doi:10.5128/ERYa12.01

EESTI KEELE ÜHENDVERBIDE KOMPOSITSIONAALSUSE MÄÄRAMINE

Eleri Aedmaa

Ülevaade. Keele automaattöötluse jaoks on püsiühendite tuvastamine oluline ülesanne, mille lahendamiseks on püütud ühendeid eri meetodeid rakendades automaatselt klassifitseerida ning nende kompositsionaalsust määrata. Artiklis rakendatakse sõnadevahelise seose tugevuse mõõtmise statistilisi meetodeid eesti keele ühendverbide automaatseks klassifitseerimiseks nende tähenduse moodustamise viisi alusel ning vaadeldakse, millise meetodi tulemused on kõige paremad ja kas need on piisavalt head, et ühendverbide jaotus võiks sellele meetodile tugineda. Uurimuse põhieesmärk on välja selgitada, kas distributiivse semantika vahendeid rakendades on võimalik automaatselt kindlaks määrata eesti keele püsiühendite kompositsionaalsuse taset. Selleks tutvustatakse ja rakendatakse distributiivsel semantikal põhinevat tarkvara word2vec.*

Võtmesõnad: distributiivne semantika, keeletöötlus, püsiühend, ühendverb, eesti keel

1. Sissejuhatus

Püsiühendit (ingl *multiword expression*) on arvutilingvistilistes uurimustes määratletud erinevalt, näiteks idiosünkraatilise sõnaühendina (Sag jt 2002: 2) või igasuguse sõnaühendina, mis mingi tähenduse väljendamiseks koos esineb (Kühner, Schulte im Walde 2010: 47). Püsiühendi mõiste alla kuuluvad sõnaühendite rühmad erinevad üksteisest püsivuse astme, tähenduse moodustamise viisi ja süntaktilise struktuuri poolest (Kaalep, Muischnek 2009: 158). Tähenduse moodustamise viisi poolest peavadki mõned uurijad püsiühendeid ainult idiosünkraatilisteks ehk mittekompositsionaalseteks (Sag jt 2002), mis tähendab, et ühendi tähendus ei ole tema osade tähendustest otseselt tuletatav (Manning, Schütze 1999: 184). Teised autorid (nt McCarthy jt 2003) aga leiavad, et püsiühendid paigutuvad skaalale, mille ühes otsas on kompositsionaalsed ühendid, mille tähendus on tuletatav komponentide

* Kirjutise valmimist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute tippkeskus), see on seotud Eesti Haridus- ja Teadusministeeriumi uurimisprojektiga IUT20-56 "Eesti keele arvutimudelid".

tähendustest, ning teises otsas mittekompositsionaalsed ühendid (Kühner, Schulte im Walde 2010: 47).

Püsiühendite ja nende kompositsionaalsuse tuvastamine on oluline nii leksikograafide jaoks, otsustamaks, kas ühend esitada sõnastikukirjena või mitte, kui ka keeletehnoloogia rakenduste tarvis, et teada, missuguseid sõnu kohelda koos ja milliseid eraldi (Kühner, Schulte im Walde 2010: 47). Eesti keele põhjal on tehtud mitmeid ühendverbide automaatset tuvastamist puudutavaid uurimusi. Näiteks Heiki-Jaan Kaalep ja Kadri Muischnek (2002) tuvastasid eestikeelsest tekstikorpusest ühend- ja väljendverbe, kombineerides lingvistilisi ja statistilisi meetodeid. Selgus, et tekstikorpusest verbiühendite leidmine ei ole triviaalne ja väljund vajab käsitsi toimetamist, kuid probleemidele vaatamata sobib rakendatud tööriist SENVA hästi vaba sõnajärje ja keeruka morfoloogiaga eesti keele töötlemiseks. Kristel Uiboaed (2010) tuvastas automaatselt kaheliikmelisi ühendverbe murdekorpuse materjalist ning leidis, et erinevat tüüpi statistilised mõõdikud sobivad erinevat tüüpi ülesannete lahendamiseks. Lisaks loodi 2010. aastal TÜ tasakaalus korpuse jaoks veebis kasutatav automaatne kollokatsioonide leidja,[1] mis rakendab kolme mõõdikut: log-tõepära funktsiooni, vastastikuse informatsiooni väärtust ja minimaalset tundlikkust. Jelena Kallas (2013) kasutas ühendverbide tuvastamiseks lisaks statistilistele meetoditele ka reeglipõhist lähenemist. Käesolev uurimus on järg Eleri Aedmaa (2015) uurimusele ühendverbide automaatse tuvastamise kohta, kus rakendati tekstikorpusest ühendverbide tuvastamiseks statistilisi mõõdikuid. Ühendverbide tuvastamiseks teistes keeltes on rakendatud mitmeid meetodeid: näiteks Stefan Evert ja Brigitte Krenn (2001) kasutasid erinevaid mõõdikuid saksa keele ühendverbide tuvastamiseks, Timothy Baldwin ja Aline Villavicencio (2002) tuvastasid parserite abil inglise keele verbiühendeid ning Don Blaheta ja Mark Johnson (2001) töötasid sama ülesande lahendamiseks välja log-lineaarse mudeli.

Palju on tehtud katseid automaatselt klassifitseerida erinevate keelte püsiühendeid eri tüüpidesse. Näiteks Graham Katz ja Eugenie Giesbrecht (2006) proovisid mitmesuguseid meetodeid kasutades automaatselt eristada kompositsionaalseid ühendeid mittekompositsionaalsetest. Nende uurimus aga kinnitas varasemat Colin Bannardi, Timothy Baldwini ja Alex Lascaridese (2003) väidet, et püsiühendeid ei ole mõtet jagada kaheks selgeks rühmaks – kompositsionaalseteks ja mittekompositsionaalseteks –, vaid need moodustavad kontiinuumi kompositsionaalsuse-mittekompositsionaalsuse teljel. Selle väite järgi ei peaks ühendeid jagama klassidesse, vaid määrama nende kompositsionaalsuse taset. Püsiühendite kompositsionaalsuse tuvastamiseks inglise keeles (näiteks Lin 1998, Schone, Jurafsky 2001) on rakendatud distributiivset (ingl *distributional semantics*) ehk tõenäosuslikku (*probabilistic semantics*) ehk vektorsemantikat (*vector semantics*) (vt lähemalt Harris 1954), mille abil saab mõõta suurte keeleandmete põhjal sõnade, tekstide jms keeleüksuste tähenduste sarnasust.

Kuna eesti keele püsiühendeid pole seni püütud automaatselt tüüpidesse jagada ega nende kompositsionaalsust määrata, siis vaadeldaksegi siinses artiklis esmalt, kas ja kuidas on võimalik eesti keele ühendverbe jagada kahte klassi – ainukordseteks ja korrapärasteks –, ning seejärel otsitakse võimalusi ühendverbide kompositsionaalsuse taseme esitamiseks. Esimese ülesande lahendamiseks kasutatakse sõnadevahelise seose tugevuse mõõdikuid (vt lähemalt Aedmaa 2015) – statistilisi

[1] https://korpused.keeleressursid.ee/clc/ (30.9.2015).

valemeid, mille abil saab välja arvutada kahe sõna vahelise seose tugevuse väärtuse. Teise ülesande jaoks kasutatakse aga distributiivse semantika mudelit, mida rakendatakse tööriistaga word2vec.[2]

Artikkel on üles ehitatud järgnevalt. Esmalt antakse ülevaade eesti keele ühendverbist, selle liigitamise alustest ning kompositsionaalsusest, millele järgneb distributiivset semantikat ja selle mudeleid kirjeldav osa. Seejärel tutvustatakse varasemaid uurimusi, mis puudutavad püsiühendite automaatset klassifitseerimist ja kompositsionaalsuse tuvastamist. Viimaks kirjeldatakse materjali ning selle põhjal tehtud ühendverbide liigitamist ning nende kompositsionaalsuse taseme tuvastamist.

2. Eesti keele ühendverb

Ühendverbi moodustavad verb ja sellega kokku kuuluv afiksaaladverb, kusjuures ühendi põhisisu kannab verbiline osis ja afiksaaladverb lisab mingi tähendusnüansi (Erelt 2013: 62). Vastavalt sellele, millise tähenduse adverb ühendverbile lisab, jagab Huno Rätsep (1978: 29–33) adverbid kolme rühma: orientatsiooni-, perfektiivsus- ja seisundiadverbid. Sama liigitust järgib ka "Eesti keele grammatika" (EKG II: 20–22), mis esitab lisaks modaalsust väljendavad adverbid. Niisiis võib afiksaaladverbiks olla lokaaladverb ehk kohamäärsõna, perfektiivsus-, seisundi- ja modaalmäärsõna. Lokaalsed afiksaaladverbid osutavad üldistatud kujul suhtelist suundumis-, paiknemis-, eemaldumis- või kulgemiskohta, näiteks *alla*, *eemale*, *ette*, *juurde*, *järel(e)*, *kaasa*, *kõrvale*, *külge*, *ligi*, *läbi*, *maha*, *mööda*, *otsa*, *peal(e)*, *ringi*, *sisse*, *taga(si)*, *taha*, *vahele*, *vastu*, *välja*, *üle(s)*, *ümber* jne. Perfektiivsust väljendavad afiksaaladverbid märgivad tegevuse lõpetatust, resultatiivust või vähemalt piirivõimaluse olemasolu, näiteks *läbi*, *maha*, *otsa*, *täis*, *valmis*, *välja*, *ära*. Seisundiadverbid kuuluvad afiksaaladverbidena ühendverbide koosseisu vaid siis, kui nad moodustavad koos verbiga uue tähendusliku terviku ja tingivad lausemalli. Sellisena esinevad adverbid *kinni*, *lahti*, *laiali*, *kokku*, *püsti*, *viltu* jne. Modaalsust väljendavad afiksaaladverbid on *vaja* ja *tarvis* sellistes ühendverbides nagu *vaja minema/olema*, *tarvis minema/olema*. (Erelt 2013: 62)

Lokaalsete afiksaaladverbide juures ilmneb kõige selgemini ühendverbide jagunemine seeriatena esinevateks korrapärasteks ja semantilise terviku moodustavateks ainukordseteks ühendverbideks. Korrapärase ühendi osised säilitavad tähendusliku iseseisvuse, kuid moodustavad süntaktiliselt lahutamatu terviku. Nad kujunevad mingi tähendusrühma verbide suhteliselt regulaarsel kombineerumisel kindlasse rühma kuuluvate afiksaaladverbidega: näiteks sellised liikumisverbide ja lokaalsete afiksaaladverbide kombinatsioonid nagu *alla/eemale/juurde/kohale/ligi/pärale/tagasi + minema/jooksma/astuma/kihutama/sõitma*. (EKG II: 21) Ainukordsete ehk idiomaatiliste ühendverbide adverbiline komponent ei ole verbi seotud laiend (Rätsep 1978: 28). Selle osised moodustavad semantiliselt ja süntaktiliselt liigendamatu terviku ning ühend on omandanud uue tähenduse, näiteks *peale käima*, *üle ajama*, *maha võtma*, *juurde lõikama*, *üle pakkuma*, *üles lööma* jne (EKG II: 21).

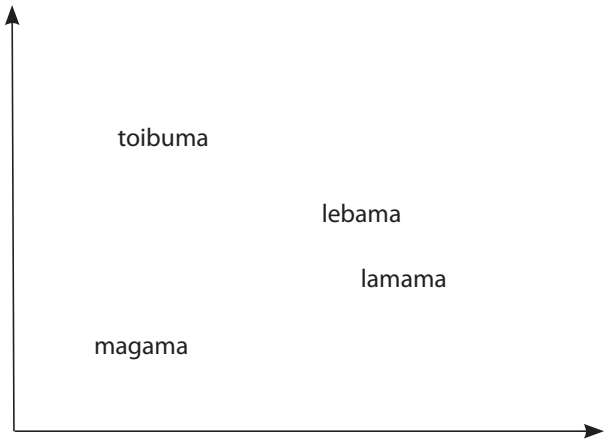
[2] https://code.google.com/p/word2vec/ (8.12.2015).

3. Distributiivne semantika

Distributiivne semantika võimaldab mudelites rakendatuna mõõta sõnatähenduste sarnasusi. Lähenemise aluseks olevat oletust nimetatakse distributiivse semantika hüpoteesiks, mille järgi esinevad sarnase tähendusega sõnad sarnases kontekstis. Sellel väitel on viimaste kümnendite jooksul olnud arvutilingvistikale suur mõju (Garvin 1962: 388). Distributiivsel semantikal põhinevate mudelitega tuvastatakse nii sarnase sisuga dokumente kui ka otsitakse suurtest tekstihulkadest sarnaste tähendustega sõnu ja püsiühendeid (Bruni jt 2014: 2).

3.1. Distributiivse semantika mudelid

Distributiivse semantika ehk vektorruumi (ingl *vector space*) ehk semantilise ruumi (*semantic space*) ehk sõnaruumi (*word space*) mudelid tuletavad sõnade tähenduse samas kontekstis koosesinemise põhjal (Bruni jt 2014: 1–2). Neid mudeleid peetakse sõna tähenduste ruumilisteks esitusteks, mis põhinevad arusaamal, et tähenduslikku sarnasust saab esitada kui sõnade lähedust *n*-mõõtmelises ruumis, kus *n* võib tähistada iga täisarvu alates ühest (Sahlgren 2006: 18). Joonisel 1 on esitatud näide kahemõõtmelisest sõnaruumist (*word space*), mis on selline esitus, kus sõnadevaheline ruumiline lähedus või kaugus tähistab nende tähenduslikku sarnasust või erinevust.



Joonis 1. Kahemõõtmeline sõnaruum

Jooniselt 1 on näha, kui sarnased sõnade tähendused üksteisele on.[3] Nii näiteks on sõna *lebama* joonisel lähemal sõnale *lamama* kui sõnale *magama*, mis tähendab seda, et verbi *lebama* tähendus on sarnasem verbi *lamama* tähendusega kui verbi *magama* tähendusega.

Sõnaruumi mudelitega leitud sõnadevahelised sarnasused saadakse automaatselt keeleandmestikust. Mudelite abil esitatakse sõnade tähendused vektoritena, mis vaikimisi mõõdavad lemmade koosesinemist samas kontekstis. Vektorid leitakse sõnale eelnevate ja järgnevate sõnade ehk konteksti abil. Kui konteksti kuulub üks eelnev ja üks järgnev sõna (ehk akna suurus on üks), siis lauses *Jagatud rõõm on suurem rõõm* on sõna *jagatud* kontekst sõna *rõõm* ning sõna *rõõm* kontekst

[3] Joonisel 1 esitatud sõnadevahelised kaugused põhinevad sõnade vektoresituse vahelisel koosinuskauguse väärtustel, mis on leitud sama korpuse põhjal, millel põhineb kogu uurimus (vt lähemalt 5. osast).

jagatud, *on* ja *suurem*. Seega sõnad *jagatud* ja *rõõm* esinevad samas kontekstis, kuid sõnad *jagatud* ja *on* ei esine. Tabelis 1 on esitatud näitelausesse kuuluvate sõnade seosed: 0 tähistab, et sõnad ei esine koos, 1 märgib sõnade koosesinemist.

Tabel 1. Sõnade koosesinemise loendid (lause *Jagatud rõõm on suurem rõõm* näitel)

	jagatud	*rõõm*	*on*	*suurem*
jagatud	0	1	0	0
rõõm	1	0	1	1
on	0	1	0	1
suurem	0	1	1	0

Tabelist 1 on näha, et sõna *rõõm* koosesinemise loend on (1, 0, 1, 1) ning sõna *suurem* loend on (0, 1, 1, 0). Selliseid loendeid nimetatakse kontekstivektoriteks. (Sahlgren 2006: 26–27). Vektoreid saab rakendada ka teistes (mitte ainult sõna) semantilise sarnasuse mudelites ja neid on kasutatud nii keele automaattöötluses kui ka kognitiivteaduses (Erk, Padó 2008: 897).

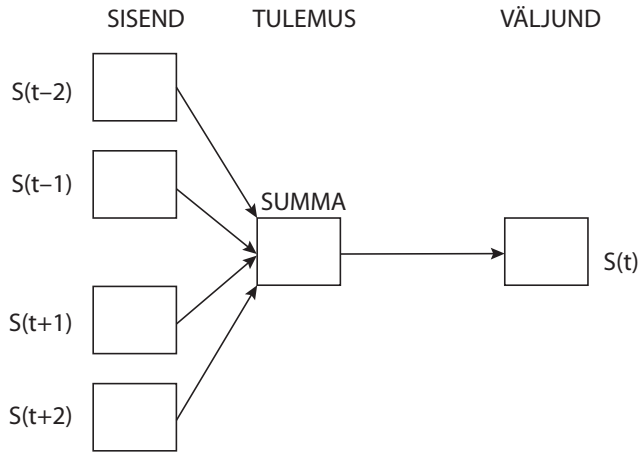
Semantilise sarnasuse väärtus, mida leitud kontekstivektorite vahel mõõdetakse, väljendubki geomeetrilise kaugusena sõnu esitavate vektorite vahel (Bruni jt 2014: 2). Julie Elizabeth Weeds (2003) on võrrelnud mitmeid mõõdikuid, mille abil on võimalik vektorite vahelist kaugust välja arvutada, näiteks koosinuskaugus (ingl *cosine similarity*), eukleidiline kaugus (*Euclidean distance*), Jensen-Shannoni divergents (*Jensen-Shannon divergence*), Jaccardi koefitsient (*Jaccard similarity coefficient*) jne. Siinses töös mõõdetakse vektoritevahelist kaugust koosinuskaugusega, sest see mõõdik on saavutanud häid tulemusi just sõnade tähendusi puudutavates uurimustes (näiteks Bullinaria, Levy 2007, Padó, Lapata 2007). Mida suurem on koosinuskauguse väärtus ühendverbi ja sellesse kuuluva verbi vahel, seda kompositsionaalsem on ühendverb (Bott, Schulte im Walde 2014: 511) ehk ühendverbi kompositsionaalsuse aste on seotud sellega, kui sarnased on ühendverbi kuuluva verbi esinemise ja ühendverbi tervikuna esinemise kontekstid.

3.2. Tööriist word2vec

Siinse töö materjali analüüsimiseks kasutati neurovõrgu mudelil[4] põhinevat tööriista word2vec, mis võimaldab automaatselt hinnata semantilise sarnasuse väärtusi. Neurovõrgu mudelid on masinõppes ja kognitiivteadustes kasutatavad mudelid, mis püüavad jäljendada bioloogiliste närvivõrkude, eelkõige aju ehitust ning on võimelised töötlema suuri andmehulki. Kuna word2vec on vabavaraline tarkvara, mida praegu arvutilingvistilistes uurimustes rohkesti rakendatakse, siis otsustati ka siinses töös selle kasuks. Word2veci sisendiks on tekstikorpus, kus iga lause on eraldi real ja väljundiks kõigi sõnade vektoresitused, mille leidmiseks koostab programm kõigepealt korpuses olevatest sõnadest sõnastiku ning seejärel õpib sõnade vektoresitused vastavalt valitud mudelile. Word2veci kasutades on võimalik rakendada kahte pidevat mudelit: järjestikuste sõnade esinemissageduse mudel (*bag-of-words*, CBOW) ja *skip*-gramm (*skip-gram*). (Mikolov jt 2013a: 4)

[4] Esimese neurovõrgu mudeli koostasid Warren S. McCulloch ja Walter Pitts (1943).

Siinses töös on vektoresituste õppimiseks kasutatud CBOW mudelit, sest see töötab paremini ning kiiremini suurte tekstikorpuste peal. Joonisel 2 on esitatud CBOW tööpõhimõte: mudel on treenitud ennustama sõna tema konteksti (ehk kahe talle eelneva ja järgneva sõna) järgi. (Mikolov jt 2013b: 3)



Joonis 2. CBOW tööpõhimõte (Mikolov jt 2013b põhjal). Lühendid:
S(t – 2) ja *S(t – 1)* = kaks keskmisele sõnale vahetult eelnevat sõna,
S(t + 1) ja *S(t + 2)* = kaks keskmisele sõnale vahetult järgnevat sõna,
SUMMA = keskmise sõna kontekst, *S(t)* = keskmine sõna, millele vektoresitust otsitakse

Jooniselt 2 on näha, et CBOW mudel otsib vektoresitusi, mis aitavad kontekstisõnade *S(t – 2)*, *S(t – 1)*, *S(t + 1)*, *S(t + 2)* järgi ennustada keskmist sõna *S(t)*. Pärast matemaatilisi arvutusi ja teisendusi on väljundiks sõna *S(t)* vektoresitus. Akna suurust ehk seda, kui palju eelnevaid ja järgnevaid sõnu konteksti arvatakse, on uurijal võimalik ise määrata.

Word2veci rakendamise jaoks on vajalik esitada (osa)laused eraldi ridadel, säilitada sõnade järjekord ja esitada sõnad algvormis. Selleks leiti morfoloogilise analüsaatori abil korpusfailidest sõnade algvormid ning seejärel kustutati sisendtekstist üleliigne (kirjavahemärgid, morfoloogiline märgendus). Seega word2veci sisendfailides pole muud informatsiooni kui eraldi ridadel esitatud (osa)laused, millesse kuuluvad sõnad on algvormidena tekstis esinemise järjekorras üksteisest tühikutega eraldatud.

Word2vec otsib vaid ühe vektoresituse igale sõnale, mis tähendab, et mudel ei arvesta polüseemsete sõnadega ega ole võimeline esitama ühe sõna erinevaid tähendusi. Kuigi siinses artiklis on põhirõhk mudeli tutvustamisel, on tulemuste analüüsimisel oluline seda fakti meeles pidada.

Word2veci on rakendatud sõnade sarnasuse, kategoriseerimise ja analoogia ülesannete lahendamisel (vt lähemalt nt Baroni jt 2014). Eesti keeles on word2veci edukalt rakendanud näiteks Tanel Pärnamaa (2015) piltide automaatse kirjeldamise algoritmi väljatöötamiseks.

4. Ühendverbide kompositsionaalsuse määramine

Kuigi eesti keele põhjal pole varem ühendverbide kompositsionaalsuse taset automaatselt tuvastatud, on seda tehtud teiste keelte peal. Inglise ja saksa keele ühendverbide kompositsionaalsust on uuritud mitmetes uurimustes. Näiteks Diana McCarthy, Bill Keller ja John Carroll (2003) püüdsid tuvastada inglise keele ühendverbide (ingl *particle verb*) kompositsionaalsust ja kasutasid selleks automaatselt koostatud tesaurust. Erinevaid mõõdikuid rakendades võrdlesid nad ühendverbide ja nendesse kuuluvate verbide lähimaid naabreid ning selgitasid välja, et teatud mudelite ja inimeste hinnangud langevad üsna hästi kokku. Bannard (2005) rakendas distributiivset lähenemist, et uurida verbi ja partikli konstruktsioone. Uurimus kinnitas kasutatud mudeli tulemuste ja inimeste (nii ekspertide kui ka mitteekspertide) arvamuste vahelist seost.

Natalie Kühner ja Sabine Schulte im Walde (2010) kasutasid saksa keele ühendverbide kompositsionaalsuse määra tuvastamiseks hägusat klasteranalüüsi. Nad võrdlesid rakendatud mudelite tulemusi inimeste arvamuste põhjal koostatud järjestusega ning leidsid, et kasutatud meetod korreleerub kuldstandardiga. Stefan Bott ja Sabine Schulte im Walde (2014) hindasid distributiivse semantika mudelit saksa keele ühendverbide kompositsionaalsuse tuvastamiseks. Selleks arvutasid nad iga ühendverbi ja sellesse kuuluva verbi vektorite vahelise kauguse väärtuse ning vastavalt sellele järjestasid nad ühendverbid nii, et kõige tugevama väärtusega ühendverb on kõige kompositsionaalsem. Lisaks tegid nad eksperimendi, kus palusid inimestel samu ühendverbe kompositsionaalsuse järgi järjestada, ning kahe järjestuse võrdlus tõestas, et distributiivse mudeli põhjal moodustunud ühendverbide järjekord korreleerub katseisikute koostatud järjekorraga.

Nimetatud uurimused kinnitavad, et testitud mudelid korreleeruvad inimeste keeletajul põhinevate hinnangutega, mistõttu saab öelda, et mudelid saavutavad oma eesmärgi ja kirjeldavad hästi inimeste semantilist intuitsiooni.

5. Materjal

Siinses uurimuses püütakse esmalt jaotada ühendverbid tähenduse moodustamise viisi alusel automaatselt rühmadesse ning seejärel määrata automaatselt ühendverbide kompositsionaalsust. Esimese ülesande lahendamiseks kasutatakse sõnadevahelise seose tugevuse mõõdikuid, teise ülesande jaoks aga distributiivse semantika mudelit. Mõlemad ülesanded põhinevad eesti keele koondkorpuse ajakirjandustekstidel, millest siinsesse töösse on kaasatud 170 miljoni sõna suurune osa. Korpus oli eelnevalt morfoloogiliselt analüüsitud ja ühestatud ning osalausestatud. Sellest korpusest õnnestus tuvastada 1676 "Eesti keele seletavasse sõnaraamatusse" (EKSS) kuuluvat ühendverbi (vt lähemalt Aedmaa 2015 ptk 3.2), mis on siinses uurimuses lähemalt vaatluse alla võetud.

Mõõdikute töö hindamiseks on vajalik kuldstandard, mis on ühendverbide jaotus kolme tüüpi. Kuldstandardi põhjal saab välja arvutada, kui suur protsent tuvastatud ühendverbidest kuulub ühte, teise või kolmandasse tüüpi. Siinses uurimuses kasutatud kuldstandard on autori hinnangul põhinev ühendverbide jaotus vastavalt eespool esitatud teooriale (vt osa 2). Kuna aga eesti keeles esineb ühendverbe, mida

võib vastavalt kontekstile liigitada kas ainukordseks (1) või korrapäraseks (2), siis binaarne jaotus polnud võimalik ja seesugused ühendverbid paigutati kolmandasse, n-ö segagruppi "ainukordsed/korrapärased".

(1) Ta *näeb* tulevikku *ette*.

(2) Vihmasaju tõttu *ei näinud* juht kaugele *ette*.

Nii jaotati 1676 eelnevalt tuvastatud ühendverbi kolme rühma. Jaotus põhineb autori keeletajul, kuid aluseks olid EKSS-is esitatud tähendused. Ainukordseteteks loeti 393 ühendit, näiteks *välja sööma*, *ära istuma*, *üles kihutama*. Korrapäraseid ühendeid on 983, näiteks *välja hingama*, *maha saagima*, *sisse kutsuma*. Kolmandasse rühma kuulub 300 ühendverbi, näiteks *läbi minema*, *sisse sööma*, *peale minema*.

Ühendverbide kompositsionaalsuse tuvastamiseks kasutati täpselt sama korpust, mida kasutati ühendverbide klassifitseerimiseks.

6. Ühendverbide liigitamine

Aedmaa (2015) uurimus tõestas, et korpuse suurus ja vaadeldavate kandidaatpaaride hulk mõjutavad rakendatud sõnadevahelise seose tugevuse mõõdikute tööd. Seega võib eeldada, et ka tuvastatavate ühendverbide liigil on mõju tulemustele. Kõikide ühendverbide tuvastamisel rakendati järgmisi mõõdikuid: t-skoor, vastastikuse informatsiooni väärtus (ingl *Mutual Information*, MI), hii-ruut-statistik, log-tõepära funktsioon ja minimaalne tundlikkus (*Minimum Sensitivity*, MS). Lisaks vaadeldi ühendisse kuuluvate osiste koosesinemise sagedust. Selgus, et parim mõõdik kõikide ühendverbide tuvastamiseks ilma neid tüüpidesse jagamata on t-skoor, kuid olenevalt vaadeldavast suurima mõõdiku väärtuse saanud ühendite hulgast töötavad hästi ka lihtne koosesinemise sagedus ja log-tõepära funktsioon. Kuna kõik nimetatud mõõdikud osutusid tulemuslikeks, siis rakendati eri tüüpi ühendverbide tuvastamiseks samu mõõdikuid. Eeldati, et ühendverbi tüüp mõjutab mõnevõrra mõõdikute paremusjärjestust.

Meetodite tulemuslikkuse hindamiseks kasutati täpsuse väärtust, mis väljendab, kui suur protsent kõikidest tuvastatud ühenditest on ainukordsed, korrapärased või n-ö segarühma kuuluvad. Eesmärk oli välja selgitada, kas mõni mõõdik tuvastab ühendverbe niivõrd edukalt, et selle abil saab ühendeid kindlatesse klassidesse jagada. Statistiline analüüs on tehtud tarkvaraga R (R Development CoreTeam 2013).

Tabel 2 esitab rakendatud mõõdikute ja koosesinemise sageduse tulemused eri tüüpi ühendverbide tuvastamisel ajakirjanduskorpusest mõõdiku kõrgeima väärtuse saanud 100, 500 ja 1000 ühendverbi seas.

Tabel 2. Mõõdikute ja koosesinemise sageduse täpsused eri liiki ühendverbide tuvastamisel.

Mõõdik	Ainukordsed			Korrapärased			Ainukordsed/korrapärased		
	100	500	1000	100	500	1000	100	500	1000
t	24,0%	22,4%	20,8%	32,0%	46,6%	55,2%	44,0%	31,0%	24,0%
MI	27,0%	21,8%	21,3%	58,0%	62,8%	60,2%	15,0%	15,4%	18,5%
hii	33,0%	22,2%	20,8%	39,0%	51,0%	56,7%	28,0%	26,8%	22,5%
log	25,0%	22,0%	20,4%	36,0%	48,2%	56,0%	39,0%	29,8%	23,6%
MS	28,0%	22,6%	23,7%	37,0%	46,6%	52,0%	35,0%	30,8%	24,3%
sag	22,0%	25,8%	23,3%	34,0%	41,6%	51,1%	44,0%	32,6%	25,6%

Lühendid: t = t-skoor, MI = vastastikuse informatsiooni väärtus, hii = hii-ruut-statistik, log = log-tõepära funktsioon, MS = minimaalne tundlikkus, sag = koosesinemise sagedus.

Tabelist 2 on näha, et 100 mõõdiku suurima väärtuse saanud paari seast tuvastab teistest paremini ainukordseid ühendverbe hii-ruut-statistik, 500 kandidaatpaari hulgast aga koosesinemise sagedus. 1000 kandidaatpaari arvestuses on mõõdikute täpsused üsna võrdsed, teistest pisut paremaks võib pidada MS-i. Kokkuvõtlikult saab öelda, et ainukordsete ühendverbide tuvastamisel töötavad teistest paremini MS, hii-ruut-statistik ja koosesinemise sagedus.

Korrapäraste ühendverbide tuvastamisel saavutab parima tulemuse MI, mille täpsus 100 kandidaatpaari seas on 19% suurem kui talle järgneval hii-ruut-statistikul. 500 ja 1000 kandidaatpaari seas on täpsuste vahe väiksem ning MI-le järgnevad hii-ruut-statistik, log-tõepära funktsioon, t-skoor ning MS. Väikseim täpsuse väärtus on koosesinemise sagedusel. Kui võrrelda neid tulemusi eelnevate uurimuste (nt Evert, Krenn 2001, Aedmaa 2015) tulemustega, siis võib öelda, et MI suur täpsus on ootamatu, sest sarnaste ülesannete lahendamisel on MI täpsus võrreldes teiste mõõdikutega olnud madalam. Siinkohal on MI kõrge täpsuse põhjuseks aga asjaolu, et MI tõstab esile harva esinevad ühendid, mille komponendid esinevad samuti harva, näiteks *takka kiitma*, *lahti korkima*, *sisse logima*, *üles paistetama*. Harva esinevad ühendid pole üldjuhul polüseemsed, mis on eesti keele puhul iseloomulik just korrapärastele ühendverbidele.

Nende ühendverbide tuvastamiseks, mis võivad vastavalt kontekstile olla kas ainukordsed või korrapärased, töötab 100 ja 500 kandidaatpaari seas kõige paremini koosesinemise sagedus ja t-skoor. MI tulemused on väikseimad. Eespool esitatud oletustele toetudes võib selle rühma tulemusi pidada ootuspärasteks: ülesandega saavad kõige paremini hakkama koosesinemise sagedus ja t-skoor, millele järgnevad MS, log-tõepära funktsioon ja hii-ruut-statistik ning kõige väiksema täpsusega MI.

Vaadeldavate ühendverbide hulga suurenemisel tulemused üldiselt muutuvad. Ainukordsete ühendverbide tuvastamisel t-skoori, hii-ruut-statistiku, log-tõepära funktsiooni ja MS-i täpsused langevad. Vaid koosesinemise sageduse täpsus kasvab. Korrapäraste ühendverbide tuvastamisel kasvavad kõikide mõõdikute täpsused, teistest vähem MI täpsus. Ainukordsete/korrapäraste ühendverbide korral t-skoori, hii-ruut-statistiku, log-tõepära, MS-i ja koosesinemise sageduse täpsused kahanevad, MI täpsus aga kasvab mõne protsendi võrra. Seega võib öelda, et tulemused sõltuvad vaadeldavate andmete hulgast.

Seega leiab kinnitust eeldus, et mõõdikud töötavad eri liiki ühendverbide tuvastamisel erinevalt, kuid ükski neist ei saavutanud niivõrd häid tulemusi, et

nende abil ühendverbe kahte või kolme kindlasse klassi paigutada. Nii võib eeldada, et ka eesti keele ühendverbid ei jagune kindlalt kompositsionaalseteks ja mittekompositsionaalseteks, vaid jaotuvad kompositsionaalsuse ja mittekompositsionaalsuse skaalale. Selle oletuse kontrollimiseks püütakse järgnevalt ühendverbid kompositsionaalsuse taset määrata distributiivse semantika võimalusi rakendades.

7. Ühendverbide kompositsionaalsuse taseme määramine

Selleks, et vaadelda, kas eesti keele ühendverbe, nii nagu erinevate keelte püsiühendeid (Bannard jt 2003: 65), on võimalik asetada kompositsionaalsuse ja mittekompositsionaalsuse skaalale, rakendati distributiivset semantikal põhinevaid vahendeid. Igasugused distributiivse semantika mudelid on juhendamata ja mõõdavad kontekstidevahelisi sarnasusi, kuid nad ei ennusta otseselt kompositsionaalsuse väärtust (Bott, Schulte im Walde 2014: 511). Nii on võimalik vastavalt leitud koosinuskaugustele asetada ühendid üksteise suhtes skaalale, mille ühes otsas on tugevalt kompositsionaalsed ning teises otsas mittekompositsionaalsed ühendid. Selle artikli eesmärk on tuvastada ühendverbide kompositsionaalsuse määra, kirjeldada vektorsemantika rakendamise võimalusi ning anda esialgne ülevaade selle rakendamise tulemustest.

Sõnadele ja fraasidele vektoresituste otsimiseks kasutati tarkvara word2vec ning nendevahelise koosinuskauguse arvutamiseks Gensimi[5] moodulit programmeerimiskeeles Python. Selleks et leida koosinuskaugus ühendverbi vektori ja ühendverbis sisalduva verbi vektori vahel, otsiti korpusest vektoresitused kõikidele uni- ja bigrammidele. See tähendab, et ühendverbide vektoresituste leidmiseks otsiti vektoresitus kahest osisest – adverbist ja verbist – koosnevale bigrammile, samas eirati ühendverbide omadust esineda lauses üksteisest kaugel ning seda, et ühendverbide osiste järjekord pole fikseeritud. See otsus on tingitud asjaolust, et word2vec võimaldab vektoresituse leida kas üksikule sõnale või üksteise kõrval esinevatele sõnadele, kuid mitte üksteisest kaugel esinevatele sõnadele. Akna suuruseks oli viis, mis tähendab, et vektoresituste leidmisel arvati konteksti viis väljundsõnale eelnevat ja viis järgnevat sõna. Kui sõnale eelnes või järgnes vähem kui viis sõna, siis arvestati konteksti sõnu vähem, mis tähendab, et vektoresituse otsimisel ei ületata (osa)lausepiire.

Pärast vektoresituste leidmist nii ühendverbidele kui ka verbidele arvutati Gensimi moodulit kasutades koosinuskaugused verbi ja seda sisaldava ühendverbi vektorite vahel. Varasemas uurimuses tuvastatud 1676 ühendverbist õnnestus koosinuskauguse väärtus leida 953 ühendverbi vektori ja ühendverbi verbilise osise vektori vahel. Kuna 723 ühendverbi kuuluvat adverbi ja verbi ei esinenud andmestikus kõrvuti, siis jäi neile esitus otsimata.

Mida lähemal on ühendverbi esitava vektori ja sellesse kuuluva verbi esitava vektori koosinuskauguse väärtus ühele, seda väiksem on vektoritevaheline nurk ja seda sarnasemad on nende ühendverbide ja verbide tähendused. Kui koosinuskauguse väärtus on lähedal nullile, siis on vektoritevaheline kaugus 90 ja koosinuskaugus väärtusega –1 osutab, et vektorid osutavad vastassuunas ja nendevaheline

[5] https://radimrehurek.com/gensim/models/word2vec.html (8.12.2015).

nurk on väga suur. Seega, mida suurem on nurk vektorite vahel, seda väiksem on koosinuskauguse väärtus ja seda idiosünkraatilisem on ühend.

Tabelis 3 on esitatud 50 ühendit, mille vektori ja millesse kuuluva verbi vektorite põhjal leitud koosinuskaugus on suurim. Tabelis eespool esitatud ühendverbid on koosinuskauguse väärtuse põhjal kompositsionaalsemad kui madalamal esitatud või tabelist välja jäänud ühendverbid.

Tabel 3. 50 kõige kompositsionaalsemat ühendverbi

Adverb	Adverbi sagedus	Verb	Verbi sagedus	Koosinus-kaugus	Ühendverbi sagedus
üles	64480	*paistetama*	168	0,6926	85
kokku	150217	*voltima*	243	0,6784	60
ära	176797	*jooma*	18903	0,6633	931
välja	297714	*sirutama*	1596	0,6433	339
välja	297714	*sirutuma*	220	0,6433	30
üle	134583	*võõpama*	555	0,6178	73
otsa	13292	*sõitma*	115057	0,6155	3339
maha	74140	*põlema*	15309	0,6087	1042
edasi	75469	*sõitma*	115057	0,5993	1597
kaasa	75124	*laulma*	22190	0,5836	857
välja	297714	*loosima*	2412	0,5802	1320
välja	297714	*sõitma*	115057	0,5796	5048
sisse	53046	*logima*	124	0,5681	81
ette	114045	*laulma*	22190	0,5674	128
ümber	26581	*ehitama*	52283	0,5618	1736
ümber	26581	*matma*	7039	0,5615	251
ümber	26581	*istutama*	3612	0,5590	132
kaasa	75124	*sõitma*	115057	0,5479	800
edasi	75469	*õppima*	52843	0,5431	908
sisse	53046	*pühitsema*	1321	0,5408	90
ära	176797	*sööma*	33929	0,5394	2525
mööda	13752	*sõitma*	115057	0,5357	1218
ära	176797	*ostma*	91098	0,5291	4312
välja	297714	*lülitama*	6938	0,5187	3463
ära	176797	*pesema*	7896	0,5154	216
läbi	91173	*põimuma*	952	0,5144	159
kokku	150217	*põrkama*	5227	0,5094	3595
ära	176797	*kulutama*	19634	0,5088	649
üle	134583	*kuumenema*	471	0,5079	263
kokku	150217	*ostma*	91098	0,5073	2349
välja	297714	*hingama*	4371	0,4992	219
lahti	34458	*mõtestama*	1558	0,4909	436
ära	176797	*lööma*	53563	0,4878	654
maha	74140	*saagima*	1271	0,4840	286
läbi	91173	*müüma*	73763	0,4834	489

Adverb	Adverbi sagedus	Verb	Verbi sagedus	Koosinus-kaugus	Ühendverbi sagedus
välja	297714	*kihutama*	9959	0,4768	607
üle	134583	*värvima*	4317	0,4758	475
sisse	53046	*laulma*	22190	0,4695	264
lahti	34458	*korkima*	92	0,4641	66
välja	297714	*maksma*	153513	0,4595	4816
välja	297714	*arendama*	12554	0,4595	1455
ära	176797	*keelama*	25132	0,4587	1964
läbi	91173	*kihutama*	9959	0,4585	139
välja	297714	*ehitama*	52283	0,4573	1821
sisse	53046	*lülitama*	6938	0,4571	1271
tagasi	100784	*sõitma*	115057	0,4555	1693
ringi	23320	*sõitma*	115057	0,4541	2169
üles	64480	*putitama*	271	0,4534	43
välja	297714	*puhkama*	10931	0,4471	514
maha	74140	*müüma*	73763	0,4465	8435

Tabelist 3 selgub, et kõige sarnasemas kontekstis esinevad ühendverb *üles paistetama* ja selle osis *paistetama*. Kui vaadelda *üles paistetama* tähendust EKSS-is, siis on see võrdsustatud *paistetama* tähendusega, lisaks on sulgudes sõna *tugevasti*. Nii võib öelda, et *üles paistetama* on pigem kompositsionaalne kui idiomaatiline ühendverb. Niisamuti võib tähenduselt sarnasteks pidada ka näiteks ühendverbi *kokku voltima* ja verbi *voltima*, ühendverbi *välja sirutama* ja verbi *sirutama* ning ühendit *üle võõpama* ja verbi *võõpama*. Suure kompositsionaalsuse väärtusega on aga ka ühend *ära ostma*, mis võib olla ka mittekompositsionaalne tähenduses 'kedagi altkäemaksuga enda nõusse, enda poole meelitama'. Selle ühendi korral on keeruline otsustada, kas tegemist on pigem kompositsionaalse või pigem mittekompositsionaalse ühendiga. Samuti on näha, et suure kompositsionaalsusega on EKG järgi (EKG II: 21) korrapäraste ühendverbide hulka kuuluvad *otsa sõitma*, *välja kihutama*, *läbi kihutama*, *tagasi sõitma* ja *ringi sõitma*.

Andmestikus on aga ka selliseid ühendeid, mille puhul võib eeldada kõrget koosinuskaugust, näiteks EKSS-i tähenduste järgi selgelt korrapärane ühendverb *läbi kõndima* tabelisse 3 koosinuskaugusega 0,1889 ei mahu. Niisamuti on selliseid ühendeid, mille koosinuskaugus jääb ootuspäraselt kõrgete ja madalate väärtuste vahele. Selliseks näiteks on *tagasi astuma* (koosinuskaugus 0,1229), mis liigitub n-ö segarühma: olenevalt kontekstist nii korrapäraseks (tähendusega 'tagasi või tahapoole siirduma') kui ka idiomaatiliseks (tähendusega 'näiteks ametist, võimust loobuma, tagasi tõmbuma') ühendverbiks. Samuti on mitme tähendusega ühendiks *tagasi minema* (koosinuskaugus 0,1586): 'lähtekohta tagasi siirduma'; 'tasemelt langema, taandarenema'; 'tagasi ulatuma, pärinema'. Koosinuskauguse väärtuse järgi saab ühendi *tagasi minema* paigutada vähem kompositsionaalsemaks kui ühendverbi *läbi kõndima*, kuid kompositsionaalsemaks kui ühendi *tagasi astuma*. Samas tuleb meeles pidada, et *minema* on väga polüseemne verb, mis mõjutab tulemusi. Selliste ühendite puhul on tulevikus tarvis rakendada mudelit, mis arvestab ka verbide polüseemsusega. Seetõttu on edaspidi vajalik kasutada mudeleid, mis

otsivad erinevad vektorid sõna kõikidele tähendustele, näiteks mudelit *Distributed Multi-sense Word Embedding* (DMWE)[6].

Tabelis 4 on esitatud 50 ühendverbi, mida esitava vektori ja ühendverbi kuuluva verbi vektori vaheline koosinuskaugus on kõige väiksem. Tabelis esitatud ühendverbe võib pidada vähem kompositsionaalsemateks kui teisi ühendverbe.

Tabel 4. 50 kõige väiksema kompositsionaalsusega ühendverbi

Adverb	Adverbi sagedus	Verb	Verbi sagedus	Koosinus-kaugus	Ühendverbi sagedus
ära	176797	*langema*	44192	–0,1637	747
tagant	2506	*sundima*	33781	–0,1433	172
edasi	75469	*jõudma*	168218	–0,1189	911
ära	176797	*kaotama*	66927	–0,1087	1121
üle	134583	*paisutama*	946	–0,1020	185
läbi	91173	*ajama*	46586	–0,1019	1480
lahku	3331	*lööma*	53563	–0,0995	543
peale	33718	*pressima*	4666	–0,0978	177
kokku	150217	*saama*	1057514	–0,0976	13199
ette	114045	*näitama*	109261	–0,0947	2507
maha	74140	*ütlema*	551017	–0,0883	127
peale	33718	*tungima*	8817	–0,0834	175
kõrvale	15717	*hoidma*	76157	–0,0825	731
üle	134583	*astuma*	52604	–0,0804	860
läbi	91173	*võtma*	322185	–0,0759	962
üles	64480	*tõusma*	61765	–0,0741	981
kinni	60871	*mätsima*	399	–0,0719	276
esile	15517	*tõstma*	46220	–0,0714	3920
üles	64480	*tähendama*	74045	–0,0678	267
edasi	75469	*kandma*	64698	–0,0672	737
esile	15517	*kutsuma*	63938	–0,0669	3074
sisse	53046	*raiuma*	4118	–0,0660	47
vastu	76337	*töötama*	87160	–0,0635	386
kokku	150217	*viima*	116301	–0,0627	1799
üle	134583	*saama*	1057514	–0,0616	8585
ette	114045	*tegema*	598768	–0,0605	1489
maha	74140	*arvama*	168766	–0,0584	1272
kõrvale	15717	*jätma*	101181	–0,0581	2321
kõrvale	15717	*tõmbuma*	3429	–0,0574	155
üles	64480	*andma*	335131	–0,0567	1585
läbi	91173	*kukkuma*	32579	–0,0540	2330
üleval	5143	*pidama*	702611	–0,0534	1082
ette	114045	*nägema*	200795	–0,0529	26025
välja	297714	*käima*	172564	–0,0494	5116
läbi	91173	*viima*	116301	–0,0493	12750

[6] http://www.dmtk.io/word2vec_multi.html (14.12.2015).

Adverb	Adverbi sagedus	Verb	Verbi sagedus	Koosinus-kaugus	Ühendverbi sagedus
vastu	76337	*käima*	172564	–0,0463	267
välja	297714	*viima*	116301	–0,0456	4450
vastu	76337	*seisma*	56244	–0,0452	1563
ära	176797	*meelitama*	9474	–0,0445	157
takka	2	*kiitma*	28533	–0,0436	1
sisse	53046	*kirjutama*	132506	–0,0419	1751
ühte	2132	*heitma*	17617	–0,0410	128
ära	176797	*eksima*	10770	–0,0387	527
külge	2997	*hakkama*	266118	–0,0365	254
välja	297714	*tõrjuma*	8584	–0,0364	842
ümber	26581	*lööma*	53563	–0,0362	64
mööda	13752	*saatma*	84129	–0,0353	426
üles	64480	*näitama*	109261	–0,0352	1740
kõrvale	15717	*heitma*	17617	–0,0327	513
ette	114045	*tooma*	142990	–0,0323	907

Tabelist 4 selgub, et idiomaatilisemad kui teised ühendverbid on näiteks *ära langema*, *tagant sundima*, *edasi jõudma*, *ära kaotama* ja *üle paisutama*. Kui vaadelda ühendverbi *ära langema* EKSS-i tähendusi, siis need on ‘olematuks muutuma, ära kaduma, kõrvale jääma, välja langema’; ‘(ära) taganema’; ‘viletsamaks, kehvemaks muutuma, normaalset olekut kaotama, ära vajuma’. Tähenduste põhjal võib öelda, et tegu on mitmetähendusliku verbiga, mille liigitamine ühte või teise rühma on keeruline. Ühendverbi *tagant sundima* tähenduseks on ‘taga sundima’, mille tähenduseks omakorda on ‘korduvalt, pidevalt sundima’. Seega ühendit *tagant sundima* võib seepärast pigem kompositsionaalseks pidada. Samas ühendverbi *ära kaotama* tähendus on ‘kaotama’ ja see ühend on selgelt kompositsionaalne.

Mati Erelt (2013: 64) toob välja, et sellised adverbid nagu *kinni*, *lahti*, *kokku* esinevad afiksaaladverbidena vaid idiomaatiliste ühendverbide koosseisus, vastasel juhul on nad iseseisvad adverbid ja lauses seisundimääruseks. Kui vaadelda ühendverbi *kokku kukkuma* tähendusi EKSS-is, siis selgub, et selle tähendus võib olla ka kompositsionaalne ‘maha langedes või ümber kukkudes koost lagunema’. Ka koosinuskauguse (0,2494) põhjal võib seda ühendit pidada n-ö segarühma kuuluvaks. Nii ühendverbil *kinni nabima* (koosinuskaugus 0,1352) kui ka ühendil *lahti saama* (koosinuskaugus verbiga *saama* on –0,0976) on EKSS-is mitu tähendust ning vaatamata koosinuskauguse väärtusele ei saa kummagi kohta öelda, et tegemist on selgelt idiomaatilise ühendverbiga.

Lõpetatust väljendavaid ühendverbe, näiteks *maha põlema*, *ära ostma*, *välja puhkama*, *ära kaotama* jne, on mõlemas tabelis – nii kõrge kui ka madala koosinuskaugusega ühendite seas. Võib oletada, et ühendisse kuuluva adverbi liik ei mõjuta ühendverbi kompositsionaalsuse taset.

Selline koosinuskauguse väärtuste järgi ühendverbide kompositsionaalsuse skaalale sättimine vajab edasist põhjalikku hindamist, mille jaoks on vaja välja töötada mitme inimese arvamusel põhinev kuldstandard, kus ühendverbid (või mingi hulk ühendverbe) on järjestatud kompositsionaalsuse järgi. Tänu inimeste

keelepädevusele peetakse keeletöötluses tihti mitmel (spetsialisti) arvamusel tehtud kokkuvõtteid sellisteks, millest automaatne keeletöötlus parem olla ei saa (Kumar 2011: 20). Sellise kuldstandardi väljatöötamine annaks piisavalt usaldusväärse andmestiku, millega siinsest uurimuses saadud tulemusi võrrelda.

8. Kokkuvõte

Artiklis vaadeldi esmalt, kas ja kuidas on võimalik eesti keele ühendverbe automaatselt liigitada ainukordseteks ja korrapärasteks, ning seejärel rakendati distributiivse semantika võimalusi leidmaks ühendverbide kompositsionaalsuse taset.

Selgus, et ühendverbide liigitamine kindlatesse rühmadesse ei ole kasutatud mõõdikuid rakendades võimalik, sest ühelgi neist ei olnud selleks piisavalt head tulemused. Korrapäraste ühendite puhul osutus parimaks vastastikuse informatsiooni väärtus (MI), sest see tõstab esile harva esinevad ühendid, mis üldjuhul pole polüseemsed, ja see on iseloomulik eesti keele korrapärastele ühendverbidele. Samuti selgus, et eri liiki ühendverbide tuvastamisel töötavad mõõdikud erinevalt.

Teiseks püüti distributiivse semantika vahendeid kasutades määrata ühendverbide kompositsionaalsuse taset. Selgus, et ühendverbid asetuvad kompositsionaalsuse taseme alusel skaalale, mille ühes otsas on kompositsionaalsed ja teises otsas mittekompositsionaalsed ühendverbid ning kompositsionaalsuse määramisel tasub rakendada koosinuskauguse väärtust. Rakendatud meetodit tuleb kindlasti täiendada mudeliga, mis võtaks arvesse sõnade polüseemsust, ning sama ülesande täitmiseks kasutada ka teisi meetodeid, näiteks klasteranalüüsi. Kõikide meetodite tulemuslikkust on võimalik põhjalikumalt hinnata peale mitme inimese keelepädevusel põhineva kuldstandardi ehk ühendverbide kompositsionaalsuse järjestuse väljatöötamist, mis ongi selle uurimuse järgmine samm.

Viidatud kirjandus

Aedmaa, Eleri 2015. Statistilised meetodid ühendverbide tuvastamisel tekstikorpusest. [Statistical methods for Estonian particle verb extraction from text corpora.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 11, 37–54. http://dx.doi.org/10.5128/ERYa11.03

Baldwin, Timothy; Villavicencio, Aline 2002. Extracting the unextractable: A case study on verb-particles. – Proceedings of the Conference on Computaional Natural Language Learning (CoNLL 2002), Taipei, Taiwan, 31 August – 1 September 2002. Association for Computational Linguistics, 1–7. http://dx.doi.org/10.3115/1118853.1118854

Bannard, Colin 2005. Learning about the meaning of verb-particle constructions from corpora. – Computer Speech & Language, 19 (4), 467–478. http://dx.doi.org/10.1016/j.csl.2005.02.003

Bannard, Colin; Baldwin, Timothy; Lascarides, Alex 2003. A statistical approach to the semantics of verb-particles. – Proceedings of the ACL 2003 workshop on Multiword expressions: Analysis, acquisition and treatment, Vol. 18. Association for Computational Linguistics, 65–72. http://dx.doi.org/10.3115/1119282.1119291

Baroni, Marco; Dinu, Georgiana; Kruszewski, Germán 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. – Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, 238–247. http://dx.doi.org/10.3115/v1/p14-1023

Blaheta, Don; Johnson, Mark 2001. Unsupervised learning of multi-word verbs. – Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations, 54–60.

Bott, Stefan; Schulte im Walde, Sabine 2014. Optimizing a distributional semantic model for the prediction of German particle verb compositionality. – Proceedings of the 9th Conference on Language Resources and Evaluation, Reykjavik, Iceland.

Bruni, Elia; Tran, Nam-Khanh; Baroni, Marco 2014. Multimodal distributional semantics. – Journal of Artificial Intelligence Research (JAIR), 49, 1–47.

Bullinaria, John A.; Levy Joseph P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. – Behavior Research Methods, 39 (3), 510–526. http://dx.doi.org/10.3758/BF03193020

EKG II = Erelt, Mati; Reet Kasik; Helle Metslang; Henno Rajandi; Kristiina Ross; Henn Saari; Kaja Tael; Silvi Vare 1993. Eesti keele grammatika II. Süntaks. Lisa: kiri. [The Grammar of the Estonian Language II: Syntax.] Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

EKSS = Eesti keele seletav sõnaraamat I–VI. [The Explanatory Dictionary of Estonian.] Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll (Toim.). Eesti keele instituut. Tallinn: Eesti Keele Sihtasutus, 2009.

Erelt, Mati 2013. Eesti keele lauseõpetus. Sissejuhatus. Öeldis. [Estonian Syntax. Introduction.] Tartu ülikooli eesti keele osakonna preprindid 4. Tartu Ülikool.

Erk, Katrin; Padó, Sebastian 2008. A structured vector space model for word meaning in context. – Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 897–906. http://dx.doi.org/10.3115/1613715.1613831

Evert, Stefan; Krenn, Brigitte 2001. Methods for the qualitative evaluation of lexical association measures. – Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 188–195. http://dx.doi.org/10.3115/1073012.1073037

Garvin, Paul L. 1962. Computer participation in linguistic research. – Language, 38 (4), 385–389. http://dx.doi.org/10.2307/410674

Harris, Zellig S. 1954. Distributional structure. – Word, 10, 146–162. http://dx.doi.org/10.1080/00437956.1954.11659520

Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Püsiühendite leidmine teksti abil. [Extraction of multiword expressions using text corpus.] – Renate Pajusalu, Tiit Hennoste (Toim.) Tähendusepüüdja: pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Catcher of the Meaning: Festschrift for Professor Haldur Õim on the occasion of his 60th birthday. TÜ üldkeeleteaduse õppetooli toimetised 3. Tartu: Tartu Ülikool, 172–184.

Kaalep, Heiki-Jaan; Muischnek; Kadri 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. [Estonian multiword expressions in computational linguistics.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 5, 157–172. http://dx.doi.org/10.5128/ERYa5.10

Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. [Syntagmatic Relationships of Estonian Content Words in Corpus and Pedagogical Lexicography.] Tallinna Ülikooli humanitaarteaduste dissertatsioonid 32. Tallinn: Tallinna Ülikool. http://www.etera.ee/zoom/2000/view?page=3&p=separate&view=0,432,2067,788 (25.2.2016).

Katz, Graham; Giesbrecht, Eugenie 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. – Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. Association for Computational Linguistics, 12–19. http://dx.doi.org/10.3115/1613692.1613696

Kühner, Natalie; Schulte im Walde, Sabine 2010. Determining the degree of compositionality of German particle verbs by clustering approaches. – Proceedings of the 10th Conference on Natural Language Processing, 47–56.

Kumar, Ela 2011. Natural Language Processing. New Delhi–Bangalore: I.K International Publishing House Ltd.

Lin, Dekang 1998. Automatic retrieval and clustering of similar words. – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Vol. 2. Association for Computational Linguistics, 768–774. http://dx.doi.org/10.3115/980691.980696

Manning, Christopher D.; Schütze, Hinrich 1999. Foundations of Statistical Natural Language Processing. Cambridge (Mass.)–London: MIT press.

McCarthy, Diana; Keller, Bill; Carroll, John 2003. Detecting a continuum of compositionality in phrasal verbs. – Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, acquisition and treatment, Vol. 18. Association for Computational Linguistics, 73–80. http://dx.doi.org/10.3115/1119282.1119292

McCulloch, Warren S.; Pitts, Walter 1943. A logical calculus of the ideas immanent in nervous activity. – The Bulletin of Mathematical Biophysics, 5 (4), 115–133. http://dx.doi.org/10.1007/BF02478259

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey 2013a. Efficient estimation of word representations in vector space. – arXiv preprint arXiv:1301.3781 (25.2.2016).

Mikolov, Tomas; Le, Quoc V.; Sutskever, Ilya 2013b. Exploiting similarities among languages for machine translation. – arXiv preprint arXiv:1309.4168 (25.2.2016)

Padó, Sebastian; Lapata, Mirella 2007. Dependency-based construction of semantic space models. – Computational Linguistics, 33 (2), 161–199. http://dx.doi.org/10.1162/coli.2007.33.2.161

Pärnamaa, Tanel 2015. Piltide automaatne kirjeldamine eesti keeles – visuaalse ja semantilise ühisesituse õppimine neurovõrkudega. [Translating pictures to Estonian – learning shared representations of images and languages using neural networks.] Magistritöö. Käsikiri Tartu ülikooli matemaatilise statistika instituudis. http://hdl.handle.net/10062/47568 (25.2.2016).

Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. [Types of Estonian Simple Sentences.] Tallinn: Valgus.

R Development CoreTeam 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Sag, Ivan A.; Baldwin, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan 2002. Multiword expressions: A Pain in the neck for NLP. – Alexander Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. Third International Conference, CICLing 2002, Mexico City, Mexico, February 17–23, 2002. Proceedings. Lecture Notes in Computer Science 2276. Springer Verlag, 1–15.

Sahlgren, Magnus 2006. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Stockholm University.

Schone, Patrick; Jurafsky, Daniel 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. – Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 100–108.

Uiboaed, Kristel 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel. [Statistical methods for phrasal verb detection in Estonian dialects.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 6, 307–326. http://dx.doi.org/10.5128/ERYa6.19

Weeds, Julie Elizabeth 2003. Measures and Applications of Lexical Distributional Similarity. Doctoral Dissertation. University of Sussex.

Võrgumaterjalid

Distributed Multi-sense Word Embedding (DMWE). http://www.dmtk.io/word2vec_multi.html (14.12.2015).

Eesti keele koondkorpus. http://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (25.2.2016).

Gensim. Topic modelling for humans. https://radimrehurek.com/gensim/models/word2vec.html (8.12.2015).

Kollokatsioonide otsing korpusest. https://korpused.keeleressursid.ee/clc/ (30.9.2015).

word2vec. https://code.google.com/p/word2vec/ (8.12.2015).

Eleri Aedmaa (Tartu Ülikool) on üldkeeleteaduse doktorant. Uurimisvaldkonnad: korpuslingvistika, püsiühendid, statistilised meetodid keeleteaduses.
Jakobi 2, 50090 Tartu
eleriaed@ut.ee

DETECTING THE COMPOSITIONALITY OF ESTONIAN PARTICLE VERBS

Eleri Aedmaa
University of Tartu

The purposes of this article are to automatically classify Estonian particle verbs and detect their degree of compositionality. In order to group particle verbs, the lexical association measures (AMs) are compared. For the detection of the degree of compositionality of Estonian particle verbs, a model based on distributional semantics is used. The experiment is carried out with the word2vec tool, using a continuous bag-of-words model which predicts the word given its context.

The analysis of the comparison of AMs revealed that none of the AMs used achieve high enough precision values to classify the particle verbs. Hence, it can be assumed that Estonian particle verbs cannot be divided cleanly into the classes of compositional and non-compositional particle verbs, but rather populate a continuum between entirely compositional and entirely non-compositional expressions.

The experiment of assessing the degree of compositionality of the particle verbs using distributional semantic model proved successful. It is demonstrated that the value of cosine similarity can predict the degree of compositionality of particle verbs. However, in order to evaluate the method introduced here, it is important to create a ranking of human judgement on semantic compositionality for a series of particle verbs and base verbs to which they correspond.

Keywords: distributional semantics, natural language processing, multiword expressions, particle verbs, Estonian