

“KEELETEADLANE RAPUTAB LAUSEMALLE KORPUSEST LOENDISSE”: LAUSEMALLIDE AUTOMAATSE TUVASTAMISE ESIMESED SAMMUD

Kertu Saul, Kadri Muischnek, Jelena Kallas

Ülevaade. Artiklis kirjeldatakse arvutilingvistilist katset, mille eesmärk on luua esmane töövoog eesti keele lausemallide korpuspõhiseks automaattuvastamiseks. Materjalina on kasutatud märgendussüsteemi *Universal Dependencies* alusel automaatselt ja käsitsi annoteeritud korpusi. Testandmestikuna on kasutatud 28 liigutamisverbi lausemalle. Pakutud meetod moodustab verbist ja selle otsestest alluvatest paarid, eemaldab need paarid, mis jäävad alla 5% sageduslävendi ning kombineerib allesjäänutest lausemallid. Meetod osutus efektiivseks verb-alluva paaride tuvastamisel, kuid terviklausemallide tuvastamise kvaliteet oli halvem. Artiklis analüüsitakse põhjalikult tuvastusvigade tüüpe ning põhjusi, pakutakse lahendusi iga konkreetse veapõhjuse kõrvaldamiseks ja määratakse edasiarenduse suundi. Meetod osutus kasulikuks ka verbiüleste ja seni kirjeldamata mallide tuvastamiseks.*

Võtmesõnad: arvutilingvistika, arvutileksikograafia, süntaks, sõltuvussüntaks, lausemall, eesti keel

1. Sissejuhatus

Arvuti- ja korpuslingvistika kiire areng loob uusi võimalusi loomuliku keele analüüsimiseks ning eri tüüpi info automaatselt tuvastamiseks. Kuigi arvutileksikograafias on tavapäraseks saanud automaatne märksõnastiku genereerimine ning mitmesõnaliste üksuste, kollokatsioonide ja näitelauseste tuvastamine, siis grammatilise ja eriti süntaktilise informatsiooni automaatne tuvastamine ei ole veel sugugi nii levinud (Tiberius jt 2024). Eesti sõnastikes on eri sõnaliikide süntaktilist käitumist traditsiooniliselt kirjeldatud reksioonimallide kaudu (Kallas 2013: 100–104), kuid näiteks verbi puhul ei anna reksioonimall kasutajale piisavalt juhiseid terviklause moodustamiseks. Katse pakkuda koondkäsitlus eesti keele lihtlauseste vormilistest

* Uurimistööd on finantseerinud Eesti Teadusagentuuri grant “Uue aja sõnastik: grammatika ja keelepädevuse kirjeldamine integreeritud multifunktsionaalses leksikograafilises ressursis” (PRG 1978) ja programmi “Eesti keeletehnoloogia 2018–2027” projekt “Semantilise analüüsi baasvahendid” (EKTb75). Täname anonüümseid retsensente väga kasulike märkuste ja tähelepanekute eest.

tüüpidest ehk lausemallidest tegi 1978. aastal Huno Rätsep, kelle materjal oli eelkõige “uurija poolt esitatud keeleainestik” (Rätsep 1978: 6–8). Käesolevas artiklis uurime aga, kas ja kuidas saaks lausemalle tuvastada sõltuvussüntaktiliselt märgendatud korpusest. Eesmärk on luua esmane töövoog lausemallide korpuspõhiseks tuvastamiseks ning saadud tulemuste põhjal välja selgitada kerkivad probleemid. Vastupidiselt Rätsepa eesmärgile kirjeldada kõiki eesti keeles võimalikke lihtlause malle (Rätsep 1978: 8) püütakse siin töö võimalikest rakendustest tulenevalt leida just neid, mis kasutuses sagedasti esinevad. Väljapakutud töövoos abil saab edaspidi koostada eesti keele verbikesksete lausemallide loendi, mida saab edasi arendada morfosüntaktilist ja semantilist infot ühendavaks leksikoniks (cf. Ruppenhofer jt 2016, Kipper-Schuler 2005, Palmer jt 2005). Töö tulemustest on kasu ka näiteks keeleõppijatele suunatud leksikograafiliste ressursside koostamisel: sõnastikus saaks verbikirjes esitada selle verbiga sageli moodustatavate lausetüüpide näited (vt nt Colman, Tiberius 2018). Verbikesksete lausemallide kasutamine parandaks tõenäoliselt ka automaatse süntaksianalüsaatori täpsust.

Siinne artikkel esitab eelkõige verbikesksete lausemallide automaatse tuvastamise esimest katset, mitte parimat võimalikku viisi nende mallide kvaliteetse leksikoni loomiseks. Viimasel puhul peaks lõpliku mallide valiku ja valideerimise tegema ikkagi inimene.

Artiklis avame kõigepealt lausemallidega seotud mõisteid ning tutvustame mõnda lausemallide ekstraheerimise meetodit. Seejärel selgitame, kuidas me lausemalle automaatselt tuvastasime ning mis materjali me selleks kasutasime. Anname ka ülevaate lausemallide tuvastamise kvaliteedist ning meetodi tüüpilistest vigadest ja nende põhjustest. Lisaks tutvustame meetodi häid külgi, mille hulka kuulub ka uute, varem kirjeldamata lausemallide tuvastamine. Kuna töö üks eesmärgi oli kaardistada lausemallide tuvastusega seotud probleemid, et neid järgnevates katsetes parandada, pakume lõpetuseks välja meetodi võimalikud edasiarendused.

2. Mis on lausemall

Termini *lausemall* tõi eesti grammatikakirjeldusse Huno Rätsep (1978: 18), kes defineeris verbikeskse lausemalli abstraktsete oluliste sõnaklasside ja grammatiliste kategooriate moodustatud lause põhistruktuurina, mis peegeldab selle lause olulisi verbi tähendusest tingitud süntaktilis-morfoloogilisi omadusi. Nn Rätsepa lausemallides on abstraheeritud nii verbid kui nende seotud laienditeks olevad lauseliikmed ja iga malli kirjelduse osaks on selle moodustada võivate verbide loend. Selline lausemallide esitusviis on ka meie töö lõppeesmärgiks, aga selle lõppeesmärgi saavutamiseks tuleb alustada konkreetsete leksikaalsete verbide mallide, mille hulgast kattuvad mallid ühendada verbiülesteks. Rätsepa teos sobib meile eeskujuks hästi ka põhjusel, et meie töö sisendiks olev universaalsõltuvuste (*Universal Dependencies*, edaspidi UD) süntaktiline märgendusskeem¹ (vt lähemalt ptk 4) põhineb sõltuvussüntaksi ja lause verbikesksuse põhimõtetel, millest lähtus ka Rätsep (1978: 11).

Lause põhistruktuur ehk lausemall koosneb seotud laienditest, mis tähistavad verbi tähendusstruktuuri kuuluvaid entiteete (Erelt 2017a: 65). Neist tähtsaimad on tavaliselt need situatsiooni osalised, kes tegevuse algatasid ja keda tegevus

kõige rohkem mõjutas. Seetõttu on tüüpiliselt kõige seotum subjekt, tähtsusest teine on objekt ja neile järgneb obliikva. (Croft 2022: 173) Vabad laiendid on seevastu verbi laiendid, mille olemasolu ega vorm ei sõltu verbist, vaid lause sisust, mistõttu saavad need esineda pea iga verbiga (Rätsep 1978: 15). Kuna lausemall kirjeldab ainult lause põhistruktuuri, tuleks lausemallide tuvastamisel eristada seotud laiendid vabadest ning malli liikmeks võtta ainult esimesed. Situatsiooni osaliste jagamine seotud ja vabadeks laienditeks ei ole paraku alati lihtne. Seetõttu ei vaatle mõned autorid seotust binaarse tunnuseks, millel on ainult kaks võimalikku väärtust, vaid ollakse seisukohal, et seotud ja vabad laiendid moodustavad pigem skaala (nt Croft 2001: 273–274). Meie testmaterjali ehk liigutamisverbidega lausete seisukohast on seotud ja vaba laiendi eristamise küsimus aktuaalne, sest nende verbidega lausetes on palju kohamäärusi, mis võivad olla aga nii seotud kui vabad laiendid (Veismann jt 2017: 300). Tendentsina on näiteks lähtekohamäärus pigem vaba ja sihtkohamäärus pigem seotud, mõne verbi puhul lausa obligatoorne laiend (Veismann jt 2017: 330–336).

Tänapäeva vahendid võimaldavad lausemalle leida automaatselt ja kasutus-põhiselt korpusdest kogutud statistika alusel. Taolisi võimalusi on eesti keele jaoks kasutatud küll verbiga seotud lauseliikmete leidmiseks (Muischnek, Sahkai 2009, Orasmaa 2013), kuid kumbki neist kahest uurimusest ei seadnud eesmärgiks tuvastada terviklikke lausemalle, millele sinne töö keskendub.

3. Lausemallide korpuspõhine tuvastamine

Lausemallide automaatsel tuvastamisel toetutakse peamiselt korpusetele ja sealt leitavatele koosinemistele (Simon 2023: 14). Need on tavaliselt sõna kujul, aga leksikaalseid üksusi saab abstraherida ka grammatika, näiteks sõnaliigi või fraasistruktuuri tasandile (Forsberg jt 2014: 120). Korpusest saab niisiis kätte statistika iga üksiku lauseliikme ja selle leksikaalse realisatsiooni sageduse kui ka lauseliikmete koosinemiste ja nende leksikaalsete realisatsioonide sageduse kohta. Saab mõõta ka nende üksuste omavahelise seose tugevust. Tüüpiliselt kasutatakse lauseliikmete ja sõnade koosinemise leidmiseks n-gramme või aknameetodit. N-gramm-meetodiga arvestatakse koosinemisteks ainult üksteise järel tulevad sõnad, aknameetodiga vaadatakse aga sõnu ümbritsevat laiemat konteksti. Pärast koosinemiste tuvastamist arvutatakse, kui tugevalt selle osad üksteisega seotud on. (Simon 2023: 14) Seose tugevuse leidmiseks kasutatakse vastavalt ülesandele erinevaid statistikuid: MI-skoor on hea haruldaste kollokatsioonide, nt idioomide leidmiseks, ja t-skoor, log-likelihood ning χ^2 on head sagedaste muustrite leidmiseks (Evert 2009: 1229–1230). Koosinevatele üksustele põhinedes ei pruugi olla tuvastatud ühend aga alati konstruktsioon, sest üksteise läheduses esinevad sõnad ei pruugi olla alati süntaktiliselt seotud (Simon 2023: 15).

Koosinemistest lähtub ka kollostruktuuriline analüüs, mis on sõnade ja grammatiliste struktuuride seose uurimiseks mõeldud statistiliste meetodite kogum. Peamiselt kasutatakse selle raamistiku meetodeid verbidega seotud konstruktsioonide uurimiseks, mille hulka kuuluvad ka lausemallid (Stefanowitsch 2013: 290–291). Siinse töö vaatepunktist on kollostruktuurilise analüüsi meetoditest kõige olulisem lihtne kollekseemi analüüs (Stefanowitsch, Gries 2003), mis

keskendub konstruktsioonide koosesinemisele ehk uurib seda, kui tihedalt on kaks konstruktsiooni omavahel seotud. Tavaliselt on üks neist leksikaalne ja ning teine grammatiline. Niisiis saab öelda, et selle meetodiga leitakse automaatselt, kui hästi sõna konstruktsiooni sobitub (Stefanowitsch 2013: 291–295). Seda meetodit saaks töö edasiarendusena kasutada leitud verbiülestes lausemallides esinevate verbide leidmiseks.

Meie loodud lausemallide tuvastajaga sarnaneb kõige rohkem Giulia Rambelli, Alessandro Lenci ja Thierry Poibeau loodud töövahend nimega *UDLex* (Rambelli jt 2017), mis põhineb varasemal *LexIt* süsteemil (Lenci jt 2012). *UDLex*-i sisendiks on, nagu meilgi, UD-märgendusega korpus, kuid meist erinevalt tuvastavad nad nii konkreetse verbiga seotud süntaktilisi lausemalle kui ka nende leksikaalseid realiseerimisi.

4. Materjal ja meetod

Lausemallide automaatseks tuvastamiseks kasutasime *Universal Dependencies* märgendussüsteemi alusel automaatselt ja käsitsi annoteeritud korpusi. UD on keelest sõltumatu raamistik teksti morfoloogiliseks ja sõltuvussüntaktiliseks annoteerimiseks. Morfoloogiatasandil kasutab UD laias laastus traditsioonilisi sõnaliike ja morfoloogilisi kategooriaid. UD süntaksikihi märgendus järgib sõltuvussüntaksi põhimõtteid: süntaktilised suhted on tekstisõnade vahel, vahesõlmi (fraase) ei moodustata. Süntaktiline suhe sõltuvussüntaksis on alati hierarhiasuhe: üks suhte osaline on pea ehk ülemus ja teine laiend ehk alluv. Süntaktilisel suhtel on nimi ehk märgend (süntaktiline funktsioon), millest osa on ka eesti traditsioonis tuttavad, nt subjekt või objekt. Osad süntaktilised funktsioonid on eesti traditsioonis aga vähem kasutatud, näiteks *external complement* (*xcomp*). UD põhimõtteliselt ei erista seotud ja vabu laiendeid (põhjendust vt nt de Marneffe jt 2021: 268), nt nii nimisõnaline rektiooni- kui temporaaladverbiaal saavad mõlemad obliikva (*obl*) märgendi. See põhimõte lihtsustab UD märgendamist, kuid raskendab märgendatud tekstist lausemallide tuvastamist. Kuid UD on ainus süntaksimärgendus, mida eestikeelse teksti automaatmärgendamisel praegu kasutada saab ja ilmselt jääb selleks ka lähemas tulevikus, nii et kui tahta luua lausemallide tuvastamise meetod süntaktilise analüüsi väljundit kasutades, on UD ainus võimalus.

Nagu peatükis 2 täheldatud, on piir seotud ja vaba laiendi vahel hajus ning kuna UD märgendus ei paku samuti mingit pidepunkti nende eristamiseks, oli ainus võimalus lähtuda alluvate esinemissagedusest. Seega oletame, et verbi sage alluv morfosüntaktilisele struktuurile üldistatuna on ka verbi tähenduse väljendamise seisukohalt oluline lause osaline ja peaks olema lausemalli osa.

Lausemallide automaatseks tuvastamiseks kasutasime andmeid Tasakaalus korpusest (TK), mis sisaldab võrdses koguses (5 miljonit sõna) tekste aja-, ilu- ja teaduskirjandusest ning on märgendatud automaatselt nii morfoloogiliselt kui süntaktiliselt. Korpuse morfoloogiliseks märgendamiseks kasutati eesti keele tööt-luseks mõeldud Pythoni teekide kogu EstNLTK versiooni 1.6.9 (Laur jt 2020) morfoloogilist analüsaatorit ja ühestajat Vabamorf. Morfoloogilise märgenduse korrektsus on analüüsides mitmesust arvesse võttes viimase hindamise seisuga ümardatult 86% (Saul 2022: 36). Süntaktiliseks märgendamiseks kasutati mudelina eesti keele peal

treenitud Stanzat (Qi jt 2020), mis annab välja UD märgenduse. TK märgendati mudeli 2023. aasta versiooniga, mille analüüside täpsus on ümardatult 85%.²

Lausemallide automaatse tuvastamise meetodi kvaliteedi hindamiseks loodi testandmestik, mis koosnes 28 liigutamisverbi lausemallidest (vt tabel 1).

Tabel 1. Testmaterjalina kasutatud liigutamisverbid ja nende sagedus Tasakaalus korpuses

Lemma	Sagedus	Lemma	Sagedus
<i>panema</i>	5534	<i>asetama</i>	477
<i>tooma</i>	3397	<i>liigutama</i>	360
<i>viima</i>	2903	<i>langetama</i>	278
<i>tõstma</i>	1814	<i>toppima</i>	196
<i>tõmbama</i>	1501	<i>poetama</i>	187
<i>viskama</i>	1243	<i>sättima</i>	178
<i>heitma</i>	1147	<i>paigutama</i>	168
<i>seadma</i>	802	<i>tõukama</i>	165
<i>keerama</i>	777	<i>laduma</i>	143
<i>vedama</i>	647	<i>paiskama</i>	142
<i>pistma</i>	636	<i>tirima</i>	125
<i>raputama</i>	606	<i>loopima</i>	105
<i>torkama</i>	520	<i>riputama</i>	95
<i>lükkama</i>	486		

Liigutamisverbid valiti testmaterjaliks kui suhteliselt selgepiiriline semantiline klass, kuhu kuulub nii väga sagedasi kui ka vähemsagedasi verbe. Lausemalle koostades ei eristatud konkreetseid liigutamistähendusi abstraktsetest tähendustest: lausete *Kiirabi viis kannatanu haigemajja* ja *Uus juht viis firma pankrotti* mall on sama. Ühe semantilise klassi verbide kasutamine testmaterjalina võimaldab ka tuvastada verbiüleseid, liigutamisverbidele üldiselt omaseid malle. Kuna loodav meetod tuvastab lausemalle sõltuvussüntaktiliselt märgendatud korpusest, otsustati ka testmaterjali lausemallid koostada korpuse, mitte olemasolevate leksikonide baasil. Vältimaks automaatse morfoloogilise ja süntaktilise märgenduse vigadest tulenevat müra, koostati testmaterjal UD EDT korpuse³ põhjal, mis on märgendatud käsitsi.

Järgmise etapina kirjutati lausemallide automaatseks tuvastamiseks ja lausemalli põhjal näitelauseste otsimiseks programmvahendid (Saul 2024), kasutades selleks programmeerimiskeele Python versiooni 3.6.10.

Automaatne tuvastamine otsustati rajada verbide ja morfosüntaktiliste märgendite paaride sagedusel põhinevale tuvastusele. Valik osutus selle lähenemise kasuks, sest see tundus kõige lihtsama viisina arvestada lausete keerulistest struktuuridest tuleneva müraga. Nimelt järgivad korpuselased täpset tervikliku lausemalli struktuuri üpris harva. Põhjuseid selleks on mitmeid: lausetes võib esineda lõpmatu hulk vabu laiendeid, eesti keeles saavad olla elliptilised nii lauseliikmed kui ka muutetunnused ning lausemallid lõikuvad teiste konstruktsioonidega, mis nende struktuuri muudavad. Lisaks sellele saab sama semantilise funktsiooniga lauseliikmeid grammatiliselt mitmel moel väljendada, mis hajutab lause struktuuride sagedust veelgi. Peale nende grammatilistest väljendusvõimalustest tulevate

² https://github.com/estnltk/estnltk/blob/main/tutorials/nlp_pipeline/C_syntax/03_syntactic_analysis_with_stanza.ipynb (24.3.2025).

³ https://universaldependencies.org/treebanks/et_edt/index.html (24.3.2025).

probleemide on, nagu juba varem mainitud, automaatselt analüüsitud korpuses 15% tekstisõnadest vale süntaktilise funktsiooni ja/või ülemuse märgendiga. Kõigi eelnimetatud põhjuste tõttu tundus paaridel põhinev lähenemine parim viis leida sageli koos esinevaid lauseliikmeid ilma, et lausete struktuurilised erinevused, lauseliikmete eri grammatilised väljendusviisid ja märgendusvead tulemusi liialt mõjutama hakkaksid.

Meetodi esimene samm on tuvastada iga verbi jaoks kõik selle otsesed alluvad ja koostada nende sagedusest andmestik, milles on kirjas iga uuritava verbi otsesed alluvad, nende sagedus konkreetse verbi alluvana ning sama verbi alluvate koosenemise sagedus.⁴ Alluvaid ei esitata andmestikus mitte sõnade või sõnavormidena, vaid süntaktiliste ja morfoloogiliste märgendite kombinatsioonidena, välja arvatud subjektid ja objektid, mille puhul käändekategooriat ei arvestatud. Juhul, kui verbile alluvale käändsõnale allus omakorda kaassõna, võeti andmestikku ka see.

Andmeid ei kogutud aga kõigi verbi otseste alluvate kohta, vaid ainult nende, mille süntaktiline märgend saab (aga ei pruugi) tähistada lausestruktuuri kohustuslikku elementi.⁵ Need on toodud tabelis 2. Samuti piirasime lausemalli tuumaks olla võiva verbi vormi: materjal koostati isikulise tegumoe kindlas, tingivas ja kaudses kõneviisis, oleviku ja lihtmineviku jaatavas kõnes verbivormide põhjal.

Tabel 2. Lausemallide automaatseks tuvastuseks otsitavad lauseliikmed

Süntaktiline märgend	Seletus*	Näide
nsubj	käändsõnaline subjekt	Hansapank seda ei uuri.
obj	käändsõnaline objekt	Hansapank seda ei uuri.
obl	nimisõnaline määrus ehk obliikva	See ei jäänud talle märkamatuks. Sõitis Saaremaale.
ccomp	komplementlause peasõna	Kui te mõistate, mida ma öelda tahan.
xcomp	ahelverbi infiniitsed osad, da-infiniitsed objektid, translatiivsed ja essiivsed predikatiivadverbiaalid, seotud laiendina toimivad otstarbemäärad	Vesi hakkab keema. Tahan süüa. See teeb ta valvsaks. Köölus toimib vedruna. See on mõeldud kõigi kaitseks.
advcl*	määrusena toimiv infiniittarind	Pasta ei lähe keetes tükki.
advmod	määrsõnaline määrus	Seadustest on alati kinni peetud. Nii võib kaugele jõuda.

* Seletused pärinevad UD dokumentatsioonist: https://github.com/EstSyntax/EstUD/blob/master/UD_dokumentatsioon_oktoobris_2022.pdf; <https://universaldependencies.org/u/dep/> (24.3.2025).

Tabeli 2 selgitamisel tuleb kõigepealt mainida, et UD märgendussüsteemis saab üks märgend tähistada erineva funktsiooniga lauseliikmeid. Tabelis on märgend *advcl* märgitud tärniga, sest selle kolmest erinevast funktsioonist – määruslause, määruslik infiniittarind ja võrdlustarind – jäetakse morfoloogilise märgenduse toel sisse ainult määrusena toimivad infiniittarindid, sest need saavad märgendi funktsioonidest ainsana olla lausemalli osa (Rätsep 1978: 226–229).

Järgmise sammuna paneb koostatud programm lauseliikmete paaridest kokku lausemallid. Kõigepealt arvutatakse selle jaoks välja iga üksiku lauseliikme ja

⁴ Avaldame tänu andmestiku koostajale Katrin Tsepelinalale.

⁵ Kõiki UD märgendeid saab näha siit <https://universaldependencies.org/u/dep/index.html> (24.3.2025).

lauseliikmete paari suhteline sagedus. See näitab verbipõhiselt, mitmes protsendis selle verbiga moodustatud lausetes lauseliige või lauseliikmete paar esines. Lausemalli osalise kandidaatideks arvestatakse ainult need lauseliikmed või paarid, mis esinevad üle 5% lausetes. Selle lävendi ületanud lauseliikmete paarid kombineeritakse potentsiaalseteks lausemallideks põhimõttel, et kui lauseliikmed esinevad kõik paaridena üksteisega koos üle 5% lausetes, esinevad nad ka kolmikute ja nelikutena koos. Abstraheerituna saab seda seletada nii, et kui lävendi ületasid paarid $a+b$, $a+c$ ja $b+c$, arvestatakse lausemalli kandidaadiks ka $a+b+c$. Kuna selle tulemusel tekib palju lausemalli kandidaate, mis üksteist hõlmavad, otsitakse terviklike lausemallide saamiseks iga lauseliikme jaoks üles pikim mall, milles see lauseliige esineb. Neist duplikaatide eemaldamisel saadakse kätte lõplikud lausemallid.

5. Tulemused

Automaatselt korpusest ekstraheeritud verbimalle käsitsi koostatutega võrreldes hinnati eraldi nii üksikute lauseliikmete, st ülemus-alluv paaride, kui terviklike lausemallide tuvastamise kvaliteeti.

Lauseliikmete automaatse tuvastuse ehk verb-alluv paaride tuvastamise saagis (kui suur osa lausemalli osaks olevatest verbi alluvatest tuvastati) oli 81,2% ja täpsus (kui suur osa tuvastatud alluvatest olid testmaterjalis määratud selle verbi alluvateks) 69,9%. Kuna testandmestik koostati suhteliselt väikese korpuse põhjal, siis esines automaatselt tuvastatud verbi alluvate ja lausemallide hulgas ka selliseid, mida kuldstandardis ei olnud, aga mis suure korpuse kontrollimisel osutusid siiski korrektseteks. Kui jätta kõrvale need verbi alluvad, mida loeti valesi tuvastatuteks kuldstandardi puudulikkuse tõttu, tõusis paaride täpsus 74,2%-ni. Kokkuvõttes tähendab see, et automaatselt jäi tuvastamata 19% lauseliikmetest ja tuvastatud lauseliikmetest oli õigeid ümardatult 74%.

Kui verb-alluv paaride tuvastamise kvaliteediga võib üldiselt rahule jääda, arvestades, et tegemist on esimese katsega, mille eesmärgiks on mh välja selgitada valitud meetodi tugevad ja nõrgad kohad, siis terviklike lausemallide automaatse tuvastamise saagis oli vaid 33,7% ja täpsus 30,7%. Käsitsi koostatud mallide puudujääke arvesse võttes oli lausemallide automaatse tuvastamise täpsus 33%. See tähendab, et tuvastamata jäi 66% käsitsi koostatud materjalis olevatest lausemallidest ja kõigist automaatselt tuvastatud lausemallidest oli õigeid 33%.

5.1. Vigade tüübid

Järgnevalt vaatleme sagedasemaid vigaselt tuvastatud lausemallide tüüpe.

Vigastest lausemallidest 59% puudub lausemallis mõni testmaterjalis lausemalli osaks määratud lauseliige, mis on ka sagedasim lausemallide tuvastamise viga. Kõige sagedamini on puudu nimisõnaline määrus illatiivis (näide 1), mis moodustas kõigist tuvastamata jäänud lauseliikmetest 50% ning mille olemasolu verbimallis ei tuvastatud 70% verbide puhul, milles see käsitsi koostatud mallis seotud laiendina esines. Sageduselt teine tuvastamata jäänud lauseliige oli *ma*-infinitiivne laiend (2–3), mis jäi tuvastamata kahel kolmandikul seda nõudvatest verbidest.

Korpusenäiteid uurides jäi silma, et viimaste puhul on tegemist küll lingvistiliselt oluliste, aga üpris perifeerse kasutusega mallidega. Ilmselt jäid need tuvastamata just selles töös rakendatava statistikal põhineva meetodi tõttu, mis tuvastab ainult kõige sagedasemaid malle.

- (1) Tükk toidust visatakse *lõkkesse*, et jumalad saaksid pidusöögist osa. (ÜS 2024)
- (2) Väljast tulnu aga heitis narile *puhkama*. (TK)
- (3) Eesav tõmbas uue tiku *põlema* ja lähendas jälle pealaele. (TK)

Üleliigse liikmega lausemallid moodustasid vigastest mallidest 29%. Pooltel juhtudel oli üleliigseks lauseliikmeks määrsõnaline määrus, mis on UD süsteemis suur ja heterogeenne hulk, mida pole jagatud ei semantilistesse (määruse liigid) ega süntaktilistesse (lauselaiend *vs.* verbi laiendav määrus) alaliikidesse. 15% üleliigsetest lauseliikmetest moodustasid nimisõnalised määrused adessiivis, mis liigutamisverbide puhul tüüpiliselt pigem ei ole lausemalli osaks või on selleks ainult teatud kindlas konstruktsioonis. Näiteks verbi *vedama* puhul on nimisõnaline adessiivikäändes määrus verbi seotud laiendiks ainult juhul, kui lauses pole objekti (4).

- (4) Samas pidi ta tunnistama, et *sõbrataril* vedas armsamaga üsna hästi. (TK)

Mõnda testmaterjali põhjal üleliigseks loetud lauseliiget võib pidada aga ka võimalikuks lausemalli osaks, sest testmaterjali koostamisel lähtuti lauseliikme lausemalli osaks arvestamisel sagedusest ja koostaja intuitsioonist, mis jätab aga lauseliikme seotuse otsustamisel ruumi vaidlusteks. Kui subjekti, objekti, predikatiiv- või reksiooniadverbiaali puhul kõhklusi ei teki, siis näiteks instrumentaaladverbiaali puhul aga küll. Nii tuvastati automaatselt viie verbi (*vedama*, *tõukama*, *tõmbama*, *lükka*, *sõidutama*) lausemalli osaks nimisõnaline määrus komitatiivis (5–6), mis käsitsi koostatud materjalis jäeti vaba laiendina lausemallist välja. Kõrge sageduse tõttu võib see aga pigem paigutuda ka vaba ja seotud laiendi piirimaile, st võib-olla oleks mõistlik see lausemalli osaks kuulutada.

- (5) Kaubandusettevõtte vedavat neid *bussiga* kohale Ida-Virumaalt. (TK)
- (6) Bronka tõukab *jalaga* väikest kappi. (TK)

5.2. Vigade põhjusi

Järgnevalt vaatleme verbimallide tuvastamisvigade võimalikke põhjusi.

Nagu juba mainitud, on pooltes üleliigse liikmega lausemallides selleks üleliigseks lauseliikmeks määrsõnaline määrus, mis on UD süsteemis suur ja heterogeenne hulk, mida pole jagatud ei semantilistesse (määruse liigid) ega süntaktilistesse (lauselaiend *vs.* verbi laiendav määrus) alaliikidesse. Seetõttu on sama märgendi alla koondunud eri funktsiooniga lauseliikmed, mis suurendab ka märgendi esinemissagedust. Kuna lausemalli liikmeid tuvastati ainult märgendite sagedusele toetudes, arvestas meetod määruse märgendiga lauseliikme tihti selle kõrge sageduse tõttu ekslikult lausemalli osaks.

Kindlasti raskendab lausemallide tuvastamist ka verbide polüseemia. Eesti keele jaoks puudub hetkel tekstis verbide tähendust eristav ressurss (sõnatähen-duste ühestaja), mistõttu ei ole automaatselt võimalik lauseid verbitähenduse

järgi jaotada ja malle tähenduste kaupa tuvastada. Verbi eri tähendustel saavad aga olla erinevad lausemallid. Näited (7–9) esindavad verbi *heitma* intransitiivset (7) ja transitiivset kasutust (8) ning transitiivse malli varianti, kus verbi *heitma* kasutatakse saatelauseverbina (9). Hetkel kasutatav tuvastamismeetod on kahest esimesest kasutusest kokku pannud subjektist, objektist, määrsõnalisest määrusest ning allatiivi käändes nimisõnalisest määrusest koosneva malli. Saatelauseverbiline kasutus oli ilmselt liiga harv oma malli moodustamiseks või ekslikult teise malli osaliseks pääsemiseks.

- (7) [---] kes tippspordist loobudes diivanile heidavad [---] (TK)
- (8) [---] peagi heitis ta põrandale münte täis kukru. (TK)
- (9) “Lase jalga, Gregg” heitis Lamont üle õla. (TK)

Lausemallid võivad lõikuda ka teiste konstruktsioonidega, mis lause struktuuri mõjutama hakkavad (Hilpert 2014: 28–29). Näiteks kui tavaliselt väljendab verb *tirima* ainult liigutamise tähendust ning kasutab selleks kohamäärusi (10), siis muutmiskonstruktsiooni sees kasutatakse verbi *tirima* hoopis seisundimuutuse tähenduses, mille tulemusel lisandub verbi laiendite hulka kohamääruse asemel translatiivne predikatiivadverbiaal (11) (Erelt 2017b: 292).

- (10) Tirisime *lavale* kontsertklaveri. (ÜS 2024)
- (11) Hiljem kuulis Anne, et sama papagoi oligi varem ühel restoranikülalisel käe *ribadeks* tirinud. (Koondkorpus)

Automaatsel tuvastamisel on keeruline selliste konstruktsioonide lõikumisega arvestada, mistõttu loetakse näites (11) verbi *tirima* otseseks alluvaks ehk lausemalli osalise kandidaadiks ka translatiivne predikatiivadverbiaal *ribadeks*. Kui selliseid juhte esineb mingil põhjusel palju, võivad need hakata verbi teiste laiendite sagedust langetama, mistõttu võis mõni tähenduselt oluline, kuid sageduselt piiripealne lausemalli liige tuvastamata jääda.

Automaatselt leitud sageduse põhjal seotud laiendite tuvastamisel peab arvestama ka sellega, et eesti keeles saab pea kõiki lauseliikmeid lausest välja jätta, kui neid saab tuletada kontekstist, teise lauseliikme grammatilisest infost või ümbritsevast situatsioonist. Elliptilised võivad olla nii subjekt, objekt, määrus kui ka verb ise. (Erelt 2017c: 590–597) Kui elliptilisi kasutusi on korpuses piisavalt palju, tuvastabki meie meetod eraldi ka elliptilise malli. Kui nt *sööma*-verbil ongi kaks malli, objektiga ja ilma, siis mittepöluseemsete liigutamisverbide puhul jääb objekt välja ainult kontekstiellipsi puhul, aga neid kahte juhtumit on automaatselt keeruline eristada.

Tuleb rõhutada, et meetodi praegune versioon tegeleb ainult lihtverbide lausemallidega. UD süntaktilise märgenduse alusel on kirjeldatavas eksperimendis ära tuntud ja kõrvale jäetud ühendverbid, aga väljendverbe ei kohelda UD märgendus-süsteemis ühe tervikliku predikaadina, vaid väljendverbi käändsõnaline osis saab märgendi vastavalt oma formaalsüntaktilisele funktsioonile. See aga tähendab, et lihtverbi alluvate sageduse hulka võib sattuda ka väljendverbidest tulenev sagedus, mille tulemusel liht- ja väljendverbide mallid segunevad. Õnneks kaldub väljendverbi osaks olev käändsõna järgima lihtverbi malli (Rätsep 1978: 21), aga muidugi mitte alati. See tähendab, et kui lihtverbist erineva lausemalliga väljendverb esineb küllalt sageli, võib selle mall sattuda ekslikult automaatselt tuvastatud lihtverbi

malliks. Näiteks tuvastas süsteem verbi *torkama* lausemalli osaks inessiivse määruse, mis osutus kontrollimisel väljendverbi *silma torkama* laiendiks (12).

(12) *Haljastuses* torkasid silma palmid ja Norfolki männid. (TK)

Üleliigse liikmega malli tekkimisele aitab pisut kaasa tuvastamismeetod, mis, nagu öeldud, võtab hetkel kõigist genereeritud mallidest iga lauseliikme jaoks kõige pikema lausemalli. Seepärast jäi leidmata näiteks *tõstma* mall subjekti ja objektiga, sest *tõstma* jaoks oli tuvastatud ka mall subjekti, objekti ja määrsõnalise määrusega, mis sisaldas mõlemat lauseliiget ja oli esimesest mallist pikem.

5.3. Uued ja verbiülesed mallid

15 miljoni sõna suuruse Tasakaalus korpuse põhjal tuvastas meetod ka selliseid liigutamisverbide malle, mida polnud ei testmaterjalis ega Huno Rätsepa lihtverbi lausemallide seas.

Verbile *liigutama* tuvastati kaks uut lausemalli, millest kummaski ei ole liigutamisverbidele tüüpilisi sihtkohta ja lähtekohta väljendavaid määrusi, vaid hoopis subjekt, objekt ja inessiivne kohamäärus või muidu tüüpiliselt vaba laiendina esinev määrsõnaline määrus. Mõlemast mallist on olemas variant, milles objekti rollis on refleksiivne pronoomen *end*, mis on huvitav, sest verbil *liigutama* on olemas ka intransitiivne kasutus, st *end* ei ole selle malli osaline pelgalt selle pärast, et *liigutama* nõuaks tingimata sihitist.

Subjekti, objekti ja inessiivis kohamäärusega malli lausenäidetest selgus, et inessiivset kohamäärust võib tõlgendada kogejana ja subjekti või objekti rollis on tunne või mõte (13–15). Tunnetele viitavates *liigutama*-lausetes muutub lause inessiivis määruse väljajätmisel arusaamatuks, mistõttu see on määratud lausemalli osaks.

(13) Mingi ähmane aimdus liigutas end *Luiges*. (TK)

(14) Väike okas liigutas end Tiia *sisemuses* [---] (TK)

(15) Muusikapalade hulgas oli tihti lemmikpalasid, mis *hinges* imelisi keeli liigutasid. (TK)

Verbile *liigutama* leitud teine uus mall subjekti, objekti ja määrsõnalise määrusega tundub rakenduvat vaid siis, kui määrsõnaline määrus väljendab viisi ja objekti rollis on kas subjekt ise (16) või mõni tema kehaosa (17). Kui viisimäärusega lauses ei väljenda objekt ja subjekt sama entiteeti (18) või enesekohase objektiga lauses ei väljenda määrus viisi (19), pole määrsõnaline määrus ka kohustuslik.

(16) Ta liigutas *end kiiresti* ja jõudis paari minutiga tee otsa kiriku ette. (TK)

(17) Ta on halvatud, *vaevu* liigutab *mõnda sõrme*. (ÜS 2024)

(18) *Mind* liigutas *sügavalt* Teie südamlilik *kiri* minu artikli kohta. (TK)

(19) *Allpool* liigutas *end* miski. (TK)

Verbile *pistma* tuvastati samuti üks uus lausemall, mis koosneb objektist, kohta väljendavast nimisõnalisest määrusest aditiivis ja saajat väljendavast nimisõnalisest määrusest allatiivis. Just saajat väljendavat lauseliiget polnud ei meie testmaterjalis ega Rätsepa lausemallides verbi *pistma* lausemalli osana varem käsitletud. Küll oli tuvastatud mallil puudusi, sest malli osaks peaks tegelikult olema ka subjekt, mis jäi ellipsi tõttu tuvastamata.

(20) Pistsin *talle* tuki näkku, lükkasin raisa eemale ja tee letini oli vaba. (TK)

(21) Pistab *mulle* paberi pihku, et näe, astu ka kuuritsasõprade seltsi. (TK)

Peale uute mallide võib positiivse tulemusena mainida verbipõhiste mallide üldistusena tuvastatud verbiüleseid malle. Sama semantilise klassi verbid jagavad tihti lausemalle ning see tendents kehtib ka liigutamisverbide puhul, millele tuvastati kokku 16 verbiülest malli. Vaatame lähemalt neist kolme kõige sagedasemat.

Esimesed kaks sagedasimat tuvastatud lausemalli olid üheteistkümmel verbiühised. Neist esimene koosneb subjektist, objektist ja määrsõnalisest määrusest (VERB nsubj obj advmod), mis tuvastati mallina verbidele *loopima*, *raputama*, *poetama*, *langetama*, *paigutama*, *lukkama*, *tõukama*, *keerama*, *tõmbama*, *tõstma* ja *liigutama*. Selle süntaktilise malli taga on küll mitu semantilise tasandi lausemalli, sest määrsõnaline määrus võib väljendada tegevuse igasuguseid asjaolusid (nagu viisi näites (22), aga ka sihtkohta või sihtolukorda (23)). Selle põhjuseks on mitu korda mainitud asjaolu, et UD süsteemis ei eristata seotud ja vabu laiendeid ega jagata määruslikke laiendeid alaliikidesse.

(22) Ivo keeras laisalt pead. (TK)

(23) Need tõukasid ta eemale. (TK)

Teine sagedasim verbiüle mall koosneb subjektist, objektist ja nimisõnalisest määrusest allatiivis (VERB nsubj obj obl_all) ning tuvastati mallina verbidele *loopima*, *paiskama*, *poetama*, *langetama*, *paigutama*, *vedama*, *lukkama*, *tõukama*, *keerama*, *tõmbama* ja *tõstma*.

(24) Tuhandete hobujõudude rammust värisev alus keerab end merele. (TK)

(25) Kohus lükkas uue istungi 8. oktoobriks. (TK)

Kümne liigutamisverbi ühine lausemall oli kahe eelnevalt kirjeldatud malli kombinatsioon, mis koosneb verbist, subjektist, objektist, määrsõnalisest määrusest ja nimisõnalisest määrusest allatiivis (VERB nsubj obj obl_all advmod). Sellise malli moodustavad verbid *sättima*, *laduma*, *asetama*, *sõidutama*, *heitma*, *seadma*, *viskama*, *viima*, *tooma* ja *panema*.

(26) Eva [---] tõukab Bronka jõuliselt põrandale. (TK)

(27) [---] sättisid peegli räpakalt seinale tagasi. (TK)

Kõik neis mallides esinevad lauseliikmed osutusid liigutamisverbide seas ka kõige sagedasemateks lausemalli osadeks. Neist tuvastati subjekt, objekt ja määrsõnaline määrus kõigi verbide lausemallide liikmeks ning nimisõnaline määrus allatiivis ei esinenud vaid kahe verbi – *raputama* ja *liigutama* – lausemallides.

6. Töö võimalikud edasiarendused

Eelneva veaanalüüsi põhjal pakume välja võimalusi meetodi edasiarendamiseks. UD süsteemi piirangutest tulenevalt koonduvad eri funktsiooni ja seotuse tasemega määrused samade märgendite alla, mistõttu võib ühe sagedase märgendi all tegelikult peituda mitu erinevat lauseliiget. Nende üksteisest eristamiseks tuleks määrused automaatselt erinevatesse semantilistesse alamliikidesse määrata.

Tulemaks paremini toime subjekti sagedase ellipsiga, tuleks lausemallidesse subjekt automaatselt lisada, v.a väheste nullvalentsete verbide puhul. Teine variant on kasutada UD täiustatud sõltuvuste (*enhanced dependencies*) märgenduskihti, milles on kontekstiellipsid märgendatud.

Nagu 4. peatükis öeldud, ei üritata selle töö raames eristada seotud ja vabu laiendeid, vaid loetakse lausemalli osaks sagedasemad laiendid. Siin artiklis kirjeldatud esimeses katses arvestati lausemalli kandidaatideks need lauseliikmed või paarid, mis esinevad üle 5% lausetes. See lävend selgus olevat madalavõitu, st ekstraheeritud lausemallikandidaadid sisaldasid verbi laiendeid, mis asuvad seotud-vaba laiendi skaala vaba laiendi poolises otsas. Teisest küljest soodustas madal lävend ka harvemesinevate lausemallide tuvastamist. Läveni optimaalne väärtus tuleks edasise töö käigus välja selgitada katsete abil. Samuti võiks proovida kasutada seotud ja vabade laiendite eristamiseks statistilisi meetodeid.

Kuna statistikapõhised meetodid toimivad paremini rohkemate andmete puhul, oleks võib-olla mõistlik keskenduda ainult andmestiku aluseks olevas korpusel piisavalt sageli esinevate verbide mallidele.

Kindlasti oleks abi väljendverbide leksikonist, et väljendverbide mallid ei seguneks lihtverbi omadega.

7. Kokkuvõte

Lausemall on verbist lähtuv lause põhistruktuuri formaalne esitus, mille loend on väärtuslik ressurss nii keeletehnoloogias kui ka keeleõppes, leksikograafias ja konstruktikograafias. Eesti keele jaoks on olemas keeleteadlase intuitsioonil põhinev lihtverbide lausemallide kogu (Rätsep 1978), aga puudub tegeliku keelekasutuse korpusandmetel põhinev loend. Selles artiklis pakkusimegi välja esialgse meetodi verbikesksete lausemallide automaatseks tuvastamiseks sõltuvussüntaktiliselt märgendatud korpusel baasil ning valideerisime väljapakutud meetodit liigutamisverbide käsitsi koostatud lausemallide abil.

Väljapakutud meetod moodustab programmi abil verbist ja selle otsestest alluvatest paarid, filtreerib neist välja alla 5% sageduslävendi jäävad paarid ning kombineerib allesjäänutest lausemallid. Kui verb-alluv paaride tuvastamise kvaliteediga võib rahule jääda (saagis 81,2% ja täpsus 69,9%), siis terviklausemallide tuvastamisel on parandamisruumi, praegu on saagis 33,7% ja täpsus 33%.

Vigade oluliste põhjustena saab välja tuua selle, et sisendiks oleva korpusel süntaktiline märgendus ei ole piisavalt detailne ning et hetkel puudub meetodis mehhanism verbi polüseemia ja elliptiliste lauseliikmetega arvestamiseks. Üks olulistest vigade põhjustest oli ka see, et kui ühend- ja lihtverbide malle saab tänu süntaktilisele märgendusele hõlpsalt eristada, siis väljendverbid on sisendis märgendamata, mille tulemusel lihtverbi ja väljendverbi mallid segunevad. Nende vigade analüüsi põhjal pakkusime välja peamised töösuunad meetodi edasiarendamiseks. Lisaväärtusena õnnestus tuvastada verbiüleseid malle ning leidsime mõne uue, eesti keele leksikograafilistes ressurssides puuduvad mallid.

Viidatud kirjandus

- Colman, Lut; Tiberius, Carole 2018. A good match: A Dutch collocation, idiom and pattern dictionary combined. – Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: University of Ljubljana Press, 233–246. <https://doi.org/doi.org/10.4312/9789610600961>
- Croft, William 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198299554.001.0001>
- Croft, William 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781316145289>
- De Marneffe, Marie-Catherine; Manning, Christopher D.; Nivre, Joakim; Zeman, Daniel 2021. Universal dependencies. – *Computational Linguistics* 47 (2), 255–308. https://doi.org/10.1162/coli_a_00402
- Erelt, Mati 2017a. Sissejuhatus süntaksisse. – Mati Erelt, Helle Metslang (Toim.), *Eesti keele süntaks. Eesti keele varamu III*. Tartu: Tartu Ülikooli Kirjastus, 53–89.
- Erelt, Mati 2017b. Öeldistäitemäärus. – Mati Erelt, Helle Metslang (Toim.), *Eesti keele süntaks. Eesti keele varamu III*. Tartu: Tartu Ülikooli Kirjastus, 289–299.
- Erelt, Mati 2017c. Ellipsis. – Mati Erelt, Helle Metslang (Toim.), *Eesti keele süntaks. Eesti keele varamu III*. Tartu: Tartu Ülikooli Kirjastus, 590–602.
- Evert, Stefan 2009. Corpora and collocations. – Anke Lüdeling, Merja Kytö (Eds.), *Corpus Linguistics: An International Handbook, Vol.2. Handbooks of Linguistics and Communication Science, 29.2*. Berlin–New York: De Gruyter Mouton, 1212–1248. <https://doi.org/10.1515/9783110213881.2.1212>
- Forsberg, Markus; Johansson, Richard; Bäckström, Linnéa; Borin, Lars; Lyngfelt, Benjamin; Olofsson, Joel; Prentice, Julia 2014. From construction candidates to construction entries: An experiment using semi-automatic methods for identifying constructions in corpora. – *Constructions and Frames* 6 (1), 114–135. <https://doi.org/10.1075/cf.6.1.07for>
- Hilpert, Martin 2014. Argument structure constructions. – *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press, 25–49.
- Kallas, Jelena 2013. *Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias*. Tallinna Ülikooli humanitaarteaduste dissertatsioonid, 32. Tallinn: Tallinna Ülikool.
- Kipper-Schuler, Karin 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD. Pennsylvania: University of Pennsylvania.
- Laur, Sven; Orasmaa, Siim; Särg, Dage; Tammo, Paul 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 7154–7162. <https://www.aclweb.org/anthology/2020.lrec-1.884>
- Lenci, Alessandro; Lapesa, Gabriella; Bonansinga, Giulia 2012. LexIt: A computational resource on Italian argument structure. – *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC12)*. Istanbul: European Language Resources Association, 3712–3718.
- Muischnek, Kadri; Sakhai, Heete 2009. Using collocation-finding methods to extract constructions and to estimate their productivity. – Magnus Sahlgren, Ola Knutsson (Eds.), *NODALIDA Workshop on extracting and using constructions in NLP*. SICS Technical Report T2009:10, 22–27.
- Orasmaa, Siim 2013. *Verb Subcategorisation Acquisition for Estonian Based on Morphological Information*. – Elmar Nöth, Aleš Horák, Petr Sojka (Eds.), *Text, Speech, and Dialogue*.

- Lecture Notes in Computer Science, 15049. Berlin–Heidelberg: Springer, 583–590. https://doi.org/10.1007/978-3-642-40585-3_73
- Palmer, Martha; Gildea, Daniel; Kingsbury, Paul 2005. The proposition bank: An annotated corpus of semantic roles. – *Computational Linguistics* 31 (1), 71–106. <https://doi.org/10.1162/0891201053630264>
- Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason; Manning, Christopher D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. <https://doi.org/10.48550/arXiv.2003.07082>
- Rambelli, Giulia; Lenci, Alessandro; Poibeau, Thierry 2017. UDLex: Towards cross-language subcategorization lexicons. – Simonetta Montemagni, Joakim Nivre (Eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Linköping University Electronic Press, 207–217.
- Ruppenhofer, Josef; Ellsworth, Michael; Schwarzer-Petruck, Myriam; Johnson, Christopher R.; Scheffczyk, Jan 2016. *FrameNet II: Extended Theory and Practice*. (Revised November 1, 2016.) https://framenet.icsi.berkeley.edu/the_book
- Rätsep, Huno 1978. *Eesti keele lihtlausete tüübid*. Tallinn: Valgus.
- Saul, Kertu 2022. EstNLTk morfoloogilise analüsaatori ja ühestaja kvaliteedi hindamine. BA. Tartu: Tartu Ülikool. <http://hdl.handle.net/10062/82639>
- Simon, Gábor 2023. Constructions, collocations, and patterns: Alternative ways of construction identification in a usage-based, corpus-driven theoretical framework. – Claire Bonial, Harish Tayyar Madabushi (Eds.), *Proceedings of the 1st International Workshop on Construction Grammars and NLP*. Association for Computational Linguistics, 12–20. <https://aclanthology.org/2023.cxgslp-1.2>
- Stefanowitsch, Anatol 2013. Collostructional analysis. – Thomas Hoffmann, Graeme Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 290–306. <https://doi.org/10.1093/oxfordhb/9780195396683.013.0016>
- Stefanowitsch, Anatol; Gries, Stefan T. 2003. Collostructions: Investigating the interaction of words and constructions. – *International Journal of Corpus Linguistics*, 8 (2), 209–243. <https://doi.org/10.1075/ijcl.8.2.03ste>
- Tiberius, Carole; Kallas, Jelen; Koeva, Sveta; Langemets, Margit; Kosem, Iztok 2024. A lexicographic practice map of Europe. – *International Journal of Lexicography* 37 (1), 1–28. <https://doi.org/10.1093/ijl/ecado23>
- Veismann, Ann; Erelt, Mati; Metslang, Helle 2017. *Määrus*. – Mati Erelt, Helle Metslang (Toim.), *Eesti keele süntaks. Eesti keele varamu III*. Tartu: Tartu Ülikooli Kirjastus.

Võrguviited

- EstNLTk = Open source tools for Estonian natural language processing. <https://github.com/estnltk> (24.3.2025).
- Koondkorpus. Tartu Ülikooli arvutilingvistika uurimisrühm. <https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et> (24.3.2025).
- Saul, Kertu 2024. Lausemallide automaatne tuvastamine. https://github.com/wertepure/Lausemallide_automaatne_tuvastamine (24.3.2025).
- TK = Tasakaalus korpus. Tartu Ülikooli arvutilingvistika uurimisrühm. <https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=et> (24.3.2025).
- UD = Universal Dependencies. <https://universaldependencies.org> (24.3.2025).
- Vabamorf. <https://github.com/Filosoft/vabamorf> (24.3.2025).
- ÜS 2024 = EKI ühendsõnastik 2024. Sõnaveeb. Eesti Keele Instituut. <https://sonaveeb.ee> (24.3.2025).

EXTRACTING VALENCY PATTERNS FROM A SYNTACTICALLY ANNOTATED CORPUS

Kertu Saul^{1,2}, Kadri Muischnek², Jelena Kallas¹

Institute of the Estonian Language¹, University of Tartu²

The article presents a computational linguistics experiment focused on automating the detection of Estonian valency patterns, specifically using caused-motion verbs as a case study. The goal is to develop an initial workflow for corpus-based automatic detection of valency patterns in Estonian, leveraging the Universal Dependencies annotation system (de Marneffe et al. 2021). A dataset comprising 28 motion verbs was employed to test the proposed method, which pairs verbs with direct dependents and filters statistically significant results to form valency patterns. The method showed high performance in identifying verb-dependent pairs (81.2% recall, 69.9% precision) but was less effective at detecting complete valency patterns (33.7% recall, 33% precision).

The analysis addresses various detection errors, offering targeted solutions to improve performance. Key obstacles included the misidentification of adverbial modifiers as arguments, the failure to detect certain oblique cases (notably the illative), and the exclusion of some arguments. Despite these challenges, the method succeeded in identifying previously undocumented valency patterns, such as a unique structure for the verb *liigutama* ('to move/to touch') with an emotional context and another for *pistma* ('to put') denoting specific syntactic roles (Goal and Recipient).

The author's evaluation underscores methodological limitations, such as filtering biases and insufficient corpus size, which impacted precision and recall. Proposed improvements include adjusting frequency thresholds, refining filters for phraseological verbs, and enhancing argument-adjunct distinctions. These advancements aim to refine automatic valency pattern detection and enhance grammatical information presentation within Estonian lexicographical resources. The study contributes significant insights for the computational processing of Estonian syntax and suggests further directions for the efficient representation of valency patterns.

Keywords: computational linguistics, computational lexicography, syntax, dependency syntax, valency pattern, Estonian

Kertu Saul (Eesti Keele Instituut, Tartu Ülikool) on uurinud arvutilingvistiliste meetoditega eesti keele morfosüntaksit.

Roosikrantsi 6, 10119 Tallinn, Estonia
kertu.saul@eki.ee

Kadri Muischneki (Tartu Ülikool) teaduslikud huvialad on korpuslingvistika, eesti keele süntaktiline struktuur ning automaatne süntaktiline analüüs.

Jakobi 2, 50090 Tartu, Estonia
kadri.muischnek@ut.ee

Jelena Kallase (Eesti Keele Instituut) põhilised uurimisvaldkonnad on korpuslingvistika, korpusleksikograafia, õppeleksikograafia.

Roosikrantsi 6, 10119 Tallinn, Estonia
jelena.kallas@eki.ee