doi:10.5128/ERYa17.16

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 17, 285–307

KÄÄNDEVORMIST SÕNAKS: MIDA NÄITAB SAGEDUS?

Ene Vainik, Geda Paulsen, Ahti Lohk

Ülevaade. Artikkel tegeleb nimisõnavormide iseseisvumise küsimusega leksikograafia vajadustest lähtudes. Eeldusel, et abstraktseid käändevorme iseloomustab korpuses üldine püsiv esiletuleku proportsioon, pakume välja statistilise mõõdiku – distributsiooniindeksi, mille abil otsustada, kas sõnavormi kasutussagedus on piisav selleks, et lugeda teda paradigmast emantsipeerunuks ning seega iseseisva märksõna kandidaadiks. Indeks arvestab vormi suhtelist sagedust korpuses, võrreldes tegelikku ja normi põhjal oodatavat kasutussagedust, ning laseb samale skaalale paigutada väga erineva absoluutsagedusega juhtumeid. Artiklis illustreerime distributsiooniindeksi toimivust tavaliste rikkaliku vormistikuga nimisõnade ning ambivorme andvate sõnade vormistike võrdlusena. Seame provisoorse indeksi lävendväärtuse, millest suurema väärtusega vormi võib pidada iseseisvaks lekseemiks. Indeksit ning lävendväärtust testitakse erinevate korpuste (EtTenTen13, ÜK 2019) andmete peal.*

Võtmesõnad: leksikograafia, korpuslingvistika, keeletehnoloogia, käändevormide iseseisvumine, sõnaliigid, eesti keel

Kvantitatiivsete meetodite rakendamine on kunst,
mida õpitakse edukate ja läbikukkunud katsete kaudu.
(Heiki-Jaan Kaalep 2018: 714)

1. Sissejuhatus

Leksikograafid soovivad keele kirjeldamisel langetada selgeid ja ühtsetel alustel põhinevaid otsustusi (Paulsen jt 2020). Keel aga pakub regulaarsete vormide ja selgesti piiritletavate üksuste kõrval ka kimbatust valmistavaid, kategooria või staatuse poolest ambivalentseid juhtumeid. Üks ambivalentsuse mõõde on sõnaliigiline mitmesus või ebamäärasus. Teine mõõde, mida traditsiooniliste sõnastike

* Uurimust on toetanud Eesti Teadusagentuur (PSG227). Täname artikli retsensente sisukate kommentaaride eest.

koostamise puhul on korduvalt esile toodud, on staatuse küsimus: kas esitada keelend sõnastikus iseseisva märksõnana või pigem allmärksõnana? (Vt Kaalep jt 2000, Karelson 2005, Muischnek, Vider 2005, Koppel 2020, Paulsen jt 2020, Vainik jt 2020) Leksikograafilises kirjanduses on osutatud vajadusele pöörata muuhulgas eraldi tähelepanu märksõna morfoloogilisele vormile, näiteks esitada eraldi märksõnana mitmuslik vorm (Atkins, Rundell 2008: 325). Nii on ka eesti keele puhul toodud eri märksõnadena *juurikas* 'taime tugev maa-alune vars' ja mitmuslik *juurikad* 'juurviljad toiduainena'; *helves* 'lendlevalt kerge ja väike aineosake' ja *helbed* 'jahust või tangust pressitud õhukesed libled' (vt EKI ühendsõnastik 2020).

Seoses sõnastike kolimisega elektroonilisse keskkonda on kadunud paberväljaannetele iseloomulikud mahupiirangud ning ilmneb kalduvus keeleüksusi käsitada ja esitada omaette märksõnadena, kui see sisuliselt vähegi põhjendatud on. Seda suundumust toetab relatsiooniliste andmebaaside kirjete tabelilaadne struktuur. Relatsiooniliste andmebaaside hulka võib liigitada ka eesti keele instituudi uue sõnastikusüsteemi Ekilex (lähemalt vt Tavast jt 2018). Lisaks on elektroonsete sõnastike üks põhiväärtusi kasutusmugavus, mis võib muu hulgas hakata tähendama keeleüksuste kättesaadavust võimalikult sellisel kujul, nagu kasutaja nendega tekstides kokku puutub. Ükskeelsete sõnastike kasutajaina nähakse üha enam mitte ainult emakeelseid isikuid, kes eeldatavasti suudavad muutevormi põhjal selle algvormi (käändsõnade puhul nimetava) oletada, vaid ka keeleõppijaid, kellele see võib raskusi valmistada (Paulsen jt 2020). Seega on sõnastikutegijad uute potentsiaalsete märksõnade tulva ees, mille puhul tuleb esmalt otsustada, kas need on väärt saama omaette sõnaartikliks Sellised sõnad või sõnavormid on näiteks *loodav*, *põhinev*, *piires*, *valitsev*, *lõppemine*, *väheks*, *varjus*, *kaasav* (uute sõnade automaatsest tuvastamisest vt nt Langemets jt 2020).

Leksikaalsete üksuste eristamisel vajab leksikograaf mitmesugust infot, nii kvalitatiivset (sõna tähendus, morfoloogiline käitumine, süntaktiline roll lauses) kui kvantitatiivset (kasutussagedus, vormi ja tähendusega seotud tendentsid). Keelekasutust peegeldavad laiad keelekorpused annavad mitmekülgset teavet keeleüksuste kasutuse ja sageduse kohta, sh erinevates allkeeltes (nt ametlik keelekasutus *vs*. sotsiaalmeedia), mida esindavad vastavad alamkorpused. Sagedusinfo sellisena annab keeleuurijale pildi suundumuste kohta, kuid tahes-tahtmata jääb õhku küsimus, kui sage on piisavalt sage, et näiteks märgendada sõna või paradigmast irduma hakanud vorm iseseisvaks lekseemiks. On juhitud tähelepanu asjaolule, et üha suurenevate andmehulkade tingimustes pole sõnaraamatutegija peamine probleem enam mitte piisaval hulgal representatiivsete näidete leidmine, vaid pigem tohutu hulga kontekstinäidetega toimetulek (Alexander 2015). Seega vajab leksikograaf mingisugust orientiiri, mis laseks samalaadseid nähtusi võrrelda kas siis omavahel või mingi standardi taustal.

Käesolevas uurimuses püüame leida statistilise orientiiri ühele paljudest sõnaliigilise kategoriseerimise probleemidest: nimisõnade käändevormide sõnaliigimuutustele. Valikut põhjendab tõsiasi, et nimisõnal on hägusate piiridega sõnaliikide seas keskne positsioon (Vainik jt 2020) ning käändsõna vormistik toimib eesti keeles grammatiseerumis- ja leksikaliseerumisprotsesside allikana. Uute adverbide ja adpositsioonide tekkimisega noomenite vormidest kaasneb ka rakenduslikust vaatenurgast üks suurimaid sõnaliigitusprobleeme eesti keeles (Kaalep jt 2000, Karelson 2005, Muischnek, Vider 2005, Habicht jt 2011, Paulsen 2018, 2019).

Sõnaliigilt mitmese tõlgendusega sõnu ja vorme käsitleme selles töös ambivormidena (Vainik jt 2020). Keskendume oma uurimuses nimisõnadele ning nende vormidest arenevatele (potentsiaalsetele) lekseemidele, nagu näiteks *õnnetuseks* [pole palju vaja] : *õnnetuseks* [väänasin jala välja]; [kahekümne] *ringis*, [pühade] *ajal*.

Esialgsete tähelepanekute järgi kalduvad käändsõnadest ambivorme andma eeskätt semantiliste käänete vormid. Uurimuse eesmärk on leida ja määratleda statistilised kriteeriumid, mis aitaksid leksikograafil otsustada, kas kahtlusalust sõnavormi tuleks käsitleda regulaarse nimisõnavormina või on ta sedavõrd iseseisvunud, et teda võiks sõnaraamatus esitada omaette märksõnana.

Artikli ülesehitus on järgmine. Teises osas kirjeldame põgusalt üldteoreetilist tausta ja analoogse probleemi lahendamist Rootsi leksikograafias ning defineerime artiklis kasutatavad põhimõisted. Kolmandas osas tutvustame enda lahendusideed eesti keele nimisõnaliste ambivormide iseseisvusstaatuse tuvastamiseks ning töötame välja käänete distributsiooni normid. Neljandas osas esitleme konkreetse ambivormi sagedusandmeid käänete standardsete jaotusandmetega peegeldavat statistilist mõõdikut – distributsiooniindeksit. Viiendas ja kuuendas osas testime distributsiooniindeksit võrdlevalt rühma ambivormide ja n-ö tavasõnade (kontrollgrupiks valitud nimisõnade) peal kolmes korpuses ning pakume välja vormi iseseisvumisele viitava lävendväärtuse. Kokkuvõtvas osas arutleme, kas ja kuivõrd on distributsiooniindeks statistilise mõõdikuna kasutuskõlblik ning millises suunas võiks vormide emantsipeerumise kindlakstegemise metoodika loomisel edasi liikuda.

2. Taust ja kesksed mõisted

Teoreetiline tagapõhi, mis leksikograafide praktilise probleemi lahendamisel arvesse tuleb, on küsimus sõnavormide emantsipeerumisest, st muutumisest teatud lekseemi muutevormist iseseisvaks leksikoniüksuseks. Lekseemiks pürgiva vormi iseseisvust iseloomustab süntaktilise kasutuse ning tähenduse nihe võrreldes lähtevormiga (Paulsen 2018, 2019, Paulsen jt 2020). Kui muutus on suunaga sisusõnast suhtesõnaks, kirjeldatakse muutust grammatiseerumisteooria raames tähenduse pleekimise või abstraheerumisena (vt nt Lehmann 1985, Heine jt 1991, Habicht, Penjam 2006). Eesti keele puhul on kirjeldatud käändevormi abstraheerumist kui olemuslikku määr- ja kaassõnade tekkemehhanismi, mida saab vaadelda keeles jätkuvalt toimuva dünaamilise protsessina (nt EKG II: 38, Grünthal 2003: 26, Veismann, Erelt 2017: 448–452; leksikograafia vaatenurgast Habicht jt 2011, Paulsen 2018, 2019).

Üks morfoloogilise vormi iseseisvumise märke on tema kõrge esinemissagedus korpusetekstides. Siinkohal on mõttekas eristada vormi absoluutset sagedust (kõik esinemisjuhtumid kokku) ning suhtelist sagedust ehk esiletuleku määra võrreldes lekseemi vormide koguhulgaga korpuses. On leitud, et vormi suur suhteline sagedus on parem iseseisvumise ning semantilise nihke ennustaja kui seda on vormi kõrge absoluutsagedus; vorme, mis esinevad kasutuses sagedamini kui nende alussõnad, kaldutakse tajuma semantiliselt jagamatute tervikutena. (Hay 2001)

Kristian Blensenius ja Monica von Martens (2019: 663) kirjeldavad, kuidas Rootsi sõnastikutegijad käsitlevad juhtumeid, kus märksõna algvorm on küll

kasutusel, kuid selle kõrval on märksa suurema sagedusega hoopis mingi spetsiifiline sõnavorm, millel ilmneb lisaks ka tähendusnihe. Selle probleemi lahendamiseks on loodud võimalus sõnavormide sagedused sõnastikuartikli loomise käigus operatiivselt üle vaadata. Tööriist on rajatud relatsioonilise andmebaasina, mis sisaldab infot morfoloogiliste paradigmade moodustamise kohta ning seob sellega info korpuspäringutest. Sama tööriistaga saab kontrollida ka sõna ortograafiliste variantide suhtelist sagedust ning omakorda nende muutevormide esinemist. Rootsi süsteem võtab morfoloogilisest andmebaasist sõna kõik võimalikud vormid ning ekstraheerib korpuste otsingusüsteemist Korp nende taha värskeimad andmed korpustes esinemise kohta. Osa päringutulemusi on kuvatud otse leksikograafi töökeskkonda; osa jaoks on aga salvestatud protseduurid, mida haldab standardne liides MySQL Workbench. Välja saab pärida näiteks n-ö kahtlaste vormide loendid, mida peaks lähemalt inspekteerima (Blensenius, von Martens 2019: 667).

Rootsi leksikograafi töölauale kuvatakse korraga vormide sagedusinfo mitme erineva korpuse kohta koos mõningate andmetega vormide normatiivsest sagedusjaotusest (näiteks info, et ainsuse ja mitmuse vahekord on rootsi keeles 75% : 25%). Eraldi kuvatakse iga sõnavormi puhul ka sellega homograafsete vormide arv. Kui vormile leidub alternatiivne tõlgendus (või mitu), vähendab see vormi sagedusandmete usaldatavust. Normaalse jaotuvuse leidmiseks rehkendatakse homograafide sagedusandmed maha. Tulemuste valideerimiseks pakutakse võrdlusalusena infot mõnede sama sõnaklassi tüüpiliste esindajate vormide sagedusest. Vormide normaalsest erinev sagedusjaotus suunab leksikograafi näitestikku üle vaatama ning võib juhtida tähelepanu vajadusele organiseerida ümber sõna tähendusjaotus või koguni lisada uus märksõna (Blensenius, von Martens 2019).

Blensenius ja von Martens osutavad, et pole lihtne ülesanne pakkuda leksikograafile informatsiooni parajal määral ja õiges tööetapis. Leksikograafi ülekoormamine statistilise infoga võib loova sõnastikutöö juures lausa takistuseks osutuda; kui aga vajalik tükk infot näiteks mingi sõnavormi hälbivalt suurest sagedusest jääb õigel ajal arvesse võtmata, on hiljem juba keerukam seda lisada või sõnaartiklit ümber teha.

Vormi emantsipeerumises on mitu tegurit: semantika või süntaktilise funktsiooni muutus, pragmaatilise erifunktsiooni olemasolu võrreldes sama sõna muude vormidega. Siinses uurimuses jätame kõrvale sõnaliigimuutust peegeldavad kvalitatiivsed kriteeriumid ja uurime eesti keele nimisõnavormide iseseisvumist üksnes statistiliste näitajate põhjal. Eeldame, et emantsipeerumise tee valinud sõnavorm, mis on leidnud endale kasutus- ning tähendusniši, kaldub ka tekstides esinema sagedamini, kui ta esineks lekseemi muuteparadigma lihtliikmena.

Üks keskseid mõisteid käesolevas uurimuses on **ambivorm** (ambivalentne vorm), mille all mõtleme leksikograafias kategoriseerimisprobleeme tekitavat sõnavormi. Ambivormid hõlmavad sõnaliigilt mitmese tõlgendusega vorme (*andeks*, *hävitav*, *äraarvamata*), aga ka kinnistunud/etableerunud sõnu, millele on omane n-ö liikuda ühest sõnaliigist teise (*haige*, *ainurakne*, *vigur*, *kübeke*).

Teine oluline mõiste on **tekstisõna** ehk **sõne** (nt *homme*, *päeval*). Muutuvate sõnade puhul iseloomustab sõnesid morfoloogiline vorm (*päeval* – ainsuse adessiiv), muutumatute sõnade puhul mitte. Kasutame terminit ***käändevorm*** abstraktselt käände ja arvu kombinatsiooni tähenduses (nt ainsuse illatiiv, mitmuse komitatiiv). **Lemma** (nt *päev*) esindab sõna ehk lekseemi selle kõigis muutevormides (*päevani*,

päevaga, *päevadest* jne). Korpuslingvistikas tähistab lekseemi vormide koguarvu tekstis termin ***lemma sagedus*** (nt Kirt 2013). Sõna ise võib koosneda nii ainult ühest kui ka mitmest tüvimorfeemist või hoopis tüvimorfeemi(de)st ja juurde lisatud tunnusmorfeemi(de)st (vt Viht, Habicht 2019: 21–23, 31–34).

3. Käändevormide normaalne jaotus

Tekstikorpuste põhjal on võimalik moodustada sagedusloetelusid nii sõnede kui lemmade kohta (nt Kaalep, Muischnek 2002, Kirt 2013), samuti morfoloogiliste vormide esinemise kohta abstraktselt (konkreetse sõnatüvega seostamata). Korpuslingvistikas on tavaks oletada, et keeleüksuste sagedusjaotused on olulised ning et leksikaalse ja/või grammatilise variandi esinemissagedus vastab mingil kombel selle juurdumusele kognitiivses protsessis või protsessiga seotud esitusviisis (Kaalep 2018). See mõtlemine on jõudnud ka leksikograafiasse, nt EKI ühendsõnastikus (2020) on kavas hakata muutevormide kõrval kuvama infot nende korpusesageduse kohta (Jelena Kallas, suuline teade). Samast sagedusjaotuse tähenduslikkuse loogikast lähtudes eeldame, et käändevormid (arvu ja käände kombinatsioonid) ei jaotu korpuses juhuslikult, vaid et neile on tänu tähendusele ja tüüpilistele süntaktilistele funktsioonidele omane enam-vähem stabiilne ja iseloomulik esiletuleku proportsioon. Kui see on nii, saab emantsipeerumiskahtlusega käändevormide sagedusi võrrelda normaalse jaotuse puhuste sagedustega ning teha järeldusi selle kohta, kas konkreetse vormi esiletulek on normikohane või hälbib sellest.

Käändevormide normaalse jaotuse kindlakstegemiseks võtsime andmed veebis saada olevast statistikast morfoloogiliste vormide esinemiste kohta tasakaalus korpuses[1] (TKK) ja morfoloogiliselt ühestatud korpuses (MÜK). Kahe korpuse mahu erinevus on kolm korda (vastavalt *ca* poolteist ja pool miljonit sõnet); erinevus on ka kategooriate määratlemise tehnikates. Väiksema korpuse määrangud on eksperdid koostanud käsitsi, probleemseid kohti üle arutades ning lingvistiliselt võimalikult kvaliteetse tulemuse poole püüeldes (Kaalep jt 2000); mahukama TKK puhul automaatselt koos ühestamise ja järelühestamisega (Kirt 2013: 28). Eeldame, et mõlemal puhul on saavutatud usaldusväärne statistika lemmade, sõnede ning käändevormide esinemistest.

Käändevormide jaotus TTK-s ja MÜK-is on esitatud tabelis 1. Tulemused osutavad, et kahe morfoloogiliselt märgendatud korpuse käändevormidel on tõepoolest väga sarnased osakaalud. Andmeridade korrelatsioon on 0,999, keskmine erinevus 0, erinevuse standardhälve 0,008. Ka kändeuvate sõnade üldine osakaal on kahes korpuses väga sarnane: MÜK-is 54,3% ja TKK-s 53,2%. Tõdeme, et käändevormidel on eestikeelsetes tekstides tõepoolest iseloomulik jaotusmuster. Selleks, et täiendavalt neutraliseerida efekte, mida võiks anda aluseks olnud korpuste erinev suurus ning vormide tuvastamise metoodika, arvutasime käändevormide osakaalude keskmised, mida nimetame edaspidi vastava käände distributsiooni normiks. Kasutame neid norme edasises analüüsis võrdlusalusena.

On iseloomulik, et 74% käänete esinemisjuhtudest katavad grammatilised käänded. Semantilised käänded katavad ülejäänud 26% (sh kohakäänded 20% ja ülejäänud 6%). Ainsuslike ja mitmuslike vormide vahekord on vastavalt

[1] Tasakaalus korpus on tekstiliigiliselt ühtlustatud tekstikogu, sisaldades ilukirjanduse, ajakirjanduse ja teaduse tekste. Morfoloogiliste vormide sagedustabel on kättesaadav aadressil https://www.cl.ut.ee/ressursid/gram-kat/ (15.8.2021)

79,76% : 20,24%. Võrreldes rootsi keelega (vastavalt 75% : 25%, vt osa 2) kasutatakse eestikeelses tekstis niisiis mitmusevorme ligikaudu 5% võrra vähem.

Käesolevas uurimuses keskendume 11 semantilise käände ainsuse- ja mitmusevormide esiletuleku vaatlusele. Jätame kõrvale grammatilised käänded, kuna emantsipeerumiskahtlust ning sõnaliigilist mitmeti tõlgendatavust esineb põhiliselt just semantiliste käänete vormide puhul; ka ei ole grammatilisi käändevorme pelgalt väliskuju põhjal paljudel juhtudel võimalik tuvastada. Meie algandmed pärinevad TKK lemmade ja vormide kombineeritud sagedusloendist.[2] Lähtusime selles analüüsiosas üksnes välisest vormist.

Tabel 1. Käändsõnade muutevormide jagunemine morfoloogiliselt ühestatud korpuses (MÜK), tasakaalus korpuses (TKK) ning keskmiselt

Arv	Kääne	MÜK		TKK		Keskmine osakaal (distributsiooni norm)
		N	Osakaal	N	Osakaal	
ainsus	nominatiiv	73363	0,2619	1821718	0,2625	0,2622
	genitiiv	66552	0,2376	1368501	0,1972	0,2174
	partitiiv	27012	0,0964	744594	0,1073	0,1018
	aditiiv	2746	0,0098	79769	0,0115	0,0106
	illatiiv	1510	0,0054	31306	0,0045	0,0050
	inessiiv	11757	0,042	293980	0,0424	0,0422
	elatiiv	7524	0,0269	199978	0,0288	0,0279
	allatiiv	8110	0,029	181676	0,0262	0,0276
	adesiiv	12527	0,0447	299092	0,0431	0,0439
	ablatiiv	1199	0,0043	27089	0,0039	0,0041
	translatiiv	7481	0,0267	188582	0,0272	0,0269
	terminatiiv	618	0,0022	13751	0,0020	0,0021
	essiiv	924	0,0033	30956	0,0045	0,0039
	abessiiv	338	0,0012	8060	0,0012	0,0012
	komitatiiv	5617	0,0201	149779	0,0216	0,0208
mitmus	nominatiiv	18416	0,0657	483310	0,0696	0,0677
	genitiiv	13550	0,0484	402965	0,0581	0,0532
	partitiiv	9508	0,0339	271522	0,0391	0,0365
	illatiiv	366	0,0013	11214	0,0016	0,0015
	inessiiv	1832	0,0065	55862	0,0080	0,0073
	elatiiv	2188	0,0078	68953	0,0099	0,0089
	allatiiv	2082	0,0074	60706	0,0087	0,0081
	adessiiv	2506	0,0089	74579	0,0107	0,0098
	ablatiiv	240	0,0009	6046	0,0009	0,0009
	translatiiv	387	0,0014	13283	0,0019	0,0017
	terminatiiv	62	0,0002	1784	0,0003	0,0002
	essiiv	117	0,0004	4470	0,0006	0,0005
	abessiiv	64	0,0002	1942	0,0003	0,0002
	komitatiiv	1532	0,0055	44883	0,0065	0,0060
	kokku	280130	1	6940350	1	1

[2] Täname Kadri Muischnekit andmete eest. Andmed ja protseduur on algselt kirjeldatud Riin Kirdi magistritöös (2013: 44): lemmade ja sõnavormide kombineeritud loend annab informatsiooni selle kohta, millistest sõnavormidest lemma sagedus koosneb.

Tabelis 1 toodud normid näitavad, kuidas jaotuvad käändsõnad kahe korpuse kokkuvõttes arvu ja käände kategooriate lõikes üldiselt. Üksiksõnade puhul tuleb kindlasti ette varieeruvust (sh oletatavasti sõltuvalt nt lemma sagedusest, tähendusest ning tüüpilisest süntaktilisest rollist, esinemisest kollokatsioonides, konstruktsioonides, püsiühendites jm). Hoolimata sellest mööndusest peaksid need normid meie uurimuse praeguses etapis sobima selleks, et ennustada näiteks lemma üldsageduse põhjal selle jaotumist käändevormide vahel. Seda muidugi tingimusel, et jaotumine on normaalne ega sisalda emantsipeerunud vorme. Emantsipeerunud vormi puhul peaks ilmnema oluline lahknevus normi põhjal ennustatud esiletuleku määrast.

Käändekategooria sagedusandmete võrdluseks võib siinkohal tuua Jüri Valge (1970) koostatud eesti keele käänete sagedusloendid, mille saamiseks oli uuritud kolme tekstitüübi – ilukirjanduse, kõnekeele (draamateoste tegelaste kõne põhjal) ja ajalehekeele – nimisõnade käändevormide sagedusjaotusi. Valge (1970: 33) statistilises analüüsis on grammatiliste käänete osakaal veelgi suurem kui käesolevas uurimuses (vastavalt 77,1% ilukirjandustekstides, 79% kõnekeelsetes tekstides ja 77,2% ajakirjanduses). Semantilistest käänetest on kõigis uuritud tekstiliikides kõrgeima sagedusega adessiiv, kohakäänete sagedusrühma tasemele pretendeerivad ka translatiiv ja komitatiiv; ülejäänud erikäänded on sagedusloendite lõpus, kuhu positsioneerub kõigis kolmes tekstitüübis ka ablatiiv. Mitmuse ja ainsuse vorme selles uurimuses ei eraldatud. Eesti keele väliskohakäänete esinemissagedust poolspontaanses kõnes on uurinud Jane Klavan, Tanel Alumäe ja Arvi Tavast (2020), kelle tulemused kinnitavad, et väliskohakäänetest sagedasim on adessiiv ning selgelt harvim ablatiiv, mis kuulub nelja eesti keele kõige väiksema sagedusega käände hulka.

4. Distributsiooniindeks

Järgnevas osas vaatleme, kas ja kuidas on käändevormide normaalse jaotuse teadmisest kasu üksikvormi esiletuleku sageduse hindamisel. Teades mingi käändevormi distributsiooni normi ning lemma sagedust korpuses, on võimalik välja arvutada konkreetse sõne oodatav esiletuleku sagedus selles korpuses. Kui võrrelda oodatavat sagedust selle sõne tegeliku esinemisega, selgub erinevus, mis võib näidata kas prognoositust suuremat või väiksemat esinemist. Selleks, et eri sõnede andmed oleksid omavahel võrreldavad, on mõistlik tulemus normaliseerida, st jagada lemma sagedusega korpuses.

Nende tehete tulemuseks on distributsiooniindeks (ehk D-indeks), mis iseloomustab konkreetse käändevormi sageduse normikohasust tekstimassiivis. Distributsiooniindeksit saab arvutada valemiga $DI = (Z - X \times Y) / X$, kus DI on distributsiooniindeks, Z sõne sagedus korpuses, Y käändevormi distributsiooni norm (tabelist 1) ning X lemma sagedus korpuses. Valemi olemus seisneb selles, et sõne oodatav sagedus (käändevormi normi ning lemma sageduse korrutis) lahutatakse sõne tegelikust korpusesagedusest ning vahe jagatakse normaliseerimise eesmärgil lemma sagedusega.

Indeksi väärtus jääb vahemikku pluss ühest ja miinus üheni. Positiivsed väärtused näitavad sõne üleesinemist ning negatiivsed alaesinemist võrreldes normiga.

Nullilähedased väärtused näitavad, et konkreetse sõnavormi sagedus vastab enamvähem täpselt normile. Indeksi üks eeliseid seisneb selles, et samal skaalal saab omavahel võrrelda eri suurusega korpuste andmeid. Näiteks tasakaalus korpuse ja uusima eesti keele ühendkorpuse (ÜK 2019) mahu erinevus on ligi tuhatkordne.[3] Juhuslikult valitud sõnavorm *munale* (ainsuse allatiiv sõnast *muna*) esineb tasakaalus korpuses ühe korra ning ÜK 2019-s 581 korda[4]; lemma *muna* vastavalt 883 ning 107 659 korda. Valemi järgi (sõna *muna* sagedus korrutatud ainsuse allatiivi normiga 0,0276 tabelist 1) oleks sõne *munale* oodatavad sagedused tasakaalus korpuses 24,4 ning ÜK 2019-s 2970,1. Mõlemas korpuses on selle sõne tegelik sagedus väiksem käändevormi distributsiooni normiga ennustatavast, erinevus on vastavalt –23,4 ja –2388,7. Sõne *munale* distributsiooniindeksi väärtuseks kujuneb tasakaalus korpuse puhul –0,026 ning koondkorpuses –0,022. Indeks on mõlemal puhul küll negatiivse väärtusega, kuid nullilähedane, mis räägib sellest, et sõne *munale* esiletulek on mõlemal puhul normist õige pisut madalam.

Indeks annab ka võimaluse võrrelda samal skaalal absoluutsageduselt vägagi varieeruvaid sõnesid. Näiteks sõne *aastal* absoluutsagedus tasakaalus korpuses on 13514 ning sõnel *hingepõhjani* 31. Lemmade *aasta* ning *hingepõhi* sageduste absoluutväärtused on samuti väga erinevad: 51581 ja 62. Indeksi väärtuseks (vastavalt ainsuse adessiivi ning ainsuse terminatiivi norme arvesse võttes) kujuneb sõne *aastal* jaoks 0,218 ning *hingepõhjani* puhul 0,498. See tulemus ütleb, et mõlemate esilduvus ületab normikohast, ning et vorm *hingepõhjani* ületab normi poolt ennustatavat sagedust märksa rohkem.

D-indeks ei selgita mõistagi kummagi ülaltoodud näite puhul esiletuleku proportsioonide põhjusi, see on lihtsalt statistiline mõõdik, mis aitab paigutada eri sõnesid esiletuleku järgi suhtelisele skaalale. Kuidas selle suhtelise skaala abil tekstimassiivist võimalikult täpselt eraldada emantsipeerumiskahtlusega sõnavormid, püüame välja selgitada artikli järgnevas osas.

5. Distributsiooniindeksi testimine

5.1. Valim ja protseduurid

Kuivõrd distributsiooniindeksi väljatöötamist motiveerib taustaoletus ambivormide ning tavaliste nimisõnavormide suhtelise esiletuleku erinevustest, siis proovime järele, kas indeks aitab üksteisest eristada regulaarseid nimisõnavorme ning emantsipeerumiskahtlusega ambivorme. Selleks moodustasime kaks testisõnade valimit: n-ö tavasõnu esindavad nimisõnad (N = 26) ning ambivormid (N = 46). Võrdlesime nende vormistike distributsiooniindekseid TKK andmete põhjal. Eeldasime, et tavaliste – parimal juhul täisparadigmaga – nimisõnade vormid jaotuvad käänete lõikes normaalselt. See tähendab, et distributsiooniindeks peaks olema kõikides vormides nullilähedane. Teiseks eeldasime, et ambivormide distributsiooniindeks saab nullist suuremaid väärtusi, st vastav vorm tuleb esile sagedamini, kui oleks oodatav vormide normaalse jaotumise põhjal.

Uurimuses kasutatud tavasõnad (algselt 39) valisime tasakaalus korpuse sõnaloendist kolme põhikriteeriumi järgi. Esiteks valisime sõnu, millel oleks esinemisi

[3] ÜK 2019 maht on 1 500 284 681 sõnet, andmed võetud korpuspäringusüsteemi SketchEngine kaudu (18.9.2020).

[4] SketchEngine esitab käändsõnavormide ülevaates ainsuse ja mitmuse vormid koos. Sõnede *munale* ja *munadele* puhul selgitati tulemuste juhuvalimi põhjal ainsusliku vormi keskmine esiletuleku osakaal (0,47) ning arvutati sõne *munale* esiletuleku arv kaudselt.

võimalikult paljudes käände ja arvu kombinatsioonides. See on vajalik selleks, et käändekategooriate normaalne jaotus saaks põhimõtteliselt esile tulla. Vormide rohkus käib korpuses üldiselt käsikäes suurema esinemissagedusega, kuid suur sagedus ei olnud eesmärk omaette (vt lisa 2). Teiseks vältisime sõnu, mille vormide rohkus tuleneb homonüümsete sõnade paralleelsetest muuteparadigmadest (*kurk* : *kurgile*/*kurgule*). Kolmandaks jälgisime, et tavasõnad esindaksid erinevaid semantilisi kategooriaid (nt konkreetsed ja abstraktsed nimisõnad, kohad ja asjad, inimest tähistavad sõnad). Valimis on nii ühetüvelisi kui ka liitsõnu. Välistasime katsesõnade seast need 13, mille mõni vorm on esindatud meie nn ambivormide andmebaasis[5]. Katsesõnade arvuks jäi 26.

Sõnaliigiliselt mitmese tõlgendusega sõnad (algselt 64) valisime ambivormide andmebaasist põhimõttel, et nende väliskuju vastaks nimisõna semantilise käände vormile ning vastav nimisõna (oletatav lemma) sisalduks tasakaalus korpuse sõnaloendis. Jälgisime, et valimis oleks esindatud kõik eesti keele semantilised käänded, lisaks püüdsime sama käänderühma seast valida võimalikult erineva sõnaliigilise potentsiaaliga ambivorme (näiteks ainsuse inessiivi laadne määrsõnakandidaat *palavikus* ning kaassõnakandidaat ([kellegi, millegi] *tuules*). Ambivormide seas on nii selliseid, mis on EKI ühendsõnastikus juba saanud omaette märksõna staatuse, kui selliseid, millel seda (veel) ei ole (vt lisa 1). Kuna tegemist on leksikaalse kategooria mõttes piiripealsete sõnadega, kuulubki ambivormide staatuse juurde teatav dünaamilisus. Eemaldasime valimist ambivormid, mis olid taandatavad mitmele eri lemmale (nt ainsuse elatiiv *peoga* < *pidu* ja *pihk*) või mille lemma oli korpuses tuvastatud käändevormikujulisena (nt *pahinal*, *hakul*). Lõplikult jäi valimisse 46 ambivormi, mida oli võimalik tõlgendada kindla nimisõna vormina.

Valimisse kuuluvate sõnade kohta tegime väljavõtte tabelist, mis koondab tasakaalus korpuse lemmasid ja nendega seonduvaid vorme koos sagedustega. Valimi 26 tavalist nimisõna esinevad kokku 688 vormis. Kuna grammatiliste käänete distributsioon ei ole käesoleva uurimuse fookuses, märgendasime üksnes semantiliste käänete vormid (arvu ja käände kombinatsioonid). Seda tegime MS Excelis sõnalõppude filtreerimise käskudega ning sissesattuvaid väärtõlgendusi manuaalselt parandades.

Ambivorme käsitlesime sama protseduuri alusel. Tegime väljavõtte nii ambivormide kui nende oletuslike lemmadega seotud sõsarvormide kohta: näiteks vorm *esmapilgul* tõlgendati ainsuse adessiivina sõnast *esmapilk*. Teised sama lemma (sõsar)vormid (nt *esmapilgule*, *esmapilguks* jt) kaasasime analüüsi selleks, et vaadelda ambivorme indeksi väärtuste alusel mitte eraldiseisvatena, vaid sarnaselt tavasõnadega kogu vormistiku taustal. Ambivormide lemmade vormide üldarvuks jäi 680.

5.2. Tulemused

Üldisi tulemusi tavasõnade ja ambivormide lemmade semantiliste käänete vormide jaotuses näitab tabel 2.

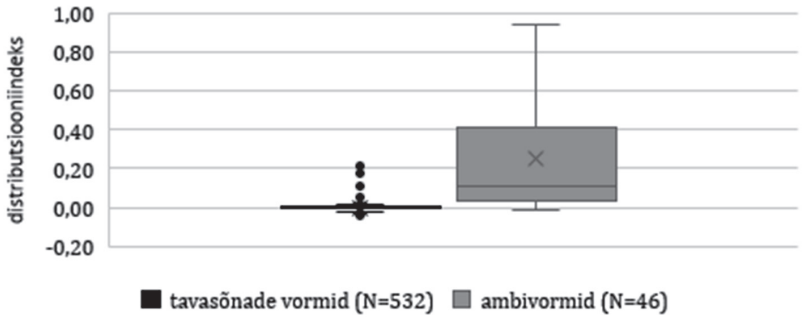
[5] Andmebaas (*ca* 3500 kirjet) sisaldab leksikograafilises töös sõnaliigilise kuuluvuse poolest probleeme valmistavaid, mitmese kuuluvusega või märgendita n.-ö allmärksõna staatuses olevaid üksusi. Lähemalt vt Vainik jt 2020.

Tabel 2. Kirjeldav statistika uuritud vormide kohta tasakaalus korpuses

Võrreldav näitaja	Tavasõnad	Ambivormide lemmad
Lemmade arv	26	46
Semantilistes käänetes olevate vormide arv	533	475
Semantilistes käänetes olevaid vorme keskmiselt	20,5	10,32
Indeksi väärtuste mediaan	–0,001	0,034
Indeksi väärtuste keskmine	0,000	0,034
Indeksi väärtuste standardhälve	0,024	0,121
Indeksi väärtuste maksimum	0,218	0,937
Indeksi väärtuste miinimum	–0,044	–0,043

Ootuspäraselt on tavasõnadel semantiliste käänete vorme keskmiselt rohkem, kuna välja valitud olidki võimalikult rikkaliku vormistikuga sõnad. See, et erinevus on kahekordne, näitab, et ambivormide aluseks olevad lemmad on korpuses esindatud vaegparadigmadega. Indeksi väärtuse mediaanid ja keskmised langevad mõlemas valimis kokku, ambivormide lemmadest moodustatud vormidel on nende väärtused aga kõrgemad, mis reedab D-indeksi suuremate väärtustega vormide mõju. Samuti iseloomustab ambivormide lemmadest moodustatud vormide indeksite väärtusi viis korda suurem standardhälve, mis tähendab, et nende vormide esiletuleku erinevus normist varieerub suuresti. Ambivormide indeksite maksimumväärtus on tavasõnade vastavast väärtusest üle nelja korra suurem, mis osutab, et indeks näitab tõepoolest teatud vormide esinemist normist tublisti rohkem. Miinimumväärtused on ambi- ja tavasõnade puhul võrdsed ning üsna väikese negatiivse väärtusega. Järelikult korpuse andmetes ei ole nende lemmade puhul vorme, mida esineks normiga võrreldes drastiliselt vähe. Need vormid, mis üldse puuduvad, statistikas ei kajastu.

Joonisel 1 on esitatud võrdlevalt tavasõnade semantiliste käänete vormide ning uuringusse valitud ambivormide jaotused D-indeksi alusel. Vertikaaltelg kujutab distributsiooniindeksi kasvavat väärtust. Diagrammi "karp" näitab vahemikku alumise ja ülemise kvartiili vahel, kuhu paigutub pool andmestikust, "vurrud" näitavad usaldusväärset varieerumisvahemikku. Eraldi märgitud andmepunktid tavasõnade graafikul näitavad üldise varieerumisvahemikuga võrreldes kahtlaselt erinevaid väärtusi.[6]



Joonis 1. Tavasõnade semantiliste käänete vormide ja ambivormide võrdlev jaotus tasakaalus korpuses

[6] Vt http://www-1.ms.ut.ee/mart/biomeetria2013/loeng1.pdf, lk 32 (5.2.2021).

Nagu oletatud, koondub enamik tavasõnade vormidest normaalse jaotuse tõttu indeksi nullväärtuse lähedusse, seevastu ambivormide D-indeksid varieeruvad skaalal, mille keskväärtus on 0,25 ning mediaan 0,107. Seega peab paika meie oletus, et tavalised nimisõnavormid järgivad normaalset jaotust ning ambivormid hälbivad sellest. Ei saa siiski öelda, et indeks eristaks ühemõtteliselt ambivormid tavasõnade vormidest. Seda esiteks seetõttu, et mitte kõiki ambivorme ei iseloomusta normaalse jaotuse põhjal ennustatust suurem sagedus ning teisest küljest leidub ka tavasõnadel vorme, mille D-indeks küündib välja normaalsest jaotusest, mõnedel kuni ambivormide keskmiseni.

Järgmiseks vaatleme võrdlevalt tavasõnade vorme, mis on saanud kõrgemaid D-indeksi väärtusi, ning ambivormide D-indeksite tulemusi. Joonis 2 esitab 46 kõige esileküündivama tavasõna vormi indeksid ning joonis 3 kõigi ambivormide vastavad näitajad.

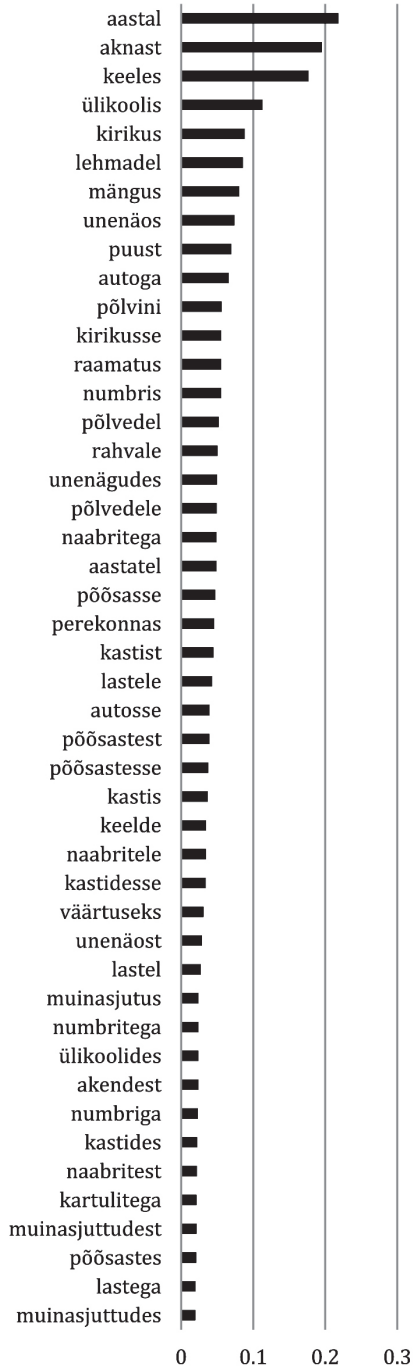
Joonisel 2 on indeksite kahanevas reas märgata selget jõnksu: vormid *aastal*, *aknast* ja *keeles* eristuvad teistest järgu võrra ning edasi kahaneb vormide esiletulek sujuvalt. Joonisel 3 on kõrgeima D-indeksiga (0,937) vorm *esmapilgul* ning sellele järgnevad vormid indeksi järsult kahanevas reas kuni vormini *ankrusse* (0,154), edasi on langustrend märksa laugem. Leidub ambivorme, mille D-indeks on nullilähedane (*lambist*, *omadega*, *tuhandest*) või koguni alla selle (*moest*). Sellest võiks järeldada, et need ambivormid ei kuulu sageduse poolest emantsipeerumas olevate vormide hulka. Ja ometi on osa neist (nt *lambist*, *omadega*, *tuhandest*) juba EKI ühendsõnastikus märksõna staatusega. See vastuoluline tulemus osutab asjaolule, et üksnes kvantitatiivsetest näitajatest (ei absoluutsest aga suhtelisest sagedusest) ei piisa ambivormi iseseisvuse üle otsustamisel. Vaja on arvesse võtta ka kvalitatiivseid tegureid nagu lemma tähendus, tüüpiline süntaktiline roll, kollokatsiooniline käitumine jne (vt arutelu osas 3); nendele kavatseme keskenduda järgmistes uurimustes.

6. Distributsiooniindeksi lävendväärtus ja selle testimine

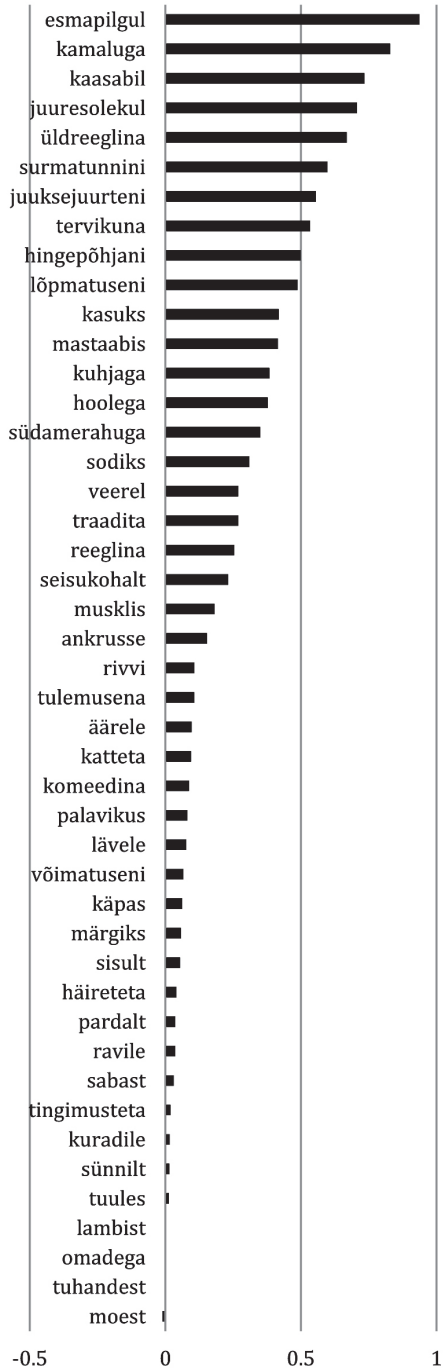
6.1. Lävendväärtuse otstarve ja protseduurid

Järgmiseks on meie eesmärk välja selgitada, kas ja milline võiks olla D-indeksi kriitiline väärtus ehk lävend, millest suurema väärtuse korral ei ole vormi iseseisvumises mingit kahtlust. Sellise lävendi olemasolu võiks leksikograafile anda statistilise argumendi, kui on tarvis otsustada, milline sõnavorm on väärt olema sõnastikus omaette märksõna. See oleks loomulikult vaid üks argumente teiste, sisulisemate hulgas.

Nagu võisime sedastada jooniste 2 ja 3 kirjeldamisel, esinesid D-indeksi kahanevates väärtustes teatavad jõnksud. Meie tavasõnade valimisse olid sattunud mõned sõnad (nt *aasta*, *keel*), mis evivad iseseisvumas olevaid vorme (nt *aastal, keeles*). Nende esilduvate vormide D-indeks jäi vahemikku 0,218–0,177; sellest selge vahe võrra väiksemate indeksi väärtuste suurus algab vormist *ülikoolis* (0,113). Ambivormide puhul nägime joonisel 3 vormi *ankrusse* (0,154) juures langustrendi selget muutust, jõnksu võrra madalamad indeksi väärtused algavad vormi *rivvi* juurest (0,107). On võimalik postuleerida provisoorne D-indeksi lävendväärtus



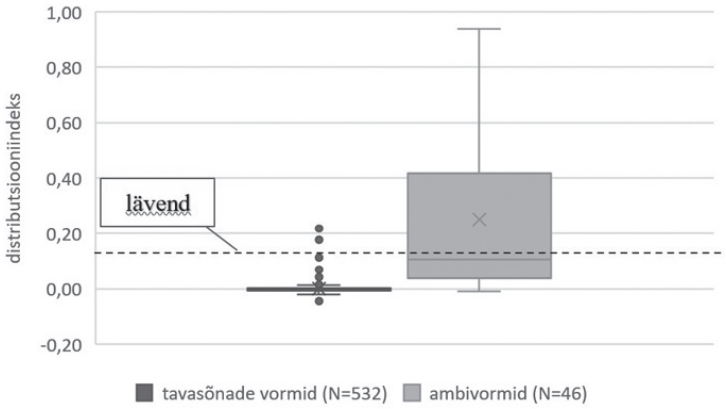
Joonis 2. Esileküündivate vormide pingerea tipp tavasõnade vormistikus tasakaalus korpuse põhjal (N = 46)



Joonis 3. Ambivormide D-indeksite väärtused tasakaalus korpuse põhjal

nende vahemike keskväärtusena (DI = 0,130), millest ülespoole jäävad kindlalt iseseisvumas olevad vormid. Lisame joonisele 4 seda lävendväärtust tähistava katkendjoone. Katkendjoonest üles jäävad D-indeksi väärtused, mille puhul sagedusnäitajast tundub vormi emantsipeerumise üle otsustamiseks piisavat.

Sellest lävendist ülespoole jääb ka 4 tavasõnade vormi (0,8% vormide üldhulgast). Tavasõnade vormidest 99,2 % jäävad allapoole seatud lävendit, mis näitab, et lävend suudab efektiivselt eristada normikohast ning normist hälbivat distributsiooni. Lävendist ülespoole jääb 22 ambivormi, mis moodustab 47,8 testitud ambivormidest. 52,2 % ambivormidest aga lävendi põhjal ei eristunud ning nende kirjeldamiseks on vaja lisaks kvantitatiivsetele ka kvalitatiivseid tunnuseid.



Joonis 4. Lävendväärtus tasakaalus korpuse andmestikul

Järgnevas vaatleme tasakaalus korpusest mahukamate korpuste andmete põhjal, millised vormid D-indeksi järgi normikohasest jaotuvusest eristuvad ja lävendväärtust ületavad. Kasutades sama valimit, mis oli aluseks TKK analüüsil (osas 5), pärisime tavasõnade ja ambivormide lemmade kohta välja vormid ÜK 2019-st ja etTenTen13-st.[7]

Tava- ja ambisõnade esinemisvormid ekstraheeriti ÜK 2019-st programselt korpustarkvara Sketch Engine rakenduse programmeerimise liidese (API) abil. Seevastu veebikorpuse etTenTen13 puhul saadi tava- ja ambisõnade esinemisvormid koos nende esinemissagedusega tekstikogu lausehaaval läbi käies. Selleks kasutati programmeerimiskeelt Python ja eestikeelse vabateksti masintöötlemiseks loodud teeki EstNLTK. EstNLTK tegi mh võimalikuks lauses esinevate sõnede algvormide leidmise ja morfoloogilise analüüsi.

6.2. Tulemused

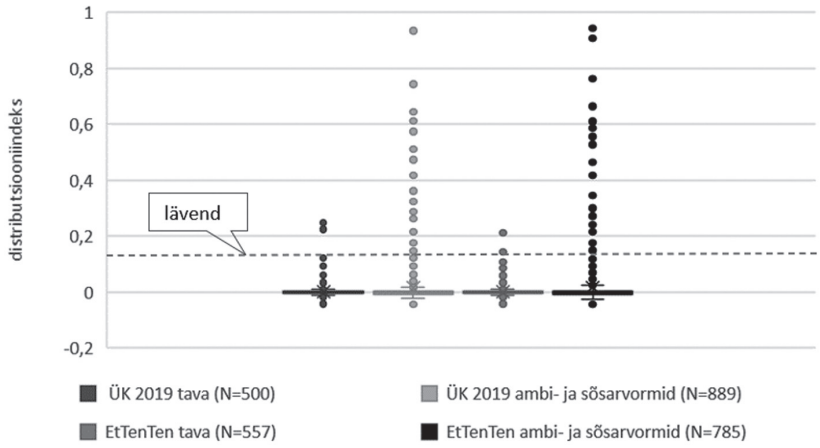
Kõigile ekstraheeritud semantilistele käändevormidele arvutasime D-indeksid, üldistatult on tulemused esitatud tabelis 3. Joonis 5 näitab tavasõnade ning ambisõnade lemmade vormide jaotust. Pildile on kaasatud mitte üksnes algsed ambivormid, vaid ka nende semantilistes käänetes olevad sõsarvormid ning seetõttu meenutavad graafikud nii tava- kui ambivormide puhul tagurpidi T-tähe kuju.

[7] Artikli kirjutamise ajaks suurim eestikeelseid tekste sisaldav korpus, ÜK 2019, koosneb eesti keele koondkorpusest (250 mln sõnet), sealhulgas tasakaalus korpusest (15 mln sõnet), ning eesti veebikorpustest 2013/2017/2019; korpuses domineerib ajakirjanduskeel. Kokku on ÜK 2019 maht poolteist miljardit sõnet. etTenTen13 on eestikeelsete veebilehtede korpus, mis lisaks formaalsematele tekstiliikidele sisaldab ka vabakeelt (foorumid ja blogid), mahuga 270 000 000 sõnet.

Diagrammi karp nullväärtuse juures näitab, et pool andmestikust asetus sellele kitsale vahemikule. Nii tava- kui ambisõnade vormide puhul näeme andmepunkte, mis ületavad mõistlikku varieeruvust, ambisõnade vormistike puhul on nende hulk ning väärtused aga kordades suuremad kui tavasõnade vormidel.

Tabel 3. Uuritavate vormide sagedused ja nende suhe lävendiga korpustes ÜK 2019 ja etTenTen13

Ülevaatlik näitaja	**ÜK 2019**	**etTenTen13**
Semantiliste käänete vorme kokku	1390	1341
Nullist suurema DI väärtusega vorme	467 (33,6%)	495 (36,9%)
Lävendit ületavaid (või võrdseid) vorme nullist suurema väärtusega vormide hulgas	42 (8,9%)	40 (8%)
Tavasõnade vorme lävendit ületavate hulgas	3	3
Algseid ambivorme lävendit ületavate hulgas	23 (53%)	27 (67%)
Ambivormide sõsarvorme lävendit ületavate hulgas	17 (42%)	14 (32%)



Joonis 5. Tavasõnade ning ambivormide lemmade vormistike võrdlev jaotus ÜK 2019 ja etTenTen13 andmete põhjal

Tabelist 3 on näha, et normaalset jaotust eiravaid vorme on kõigi vormide (nii ambi- kui tavasõnad) hulgas umbes kolmandik. Praeguste teadmiste põhjal seatud lävendväärtuse (0,130) abil eristub aga tõesti vaid kõige tipmisem tipp neist (alla 10%). Mõlema korpuse tulemustes leiduvad tavasõnade vormid, mille D-indeks ületab lävendit (nt *keeles* ja *aastal*). Algselt katsesse valitud ambivorme tuli mõnevõrra paremini esile etTenTen13 andmetest, näiteks *esmapilgul, kamaluga, juuresolekul, kaasabil, hingepõhjani, surmatunnini, kuhjaga, üldreeglina, tervikuna, mastaabis, traadita, musklis* jpt. ÜK 2019 andmetest tuli esile seevastu mõnevõrra rohkem ambivormide sõsarvorme (ambivormide lemmade vormid, mis ei olnud esialgses katsevalimis), näiteks *pardalt* asemel *pardale* ja *pardal* või *katteta* asemel *katteks*. Tuleb kindlasti märkida, et ka vastu ootusi esiletulnud vormide hulgas on selliseid, mis tegelikult juba sisalduvadki meie ambivormide andmebaasis (nt *pardal, rivis*), nii et lävendväärtuse abil ambivormide eristamise tabavuse määraks kujunes kahe korpuse keskmisena 65,9 %. Siiski jäi märkimisväärne hulk esialgses ambivormide

valikus olnud sõnesid oma indeksi poolest allapoole seatud lävendit. Sama nähtust – et ambivormi staatus ei tähenda ilmtingimata normist suuremat esinemissagedust – võisime sedastada ka tasakaalus korpuse tulemuste põhjal (joonis 3).

Põhjuseid, miks esilduvad ka n-ö ambivormide sõsarvormid, on mitmeid. Üks kaalukas põhjus on see, et ambivormide valimisse on sattunud mõned kohakäändevormid (nt *rivvi*, *pardalt*), mille puhul selgus, et need ei olegi oma lekseemi kõige domineerivamad vormid. Tähelepanu väärivad ka paar juhtumit, kus kriitilisse vahemikku langeb kaks käändevormi samast sõnast, näiteks *südamerahuga* ning *südamerahus*, mis on sünonüümsete viisimäärustena kasutuses tähenduses 'millestki hoolimata, muretult'. Teistsugune on vorm *surmatunnil,* mis sekundeerib loendis olnud ambivormile *surmatunnini*. Kui *surmatunnini* esineb sageli üldkeeles kinnistunud väljendites nagu *surmatunnini mäletama*, *jääb surmatunnini meelde*, siis otsetähenduslik vorm *surmatunnil* prevaleerib religioosse sisuga tekstides, mida korpus samuti arvestataval määral sisaldab. Mõnel puhul jääb esialgses ambivormide loendis olnud vorm sageduselt lihtsalt alla oma lemma teistele semantilise käände vormidele, millel pole märgata dekompositsionaalsuse teket ega tähendusnihet. Nii on näiteks vormi *tingimusteta* kõrval märksa levinumad *tingimustes*, *tingimustel, tingimusel*, *tingimustega* jne ning *katteta* kõrval on märksa sagedamad *katteks* ning *kattega*.

Üksikute ambivormide D-indeksite väärtuste võrdlemine eri korpustes võib anda pildi vormi iseseisvumise kulust. Näiteks kui tasakaalus korpuse andmestikus oli vormi *lambist* D-indeks alla lävendi, siis korpuses etTenTen13, mis kajastab kõnekeelsemat kasutust, on selle vormi indeks selgelt üle lävendi (0,165) ning mitmesugust keeleainest sisaldavas korpuses ÜK 2019 on väärtus samuti üle lävendi (0,138). Vormi *lambist* edasisest semantilisest emantsipeerumisest kõnekeeles annavad märku selle sõsarvormi *lampi* (lühike illatiiv) kasutamine samuti adverbilaadsena ning ka lemma *lamp* kasutamine adjektiivina, nt *täitsa lamp idee*.

Lävendväärtuse testimine näitas, et see võiks olla üheks orientiiriks, kui on vaja otsustada vormi emantsipeerumise määra üle. Vorme, mis lävendit ületasid, kuid ei ole veel ambivormidena arvele võetud, on põhjust lähemalt uurida, kuna võib selguda, et nii mõnigi neist võiks väga hästi nende hulka ka sobida. Samas ei garanteeri lävend kõikide huvipakkuvate vormide sõelalejäämist, kuna normist tuntavalt suurem esinemissagedus iseloomustab vaid osa ambivormidest.

Edasises uurimistöös ning leksikograafilistes rakendustes võiks lävendit ka ühele või teisele poole nihutada, et saada kas kitsamalt või laiemalt piiritletud tulemusi. Mõelda võiks ka lävendialuse n-ö eelkriitilise vahemiku määratlemisele, kuhu langeksid normist sagedamini esiletulevad sõnavormid, mille puhul aga nende iseseisvumine pole kindel ning vajab teiste argumentide tuge.

7. Kokkuvõtteks

Käesolevat uurimistööd on ajendanud vajadus statistilise orientiiri järele leksikaalsete üksuste eristamisel sõna morfoloogilistest vormidest. Artikli fookuses on nimisõnalised ambivormid, täpsemalt semantiliste käändevormide põhjal sisusõnast funktsioonisõnaks arenevad potentsiaalselt iseseisvumas olevad lekseemid. Kvantitatiivse analüüsi põhjal töötasime välja statistiku, mis käändevormide normaalsete osakaalude ja konkreetse ambivormi ning tema lemma sagedusel põhinevate

andmete järgi määrab vormi distributsiooniindeksi ehk D-indeksi. Viimase abil saab hinnata, kuivõrd on vorm normaalse või normaalsest hälbiva sagedusega.

Töö praktilises osas rakendasime D-indeksit kõigepealt kahe sõnavalimi (26 rikkaliku vormistikuga nimisõna ja 46 ambivormi/sõna) vormistike testimiseks tasakaalus korpuses. D-indeks võimaldab vormistikest ekstraheerida normist suurema esiletulekuga käändevormid ja paigutada need ühtsele skaalale. D-indeksi väärtuste üldiste sagedusjaotuste põhjal määratlesime provisoorse lävendväärtuse (≥ 0,130), millest suurema väärtusega vormi võib pidada iseseisvunuks. Lävendväärtust testisime omakorda kahe suurema korpuse, ÜK 2019 ja etTenTen13 tekstikogude põhjal, kasutades sama valimi tavasõnu ja ambisõnu, mida algselt olime testinud tasakaalus korpuses.

Analüüsi tulemuste põhjal järeldame, et seatud orientiir täidab teatud piirideni oma ülesannet: 65,9% vormidest, mille D-indeks tõuseb üle lävendi, on selgelt iseseisvunud; allpool lävendit paiknevate vormide staatust indeks otseselt ei näita ning vormi iseseisva lekseemi staatuse üle otsustamiseks läheb vaja kvalitatiivset analüüsi. Siinjuures võib üldistada, et korpuse sisu ja suurus ning ekstraheerimisprotseduuride omapärad mõjutavad D-indeksi väärtusi suhteliselt vähe. D-indeksi olulisim väärtus seisneb selles, et selle alusel on võimalik raputada korpusest välja vorme, mis vajavad leksikograafi tähelepanu ning otsustust märksõnastaatuse suhtes. See on üks korpustööriistu, mille järele leksikograafid on väljendanud vajadust ja mis võib tõhustada uute sõnade poolautomaatset jõudmist sõnastikku (vt Langemets jt 2020, Paulsen 2020: 194). Märksõna staatusse tõusnud nimisõnavorm omakorda tekitab küsimuse, milline on iseseisvunud lekseemi sõnaliigiline kuuluvus, ning siin ei saa ilma leksikograafi kvalitatiivse analüüsita.

Üks korpusmaterjalil D-indeksi testimise tulemusi on, et esilduvaid vorme võivad anda ka nn tavasõnad (nt *aastal*, *keeles*). Samuti selgus, et valimisse kuulunud ambisõnade lemmade vormistikus on esilduvaid vorme veelgi, osaliselt valimi ambisõnadega sarnase semantilise ja süntaktilise funktsiooniga. Edasises uurimistöös saaks seda nähtust kasutada sõnade omavaheliseks võrdluseks nende käändevormide distributsioonide alusel. Siinkohal tuleb mainida, et valimisse kuulunud tavasõnad ei ole tingimata "tavalised" – rikkaliku vormistikuga nimisõnad ei pruugigi olla tüüpilised seetõttu, et sõnad kipuvad esinema teatavates struktuurides ning kõikides võimalikes vormides esinevaid sõnu ei ole lihtne tuvastada. Tavasõnad on "tavalised" eelkõige selles mõttes, et vastupidiselt ambivormidele ei ole nende kuuluvuses nimisõnade või sõnastikumärksõnade hulka mingit kahtlust.

Töö käigus selgitati välja käändevormide normaalne jaotumine eestikeelses tekstis. Käändekategooriate esinemise normidena kasutati kahe korpuse (tasakaalus korpuse ja morfoloogiliselt ühestatud korpuse) käändevormide osakaalude keskväärtusi. Edaspidises uurimistöös on võimalik moodustada spetsiifilisemaid norme, näiteks eri tekstiliikide põhjal. Tasakaalus korpuses on esindatud kolm tekstiliigilist kategooriat (ilukirjandus-, ajakirjandus- ja teadustekstid), žanrilisi alamkorpusi saaks eraldi vaadelda ka näiteks ühendkorpuses. Põhimõtteliselt saaks n-ö globaalsete normide kõrval moodustada ka lokaalseid norme, mis hõlmaks näiteks teatud semantilistesse tüüpidesse kuuluvate nimisõnade käändevormide distributsioone. Võib näiteks eeldada, et semantilist tüüpi KOHT esindavate nimisõnade puhul prevaleerib loomulikult (sise)kohakäänete esiletulek, ning luua neile vastav norm. Statistilise analüüsi käigus selgus ka eesti keele nimisõnade üldine

ainsuslike ja mitmuslike käändevormide vahekord, mis on ümardades vastavalt 80% : 20%.

Lõpetuseks tõdeme, et leksikograafide soov on näha oma töölaual keeleandmetest esitusi, mis pakuvad ainestikust statistilisi väljavõtteid või paigutavad tulemused olulisuse skaalale (vt Paulsen jt 2020). Leksikograaf ei ole ettevalmistuselt andmeanalüütik, talle tuleb pakkuda infot võimalikult kompaktsel ning intuitiivselt haaratával moel. Seetõttu võiks olla päris teretulnud selline rakendus, mis rehkendab korpuste põhjal sõnavormidele D-indeksid ning kuvab need leksikograafi töölaual näiteks rohe-kolla-punase skaala taustal.

8. Edasisi arenguid

Käesoleva artikli avaldamise ajaks on meie meeskond loonud distributsiooniindeksi kalkulaatori prototüübi – veebipõhise liidese, mis ÜK 2019 andmetele tuginedes kogub statistika ja rehkendab huvipakkuvale käändevormile D-indeksi.[8] Kalkulaatori loomise põhimõtteid ning käändevormide sõnaks analüüsimise probleeme oleme täiendavalt lahanud artiklites (vastavalt Vainik jt 2021 ja Paulsen jt 2021).

Viidatud kirjandus

Alexander, Marc 2015. Words and dictionaries. – John R. Taylor (Ed.), The Oxford Handbook of the Word. Kindle Edition. Oxford: OUP, 37–53. https://doi.org/10.1093/oxfordhb/9780199641604.013.021

Atkins, Sue B. T.; Rundell, Michael 2008. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Blensenius, Kristian; von Martens, Monica 2019. Improving dictionaries by measuring atypical relative word-form frequencies. – Proceedings of eLex 2019 Conference. 1–3 October 2019. Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 660–675.

EKG II = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Silvi, Vare 1993. Eesti keele grammatika II. Süntaks. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Grünthal, Riho 2003. Finnic Adpositions and Cases in Change. Suomalais-Ugrilaisen Seuran toimituksia 244. Helsinki: Finno-Ugrian Society.

Habicht, Külli; Penjam, Pille 2006. Kaassõna keeleuurija ja -kasutaja käsituses ['Adpositions as viewed by a linguist and by a language user']. – Emakeele Seltsi aastaraamat, 52, 51–68.

Habicht, Külli; Penjam, Pille; Prillop, Külli 2011. Sõnaliik kui rakenduslik probleem: sõnaliikide märgendamine vana kirjakeele korpuses ['Parts of speech as a functional and linguistic problem: annotation of parts of speech in the corpus of Old Written Estonian']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 19–41. https://doi.org/10.5128/ERYa7.02

Hay, Jennifer 2001. Lexical frequency in morphology: Is everything relative? – Linguistics, 39 (6), 1041–1070. https://doi.org/10.1515/ling.2001.041

Heine, Bernd; Claudi, Ulrike; Hünnemeyer, Friederike 1991. Grammaticalization. A Conceptual Framework. Chicago: University of Chicago Press.

Kaalep, Heiki-Jaan 2018. Statistika koht keelemudelis ['Place of statistics in a language model']. – Keel ja Kirjandus, 8–9, 713–727.

Kaalep, Heiki-Jaan; Muischnek, Kadri; Müürisep, Kaili; Rääbis, Andriela; Habicht, Külli 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti keele

[8] http://teenus.eki.ee/d-index/ (15.10.2021).

testkorpuse morfosüntaktilise märgendamise kogemusest [‘Do the available morphological descriptions of Estonian work on a real text?’]. – Keel ja Kirjandus, 9, 623–633.
Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik [‘Frequency Dictionary of Written Estonian’]. Tartu: Tartu Ülikooli Kirjastus.
Karelson, Rudolf 2005. Taas probleemidest sõnaliigi määramisel [‘Once more on the issues of determining parts of speech’]. – Eesti Rakenduslingvistika Ühingu aastaraamat, 1, 53–70. https://doi.org/10.5128/ERYa1.03
Kirt, Riin 2013. Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning “Eesti keele seletava sõnaraamatu” märksõnaloendi võrdlus [‘Word Frequency Lists based on the “Balanced Corpus of Estonian” and Selective Comparison of Corpora Frequency Lists with Keywords from the “Explanatory Dictionary of Estonian”’]. Magistritöö. Tartu: Tartu Ülikool. http://hdl.handle.net/10062/39530
Klavan, Jane; Alumäe, Tanel; Tavast, Arvi 2020. Eesti keele väliskohakäänete kasutus poolspontaanses kõnes automaatse transkriptsiooni põhjal [‘Analysis of Estonian external locative cases in semi-spontaneous speech using an automatic transcription system’]. – Keel ja Kirjandus, 8–9, 757–774.
Koppel, Kristina 2020. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele [‘Corpus-based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners’]. Dissertationes linguisticae Universitatis Tartuensis 38. Tartu: Tartu Ülikooli Kirjastus.
Langemets, Margit; Kallas, Jelena; Norak, Kaisa; Hein, Indrek 2020. New Estonian words and senses: Detection and description. – Journal of the Dictionary Society of North America, 41 (1), 69–82. https://doi.org/10.1353/dic.2020.0005
Lehmann, Christian 1985. Grammaticalization: Synchronic Variation and Diachronic Change. – Lingua e stile, 20, 303–318.
Muischnek, Kadri; Vider, Kadri 2005. Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis [‘The problems of word class disambiguation in the automatic analysis of Estonian’]. – Eesti Rakenduslingvistika ühingu aastaraamat, 1, 99–112. https://doi.org/10.5128/ERYa1.05
Paulsen, Geda 2018. Manner and adverb: Fuzzy categorial boundaries in collocations. – Eesti Rakenduslingvistika Ühingu aastaraamat, 14, 117–135. http://dx.doi.org/10.5128/ERYa14.07
Paulsen, Geda 2019. Sõnaliigipiiridest kollokatsioonide vaatenurgast: erikäändelised noomenadverbid [‘Word class boundaries and collocations: The Estonian nominal adverbs in special cases’]. – Eesti Rakenduslingvistika Ühingu aastaraamat, 15, 121–137. https://doi.org/10.5128/ERYa15.07
Paulsen, Geda; Vainik, Ene; Tuulik, Maria 2020. Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias [‘On word classes in contemporary lexicography: The lexicographers’ view’]. – Eesti Rakenduslingvistika Ühingu aastaraamat, 16, 177–202. https://doi.org/10.5128/ERYa16.11
Paulsen, Geda; Vainik, Ene; Ahti, Lohk; Maria, Tuulik 2021 Catching lexemes: The case of Estonian noun-based ambiforms. – Iztok Kosem, Michal Cukr, Miloš Jakubíček, Jelena Kallas, Simon Krek, Carole Tiberius (Eds.), Proceedings of the eLex 2021 conference. Electronic Lexicography in the 21st Century, 5–7 July 2021, virtual. Brno: Lexical Computing CZ, s.r.o., 288–311.
Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina 2018. Unified data modelling for presenting lexical data: The case of EKILEX. – Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (Eds.), Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018. Ljubljana University Press, Faculty of Arts, 749–761.
Vainik, Ene; Paulsen, Geda; Lohk, Ahti 2020. A typology of lexical ambiforms in Estonian. – Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras (Eds.), Proceedings of XIX

EURALEX Congress: Lexicography for Inclusion, Vol. 1. Alexandroupolis: Democritus University of Thrace, 119–130.
Vainik, Ene; Lohk, Ahti; Paulsen Geda 2021. The distribution index calculator for Estonian. – Iztok Kosem, Michal Cukr, Miloš Jakubíček, Jelena Kallas, Simon Krek, Carole Tiberius (Eds.), Proceedings of the eLex 2021 conference. Electronic Lexicography in the 21st Century, 5–7 July 2021, virtual. Brno: Lexical Computing CZ, s.r.o., 121–138.
Valge, Jüri 1970. Eesti keele käänete sagedused kolmes funktsionaalses stiilis ['The Frequencies of the Estonian Cases in Three Functional Styles']. Tartu: Tartu Riiklik Ülikool, Eesti keele kateeder.
Veismann, Ann; Erelt, Mati 2017. Kaassõnafraas ['The adpositional phrase']. – Mati Erelt, Helle Metslang (Toim.), Eesti keele süntaks. Eesti keele varamu 3. Tartu: Tartu Ülikooli Kirjastus, 446–462.
Viht, Annika; Habicht, Külli 2019. Eesti keele sõnamuutmine ['Estonian Morphology']. Eesti keele varamu 4. Tartu: Tartu Ülikooli Kirjastus.

Võrguviited

EKI ühendsõnastik 2020 ['The EKI Combined Dictionary']. Indrek Hein, Jelena Kallas, Olga Kiisla, Kristina Koppel, Margit Langemets, Tiina Leemets, Maia Melts, Sirje Mäearu, Tiina Paet, Peeter Päll, Maire Raadik, Mai Tiits, Katrin Tsepelina, Maria Tuulik, Udo Uibo, Tiia Valdre, Ülle Viks, Piret Voll (Koost. Toim.). Eesti Keele Instituut. Sõnaveeb 2020. https://sonaveeb.ee (14.2.2020).
EstNLTK. https://github.com/estnltk (15.10.2021).
etTenTen13 = Estonian Web corpus etTenTen13. https://the.sketchengine.co.uk/auth/corpora (16.8.2021).
Muischnek, Kadri 2016. Eesti veeb 2013 (etTenTen13) korpus, morfoloogiliselt ühestatud. Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-0012EL
MÜK = Morfoloogiliselt ühestatud korpus. https://www.cl.ut.ee/korpused/morfliides (10.9.2020).
MySQL Workbench. https://www.mysql.com/products/workbench/ (15.8.2021).
Sketch Engine. https://www.sketchengine.eu (20.9.2020).
TKK = Tasakaalus korpus. https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=et (10.2.2020).
ÜK 2019 = Kallas, Jelena; Koppel, Kristina 2020. Eesti keele ühendkorpus 2019. Center of Estonian Language Resources. https://doi.org/doi.org/10.15155/3-00-0000-0000-0000-08565L

Ene Vainik (Eesti Keele Instituut) tunneb huvi semantika ja keele psühholoogiaga piirnevate aspektide vastu (nt tundesõnad, emotsioonide väljendamine kirjalikus ja suulises tekstis, piltlik keelekasutus, sõna-assotsiatsioonid). Hetkel tegeleb sõnaliigipiiride uurimisega leksikograafia vaatenurgast.
Roosikrantsi 6, 10119 Tallinn, Estonia
ene.vainik@eki.ee

Geda Paulsen (Eesti Keele Instituut, Uppsala Ülikool) tegeleb leksikaalse semantika, morfoloogia ja korpuslingvistikaga ning sõnaliigipiiride uurimisega leksikograafia vaatenurgast.
Roosikrantsi 6, 10119 Tallinn, Estonia
geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se

Ahti Lohk (Eesti Keele Instituut, Tallinna Tehnikaülikool) uurimisvaldkondadeks on wordneti semantiliste hierarhiate valideerimine graafipõhiste meetoditega ja tekstikaeve algoritmid kasuliku, uue ja rakendatava teabe ekstraheerimiseks struktureerimata tekstidest.
Roosikrantsi 6, 10119 Tallinn, Estonia
ahti.lohk@taltech.ee

Lisa 1. Testisõnade tabel

Tähised: MS = märksõna, SL = sõnaliik, D = adverb; + olemas, – puudu

Vorm	Ambivormid	Leksikoloogiline staatus ühendsõnastikus artikli kirjutamise ajal (23.9.2020)	Oletatav lemma
illatiiv	*algupoole*	+MS/–SL	*algupool*
illatiiv	*ankrusse*	–MS/–SL	*ankur*
adessiiv	*esmapilgul*	+MS/–SL	*esmapilk*
terminatiiv	*hingepõhjani*	+MS/+SL (D)	*hingepõhi*
komitatiiv	*hoolega*	+MS/+SL (D)	*hool*
abessiiv	*häireteta*	–MS/–SL	*häire*
terminatiiv	*juuksejuurteni*	+MS/–SL	*juuksejuur*
adessiiv	*juuresolekul*	+MS/–SL	*juuresolek*
adessiiv	*kaasabil*	+MS/–SL	*kaasabi*
komitatiiv	*kamaluga*	+MS/–SL	*kamal*
translatiiv	*kasuks*	+MS/–SL (ainult oskussõnastikes)	*kasu*
abessiiv	*katteta*	–MS/–SL	*kate*
essiiv	*komeedina*	+MS/–SL	*komeet*
komitatiiv	*kuhjaga*	+MS/–SL	*kuhi*
allatiiv	*kuradile*	–MS/–SL	*kurat*
inessiiv	*käpas*	+MS/–SL	*käpp*
elatiiv	*lambist*	+MS/+SL (D, kõnekeelne)	*lamp*
terminatiiv	*lõpmatuseni*	+MS/+SL (D)	*lõpmatus*
allatiiv	*lävele*	–MS/–SL	*lävi*
inessiiv	*mastaabis*	+MS/–SL	*mastaap*
elatiiv	*moest*	–MS/–SL	*mood*
inessiiv	*musklis*	+MS/–SL	*muskel*
translatiiv	*märgiks*	–MS/–SL	*märk*
komitatiiv	*omadega*	–MS/–SL	*oma*
inessiiv	*palavikus*	–MS/–SL	*palavik*
ablatiiv	*pardalt*	–MS/–SL	*parras*
allatiiv	*ravile*	–MS/–SL	*ravi*
essiiv	*reeglina*	+MS/–SL	*reegel*
illatiiv	*rivvi*	–MS/–SL	*rivi*
elatiiv	*sabast*	–MS/–SL	*saba*
ablatiiv	*seisukohalt*	+MS/–SL	*seisukoht*
ablatiiv	*sisult*	–MS/–SL	*sisu*
translatiiv	*sodiks*	+MS/–SL	*sodi*
terminatiiv	*surmatunnini*	+MS/–SL	*surmatund*
komitatiiv	*südamerahuga*	+MS/–SL	*südamerahu*
ablatiiv	*sünnilt*	+MS/–SL	*sünd*

Vorm	Ambivormid	Leksikoloogiline staatus ühendsõnastikus artikli kirjutamise ajal (23.9.2020)	Oletatav lemma
essiiv	*tervikuna*	+MS/–SL	*tervik*
abessiiv	*tingimusteta*	+MS/–SL (ainult oskussõnastikes)	*tingimus*
abessiiv	*traadita*	+MS/–SL	*traat*
elatiiv	*tuhandest*	+MS/–SL	*tuhat*
essiiv	*tulemusena*	+MS/–SL (seletuseta, tõlkevaste vene keeles)	*tulemus*
inessiiv	*tuules*	+MS/–SL	*tuul*
adessiiv	*veerel*	–MS/–SL	*veer*
terminatiiv	*võimatuseni*	+MS/–SL	*võimatus*
allatiiv	*äärele*	–MS/–SL	*äär*
essiiv	*üldreeglina*	+MS/–SL	üldreegel

Lisa 2. Tasakaalus korpusest võrdluseks valitud tavasõnad (rikkaliku vormistikuga sõnad)

Sõna	Eri vormide arv	Sõna sagedus (= lemma sagedus)
aasta	34	51581
aken	31	3295
auto	28	6117
jumal	25	3852
kartul	21	738
kass	18	1136
kast	14	482
keel	29	12442
kirik	29	3209
laps	29	16904
lehm	25	2633
muinasjutt	29	636
mäng	21	3968
naaber	27	1019
number	31	2047
perekond	28	2220
puu	32	2674
põlv	23	1084
pöösas	30	462
päike	28	2289
raamat	26	6000
rahvas	26	6283
unenägu	27	724
väärtus	32	5937
õpetaja	24	4480
ülikool	34	5263

FROM INFLECTED FORM TO A WORD: THE ROLE OF FREQUENCY

Ene Vainik[1], Geda Paulsen[1,2], Ahti Lohk[1,3]

Institute of the Estonian Language[1], Uppsala University[2], Tallinn University of Technology[3]

This study is motivated by the need for a statistical benchmark that would help the lexicographer to judge a morphological form for its grammaticalization stage to the degree of an independent lexeme. The focus of this article is on Estonian substantives and in particular their forms in the 11 semantic cases. The choice of selection is based on the observation that the noun has a special position among the word classes with fuzzy categorial borders (Vainik et al. 2020) – the means of nominal morphology function as a source for ongoing processes of grammaticalization in Estonian. The case forms typically yield adverbs and adpositions – a phenomenon forming a part of *ambiforms* (words or forms that can be interpreted to belong to more than one word class).

The main research question of this study is: is there a statistical sign indicating that a case form of a noun is emerging as a potentially independent lexeme? Based on the normal distribution of nominal case form frequencies, we established a statistic that determines a case form's elicitation in a corpus – the distribution index (D-index). The D-index can be used as an indicator of the correspondence of a particular form's actual frequency with the predicted elicitation degree.

The D-index was tested by a sample of ambiforms (N = 46) and "ordinary" nouns (N = 26), the last group including nouns that display an abundant range of semantic case forms. This sample was used to extract all semantic case forms from altogether three Estonian corpora: the Balanced Corpus of Estonian, the National Corpus of Estonian 2019, and the web corpus etTenTen13. Based on the analysis, we defined a threshold value (≥ 0,130), indicating that the forms with higher D-indexes than this value can be regarded as independent lexemes.

We conclude that the threshold value functions as a benchmark to a certain degree: an ambiform with a D-index over the threshold value is a distinctly independent lexeme. The forms with D-indexes below the threshold value may or may not be candidates of a lexical entry in a dictionary – the statistical parameters are not sufficient to make a waterproof decision. A lexicographer's qualitative analysis will be needed in those cases.

Keywords: lexicography, corpus linguistics, language technology, word form emancipation, parts of speech, Estonian