ESUKA – JEFUL 2025, 16–2: 147–176

TÄHENDUSTE SELETAMINE LEKSIKOGRAAFIAS: KUIVÕRD ON ABI SUURTEST KEELEMUDELITEST?

Maria Tuulik[a], Ene Vainik[a], Esta Prangel[a], Margit Langemets[a], Eleri Aedmaa[a], Kristina Koppel[a], Lydia Risberg[a, b]
[a] *Eesti Keele Instituut, EE*
[b] *Tartu Ülikool, EE*
maria.tuulik@eki.ee, ene.vainik@eki.ee, esta.prangel@eki.ee, margit.langemets@eki.ee, eleri.aedmaa@eki.ee, kristina.koppel@eki.ee, lydia.risberg@eki.ee

Kokkuvõte. Artiklis uurime, kas ja mil määral saavad suured keelemudelid (SKMid) leksikograafi aidata sõnastikuseletuste loomisel ja sõnatähendus(t)e esitusel. Katsetatud SKMidest jäid lõplikku hindamisküsitlusse GPT-4o ja Gemini 1.5 Pro tähendusseletused 60 sõna kohta. Kasutatud näitetu promptimise (*zero-shot*) meetod võimaldas selgitada leksikograafide eelistusi ja nõudmisi seletustele ilma nende vormi ennetavalt piiramata. Tähendusseletusi hindas 21 leksikograafi.

Vastajate üldine rahulolu oli suhteliselt kõrge: leksikograafid hindasid oma töö jaoks väga kasulikuks või mõnevõrra kasulikuks 83% seletustest. Mudelite sooritus eesti keeles enamasti rahuldas vastajaid: 75% juhtudel hinnati seletuste keelt grammatiliselt korrektseks. Rahulolu semantilise esitusega oli mõnevõrra madalam: väljatoodud tähendusi peeti õigeks 56% juhtudel. Kahest SKM-ist hinnati kõrgemalt GPT-4o tähendusseletusi. Leksikograafide hinnangud varieerusid. Hinnangutele lisatud kommentaaride analüüs selgitas leksikograafide rahulolu/rahulolematuse sisemist struktuuri, millele toetudes on edaspidi võimalik luua täpsemaid prompte eri tüüpi sõnastikuseletuste tarvis.

Märksõnad: suured keelemudelid, eesti keel, leksikograafia, leksikaalne semantika, tähenduse seletused, polüseemia

DOI: https://doi.org/10.12697/jeful.2025.16.2.05

1. Sissejuhatus

Tänapäeva leksikograafias on tööprotsessid muutunud: paljudes maades on võetud suund ühtsele struktureeritud andmebaasile, mis on

kaasa toonud leksikaalse info koostamise kihiti, rööpsete töövoogudena (vrd varem koostati eri tüüpi sõnastikke omaette sõnastikubaasides) (Klosa-Kückelhaus & Tiberius 2024; Langemets jt 2021; Steurs jt 2020; Tavast jt 2018). See on pannud leksikograafe otsima võimalusi automatiseerida oma tööprotsessi erinevaid lülisid (märksõnade tuvastamist, sõnatähenduste seletamist, grammatika ja kasutusmustrite kirjeldamist jmt). Samal ajal on oluline säilitada kvaliteet – automaatselt või poolautomaatselt genereeritud teave peab olema vähemalt samal tasemel leksikograafi töötulemusega või isegi parem (nt süstemaatilisem).

Metaleksikograafilisi uurimusi eesmärgiga välja selgitada, millised on leksikograafilised protsessid ja millised on selle töö tegijate ootused ja vajadused erinevate tehniliste abivahendite suhtes, on tehtud rahvusvaheliselt (Kallas jt 2019), Eestis aga seni vähem (vt nt Paulsen, Vainik & Tuulik 2020). Korpuste kasutamine on (mh Eesti) leksikograafias igapäevane, aga kuivõrd korpustööriistadega genereeritud andmestike järeltoimetamine on samuti töömahukas, otsitakse ka leksikograafias agaralt võimalusi suurte keelemudelite (SKMid) kasutamiseks. Sellealaseid eksperimente on raporteeritud viimastel leksikograafiakonverentsidel (nt Evert, Ganslmayer & Rink 2024; Kosem, Gantar jt 2024; Kosem, Kuhn jt 2024; Tadić 2024; Tiberius jt 2024). Uute tehniliste võimaluste (nagu SKMid) avanedes soovitakse välja selgitada, kuivõrd hästi või halvasti saaks neid mõnes töölõigus rakendada (nt De Schryver 2023; Jakubíček & Rundell 2023). Parimaid tulemusi inglise keele jaoks on saadud ChatGPT abil just sõnaseletustega (De Schryver & Joffe 2023; Lew 2023). SKMide võimekust eesti keele sõnatähendusi seletada varem uuritud ei ole, ent eesti keele kohta on leitud näiteks seda, et SKM on võimeline suhteliselt väikese konteksti põhjal mõistatama kunstliku tehissõna (*placeholder*) tähenduse (Karjus 2023: 23) ning praegu on käsil esimesed leksikograafiaalased katsetused (Jürviste, Paet & Soosaar 2025).

Artiklis uurime, kas eesti keele kui väikese kõnelejaskonnaga ning mudelites tekstimahult alaesindatud keele puhul saavad keelemudelid olla leksikograafile abiks sellises olulises töölõigus nagu sõnatähenduste seletamine ning kuidas hindavad leksikograafid kui kvaliteedi kontrollijad ja tagajad selliselt genereeritud sisu. Kirjeldame katset, mille käigus valisime välja kaks keelemudelit – Gemini 1.5 Pro ja GPT-4o –, millel

lasksime genereerida seletused 60-le EKI ühendsõnastiku (ÜS) märksõnale. Seejärel hindasid loodud seletusi leksikograafid.[1]

Kuna SKMide kasutamine leksikograafias on uus suund ja puuduvad varasemad uurimused eesti keele tähenduste seletuste kohta, siis kaardistab siinne töö leksikograafide hinnangud, selgitab SKMide kasutuspotentsiaali ja usaldusväärsust eesti sõnade seletamisel ja pakub võrdlusalust tulevastele töödele.

1.1. Suured keelemudelid

SKMid on keeletehnoloogia ja laiemalt tehisintellekti üks märkimisväärsemaid saavutusi, mis on fundamentaalselt muutnud meie arusaama arvutuslikust keeletöötlusest (Ignat jt 2024: 8050). Need mudelid põhinevad transformeri arhitektuuril (Vaswani jt 2017), mis esindab olulist paradigmamuutust varasemate lähenemistega võrreldes. Transformeri arhitektuuri keskmes on enesetähelepanumehhanism, mis võimaldab mudelil dünaamiliselt analüüsida teksti sisemisi seoseid. Erinevalt traditsioonilisest järjestikusest töötlusest suudab see mehhanism haarata teksti terviklikult, arvestades kõigi elementide omavahelisi suhteid samaaegselt (Vaswani jt 2017). See on võrreldav inimese võimega mõista lause tähendust, võttes arvesse kõiki selle komponente korraga, mitte järjestikku sõnahaaval.

SKMide toimimisprintsiip lähtub distributiivse semantika põhimõttest, mille kohaselt tuleneb sõnade tähendus nende kasutusmustrist (Grindrod 2024: 3–4). See avaldub kontekstuaalses kodeerimises, kus näiteks sõna *pank* tähendus lauses „Pank asub jõe kaldal“ kodeeritakse seoses sõnadega *jõgi* ja *kallas*, samas kui lauses „Lähen panka raha viima“ saab see erineva numbrilise esituse ehk vektoresituse. SKMi vektorruumis paiknevad sõnad vastavalt nende distributiivsetele omadustele – sarnases kontekstis esinevad sõnad (nt *arst*, *õde*, *haigla*) asuvad ruumis lähestikku. See ruumiline organiseeritus võimaldab mudelil mõista semantilisi suhteid ja analoogiaid.

1 Küsimustik ja katsesõnade nimekiri on kättesaadavad uurimisrühma repositooriumis: https://github.com/keeleinstituut/EKKD-III1/tree/main/tähendused_ja_seletused (20.01.2025).

OpenAI (Chat)GPT[2] polnud esimene suur keelemudel, kuid selle kasutamise lihtsus tõi kaasa konkurentide soovi ka enda keelemudelid kasutajateni tuua. Sellest on välja kujunenud olukord, kus regulaarselt luuakse üha paremaid mudeleid ning SKMide turg on muutunud mitmekesiseks, eriti ingliskeelsete mudelite seas. Ka eesti keelt (heal tasemel) töötlevaid ja genereerivaid mudeleid on mitmeid, sh OpenAI GPT-d, nt GPT-4o (OpenAI 2024), Google'i Gemini mudelid, nt Gemini 1.5 (Gemini Team 2024) ja Anthropicu Claude (Anthropic 2024) mudelite erinevaid versioonid. Täpse versiooni valiku määrab suures osas rakendamise aeg, sest üldjuhul on uuemad mudelid paremad kui eelnevad. Suurtootjad oma mudelite kohta väga palju detaile ei jaga, kuid võib arvata, et eesti keel on neisse sattunud juhuslikult. Seda peab mudelite rakendamisel kindlasti silmas pidama.

SKMide, eriti baasmudelite rakendamisel ja genereeritud tulemuste analüüsil tuleb arvestada ka mudelite muude omapäradega. Põhimuredena tuuakse välja genereeritud sisus esinevaid potentsiaalselt kahjustavaid üldistusi mõne inimgrupi suhtes, väljamõeldud või ebatäpse sisu tootmist (nn hallutsinatsioone), läbipaistmatust (tulemusi on keeruline seletada), loogikareeglite eiramist, treeningandmete ühekülgsust, tulemuste muutlikkust, inglise keele ja kultuuriruumi eelistamist teistele keeltele ja kultuuriruumidele, arvutusvõimsuse tagamisega seotud keskkonnakahjud jms (vt nt Bender jt 2021; Brown jt 2020; De Schryver 2023; Gallegos jt, 2024; Hadi jt 2023; Liang jt 2021; Navigli jt 2023; Tao jt 2024).

Kui varasemate mudelite rakendamine eeldas nende treenimist spetsiifilise ülesande täitmiseks, siis tänapäevaste keelemudelite hea keeleoskus ja üldistusvõime võimaldab neid paljudeks ülesanneteks rakendada vaid ülesande kirjeldamise abil (Brown jt 2020). Olenevalt ülesande keerukusest võib mudel parema tulemuse saavutamiseks vajada ka näiteid. Vastavalt näidete arvule jagatakse promptimistehnikad üldjoontes kolmeks: näitetuks (*zero-shot*), ühe näitega (*one-shot*) ja mõne näitega promptimiseks (*few-shot*).

Mudelitel on piisavalt tõenäosuslikke teadmisi loomulike keelte sõnade tähendustest, kuivõrd nad suudavad mõista ja moodustada mõtestatud teksti. Need keele töötlemises ja produtseerimises avalduvad teadmised on mingis mõttes võrreldavad inimese keeleoskusega.

2 https://chat.openai.com/chat (20.11.2024).

Ülesanne genereerida sõnatähenduste selgitusi eeldab ka metalingvistilisi teadmisi, st võimet formuleerida seisukohti keeleüksuste endi kohta. Tõenäosuslikud teadmised tähendustest tuleb kombineerida ootuspärase seletusviisiga ja just seda võimet paneb proovile meie eksperiment sõnatähenduste seletuste genereerimiseks.

1.2. Sõnatähendused leksikograafias

Mahuka üldkeele sõnaraamatu siht on kirjeldada (kõiki) keeles esinevaid sõnu koos (kõigi) tähendustega. Loomulikus keeles esinevad sõnad alati kontekstis, nii on väidetud, et „mitte ükski sõna ei tähenda kaks korda täpselt üht ja sama“ (Hayakawa & Hayakawa 2022: 79; vt ka Alexander 2015; Langemets & Risberg 2022), samuti on teada, et pole olemas tõsikindlat või vähemalt üldaktsepteeritud meetodit, kuidas tähendusi määratleda. Sellest hoolimata on olemas sõnaraamatud, mis tähendusi kirjeldavad – üldistades, millistes tähendustes (kontekstides) on sõnu seni teadaolevalt kasutatud. See ülesanne sarnaneb paljuski SKMide toimimisprintsiibiga (ptk 1.1), kus sõnatähendus sõltub kasutusmustrist, kontekstist.

Tähenduste seletamisel on leksikograaf alati tuginenud keeleandmetele, mingil viisil dokumenteeritud kasutusjuhtumitele, olgu siis sedelkartoteegile või tekstikorpusest päritud konkordantsiridadele, kust kummastki aimub ka kontekst. Enamasti on kasutada olnud ka varasemad sõnastikud, milles esitatud tähendusi on vajadusel ajakohastatud. Töös oleva sõnaraamatu põhimõtted on tavaliselt kirja pandud stiiliraamatusse või koostamisjuhendisse. Tähenduse seletamisel on oluline nii sisu (nt sihtrühma valik, toetumine mingile lingvistikateooriale, tähenduste eristamise määr) kui ka vorm, s.t tähenduste seletamise (sõnastamise) ja esitamise viisid.

Üldkeele sõnaraamatute puhul on eelistatud terminit *tähenduse seletus* (mitte *definitsioon*), kuivõrd see vastab paremini sellele, mida leksikograafid teevad: seletavad keelekasutajale keelenäidete põhjal, mida sõna mingis kontekstis tähendab või kuidas seda mingis kontekstis on tava kasutada. Leksikograafia eripära on, et *sõnu* seletatakse (teiste) *sõnade* abil: tähenduse seletus on kirjeldatava sõna (leksikaalse üksuse) vähim, teatud tingimusi rahuldav ümbersõnastus sama keele teiste sõnade abil (Mel’čuk & Polguère 2018). „Teatud tingimused“ on

kokkulepped, mis ühe või teise sõnaraamatu koostamiseks ja toimetamiseks on tehtud.

Ükskeelses leksikograafias on kasutusel mitmeid seletuse tüüpe ja malle:

a) traditsiooniline tähenduse seletus, mis võimalikult ammendavalt määratleb mõiste sisu (näited 1–2): hrl ülemmõiste koos lisainfoga, mis võimaldab eristada teistest sama kategooria mõistetest. (Terminibaasides nimetatakse seda *definitsiooniks*.) Sobib selgemini piiritletavate kategooriate puhul (nt nimi-, omadus-, tegusõnad), aga ei sobi nt funktsioonisõnade puhul, kus ülemmõistet pole võimalik määratleda (nt partiklid);
b) sünonüümseletus (näited 3–4). Sünonüümseletusi pole üldiselt pooldatud (vt nt Atkins & Rundell 2008: 421), kuna sünonüümid on ise sageli polüseemsed (näited 4–5). Sünonüümseletuse puhul võivad kergesti tekkida ringdefinitsioonid ja tähendus jäädagi seletamata. Traditsioonilise (sõnapõhise) koostamise korral võib kaotsi minna sünonüümsuhte vastastiksuhe (kui A = B, siis B = A);
c) täislausega seletus (näited 6–7). Sõnatähendus on seletatud täislause, mitte fraasi abil. Seda tüüpi seletusi on kasutatud nt õppesõnastikes, kus metakeel peab olema lihtsam;
d) tähendusvihje (näited 8–9). Tähendust pole seletatud, vaid üksnes ligikaudu piiritletud. Seda tüüpi seletused on sageli kasutusel mitmekeelsetes sõnaraamatutes, Eestis ka õigekeelsussõnaraamatus.
   (1) **kupee** eraldatud osa vagunist, kus hrl lisaks istmetele on ka magamisasemed (ÜS)
   (2) **vanduma** midagi kindlalt lubama või tõotama, vandega kinnitama (ÜS)
   (3) **ajaldine** õigeaegne (EKSS)
   (4) **toekas** suur, tugev, kogukas, kaalukas (ÕS 2018)
   (5) **elavdama** elava(ma)ks (5., 6. täh.) muutma (EKSS)
   (6) **school** A **school** is a place where children are educated. (Collins Cobuild)
   (7) **pilvine** kui ilm on **pilvine**, siis on taevas palju pilvi (Keeleõppija Sõnaveeb)
   (8) **muffin** keeksike (ÕS 2018)
   (9) **veelaud** (majal); (spordivahend) (ÕS 2018)

Keelemudelite vastustes kasutatud seletuste tüüpe vaatleme ptk 3.6.

2. Katse: keelemudelite loodud seletuste hindamine

Katse peamine eesmärk oli hinnata, kas ja mil määral saavad SKMid aidata leksikograafidel sõnatähendusi seletada. Pöörame tähelepanu ka polüseemia esitamisele. SKMide ülesanne oli genereerida seletused valitud sõnadele (2.1). Eeltööna valisime paljudest SKMidest välja kaks keelemudelit, seejärel seadistasime valitud mudelid ja koostasime prompti (2.2). SKMide loodud seletusi palusime hinnata leksikograafidel (2.3) anonüümses veebiküsimustikus, vastates kolmele küsimusele: a) „Kas antud tähendus(ed) on sinu arvates või sõnastiku põhjal õige(d)/olemasolev(ad) tähendus(ed)?“ (tähenduste tabamine), b) „Kas tähenduse seletus on grammatiliselt korrektses keeles?“ (keeleline korrektsus), c) „Kui kasulikuks hindad mudeli pakutud tähenduse seletust sinu kui leksikograafi jaoks?“ (kasulikkushinnang).

2.1. Sõnavalim

Kuna katse üks eesmärke oli uurida, kuidas tulevad SKMid toime polüseemsete sõnadega, sõelusime välja märksõnad, mil oli 1–4 tähendust (arvesse läksid nii põhi- kui ka alltähendused). Valimis oli 60 sisusõna (nimisõnad, omadussõnad, tegusõnad ja määrsõnad): igast sõnaliigist 15 sõna, millest 5 ühe tähendusega, 5 kahe tähendusega, 3 kolme tähendusega ja 2 nelja tähendusega. Katses esindatud sõnad valisime juhumeetodil EKI ühendsõnastikust (ÜS). Võõrsõnu sattus valimisse 15 ning omasõnu 45.

2.2. Keelemudelite valik ja promptimine

SKMid, mida tähenduse seletuste loomiseks kasutada, valisime eelkatsetuste põhjal välja 2024. aasta aprillis ja mais. Vabavaralistest SKMidest testisime Llama 3, mittevabavaralistest GPT-4, GPT-3.5-turbo, GPT-4o, Gemini 1.5 Pro ja Claude 3 Sonnet’d. Leksikograafitöö kogemusega uurimisgrupi liikmed tegid nendega esmased katsetused ja valisid edasiseks tööks välja GPT-4o (edaspidi GPT) ja Gemini 1.5 Pro (edaspidi Gemini), kuna need osutusid eesti keele sõnade tähendusseletuste genereerimisel kõige edukamaks. Täpsemaks häälestamiseks määrasime GPT temperatuuriks 0,2 (skaalal 0,0–1,0) ja Gemini omaks 1,0 (skaalal 0,0–2,0). Temperatuuriga saab mõjutada mudeli loovust –

mida kõrgem temperatuur, seda erinevamaid vastuseid võib sama prompti vastuseks saada (Agarwal jt 2024). Madalama temperatuuriga ei andnud Gemini häid vastuseid ja soovis konteksti juurde. Väljundi pikkuseks määrasime mudelite maksimaalsed lubatud pikkused: GPT jaoks 4086 ja Gemini jaoks 8192 sõnet. Kahe eri mudeli puhul ei pruugi üks ja sama temperatuuri väärtus tähendada samaväärseid tulemusi. Eri väärtuste mõju SKMide tulemustele vajab eraldi uurimustööd.

Prompt SKMidele oli järgmine: „Mida tähendab eesti keeles … (katsesõna)? Mitmetähenduslikele sõnadele võid anda mitu tähendust. Kui sa ei tea, siis ütle, et sa ei tea.“ Prompt ei sisaldanud näiteid (*zero-shot* meetod) ega täpsemaid juhiseid tähendusseletuse koostamiseks, kuna soovisime uurida leksikograafide eelistusi ilma tähendusseletuse sisu ja vormi ennetavalt piiramata.

Kahe keelemudeli genereeritud tähenduste seletused olid esitatud küsimustikus alati kõrvuti, kuid juhuslikus järjekorras, et vältida ühe vastuse eelistamist positsiooni põhjal. Kõik vastajad said sama küsimustiku.

2.3. Hindajad

SKMide pakutud vastuseid hindas 21 Eesti Keele Instituudi (EKI) leksikograafi, kelle töökogemus varieerus 3 aastast 43 aastani (keskmine 19, standardhälve 13,1). Hindajad vastasid keelemudelite (GPT ja Gemini) vastuseid puudutavale kolmele küsimusele (tähenduste tabamine, keeleline korrektsus, kasulikkushinnang). Lisaks küsisime nende sõnastikukogemuse ja seletusloome praktika kohta (st milliste sõnaraamatutega ja mis tüüpi tähendusseletustega on nad varem töötanud), kuna eeldasime, et see võib mõjutada genereeritud tähendusseletustele antavaid hinnanguid. 14 vastajat on töötanud ükskeelsete sõnastikega ning 7 vastajat terminibaasidega. Pooltel vastajatest oli kogemus ise tähendusseletuste loomisega ja pooltel seletuste otsimisega muudest allikatest.

2.4. Analüüsi meetodid

Küsimustikuga kogutud skalaarseid (valikvastustega) andmeid analüüsisime kvantitatiivselt erinevate vastuste osakaaludena ja viisime läbi korrelatsioonianalüüsi hinnanguküsimuste ja mitmesuguste tegurite suhtes, mis iseloomustasid sõnavalimit ja vastajaid; vastajatevahelist üksmeelt mõõtsime Fleissi kappa abil. Tähenduste tabamise küsimuse

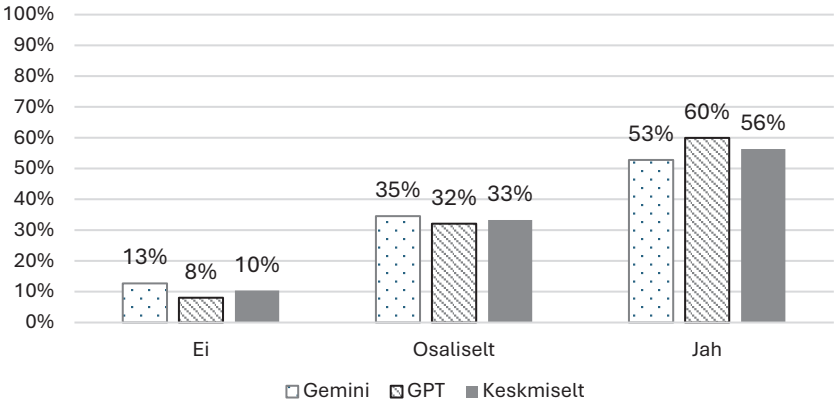
vastuste tõlgendamisel viisime lisaks läbi võrdleva analüüsi ühendsõnastiku tähendusjaotusega. Kvalitatiivse analüüsi osas võrdlesime keelemudelite tähendusseletusi ka sõnastikes üldiselt levinud seletusmallidega ning vabade kommentaaride lahtris olnud vastuseid analüüsisime sisuanalüüsi meetodiga.

3. Tulemused

Küsimustik sisaldas nii valikvastustega küsimusi kui lahtrit vabade kommentaaride jaoks (vt 1. allmärkust). SKMide seletusi võisid leksikograafid vajadusel võrrelda EKI ühendsõnastiku esitusega. Esmalt teeme ülevaate keelemudelitele antud skalaarsetest hinnangutest: leksikograafid hindasid tähenduste tabamist (3.1), keelelist korrektsust (3.2) ja seletuse kasulikkust (3.3). Seejärel kommenteerime hinnanguküsimuste vastuste varieerumist (3.4) ja analüüsime uuritud tegurite (skalaarsed hinnangud, seletatava sõna/üksuse mitmesõnalisus, polüseemia, korpusesagedus ja võõrsõnalisus) korrelatsiooni, et selgitada tegureid, mis võivad SKMide väljundi kvaliteeti mõjutada (3.5). Kirjeldame ka keelemudelite pakutud seletuste tüüpe (3.5) ning viimaks analüüsime leksikograafide jäetud vabas vormis kommentaare (3.6). Hinnangutele lisatud kommentaarid võimaldavad vaadelda täpsemalt, mis just tähendusseletuste puhul on need aspektid (nt puuduv või üleliigne tähendus, kirjelduse pikkus, toodud näidete prototüüpsus või ebaloomulikkus), mis mõjutavad leksikograafide rahulolu uue tööriistaga.

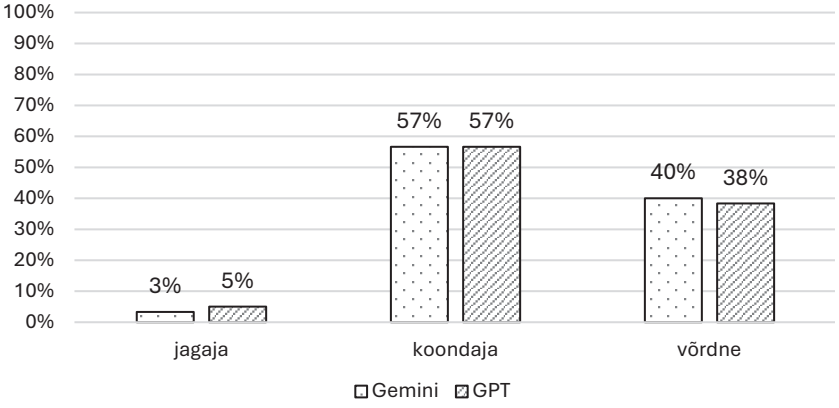
3.1. Tähenduste tabamine

Tähenduste tabamise küsimusele antud vastuste jaotumus on esitatud joonisel 1. Nagu näha, on rohkem kui pooltel kordadel hinnatud SKMe tähendusi tabavaks: keskmiselt 56% juhtudest on leitud, et pakutud tähendused on õiged (tuginedes kas leksikograafi kogemusele või siis võrreldes sõnastikuesitusega) ja 33% juhtudest leiti, et tähenduste esitus oli osaliselt õige. Pisut suurema rahulolu (*ca* 60% juhtudest) pälvisid selles suhtes GPT poolt antud vastused; samal mudelil on ka väiksem protsent juhtumeid (8%), mispuhul osutati, et tähendused ei ole õiged. Võrdluses olnud Gemini sai valede tähenduste protsendiks 12,7%; pisut GPT-st rohkem pakkus ta ka vastuseid, mispuhul leiti, et tähendused on esitatud osaliselt õigesti.



Joonis 1. Vastused küsimusele „Kas antud tähendus(ed) on sinu arvates või sõnastiku põhjal õige(d)/olemasolev(ad) tähendus(ed)?“

Võrdlesime SKMide pakutud ja sõnastikus esile toodud tähenduste arvu, mis võimaldas vaadelda, kas mudelid võrreldes sõnastikuga (ÜS) pigem koondavad tähendusi või jagavad neid peenemalt (vt joonis 2). GPT pakkus tähenduste arvu ÜSiga võrdselt 23 sõnale (38,3% juhtudest) ja Gemini 24 sõnale (40% juhtudest). Enamasti olid need monoseemsed sõnad (nt *sutureerima*, *kommentaarituba*, *natukese*). Polüseemsete sõnade puhul oli võrdne arv tähendusi Geminil 7 kahetähendusliku ning 1 kolmetähendusliku sõna puhul, GPT-l vastavalt 6 ja 0.



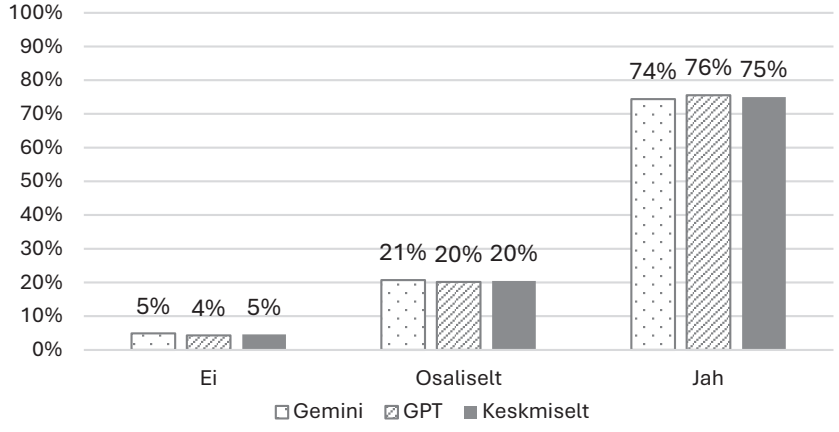
Joonis 2. SKMide ja sõnastiku (ÜS) tähenduste arvu võrdlus.

Nagu joonisest 2 nähtub, kipuvad SKMid tähendusi pigem koondama. Nii GPT kui ka Gemini pakkusid võrdselt ÜSi 34 mitmetähenduslikule sõnale vaid ühe tähenduse, võttes kokku nii kahe (nt *kostitama*), kolme (nt *kehv*), nelja (nt *hapu*) kui ka viie (nt *kaugelt*) tähendusega sõnade selgitused. Sõnastikust detailsemalt selgitas GPT 3 ja Gemini 2 sõna tähendust. GPT pakkus ühetähenduslikule sõnale *käitama* 2 ning kahetähenduslikele sõnadele *sardiinlane* ja *laetud* vastavalt 3 ja 4 tähendust; Gemini pakkus 2 tähendust ühetähenduslikele sõnadele *käitama* ja *teada olema*.

Testsõnade seas oli ka kolm homonüümset sõna (*mure*, *salk* ja *jõletu*, igaühel ÜSis 2 homonüümi). Sõna *salk* puhul tuvastasid mõlemad SKMid mõlema homonüümi põhitähendused (’inimeste hulk’, ’(juukse) kimp’). Mõlemad mudelid pakkusid sõnale *jõletu* vaid ühe tähenduse, *murele* pakkus Gemini ühe ja GPT kaks tähendust (jagades ühe homonüümi tähendusi peenemalt, samas jäi teine homonüüm tuvastamata).

3.2. Keeleline korrektsus

Keelemudelite loodud seletuste keelelist korrektsust hindasid leksikograafid nii üldiselt kui mõlema mudeli osas kõrgemalt kui tähenduste tabamist (vt 3.1): keskeltläbi kolmveerand vastustest (75%) loeti grammatiliselt korrektseks (vt joonis 3).

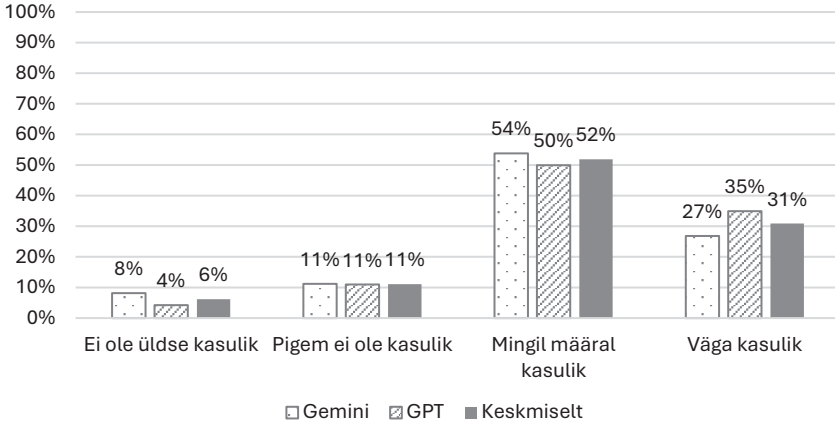


Joonis 3. Vastused küsimusele „Kas tähenduse seletus on grammatiliselt korrektses keeles?“

Selle küsimuse otstarve oli teada saada, kuivõrd suudavad valdavalt inglise keele peal treenitud keelemudelid anda eesti keeles veatuid ja loetavaid vastuseid. Protsent on kõrge ja teadaolevalt tehakse ka pingutusi eesti keele andmete ulatuslikumaks kaasamiseks.[3] Seetõttu on loota, et mistahes promptidele antavate vastuste keeleline korrektsus paraneb edaspidi veelgi. Praeguse küsimise mõte oli fikseerida hetkeseis 2024. aasta sügisel.

3.3. Kasulikkushinnang

Küsimustiku olulisim eesmärk oli välja selgitada, kas ja kuivõrd kasulikuks peavad leksikograafid keelemudelite abil genereeritud tähendusseletusi. Neljapalline kasulikkuse skaala sundis vastajaid otsustama kas pigem või täielikult kahe äärmuse suunas (st neutraalset keskmist polnud, vt joonis 4). Mõlema keelemudeli puhul ja ka üldiselt kaaluvad kasulikuks peetavad vastused (summaarselt *ca* 80%) kaugelt üles mittekasulikuks hindamise (summaarselt *ca* 20%). Kokku hindasid vastajad enda kui leksikograafi jaoks väga või mingil määral kasulikuks 83% mudelite pakutud tähendusseletustest ja pigem mitte või üldse mitte kasulikuks 17% tähendusseletust. Rohkem hinnati kasulikuks GPT-4o pakutud seletusi (85% *vs.* Gemini 81%).



Joonis 4. Vastused küsimusele „Kui kasulikuks hindad mudeli pakutud tähenduse seletust sinu kui leksikograafi jaoks?“

3 EKTB104 „Eesti keele toetus suurtes generatiivsetes vabavaralistes keelemudelites“; https://www.etis.ee/Portal/Projects/Display/a420f147-a693-4e0e-ad9f-0570862d6a9f

3.4. Hinnangute varieerumine

Hindajatevahelise kooskõla näidik Fleissi kappa4 oli kõigi hinnanguküsimuste (3.1, 3.2, 3.3) puhul keskmist arvestades suhteliselt madal (K1: $\kappa = 0{,}19$; K2: $\kappa = 0{,}15$; K3: $\kappa = 0{,}12$), kusjuures GPT vastustele antud hinnangud ühtisid vaid pisut enam (GPT; $\kappa = 0{,}22$ *vs.* Gemini; $\kappa = 0{,}16$).5 Hinnangute varieerumine 60 eri sõnaliikidest ja eri karakteristikutega sõnale genereeritud tähendusseletustele võis sõltuda nt mudelile antud ülesande keerukusest, konkreetse tulemuse kvaliteedist, hindajate ootustest jne.

Kuigi ainsana esines keelelise korrektsuse küsimuse (3.2) vastustes olukord (8 sõna puhul), kus kõik 21 vastajat vastasid samamoodi, siis paradoksaalselt varieerusid vastajate hinnangud ka keelelise korrektsuse määramisel üsna palju (madalaim $\kappa = -0{,}76$ ja kõrgeim $\kappa = 1$). Analüüsist ilmnes, et variatiivsust tekitas see, mida leksikograaf korrektseks keeleks pidas: kas ta hindas ainult ortograafiat ja morfoloogiat või vaatas ka sõnastust ja sõnavalikut. Ebakõlasid korrektsuse hindamisel võis põhjustada ka see, et küsiti seletuse korrektsuse kohta, kuid leidus juhtumeid, kus tähenduse seletus oli korrektses keeles, kuid mudeli poolt pakutud kasutusnäide mitte.

Vähim oldi ühel nõul GPT vastusega sõnale *maruhoog*, kus esineb ühildumisprobleem: „Tugev ja äkiline tuuleiili või tormi. Sõna võib kasutada ka metafoorselt, et kirjeldada äkilist ja intensiivset tegevust või emotsiooni.“ Nimelt, 6 leksikograafi oli märkinud selle keeleliselt vigaseks, 7 pidas seda korrektseks ja 8 osaliselt korrektseks. Teistel sarnastel juhtudel (nt „kindlustama auto“, „haigushoogule“) tundus, et leksikograaf pole tähele pannud, et seletuses on viga, või oli ta leidnud, et üks viga ei vääri vastust „ei“. Üldiselt ei paistnud keeleliselt (eba) korrektseks hindamise taga väga suurt süsteemsust.

4 Mõõdik, millega uuritakse rohkem kui kahe hindaja vahelise kooskõla usaldusväärsust kategooriaalsete tunnuste hindamisel, näitab, mil määral hindajate omavaheline kooskõla ületab juhusliku taseme. Kui κ väärtus on 1, tähendab see täielikku kooskõla; väärtus 0 viitab juhuslikule kooskõlale; negatiivne väärtus näitab, et kooskõla on halvem kui juhuslikult oodatav.

5 Madalad kappa väärtused viitavad küsimuste keerukusele ning vastuste subjektiivsusele tingimustes, kui kindlaid hindamiskriteeriume pole ette antud.

3.5. Tegurite korrelatsioonianalüüs

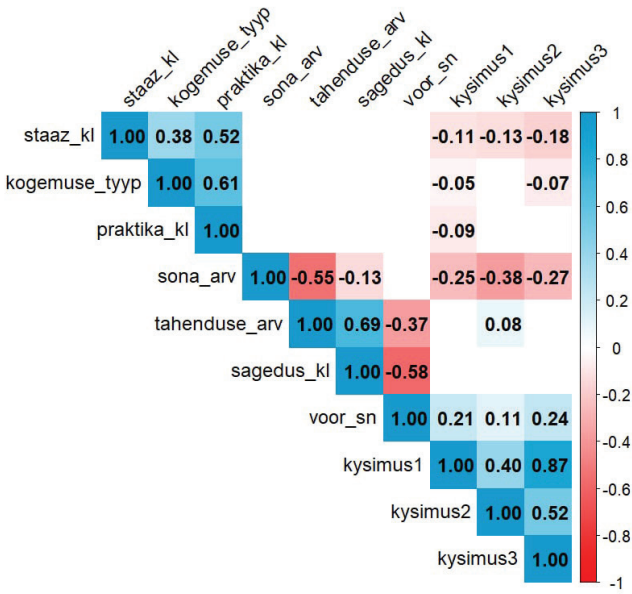
Kvantitatiivselt töödeldavaid andmeid kogunes maatriksisse 7560 andmepunkti (= 21 leksikograafi × 2 mudeli vastust × 60 sõna × 3 küsimust); lisaks koguti kolm iseloomustavat parameetrit osalejate kohta (staaž, sõnastikukogemuse tüüp, seletuseloome praktika) ning mitmeid parameetreid testsõnade kohta (sõnaliik, tähenduste arv, mitmesõnalisus, korpusesagedus, võõrsõnalisus), mille puhul meid huvitas nende võimalik seos SKMide genereeritud tähendusseletuste kvaliteedi ja seeläbi ka leksikograafide hinnangutega.

Joonisel 5 on kujutatud statistiliselt olulisi seoseid.[6] Uuritud tegurid korreleeruvad omavahel järgmiselt: 1) positiivselt ja mõõdukalt on seotud staaž, töökogemuse tüüp (st milliste sõnastikega on tegeldud) ja seletusloome praktika (st kas on tähendusi pigem eri allikatest kopeeritud, kombineeritud või ise keeleandmete pealt sõnastatud); 2) positiivselt on omavahel seotud kolm hinnanguküsimust (tähenduste tabamine (3.1), keeleline korrektsus (3.2) ja kasulikkushinnang (3.3), joonisel 5 vastavalt küsimus 1, 2 ja 3): tugev positiivne seos (joonisel tumesinine) on seos tähenduste tabamise ja kasulikkushinnangu vahel, teised kaks vastastikseost on mõõdukad; 3) märgata võib ka seoseid (nii positiivseid kui negatiivseid) testsõnu iseloomustavate parameetrite vahel (mitmesõnalisus, polüseemia, korpusesagedus ja võõrsõnalisus).

Kolme hinnanguküsimuse seos vastajate tausta ja/või sõnade iseloomuga seotud teguritega on vähene, ent mitte täitsa olematu. Need avalduvad nõrgemalt, kuid siiski olulisel määral.

Tähenduste tabamise (kysimus1) õigekstunnistamise määr on seotud nõrgalt, kuid positiivselt ($\rho = 0{,}21$; $p < 0{,}05$) sellega, kas sõna oli võõrsõna. See tähendab, et võõrsõnade tähenduste esitust hinnati veidi sagedamini õigeks kui omasõnade omi. Tähenduste tabamist hinnati madalamalt, kui märksõna oli mitmesõnaline ($\rho = -0{,}25$; $p < 0{,}05$); veel nõrgem negatiivne seos on leksikograafide staaži ning viljeldud seletusloome praktikaga: kogenumad ja tähendusesitusi ise loonud vastajad on hinnanud mudelite pakutud esitusi veidi kriitilisemalt. Üks põhjuseid võib olla, et oldi põhjalikumad ja võrreldi mudelite genereeritud seletusi sõnastikes leiduvate seletuste ning sealse tähendusjaotusega.

6 Täpsemalt: ordinaalsete variaablite omavahelised seosed määrati R-s polühooriliste seoste määramise metoodikaga. Selgelt kategoriaalsed variaablid nagu sõnaliik ja SKM jäid sellest vaatlusest välja. Nende seotust katse peamiste küsimuste vastustega kontrolliti Spearmani korrelatsioonikordajaga ning statistiliselt olulisi korrelatsioone ei ilmnenud.



Joonis 5. Statistiliselt olulised korrelatsioonid *(p* < 0,05) uuritud tegurite vahel. Positiivsed seosed on sinisega, negatiivsed punasega.

Keelelise korrektsuse hinnangud (kysimus2) seostusid tugevalt ja positiivselt teiste küsimustega: tähenduste tabamise õigsusega ($\rho = 0{,}40$; $p < 0{,}05$) ja kasulikkusega ($\rho = 0{,}52$; $p < 0{,}05$). Mõõdukas ja negatiivne on seos märksõna mitmesõnalisusega ($\rho = –0{,}38$; $p < 0{,}05$).

Kasulikkushinnanguga (kysimus3) on enim positiivselt seotud tähenduste tabamine ($\rho = 0{,}87$; $p < 0{,}05$), keeleline korrektsus ($\rho = 0{,}52$; $p < 0{,}05$) ning nõrgalt märksõna võõrsõnalisus ($\rho = 0{,}24$; $p < 0{,}05$). Kuna katsesõna võõrsõnalisus oli kasulikkushinnanguga positiivses seoses, vaatlesime täpsemalt ka võõrsõnade paiknemist keskmise kasulikkushinnangu alusel. 60 katsesõnast oli võõrsõnu ainult 15, kuid 8 neist paigutusid katses kõige kõrgemalt hinnatud kirjeldusega sõnade esikümnesse – asjaolu, mis kinnitab meie tulemust, et SKMid saavad eesti keeles paremini hakkama võõrsõnade tähenduste kirjeldamisega.

Negatiivne, kuid nõrk on kasulikkushinnangu seos märksõna mitmesõnalisusega ($\rho = –0{,}27$; $p < 0{,}05$), millest võib järeldada, et mudelitel oli raskusi mitmesõnalistele märksõnadele vastuvõetavate seletuste loomisega.

Statistiliselt oluline, kuid väga nõrk, oli seos leksikograafide staaži ja mudelite väljundile antava kasulikkushinnangu vahel ($\rho = –0{,}18$;

$p < 0{,}05$). Pikema töökogemusega leksikograafid andsid tähendusseletustele veidi madalamaid hinnanguid.

Eeltestimise faasis, kui katsetasime suuremat arvu SKMe, märkasime, et SKMid genereerisid madala sagedusega sõnadele vigasemaid tähendusseletusi. Lõpliku katse küsimustiku jaoks loodud tähendusseletused teenisid aga üllataval kombel sarnase keskmise kasulikkushinnangu sõltumata sellest, kui sagedased olid kirjeldatavad sõnad eesti keele ühendkorpuses 2023 (Koppel ja Kallas 2022) või kui mitu tähendust neil ÜSis oli. Nii et keskmine kasulikkushinnang meie katses oli sarnane polüseemsetel ja monoseemsetel sõnadel ning harvadel ja sagedastel sõnadel (samale tulemusele viitas puuduv korrelatsioon kasulikkushinnangu ja sageduse/polüseemsuse vahel).

3.6. Seletuste tüübid

Vastuse küsimusele „Mida tähendab eesti keeles ... (katsesõna)?“ vormistasid keelemudelid erinevalt. Gemini vastas jutustavas stiilis täislausega kõigis vastustes (100%, näide 1; vrd ptk 1.2 näited 6–7). GPT pakutud seletused olid pigem traditsioonilised tähenduse seletused (45%, näide 2; vrd ptk 1.2 näited 1–2), mis paljudel juhtudel esinesid koos täislausetes selgituste ja näidetega (55%, näide 3); esines ka ainult täislausetega seletusi (näide 4).

(1) **vanduma** Sellel sõnal on samuti kaks peamist tähendust. Esiteks võib see tähendada roppuste või vandumise kasutamist kõnes. Näiteks: „Ta vandus vihast.“ Teiseks võib see viidata vande andmisele või pühalikku tõotust andmisele. Näiteks: „Ta vandus truudust oma riigile.“ (Gemini)

(2) **vanduma** 1. Ropendama või kasutama ebaviisakaid, vulgaarseid väljendeid. 2. Tõotama või andma vande, lubadust pidulikult kinnitama. (GPT)

(3) **kuhi** Eesti keeles tähendab sõna „kuhi“ järgmist: 1. Kogum või hunnik esemeid, mis on üksteise peale või kõrvale kuhjatud. Näiteks: „puukuhi“, „lumekuhi“. 2. Suur hulk või kogus midagi. Näiteks: „töökuhi“, „paberikuhi“. (GPT)

(4) **sutureerima** Sutureerima tähendab meditsiinis haava õmblemist või sulgemist õmblusniidi abil. See protseduur tehakse tavaliselt kirurgiliste haavade või lõigete korral, et soodustada paranemist ja vähendada nakkusohtu. (GPT)

3.7. Leksikograafide kommentaarid

Tähendusseletuse kasulikkushinnangule (3.3) lisatud kommentaarid võimaldavad selgitada leksikograafide rahulolu/rahulolematuse sisemist struktuuri: valdavalt põhjendati, et tuua välja (madalamat) hinnangut mõjutanud puudusi. Arvustati tähenduste tabamist (eristamist) (3.7.1), tähenduste seletusi (3.7.2) ja näiteid (3.7.3). Positiivseid kommentaare oli kirja pandud vähem üksikasjalikult (3.7.4).

3.7.1. Tähenduste tabamine (eristamine)

Kõige sagedamini heideti esitusele ette puuduvat tähendust (vastajatel oli võimalus kasutada EKI ühendsõnastikku), nt sõna *ogaline* puhul häiris leksikograafe mõlema mudeli esituses, et välja ei toodud piltlikku tähendust ’terav, salvav’, ja verbi *jõudma* puhul märgiti, et Gemini on kirjeldamata jätnud olulise tähenduse ’jaksama, suutma’.

Kommentaarides toodi esile ka olukordi, kus SKM pakkus ootamatut lisatähendust. Suhtumine lisatähendustesse varieerus. Lisatähendust võidi näha SKMi enda leiutisena, mis lugejat eksitab, näiteks GPT kirjeldus *sardiinlase* juures: „Sõna „sardiinlane“ võib mõnikord kasutada ka metafooriliselt, viidates tihedalt koos olevatele inimestele, sarnaselt sardiinidele konservikarbis.“ Samas oli näiteid, kus toodud lisatähendus võis juhtida leksikograafi korpusesse vastavat tähendust kontrollima ja leidma kasutusnäiteid tähenduse kohta, mis sõnastikus (veel) kirjeldamata, nt *kompu* kirjeldus Geminil: „See sõna on lühend sõnast „komputer“ ja tähendab arvutit.“ Ühendkorpuses leidub kasutusi auto-kompuutri tähenduses. Sõna *maruhoog* kirjelduses puudus Geminil tuuleiili tähendus, ent välja oli toodud emotsionaalse purske tähendus, mida ei esinenud ÜSis, seetõttu kommenteeris vastaja: „Paneb mõtlema, kas ÜSis on üks tähendus puudu.“

Vale või kahtlast tähendust heideti ette oluliselt rohkem Geminile (30 korral) kui GPT-le (11 korral). Sageli said etteheiteid samad esitused, näiteks GPT seletus verbile *segast panema*: „Segadust tekitama või segadusse ajama [...]“, sai mitmelt vastajalt kriitilisi kommentaare, sealhulgas: „Segadusse ajama ei ole tegelik tähendus, pigem on see segast panemise tagajärg, mis avaldub teise isiku juures.“

Ühe sõna (*sale*) juures kirjeldas GPT otseselt ingliskeelse sõna tähendust: „„Sale“ – nimisõna, mis tähendab müüki või allahindlust. Näiteks:

„Kaupluses on suur sale.““ Eesti keeles sellist tähendust ei esine. Kuna *sale* ongi keeltevaheline homograaf, oleks huvitav tulevikus täpsemalt uurida, kuivõrd keeltevaheline homograafia SKMide väljundit mõjutab.

Kui puuduvad/liigsed ja valed/kahtlased tähendused häirisid paljusid vastajaid ja konkreetsed esitused said mitmeid kriitilisi kommentaare, siis järgmisi aspekte – liiga üldine või liiga peen tähendusesitus – mainiti kommentaarides vähem ja märkused jagunesid eri esituste vahel laiali. Tähendust peeti liiga üldiseks/üldistatuks näiteks verbi *käitama* (GPT) kirjelduses: „.. Sõna võib kasutada ka laiemas tähenduses, näiteks ettevõtte või süsteemi juhtimise kohta.“ Keelekasutust korpusest uurides esildub ettevõtte kontekstis pigem *käivitama*. Liiga kitsa tähenduse näiteks sobib vastaja kommentaar *kostitama* (Gemini) tähendusseletuse kohta „Siin on ainult ülekantud tähendus, aga liiga kitsalt – ei ole ainult löömine, kostitada saab ka sõimuga või muu sellisega“.

Leiti ka, et tähendus on eritletud segasel viisil ehk ilmaasjata liiga peeneks aetud, nt (Gemini): *teada olema* kirjeldus: „Sellel väljendil on kaks peamist tähendusvarjundit. Esiteks võib see viidata millegi teadmisele või teadlik olemisele. Näiteks võib öelda: „Mul on teada, et ta on hea õpilane.“ Teiseks võib see tähendada millegi olemasolust teadlik olemist või selle kohta informatsiooni omamisest. Näiteks võib öelda: „Kas teile on teada, millal koosolek algab?““ Esitus häiris mitmeid vastajaid, näiteks kommenteeriti: „Ei saa aru tähenduste vahest, võiks olla üks tähendus.“

Kitsaskohana toodi esile homonüümia esitamata jätmist, nt sõna *salk* juures on esitatud *salgu* ja *salga* tähendused reas polüseemiana, ent vastaja toob esile, et „Leksikograafile on tähtis ka tüvevokaal, sest tegu ei ole üheainsa märksõnaga“. Lisaks ei suutnud GPT ära tunda määrsõna *loogu*, pidades seda kas tegusõna *looma* käskiva kõneviisi vormiks (nt „Loogu ta midagi ilusat“) või nimisõna *look* mitmuse omastava vormiks (nt „Jõe loogud“). Aga kui klassikalisest homonüümia esitusest tundsid leksikograafid puudust, siis vormihomonüümiat peeti üleliigseks. Homonüümia ja polüseemia esitus võibki sõnastikuti erineda ja sageli on otsus, kas tähendusi esitada ühel või teisel kujul, ka spetsialistile keeruline.

3.7.2. Tähenduste seletused

Kuigi statistilist korrelatsiooni esituse tähemärkide arvu ja kasulikkushinnangu vahel ei esinenud, siis üks põhilisi mainitud puudusi oli ikkagi see, et tähendusseletust peeti liiga pikaks, liiga detailirohkeks ja paljusõnaliseks (pikim seletus küsitluses oli 699 tähemärki). Pisut rohkem heideti pikkust ette GPT-le, ent selleteemalisi märkusi jätsid leksikograafid mõlema mudeli esituste juures, nt „Kantseliitlik või pigemini toortõlkemaiguline sõnastus, tarbetult vahutav“ (*sutureerima*). Nii kommentaaride kui ka hindamisskaalade vastustest nähtus, et vastajate suhtumine esituse pikkusesse varieerus. Leidus vastajaid, kes jätsid järjepidevalt kommentaare liigse pikkuse kohta, kuid samas võisid teised vastajad hinnata sama tähendusseletust kõrgelt.

Märkusi said seletused, kuhu vastaja arvates oli kaasatud liigselt ebavajalikku infot, nt *asuursinine* (Gemini pakutud seletus): „See on erksinise värvi varjund, mis meenutab selget taevast või sügavat vett. Asuursinine on sageli seotud rahu, vaikuse ja avarusega“. Ebavajalikuks võidi pidada tähendusseletuse viimast lauset, kuid kaheldi ka selles, kas just „sügav vesi“ on sobilik kirjeldama asuursinist värvitooni. Häiris, et seletust oli ilma erialateadmisteta keeruline tabada: „Seletus on liiga keeruline ja terminiline, ootaks lihtsustatud varianti“ ja „See ei ole seletus, vaid on entsüklopeediline info“ (*Hodgkini lümfoom*).

Oluliselt vähem, kuid siiski, heideti ette ka liiga nappi tähendusseletust. (Lühim kirjeldus küsitluses oli 92 tähemärki.) Kommenteeriti, et seletus on liiga napp või et seletus ei ole piisav. Lühidust heideti enam ette Geminile, nt *tundmatu* seletusele: „See sõna tähistab midagi või kedagi, keda ei tunta või kes pole teada. Tundmatu võib olla seotud nii inimeste, kohtade, asjade kui ka mõistetega.“ ÜSi seletuste pikkusega võrreldes ei ole tegemist lühikese kirjeldusega, kuid pikkuse hinnangut võis mõjutada ka võrdlusmoment teise SKMi seletusega ja samuti kõrvale vaadatud ÜSi esitus, mis oli põhjalikumalt liigendatud (kaks põhitähendust ja üks alltähendus).

Vastajad tõid kommentaarides välja kahtlust tekitavaid või otseselt valesid väiteid. See on oluline kitsaskoht töös SKMidega, kuna sellised valeväited võivad jääda tähelepanuta, kui inimesel puuduvad konkreetsel teemal piisavad erialased teadmised, et väidet kahtluse alla seada. Hilisemal kontrollimisel võib selguda, et SKM tegelikult ei eksi, aga võis ka olla, et väide osutus veaks: nt sõna *kõrkjas* tähendusseletust

kommenteeriti, et „väide kõrrelistest tundub eksitav“, ja kontrollimine kinnitas kahtlust. *Kõrkja* kasutustena tõi SKM välja ka korvipunumise ja katuste ehitamise, kuid kontrollimisel selgus, et Eestis kasvavad kõrkjaliigid ei sobi kummakski. Tekib kahtlus, et SKMide pakutud kirjeldused on pigem üldised ja konkreetse koha eripärasid arvesse ei võta.

Teatud sõnade juures tundsid vastajad puudust sõna kasutusvaldkonna paremast tutvustamisest, nt *pseudonüümima* juures tõi Gemini valdkondadena välja kirjanduse ja kunsti, ent vastajad mainisid ka andmekäitlust. *Degradatsiooni* seletuses pakkus GPT mitmeid lisavaldkondi võrreldes sõnastikuga, kuid vastajad seadsid toodud valdkonnad kahtluse alla.

Eraldi üksuse moodustasid keelelised etteheited, mis varieerusid kohmakast seletuse sõnastusest ebagrammatilisuse ja puuduvate komadeni. Korduvalt kritiseeriti kommentaarides tautoloogiat. Kuigi küsimustiku peale kokku esines märksõna või tüve kasutust seletuses paaril korral, häiris see paljusid vastajaid, nt *ogaline* (Gemini): „See omadussõna kirjeldab midagi, millel on ogad või mis on ogaline.“ Miinusena toodi välja sünonüümide kaudu seletamist, nt *jõletu* (Gemini): „See omadussõna kirjeldab midagi, mis on väga halb, kohutav või jälk.“

Nii tähenduste eristamise kui ka tähenduste seletuse kriitika alla paigutuvad sõnaliigiküsimused. Toodi esile, kui definitsioon viitas valele sõnaliigile, nt *jõletu* esitust (GPT) „Midagi väga inetut, koledat või vastikut“ kommenteeris vastaja „hrl viitab nimisõnale seda tüüpi seletustes“. Leiti üles ka esitusi, kus seletus ühele sõnaliigile vastavale tähendusele puudus. Sarnane probleem on nii SKMide vastustega kui ka korpusmaterjali analüüsides: kui materjali on palju ning üks tähendus või sõnaliik domineerib selles materjalis tugevalt, on suur oht, et vähemesindatud tähendus või sõnaliik jääb esile toomata (vt ka Risberg jt 2025).

3.7.3. Näited

Me ei küsinud SKMidelt promptis eksplitsiitselt näiteid, sest näidete hindamine ei olnud meie algne eesmärk. Sellegipoolest lisasid SKMid paljude sõnade tähendusseletuste juurde näiteid nii kollokatsioonide, näitelausete kui ka liitsõnade kujul. Näited võisid olla integreeritud juba seletuse ossa või toodud eraldi seletuse järel (vt ptk 3.6).

Peamine etteheide näidetele oli sama, mida mainiti ka seletuse juures, nimelt et nad tunduvad kahtlased või valed. Näiteks sõnale *kostitama* pakkus Gemini näitelauseks „Ta kostitas teda kõvasti näkku“. Lause on elliptiline ja tekib tunne, et *kostitama* on kasutatud otseselt ’lööma’ tähenduses, mida sellel sõnal ei ole. Üksmeelset kriitikat sai ka Gemini näitelause verbile *segast panema*: „Ta pani kõik oma asjad segast, nii et ma ei leia midagi üles.“

Samuti heideti mõlema mudeli puhul ette keelelist ebakorrektsust, nt *linnulennult* (Gemini): „Linn lennult on vahemaa lühem“; *teada olema* (GPT): „Kas sa tead, mis juhtus? Jah, ma olen sellest teada.“ Paaril üksikul korral (aga palju vastajaid tegi märkuse) puudus näitest katsesõna ise, nt *painduma* näiteks on Geminil lause: „Oks painutas tuule käes.“ Näide ja katsesõna võisid olla ka erinevast sõnaliigist, nt adverbi asemel näites adjektiiv, *rööpselt* (Gemini): „Need kaks joont on rööpsed.“ Kui sõna oli küll tabatud, võis olla probleem ka tähenduse ja näite sobimatuses, näiteks oli Geminil *ogalise* juures välja toodud ainult füüsiline ’ogadega’ tähendus, ent esines näide: „Tema sõnad olid ogalised.“

SKMid esitasid näidetena näitelauseid, kollokatsioone ja liitsõnu, mida vastajad pidasid haruldaseks või ebaloomulikuks, nt sõnale *sale* „saledad puutüved“ või *hapu* maitse seletuseks „hapu piim“. Ka lause tasandil esines näiteid, mida peeti ebaharilikuks, nt *läikima* näide (Gemini): „Päike läikis veepinnal“; *hapu* (Gemini): „Ta oli täna hommikul väga hapu.“ Viimase näite kohta kommeneeris vastaja, et meeleolule ülekantult ei kasutata omistust otse inimesele. (Pigem võiks öelda „hapu näoga“ või „hapus meeleolus“.) See näitab, et leksikograafid eelistavad prototüüpseid näiteid.

3.7.4. Positiivsed kommentaarid

Positiivsena toodi esile, et SKMid on esitanud otsese, ülekantud või abstraktse tähenduse. Kiideti tähendusesituse põhjalikkust, detailsust ja sõnastust, nt „Suurepärane, adverbi tähendused tabatud!“, „Ilus klaar sõnastus“. Väärtustati seletusvõimet, isegi kui seletuse muud tahud jätsid vastaja arvates soovida, nt *sutureerima* (Gemini): „Mudeli pakutu autentsena sõnaartiklisse tähenduseks ei sobi, aga minu jaoks tundmatu sõna seletas edukalt ära.“ Eraldi toodi välja ka SKMide tugevust registrite tuvastusel, nt sõna *perselakkuja* esituses märkis GPT ära, et tegemist on vulgaarse ja halvustava väljendiga.

SKMide välja pakutud tähendusi ja näiteid, mida sõnastikus esitatud ei olnud, peeti samuti vahel tugevuseks, nt kommenteeriti „Paneb mõtlema, kas ÜSis on üks tähendus puudu“ või sõna *salk* juures: „.. pani mõtlema, kas ÜSi peaks täiendama näitelausega „Ta lõikas juustest väikese salgu““. Seega sai SKMide tähendusseletustest uusi ideid nii tähenduste kui näitelausete osas.

Üldjoontes avaldusid kiitmisel samad aspektid, mis kriitika osas, kuid mitte nii üksikasjalikult. Positiivne hinnang ise võiski vastajatele tunduda vähem selgitust vajav kui negatiivne.

4. Kokkuvõte ja arutelu

Vastajate üldine rahulolu SKMide genereeritud tähendusseletustega oli suhteliselt kõrge: leksikograafid hindasid oma töö jaoks väga kasulikuks või mõnevõrra kasulikuks 83% tähendusseletustest. Keelemudelite sooritus eesti keeles enamasti rahuldas vastajaid: 75% juhtudel hinnati kasutatud keelt korrektseks, 20% osaliselt korrektseks ja ainult 5% seletuste puhul täiesti ebakorrektseks. Rahulolu semantilise esitusega oli mõnevõrra madalam: väljatoodud tähendusi peeti õigeks 56% juhtudel, osaliselt õigeks 33% juhtudel ja valeks 10% juhtudel. Suurim positiivne korrelatsioon esines tähenduste tabamise ja kasulikkushinnangu vahel: tähendusseletusest saadavat kasu hinnati kõrgemaks pigem selle põhjal, kui tähendused olid õigesti tabatud, kui keelelise korrektsuse kaalutlusel.

GPT-4o ületas eesti keele sõnatähenduste kirjeldamisel Gemini 1.5 Pro-d ja sai kõrgemad keskmised hinnangud kõigis põhilistes uuritud küsimustes: tähenduste tabamine, keeleline korrektsus ja pakutud tähendusseletuse kasulikkus leksikograafi jaoks.

Kaasasime analüüsi lisaparameetrid, mille võimalik seos SKMide tähendusseletuste kvaliteediga meile huvi pakkus: sõnaliik, tähenduste arv, mitmesõnalisus, korpusesagedus, võõrsõnalisus. Seletuse kasulikkushinnangu ja sõna sõnaliigi vahel statistiliselt olulist seost ei esinenud. Mitmesõnalisus oli negatiivselt seotud nii grammatilise korrektsuse hinnangu kui ka üldise kasulikkushinnanguga, millest võib järeldada, et SKMidel oli raskusi mitmesõnalistele märksõnadele sobivate seletuste loomisega. Võõrsõnalisus oli kasulikkushinnanguga nõrgas positiivses seoses. Samuti paigutus katse 15 võõrsõnast 8 kõige kõrgemalt hinnatud kirjeldusega sõnade esikümnesse, mis lisab tuge

meie tähelepanekule, et SKMid saavad eesti keeles veidi paremini hakkama võõrsõnade tähenduste kirjeldamisega.

Katse algfaasis erinevate SKMide katsetamisel märkasime, et madalama sagedusega sõnadele genereeriti vigasemaid tähendusseletusi. Eeldasime ka, et SKMidel võib olla rohkem probleeme polüseemse sõna kirjeldamisel kui monoseemse sõna esitamisel. Lõpliku küsimustiku jaoks loodud tähendusseletused olid aga sarnase kasulikkushinnanguga sõltumata sellest, kui sagedad olid kirjeldatavad sõnad korpuses või kui mitu tähendust neil EKI ühendsõnastikus (ÜS) oli. Kasulikkushinnang ei olnud vastastikuses seoses sõna polüseemsuse ega korpussagedusega. Kuigi võiks eeldada, et on raskem luua seletust harvale sõnale või väga polüseemsele sõnale, siis tuleb arvessse võtta, et tähendusseletusi võivad mõjutada vastandlikud tendentsid: monoseemsed sõnad on sageli harvemad kui polüseemsed sõnad, kuid samas on monoseemse sõna tähendusesitusega vähem võimalikke probleeme. Samuti võis materjali uurides täheldada meie katsesõnade hulgas ainult üksikute korpusesinemistega (võõr)sõnu, mis olid saanud kõrge kasulikkushinnangu, ja samuti leidus suure sagedusega (polüseemseid) sõnu, mille esitusega paljud vastajad rahul ei olnud.

Võrreldes ÜSi tähenduste esitusega, kaldusid SKMid tähendusi pigem koondama. SKMide pakutud tähenduste ja ÜSis esitatud tähenduste arvu kõrvutamisel selgus, et nii GPT kui ka Gemini esitasid sõnastikust väiksema arvu tähendusi 57% juhtudest, sõnastikuga võrdse arvu tähendusi esitas GPT 38% juhtudest ja Gemini 40% ning kõige vähem esines olukordi, kus SKMid pakkusid välja rohkem tähendusi kui sõnastik (GPT 5%, Gemini 3%). Huvitav leid oli tähenduste arvu analüüsis asjaolu, et tulemused olid mõlema mudeli puhul suhteliselt võrdsed ehk SKMide esitatud tähendusjaotuse peensus oli kahel testitud mudelil väga sarnane.

Nii kvantitatiivsest kui ka kvalitatiivsest analüüsist järeldus, et leksikograafide eelistused tähendusseletuse sisu ja vormi suhtes erinevad. Kolmest peamisest küsimusest (tähenduste tabamine, keeleline korrektsus ja kasulikkushinnang) varieerusid vastused enim (keskmine $\kappa = 0{,}12$) kasulikkushinnangu andmisel („Kui kasulikuks hindad mudeli pakutud tähenduse seletust sinu kui leksikograafi jaoks?“), mis oli ka kõige subjektiivsemat vastust eeldav küsimus. Täielikul üksmeelel ($\kappa = 1$) oldi ainult 8 tähendusseletuse keelelise korrektsuse hindamisel. Ka selle küsimuse vastuste niivõrd suur varieerumine oli mõnevõrra

üllatav, kuna kolmest esitatud küsimusest on keeleline korrektsus kõige otsesemalt vaadeldav ja hinnatav. Vastuseid analüüsides ilmnes, et variatiivsust tekitas see, mida leksikograaf korrektseks keeleks pidas – kas ta hindas ainult ortograafiat ja morfoloogiat või vaatas ka sõnastust ja sõnavalikut.

Kasulikkushinnangule lisatud kommentaaride kvalitatiivne analüüs võimaldas leksikograafide ootusi täpsemalt tuvastada. Joonistus välja leksikograafide rahulolu/rahulolematuse sisemine struktuur, mida mõjutasid tähenduste esitust (sh nii sisu kui vormi) puudutavad aspektid (nt puuduv, üleliigne või osaliselt vale tähendus, homonüümia käsitlemata jätmine); tähenduse kirjeldustega seotud aspektid (nt kirjelduse liigne pikkus, tautoloogia, sünonüümide kaudu seletamine, sõnaliigiküsimused) ja tähenduste näitlikustamisega seotud aspektid (nt näite prototüüpsuse ootus, ebaloomulikud näited, näite sobimatus konkreetse tähendusega).

Märkustest nähtus ka see, et SKMide tähendusseletusi võidi võrrelda ideaaliga, mis vastajati erines. Näiteks ei meeldinud osale vastajatest sünonüümide kaudu seletamine, ent teised võisid sama seletust hinnata kõrgelt. Samuti leidus vastajaid, kes jätsid korduvalt kommentaare tähendusseletuse liigse pikkuse (või liigse lühiduse) kohta, kuid samas võisid teised vastajad hinnata sama esitust kõrgelt. Lähenedes genereeritud tähendusseletustele kui alusmaterjalile, võib pikk selgitus olla ka positiivne, kuna võimaldab leksikograafil tutvuda suurema hulga infoga, millele seletuse koostamisel toetuda. Samas kui vastaja ootas SKMidelt juba valmis seletust, mida saaks kasutada suuremate muudatusteta, võis pikem esitus saada negatiivse hinnangu. Suhtumist võib mõjutada viis, kuidas senini leksikograafilist tööd on tehtud: kas seletusi on ise loodud keelenäidete põhjal või on neid otse üle võetud teistest sõnastikest/terminibaasidest. Töökogemus mõjutab leksikograafi ootusi SKMidele ja võib määrata selle, millist seletust või näidet eelistatakse.

SKMide tugevusena nähti tähendusseletuste põhjalikkust, detailsust ja sõnastust. Kiideti ka teavet registri kohta ja SKMide välja pakutud tähendusi ja näiteid, mida sõnastikus esitatud ei olnud. Uued üksused võisid aga olla nii tugevuseks kui nõrkuseks, olenevalt sellest, kas nad (hilisemal kontrollil) osutusid tõepäraseks või mitte. Üks põhilisi kitsaskohti töös SKMidega ongi see, et nende väited vajavad täiendavat kontrolli ja valeväited võivad olla väga raskesti tuvastatavad, kui konkreetsel teemal puuduvad ekspertteadmised. Uute pakkumiste

eelis oli jällegi see, et leksikograaf võis SKMide tähendusseletustest saada väärtuslikke ideid nii tähenduste kui ka näitelausete lisamiseks sõnastikku.

Tulevikus saame kasutada väljaselgitatud ootusi üksikasjalike promptide loomiseks, näiteks kohandada seletuse pikkust, sõnastuse keerukust, vältida tüve või märksõna ennast, vältida vormihomonüümiat (andes ette, et tegemist on lemmaga), vältida üksnes sünonüümidest koosnevaid definitsioone ja lisada või välistada näiteid. Samuti saab mudeleid esinduslike näidete varal treenida andma vastuseid, mis vajavad vähem ületoimetamist.

Kindlasti on väljakutse koostada prompte, mis rahuldaks kõiki nii erinevatele ootuste ja eelistustega spetsialiste, nagu uuringus osalenud leksikograafid. Üks lahendus oleks luua erinevad alused eri tüüpi seletuste genereerimiseks. Plaanimegi arendada eraldi prompte eri keeleoskustasemetele ja erinevate sõnastike vajadusi silmas pidades, nt tähendusvihjed ÕSile, lihtsamad seletused keeleõppijatele ja detailsed seletused EKI ühendsõnastikule, ning hinnata saadud tulemust. Samuti plaanime tulevikus hinnata SKM-ide seletusi kõrvutavalt leksikograafi koostatud seletustega. Inglise keele uurimisel (Lew 2023) avaldus, et keelemudeli definitsioone hinnati sarnaselt heaks sõnastikudefinitsioonidega (COBUILD). Kavas on katsetada ka uusi mudeleid, kuhu eesti keel on sihipäraselt kaasatud, näiteks EuroLLM, TildeLLM ja SiloGen. Kuid juba siinses katses rakendatud SKMide tulemused näitavad, et kuigi keelemudelite pakutu vajab kontrollimist ja kohandamist, saavad nad leksikograafile märkimisväärselt abiks olla nii tähenduste esitamisel/jagamisel kui ka seletuste sõnastamisel.

Tänusõnad

Artikli valmimist on toetanud teadusprojekt EKKD-III1 „Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed“ ja osaliselt PRG1978 „Uue aja sõnastik: grammatika ja keelepädevuse kirjeldamine integreeritud multifunktsionaalses leksikograafilises ressursis“. Täname EKI leksikograafe, kes leidsid aega, et vastata meie pikale küsimustikule!

Kirjandus

Agarwal, Arav, Karthik Mittal, Aidan Doyle, Pragnya Sridhar, Zipiao Wan, Jacob Arthur Doughty, Jaromir Savelka & Majd Sakr. 2024. Understanding the Role of Temperature in Diverse Question Generation by GPT-4. *arXiv*. https://doi.org/10.48550/arXiv.2404.09366.

Alexander, Marc. 2015. Words and Dictionaries. In John R. Taylor (toim), *The Oxford Handbook of the Word*. Oxford University Press, 37–52. https://doi.org/10.1093/oxfordhb/9780199641604.013.021.

Anthropic. 2024. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. https://api.semanticscholar.org/CorpusID:268232499.

Atkins, Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Bender, Emily, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Grechen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter etc. 2020. Language Models are Few-Shot Learners. Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan & Hsuan-Tien Lin (toim), *Advances in Neural Information Processing Systems* 33, 1877–1901. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

De Schryver, Gilles-Maurice. 2023. Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography* 36(4). 355–387. https://doi.org/10.1093/ijl/ecad021.

De Schryver, Gilles-Maurice & David Joffe. 2023. *The end of lexicography, welcome to the machine: On how chatGPT can already take over all of the dictionary maker's tasks*. Universiteit Gent. https://biblio.ugent.be/publication/01GWSGV0M8HYZVNXV8SSZ0VXBG.

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

Evert, Stephanie, Christine Ganslmayer & Christian Rink. 2024. Multi-Level Analysis as a Systematic Approach to Evaluating the Quality of AI-Generated Dictionary Entries. *Proceedings of the XXI EURALEX International Congress*, 317–334. https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralax-XXI-final-web.pdf.

Gallegos, Isabel, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang & Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50(3). 1097–1179. https://doi.org/10.1162/coli_a_00524.

Gemini Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, S. Mariooryad, Y. Ding, X. Geng, F. Alcober, R. Frostig, M. Omernick, L. Walker, C. Paduraru, C. Sorokin & etc. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.2403.05530.

Grindrod, Jumbly. 2024. Large language models and linguistic intentionality. *Synthese* 204(2). 71. https://doi.org/10.1007/s11229-024-04723-8.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu & S. Mirjalili. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hayakawa, Samuel Ichiye & Alan R. Hayakawa. 2022. *Keel mõttes ja teos*. Postimees Kirjastus.

Ignat, O., Z. Jin, A. Abzaliev, L. Biester, S. Castro, N. Deng, X. Gao, A. E. Gunal J. He, A. Kazemi, M. Khalifa, N. Koh, A. Lee, S. Liu, D. J. Min, S. Mori, J. C. Nwatu, V. Perez-Rosas, S. Shen, Z. Wang, W. Wu & R. Mihalcea. 2024. Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models. N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (toim), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 8050–8094. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.708.

Jakubíček, Miloš & Michael Rundell. 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*, 508–523. https://www.researchgate.net/publication/387009373_The_end_of_lexicography_Can_ChatGPT_outperform_current_tools_for_post-editing_lexicography.

Jürviste, Madis, Tiina Paet & Sven-Erik Soosaar. 2025. Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 21. 63–84. https://doi.org/10.5128/ERYa21.04.

Kallas, J., S. Koeva, Margit Langemets, C. Tiberius & I. Kosem. 2019. *Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs*. In *Proceedings of eLex* 19, 519–536. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_30.pdf.

Karjus, Andres. 2023. Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. *arXiv*. https://doi.org/10.48550/arXiv.2309.14379.

Klosa-Kückelhaus, Anette & Carole Tiberius. 2024. The Lexicographic Process Revisited. *International Journal of Lexicography* 38. 1–12. https://doi.org/10.1093/ijl/ecae016.

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.

Kosem, Iztok, Polona Gantar, Špela Arhar Holdt, Magdalena Gapsa, Karolina Zgaga & Simon Krek. 2024. AI in Lexicography at the University of Ljubljana: Case Studies. *Workshop 1: Large Language Models and Lexicography XXI EURALEX*

International Congress. Lexicography and Semantics, Cavtat, Croatia. https://videolectures.net/videos/euralex2024_kosem_ai_lexicography_studies.

Kosem, Iztok, Tanara Zingano Kuhn, Špela Arhar Holdt, Kristina Koppel, Carole Tiberius, Rina Zviel Girshin, Vivien Waszink & Karolina Zgaga. 2024. Can AI Assist in Selecting Dictionary Examples? A Case Study in Four Languages. *XXI EURALEX International Congress. Lexicography and Semantics, Cavtat, Croatia.* https://euralex.jezik.hr/wp-content/uploads/2021/09/EURALEX_2024_Full_Programme_chairs_FINAL.pdf.

Langemets, Margit, Kristina Koppel, Jelena Kallas & Arvi Tavast. 2021. Sõnastikukogust keeleportaaliks. *Keel ja Kirjandus* 64(8–9). 755−770. https://doi.org/10.54013/kk764a6.

Langemets, Margit & Lydia Risberg. 2022. Ood keelele ja teaduslikule meetodile. *Sirp* 48(3920). 20−21.

Lew, Robert. 2023. ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications* 10(1). 1–10. https://doi.org/10.1057/s41599-023-02119-6.

Liang, Paul Pu, Chiyu Wu, Louis-Philippe Morency & Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. *International Conference on Machine Learning*, 6565–6576.

Mel'čuk, Igor, & Alan Polguère. 2018. Theory and practice of lexicographic definition. *Journal of Cognitive Science* 19(4). 417–470. https://doi.org/10.17791/jcs.2018.19.4.417.

Navigli, Roberto, Simone Conia & Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality* 15(2). 1–21. https://doi.org/10.1145/3597307.

OpenAI. OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A.J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol & etc. 2024. GPT-4o System Card (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2410.21276.

Paulsen, Geda, Ene Vainik & Maria Tuulik. 2020. Sõnaliik leksikograafi töölaual: Sõnaliikide roll tänapäeva leksikograafias. *Eesti Rakenduslingvistika Ühingu aastaraamat* 16. 177–202. https://doi.org/10.5128/ERYa16.11.

Steurs, Frieda, Tanneke Schoonheim, Kris Heylen & Vincent Vandeghinste. 2020. *The Future of Academic Lexicography—A White Paper*. Leiden: Instituut voor de Nederlandse Taal. https://ivdnt.org/wp-content/uploads/2021/02/The-Future-of-Academic-Lexicography-A-White-Paper.pdf.

Risberg, Lydia, Maria Tuulik, Margit Langemets, Kristina Koppel Ene Vainik, Esta Prangel & Eleri Aedmaa. 2025. Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel. *Keel ja Kirjandus* 68(7). 605−624. https://doi.org/10.54013/kk811a3.

Tadić, Marko. 2024. Can LLMs Really Generate New Words? *XXI EURALEX International Congress. Lexicography and Semantics, Cavtat, Croatia.* https://euralex.jezik.hr/wp-content/uploads/2021/09/EURALEX_2024_Full_Programme_chairs_FINAL.pdf.

Tao, Yan, Olga Viberg, Ryan S Baker & Rene F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3(9). 1–9. https://doi.org/10.1093/pnasnexus/pgae346.

Tavast, Arvi, Margit Langemets, Jelena Kallas & Kristina Koppel. 2018. Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018*, 749−761. Ljubljana University Press, Faculty of Arts.

Tiberius, Carole, Kris Heylen, Bram Vanroy, Vincent Vandeghinste, Jesse de Does & Job van Doeselaar. 2024. LLMs and Evidence-Based Lexicography: Pilot Studies at INT. *XXI EURALEX International Congress. Lexicography and Semantics, Cavtat, Croatia*. https://euralex.jezik.hr/wp-content/uploads/2021/09/EURALEX_2024_Full_Programme_chairs_FINAL.pdf.

Vaswani, Asish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is All you Need. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (toim), *Advances in Neural Information Processing Systems* 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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

ÜS = Eesti Keele Instituudi ühendsõnastik 2025. Eesti Keele Instituut, Sõnaveeb. https://sonaveeb.ee.

Abstract. Maria Tuulik, Ene Vainik, Esta Prangel, Margit Langemets, Eleri Aedmaa, Kristina Koppel, Lydia Risberg: Describing senses for lexicography: how helpful are large language models? In this article, we investigate whether and to what extent large language models (LMMs) can assist lexicographers in writing definitions and presenting word meaning(s). Of the LLMs tested, the GPT-4o and Gemini 1.5 Pro meaning descriptions for 60 words were selected for the final evaluation. The zero-shot prompting method used made it possible to identify the lexicographers' preferences and demands for the descriptions without preemptively limiting their form. The meaning descriptions were evaluated by 21 lexicographers.

The overall satisfaction of respondents was relatively high: lexicographers rated 83% of the meaning descriptions as very or somewhat useful for their work. The performance of the models in Estonian was mostly satisfactory for the respondents: in 75% of the cases, the language of the meaning descriptions was judged to be grammatically correct. Satisfaction with the semantic presentation was somewhat lower: the meanings given were considered correct in 56% of cases. Of the two LLM-s, GPT-4o's output was rated higher.

Lexicographers' evaluations varied. The analysis of the comments accompanying the evaluations clarified the internal structure of the lexicographers' satisfaction/dissatisfaction and will allow the findings to be used in the future to create detailed prompts providing basis for different types of dictionary definitions.

Keywords: large language models, Estonian language, lexicography, lexical semantics, meaning descriptions, polysemy