märgendite kattumisest. Hindasime siin käsitsi, kas märgendid, mille Claude õigetele tähendustele määras, kattusid ÜS-is olevatega.

Claude pakkus 68-le ÜS-is kirjeldatud adekvaatsele tähendusele kokku 84 märgendit. Täpselt samad märgendid, mis tähendusel on ÜS-is, leidis Claude 31 juhul ehk ligi pooltele adekvaatsele tähendusele (nt sõnal *pepu* on ÜS-is 2 märgendit, mis on ka Claude'il samamoodi määratud: KÕNEKEELNE, LASTEKEELNE). Osaliselt samad ÜS-i märgendid leidis see 11 tähendusele (nt sõnal *punnu* on 2 märgendit ÜS-is, aga SKM leidis neist 1). Claude ei leidnud ühtegi ÜS-is olevat märgendit 26 tähendusele (nt on ÜS-is sõna *turm* tähendusel 2 märgendit, SKM andis ainult 1 ja valesti).

Seega näitas meie esimene vaatlus, et SKM-id on võimelised leidma suure osa märgendeid, mis ÜS-is tähendusele lisatud on, aga genereerivad palju ka selliseid, mis ÜS-is esitatuga kokku ei sobi. Teema vajab kindlasti süvitsi edasiuurimist.

4.4. Muid analüüsitahke

Kuna meie andmestik oli väike, ei saanud selle põhjal teha kaugeleulatuvaid järeldusi, kas korpuses sagedamini esinevad sõnad olid saanud rohkem või adekvaatsemaid hinnanguid. Ka sõna konteksti eelneval vektoriseerimisel (ja sealt kõige

[13] Eemaldasime märgendi, sest selle taustal olid olnud muud leksikograafilised valikud.

relevantsema konteksti SKM-ile edastamisel) polnud tulemustele statistiliselt olulist mõju. Samad analüüsid tuleks teha suurema andmestiku peal.

4.4.1. Näitelaused

Promptis instrueerisime, et SKM tooks iga tähendust illustreerima 5 lauset. Märkasime, et kui korpusest olid SKM-id toonud autentselt väga erisuguseid lauseid, siis eeltreenitud SKM-ide laused olid alati n-ö kaunid ja eeskujulikud: kõik olid täislaused, mis algasid suure tähega ja lõppesid lauselõpumärgiga. Ühtegi kiillauset vms keerulisemat konstruktsiooni ei loodud.

Kui eeltreenitud SKM-ide puhul pole meil võimalust teada saada, kas need laused päriselt eksisteerivad või on need täielikult SKM-i enda genereeritud, siis korpuse kontekstide lauseid saime kontrollida (vt tabel 9). Enamjaolt olid kõik SKM-id lauseid originaalkujul esitanud, sealjuures on tähelepanuväärne, et enim lauseid toonud Gemini esitas vähim vigaseid lauseid (1 502 lausest 13 vigast).

Tabel 9. Olemasolevate ning väljamõeldud või vigaste lausete osakaal SKM-ide väljundis

SKM	Lauseid kokku	Olemasolevad laused	Väljamõeldud või vigased laused
Claude	742	725	17
Gemini	1502	1489	13
GPT	965	929	36

Lähemalt vaadates selgus, et GPT hallutsineeris ainsana lauseid juurde, ent seda üksnes sõna *kobima* tähendustele. Korpusest võetud lause "Poiss kobib kitse surnuks" malli järgi genereeris GPT viie näitelause kokkusaamiseks neid juurde, nt "Ta kobis vaenlase maha ühe hoobiga". Osade tähenduste juures kordasid SKM-id samu lauseid. Oleksime pidanud jääma pilootkatse promptis olnud sõnastuse juurde, et kui korpuses on vähem näiteid, siis toogu neid vähem.

Vigaste lausete põhjus seisnes enamasti selles, et SKM jättis originaalis sulgudes olnud konteksti vahelt ära või sulatas erinevad laused üheks, aga iseenesest oli seda tehes põhinenud korpuses olemas olnud lausetel. Samuti leidus pisimuudatusi, nagu sõnajärje muutmine, sõna lisamine või ärajätmine; osadel juhtudel SKM korrigeeris korpuselauseid (lisas või kustutas tühikuid, lisas lauselõpumärke).

4.4.2. Tsirkulaarne analüüs

Nii näitelausetest kui ka registri- ja märgendivaliku põhjendustest jäi meile silma, et korpusesse ja küllap ka SKM-ide treeningandmetesse on sattunud EKI avalikud materjalid, samuti muid tekste, kus arutletakse keelekasutuse üle. See on problemaatiline, sest muudab analüüsi tsirkulaarseks. Kuna API kaudu lähenedes näeb SKM ainult seda, mis selle treeningandmetes sisaldub (vt ka Jaama 2025), ei hinda SKM ilma prompti lisatud konkreetsete juhisteta seda, kui mõni tekst keelekasutuse kohta annab talle sõna kohta aegunud infot või kui ette satub EKSS-i materjal, mille näitelaused ei põhine tänapäeva keelekasutusel (EKSS 2009: 5). Inimene saab korpust analüüsides sellise info kõrvale jätta, ehkki ta ei saa täielikult välja lülitada oma keelekogemust ja eelteadmisi (Risberg jt 2025b: 345).

Niisiis võis selles katses juhtuda, et kui tahtsime uurida sõna tegelikku kasutust, uurime kogemata hoopis mõnda EKI varasemat sõnaraamatut või neile viitavat teksti. See tuli selgelt esile nt sõnade *turm*, *elik*, *kobima* puhul. Näiteks oli sõna *elik* näitelausete hulgas "ÕS oli määranud sõna 'elik' vananenud stiiliga sõnade hulka" ning kõik SKM-id olidki pakkunud tähendusele 'ehk; või' märgendit VANANENUD – nii on see ka EKSS 2009-s ja ÕS 2018-s, samas kui ÜS 2025-s ja ÕS 2025-s on märgend STIILITUNDLIK. Kuna sõna *elik* senini kasutatakse, ei saa seda vananenuks pidada.

Ühegi katsesõna kohta ei leidu korpuses aadressiga sonaveeb.ee vasteid. See tähendab, et päris sedasama, mis on katsesõnade kohta öeldud ÜS-is, kust need võeti, ei saanud SKM-id vähemalt korpusest leida (SKM-ide treeningandmetele ei pääse me ligi).

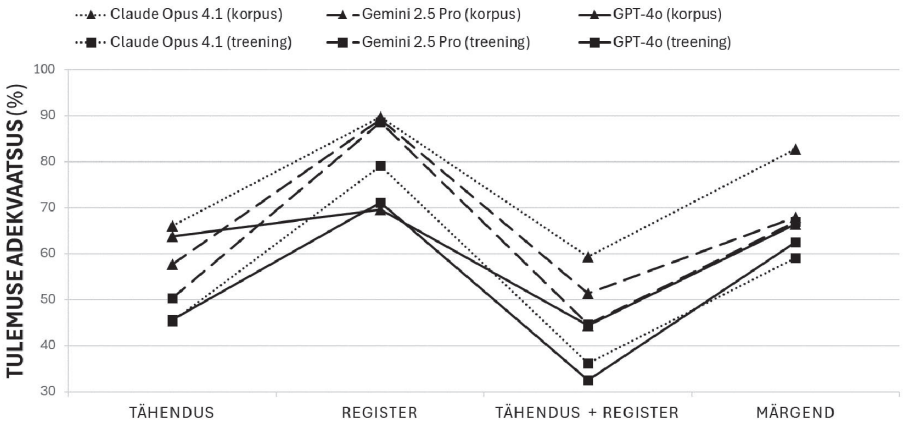
4.4.3. Keelekasutus

Eestikeelse teksti genereerimise kvaliteet on SKM-i väljundi juures tähelepanu vääriv tahk. Kvalitatiivselt oli üldpilt üsnagi hea, ehkki silma paistis mõningaid emakeelsele kõnelejale ebaloomulikke sõnastusi (nt "pikaaegses rahvasuus", "juustest, habemest või muudest sarnastest kogumikest"). Konkreetseid ortograafia-, morfoloogia- või ühildumisvigu leidus kogu andmestiku peale kokku sadakond (nt *informaale*, mitte *informaalne*; *jutustava*, mitte *jutustavat laadi*). Samuti oli üksikuid komavigu. Teisest küljest saab välja tuua, et näiteks sõna *tagaplaanile* (vrd *tahaplaanile*), *privilegeeritud* (vrd *priviligeeritud*) ja *taotlema* pöördevorme (*taotletakse*, vrd *taodeldakse*) kasutasid SKM-id kogu aeg (varasema) normingu päraselt.

5. Kokkuvõtvaid järeldusi

Sõnastikumärgendid peavad tuginema keeleandmete üldistusele. Viimastel aastakümnetel on saanud seda teha eesti keele korpust analüüsides, nüüd aga on võimalik katsetada, kas mahukate korpusandmete läbitöötamiseks saab kasutada SKM-e. Muuhulgas selleks, et vähendada subjektiivseid märgendiotsuseid, mis leksikograafi töö iseloomu arvestades paratamatult aeg-ajalt sisse libisevad. Artiklis kirjeldatud uurimusega tahtsimegi teada, milles ja mil määral saab SKM-ide väljundi adekvaatsuses kindel olla, kui SKM peab tuvastama sõnatähendusi ning määrama neile registri ja sõnastikumärgendi(d). Eesmärk oli siit katsest ammutada teadmisi SKM-ide võimekuse kohta ning leida aspektid, mida järgmise katse promptis täiendada.

Katses võrdlesime Claude Opus 4.1, Gemini 2.5 Pro ja GPT-4o väljundeid, mille küsisime nende API-de kaudu. Uurisime, kas eesti keele ühendkorpuse (2023) materjali etteandmine aitab SKM-ide väljundit parandada võrreldes üksnes eeltreeningus nähtud tekstidele toetumisega. Väljundi adekvaatsust tähenduste tuvastamisel ning neile registri ja märgendite määramisel hindasid meie töörühma liikmed.



Joonis 1. SKM-ide võrdlus sõnatähenduste, registrite ja märgendite adekvaatselt pakkumises

Selgus tõesti (vt joonis 1), et kõik SKM-id tuvastasid sõnatähendusi adekvaatsemalt, kui need said analüüsida korpuseandmeid, parima tulemuse andis Claude. Korpusmaterjalist sõna erisuguse (sh eesti keele sõnaraamatusse sobimatu) kasutuse tuvastamisega said SKM-id päris hästi hakkama. Nähtus siiski, et SKM-idel oli raskusi üldistamisega, tähendused aeti kohati liiga peeneks. Lisaks näitas meie eksperimentaalne katse, et SKM-idel (täpsemalt Claude'il) on tulevikus potentsiaali automaatselt hinnata tuvastatud tähenduste kattuvust ÜS-is esinevate tähendustega.

Eeltreenitud SKM-id genereerisid rohkem ebaadekvaatseid tähendusi, mis viitab SKM-ide kalduvusele luua ebakindlas olukorras oletuslikke vastuseid. Korpusandmetega rikastamine pakkus SKM-idele seega täiendavat konteksti ja empiirilist tuge. Võrreldes eelmise katsega (Risberg jt 2025b), kus SKM-id hindasid etteantud sõnatähenduste informaalsust, olid parimad SKM-id siinses katses, kus nad pidid tähendused ise määrama, märgatavalt adekvaatsemad.

Adekvaatselt pakutud sõnatähenduste registreid ehk (in)formaalsust tuvastasid SKM-id paremini kui tähendusi, eriti edukad olid Claude ja Gemini, sealjuures olenemata algandmetest. Samas määras neist parim, Claude, vaid 59% juhtudel adekvaatselt mõlemad, nii tähenduse kui registri. Adekvaatseid märgendeid sõnatähendustele pakkus Claude korpusandmete põhjal märkimisväärselt paremini kui GPT ja Gemini. Enim oli kõigil raskusi märgenditega VANANENUD ja HARV. Üldiselt aga näitas meie esimene vaatlus, et SKM-id on võimelised leidma suure osa märgendeid, mis ÜS-is tähendusele lisatud on, ehkki nad neid paljuski ka teisiti genereerivad.

Lisaks said kõik SKM-id päris hästi hakkama näitelausete korpusmaterjalist toomise ülesandega, kuigi vähesel määral leidus vigaselt esitatud ja ka hallutsineeritud lauseid. Eesti keelt genereerisid kõik SKM-id samuti üldiselt heal tasemel, ent leidus mõningaid vigu ja ebaloomulikke sõnastusi. Seega tuleb SKM-ide väljundit alati kontrollida. Tähelepanu väärib katsest selgunud probleem: kui SKM-id näevad algandmetes keelekasutust kommenteerivaid tekste või sõnaraamatute materjali, ei oska need hinnata, et see tuleks analüüsist kõrvale jätta (nagu jätab inimuurija). Promptis võib küll edaspidi anda juhise, et SKM nendele materjalidele ei toetuks, aga ei saa olla kindel, kas SKM seda juhist järgiks.

Kuigi mitmes aspektis tõdesime, et katses kasutatud andmestik on liiga väike kaugeleulatuvate järelduste tegemiseks, paistis 2025. a oktoobris, et leksikograafile võiks parimaks potentsiaalseks abiliseks märgenditetöös olla Claude – ja seda just koos korpuse kontekstiga ehk analüüsides autentseid kasutusandmeid.

Viidatud kirjandus

Anthropic. 2025. *System card addendum: Claude Opus 4.1*. https://www.anthropic.com/claude-opus-4-1-system-card (5.1.2026).

Biber, Douglas & Jesse Egbert. 2023. What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies* 5(1). 1–22. https://doi.org/10.1075/rs.00004.bib

Davies, Mark. 2025. Corpora and AI / LLMs: Genres. *Integrating AI / LLMs into English-Corpora.org*. https://www.english-corpora.org/ai-llms/genres.pdf (5.1.2026).

Eiche, Sandra, Reet Hendrikson, Eleri Aedmaa, Esta Prangel & Mari-Liis Tikerperi. 2025. Tehisintellekti rakendamine erialakeele arendamisel riigikaitseterminoloogia näitel. *Eesti Rakenduslingvistika Ühingu aastaraamat* 21. 29–45. https://doi.org/10.5128/ERYa21.02

EKSS 2009 = *Eesti keele seletav sõnaraamat* I–VI. "Eesti kirjakeele seletussõnaraamatu" 2., täiendatud ja parandatud tr. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks & Piret Voll (toim.). Tallinn: Eesti Keele Sihtasutus.

Fu, Tairan, Raquel Ferrando, Javier Conde, Carlos Arriaga & Pedro Reviriego. 2024. Why do large language models (LLMs) struggle to count letters? *arXiv*:2412.18626. https://doi.org/10.48550/arXiv.2412.18626

Github: EKKD-III1 registrite töörühma katsete materjalid. https://github.com/keeleinstituut/EKKD-III1/tree/main/registrid (5.1.2026).

Jaama, Meri-Kris. 2025. *Tehisintellekt, keelemudel, agent*. TI-hüppe koolitus, 22.8. https://www.youtube.com/watch?v=KIExl1p6b2k (5.1.2026).

Jürviste, Madis & Joonatan Jakobson. 2025. Vision-enabled LLMs in historical lexicography: Digitising and enriching Estonian-German dictionaries from the 17th and 18th centuries. *arXiv*:2510.07931. https://doi.org/10.48550/arXiv.2510.07931

Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala & Edwin Zhang. 2025. Why language models hallucinate. *arXiv*:2509.04664. https://doi.org/10.48550/arXiv.2509.04664

Karelson, Rudolf. 1990. "Eesti kirjakeele seletussõnaraamat" tegija pilgu läbi. *Keel ja Kirjandus* 32(1). 24–34.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1(1). 7–36. https://doi.org/10.1007/s40607-014-0009-9

Klosa-Kückelhaus, Annette & Carole Tiberius. 2024. The lexicographic process revisited. *International Journal of Lexicography* 38(1). 1–12. https://doi.org/10.1093/ijl/ecae016 (5.1.2026).

Koppel, Kristina & Jelena Kallas. 2022. Eesti keele ühendkorpuste sari 2013–2021: Mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat* 18. 207–228. https://doi.org/10.5128/ERYa18.12

Koppel, Kristina, Jelena Kallas, Madis Jürviste & Helen Kaljumäe. 2023. *Estonian National Corpus 2023*. Lexical Computing Ltd., Eesti Keele Instituut. https://doi.org/10.15155/3-00-0000-0000-0000-08C04M

Langemets, Margit, Mai Tiits, Udo Uibo, Tiia Valdre & Piret Voll. 2018. Eesti keel uues kuues. Eesti keele sõnaraamat 2018. *Keel ja Kirjandus* 60(12). 942–958. https://doi.org/10.54013/kk733a2

Lappin, Shalom. 2024. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information* 33. 9–20. https://doi.org/10.1007/s10849-023-09409-x (5.1.2026).

Laskar, Md Tahmid Rahman, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Md Amran Hossen Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty & Jimmy Xiangji Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13785–13816. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.764

Lillepalu, Helena Grete & Tanel Alumäe. 2025. Estonian native large language model benchmark. *arXiv*:2510.21193. https://doi.org/10.48550/arXiv.2510.21193

McIntosh, Timothy R., Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters & Malka N. Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv*:2402.09880. https://doi.org/10.48550/arXiv.2402.09880

OpenAI 2024 = GPT-4o system card. *arXiv*:2410.21276. https://doi.org/10.48550/arXiv.2410.21276

Pullum, Geoffrey K. 2023. Why grammars have to be normative – and prescriptivists have to be scientific. Joan C. Beal, Morana Lukač & Robin Straaijer (eds.), *The Routledge handbook of linguistic prescriptivism* (Routledge Handbooks in Linguistics), 3–16. London, New York: Routledge. https://doi.org/10.4324/9781003095125-2

Risberg, Lydia, Maria Tuulik, Margit Langemets, Kristina Koppel, Ene Vainik, Esta Prangel & Eleri Aedmaa. 2025a. Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel. *Keel ja Kirjandus* 67(7). 605–624. https://doi.org/10.54013/kk811a3 (5.1.2026).

Risberg, Lydia, Eleri Aedmaa, Maria Tuulik, Margit Langemets, Ene Vainik, Esta Prangel, Kristina Koppel & Hanna Pook. 2025b. The role of subjectivity in lexicography: Experiments towards data-driven labeling of informality. Iztok Kosem, Miloš Jakubíček, Marek Medveď, Karolina Zgaga, Špela Arhar Holdt, Tina Munda & Ana Salgado (eds.), *Electronic lexicography in the 21st century: Intelligent lexicography. Proceedings of the eLex 2025 conference*, 336–356. Bled: Lexical Computing CZ s.r.o. https://elex.link/elex2025/wp-content/uploads/eLex2025-22-Risberg_etal.pdf (5.1.2026).

Rundell, Michael. 2002. Good old-fashioned lexicography: Human judgement and the limits of automation. Marie-Hélène Corréard (ed.), *Lexicography and natural language processing: A festschrift in honour of B. T. S. Atkins*, 138–155. Stuttgart: EURALEX.

Salgado, Ana. 2025. A decade of lexicographic innovation at the Lisbon Academy of Sciences: Still from paper to digital platform… and then AI came. Paper presented at *1st International Conference on Lexicology and Lexicography*, Budapest, 29.9.–1.10.

Tehisaru baromeeter. https://baromeeter.tartunlp.ai/ (5.1.2026).

Trap-Jensen, Lars. 2002. Descriptive and normative aspects of lexicographic decision-making: The borderline cases. Anna Braasch & Claus Povlsen (eds.), *Proceedings of the 10th EURALEX International Congress*, 503–509. København: Center for Sprogteknologi.

Trap-Jensen, Lars. 2024. The best of two worlds: Exploring the synergy between human expertise and AI in lexicography. *Proceedings of the International Conference Lexicography in the XXI Century*. https://lexicography21.iliauni.edu.ge/wp-content/uploads/2024/06/03_Lars-Trap-Jensen.pdf (5.1.2026).

Tuulik, Maria, Ene Vainik, Esta Prangel, Margit Langemets, Eleri Aedmaa, Kristina Koppel & Lydia Risberg. 2025. Tähenduste seletamine leksikograafias: Kuivõrd on abi suurtest keelemudelitest? *ESUKA–JEFUL* 16(2). 147–176. https://doi.org/10.12697/jeful.2025.16.2.05

Vaik, Kristiina. 2024. *Beyond genres: A dimensional text model for text classification* (Dissertationes linguisticae Universitatis Tartuensis 47). Tartu: Tartu Ülikooli Kirjastus.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder & Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv*:2402.05672. https://doi.org/10.48550/arXiv.2402.05672

Wang, Ruiqi, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong & Xin Xia. 2025. Can LLMs replace human evaluators? An empirical study of LLM-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering* 2 (ISSTA), 1955–1977. New York: Association for Computing Machinery. https://doi.org/10.1145/3728963

ÕS 2018 = *Eesti õigekeelsussõnaraamat ÕS 2018*. 2018. Maire Raadik (toim.). Tiiu Erelt, Tiina Leemets, Sirje Mäearu, Maire Raadik (koost.). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

ÕS 2025 = *Eesti õigekeelsussõnaraamat ÕS 2025*. 2025. Eesti Keele Instituut. Tallinn: EKSA.

ÜS 2025 = *EKI ühendsõnastik 2025*. Eesti Keele Instituut, Sõnaveeb. https://sonaveeb.ee/ (5.1.2026).

WITH OR WITHOUT CORPUS DATA? LARGE LANGUAGE MODELS' CAPABILITY TO DETECT AND LABEL WORD SENSES IN ESTONIAN

Lydia Risberg[1, 2], Eleri Aedmaa[1], Hanna Pook[1], Kristina Koppel[1], Maria Tuulik[1], Esta Prangel[1], Margit Langemets[1]

Institute of the Estonian Language[1], University of Tartu[2]

Lexicographers occasionally face difficulties in deciding whether a particular word sense requires a usage label (e.g., COLLOQUIAL). In descriptive lexicography, labels are based on empirical language data rather than on intuition. In recent decades, corpus data has become a reliable foundation for such decisions, replacing earlier reliance on individual judgment or small card indexes. However, the potential of large language models (LLMs) to assist in labeling task has only recently begun to be explored.

This study investigates whether and how LLMs can support Estonian lexicographers in assigning dictionary labels to word senses. We compared the performance of three models – Claude Opus 4.1, Gemini 2.5 Pro, and GPT-4o – when asked to identify Estonian word senses and assign register and label information, using either selected contexts from the *Estonian National Corpus* (2023) that contained the target word or in a zero-shot setting. The LLMs' outputs were evaluated both manually and through automated analysis.

All LLMs performed more accurately when corpus data were provided, with Claude yielding the best overall results. LLMs generally handled register assignment (e.g., informal vs. neutral/formal) more effectively than sense identification. Although the adequacy of generated meanings varied, LLMs demonstrated a promising ability to detect usage variation in authentic Estonian corpus material. Claude also showed potential for automatically matching its identified senses with those in the *EKI Combined Dictionary*.

While LLM outputs still require expert supervision, the results suggest that LLMs can assist lexicographers in reducing subjectivity and workload when determining usage labels. As of October 2025, Claude appears to be the most promising tool for Estonian.

Keywords: large language models, Estonian National Corpus, lexicography, register, usage labels, Estonian

Lydia Risberg (Eesti Keele Instituut, Tartu Ülikool) uurib suuri keelemudeleid eesti keele registrite aspektist ja tegeleb keeleteaduse populariseerimisega. Varem on uurinud korpuspõhiselt eesti keelekorralduses antud tähendussoovitusi, toetudes kasutuspõhisele keeleteooriale.
Munga 18, 50088 Tartu, Estonia
lydia.risberg@eki.ee

Eleri Aedmaa (Eesti Keele Instituut) on juhtiv-keeletehnoloog, kelle põhitegevused on seotud keeleandmete töötlemise ja suurte keelemudelitega. Varem uurinud statistiliste meetodite rakendamist keeleteaduses.
Roosikrantsi 6, 10119 Tallinn, Estonia
eleri.aedmaa@eki.ee

Hanna Pook (Eesti Keele Instituut) on keeletehnoloog, varasemalt töötanud ka leksikograafina murdesõnastike koostamisel. Peamisteks huvialadeks on dialektoloogia, süntaks ja keele varieerumine.
Munga 18, 50088 Tartu, Estonia
hanna.pook@eki.ee

Kristina Koppel (Eesti Keele Instituut) on EKI ühendsõnastiku töörühma liige ning kahe EKI teadusprojekti (PRG1978, EKKD-III1) põhitäitja. Varem vastutanud eesti keele ühendkorpuse sarja koostamise ja ilmumise eest. Põhilised uurimisvaldkonnad: korpuslingvistika, korpusleksikograafia.
Roosikrantsi 6, 10119 Tallinn, Estonia
kristina.koppel@eki.ee

Maria Tuulik (Eesti Keele Instituut) on EKI teadusprojekti EKKD-III1 vastutav täitja ja EKI ühendsõnastiku töörühma liige. Põhilised uurimisvaldkonnad on semantika, korpuslingvistika ja leksikograafia.
Roosikrantsi 6, 10119 Tallinn, Estonia
maria.tuulik@eki.ee

Esta Prangel (Eesti Keele Instituut) on Ekilexi tehniline tootejuht, kes tarkvaraarenduse juhtimise kõrvalt varustab keeleteadlasi leksikograafiliste andmetega. Varem tegutsenud soomeugrinduse, kirjastamise ja tarkvaraarenduse vallas.
Roosikrantsi 6, 10119 Tallinn, Estonia
esta.prangel@eki.ee

Margit Langemets (Eesti Keele Instituut) on õigekeelsussõnaraamatu ÕS 2025 ja EKI ühendsõnastiku töörühma juht. Peamised huvialad on leksikoloogia ja korpuslingvistika.
Roosikrantsi 6, 10119 Tallinn, Estonia
margit.langemets@eki.ee