doi:10.5128/ERYa18.16

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 18, 279–302

KUIDAS ÄRA TUNDA ADJEKTIIVI? KORPUSKÄITUMISE MUSTRITE ANALÜÜS

Maria Tuulik, Ene Vainik, Geda Paulsen, Ahti Lohk

Ülevaade. Artiklis uurime adjektiivi morfosüntaktilisi tunnuseid ja selgitame, kuivõrd on prototüüpsele adjektiivile omistatavad tunnused (nt ühildumine, võrrete moodustamine) adjektiiviklassile eriomased. Loome neile tunnustele tuginedes parameetrid, mille abil eristame korpuse andmete põhjal adjektiive teistest sõnaliikidest. Tüüpilise adjektiivi korpusprofiili tuvastamise kaugem eesmärk on rakenduse loomine, mis võimaldaks leksikograafidel ebaselgete juhtumite puhul kontrollida sõna adjektiviseerumise astet. Tutvustame kuue parameetri testimise tulemust 12 sõnarühma peal, millest igaühte kuulub 10 sõna. Sõnavalikul arvestame adjektiiviklassi piiripealseid juhtumeid ja leksikograafilisi kitsaskohti. Analüüsime, mil määral hälbivad erinevad testrühmad testitud parameetrite põhjal prototüüpsest adjektiiviklassi esindajast ning vaatleme ka variatsiooni adjektiiviklassi sees. Kõrvalekaldeanalüüs võimaldab välja selgitada parima eristusvõimega parameetrid. Eukleidilise kauguse mõõtmine eristab hästi adjektiivisarnased sõnad ja rühmad nendest, mis sarnanevad prototüüpsele adjektiivile vähem.*

Võtmesõnad: sõnaliigid, morfosüntaks, leksikograafia, keeletehnoloogia, eesti keel

1. Sissejuhatus

Sõnaraamatute koostamisel on üks lahendamist nõudvaid küsimusi kahtlemata märksõnade sõnaliigiline mitmesus[1] ehk võimalus tõlgendada sõna või sõnavormi rohkema kui ühe sõnaklassi liikmena (Karelson 2005, Habicht jt 2011). Eesti leksikograafide uuring (Paulsen jt 2019) osutas, et sõnastikes, kus sõnaklassi määratlemine on vajalik (nagu näiteks eesti keele ühendsõnastikus), leidub hulk sõnu, mille piiritlemine on keerukas. Nende analüüsiks soovisid uuringus osalenud leksikograafid toetuda võimalikult laiale andmehulgale ning andmestiku liigitamiseks

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

[1] Korpusleksikograafias on sõnaliigituse probleemid tihedalt seotud korpuse märgendamisega, täpsemalt morfoloogilise ühestamisega. Sõnaliigilise mitmesusega seotud märgendusprobleeme korpuste automaatanalüüsis on käsitlenud Külli Habicht jt (2000), Kadri Muischnek ja Kadri Vider (2005), Liina Lindström jt (2006), Kristina Koppel (2020).

abiks korpuspõhist rakendust, mis tooks kokku asjakohase eeltöödeldud info. (*ibid.* 2019) Ka Margit Langemets jt (2021: 760) on leidnud, et tegelikku keelekasutust tekstikorpuse põhjal analüüsides annab automaatne eeltöötlus sõnaraamatutegijale sisukama ning ka objektiivsema ülevaate sõna käitumisest; andmete hindamisel ja kokkuvõtete tegemisel jääb lõplik kvalitatiivne hinnang mõistagi leksikograafide teha.

Uuringus (Paulsen jt 2020) osalenud leksikograafid tõid näiteid sõnaliigitusprobleemidest. Tervelt veerandi (26%) problemaatilistest juhtudest moodustasid adjektiiviklassiga seotud sõnad. Erilise murekohana toodi välja küsimus, kuidas teha kindlaks, kas verbi partitsiibil (nt *austatud*) on juba piisavalt adjektiivset kasutust, et oleks õigustatud sõna lisamine sõnastikku/andmebaasi ka adjektiivina[2]. Eesti keele mitmese sõnaliigimärgendiga vormide (nn ambivormide) lähem uuring (Vainik jt 2020) näitas, et adjektiiv võib anda piirnevaid juhtumeid verbidega (verbi käändelised vormid nagu mineviku ja oleviku partitsiibid, nt *armunud*, *joobnud*, *surnud*, *austatud*, *suletud*; *siduv*, *lööv*, *hävitav*, ning supiinid, nt *rääkimata*, *värvimata*); nimisõnadega (nt *vaimulik*, *rase*, *loll*, *pull*, *räbal*), pronoomenitega (nn proadjektiivid nagu *niisugune*, *samane*, *teistsugune*) ja adverbidega (nt *kiivas*, *krussis*, *laokil*, *purjus*, *lömmis*). Adjektiivide piirnemist nende sõnaliikidega on käsitletud ka varasemalt: adjektiivi ja nimisõna vahekorda on uurinud Silvi Vare (2006), adjektiivi ja adverbi kokkupuuteid Mai Tiits (1982) ning adjektiivi ja verbivormide kattuvaid alasid Krista Kerge (1998) ja Elo Allemann (2002).

Siinses artiklis keskendume küsimusele, kas kvantitatiivsetele korpusandmetele tuginedes on võimalik luua automaatset sõna(vormi) tekstis käitumise hindajat, mis toetaks leksikograafi sõnade liigitamisel. Uurimuse fookuses on adjektiiv ja adjektiiviga mingis aspektis kattuvad sõnaliigid. Küsime, kas ja mil määral iseloomustavad teoreetilistes käsitlustes esitatud adjektiivi tunnused prototüüpsete adjektiivide käitumist tekstides. Kuidas hinnata kvantitatiivselt sõna kõrvalekaldumist prototüüpse adjektiivi profiilist? Teisisõnu uurime, kas on võimalik korpusmaterjalis ilmnevate morfosüntaktiliste mustrite abil automaatselt määratleda sõnaliigilist kuuluvust. Selleks määratleme korpuseotsingu skeemid, mis eeldatavasti peegeldavad adjektiividele omaseid lausestruktuure, kasutades ülalt-alla (*top-down*) lähenemisviisi. Püüame niisiis kvantitatiivselt jäljendada otsustusi, mida eksperdid (keeleteadlased ja leksikograafid) sõnade klassikuuluvuse kohta teeksid. Uurimuse kaugem eesmärk on arendada automaatne korpusrakendus (vt Paulsen jt 2019: 332–333; Vainik jt 2021: 136), mis kiirendaks leksikograafi tööprotsessi sõnaliigimitmesustega tegeledes (näiteks hinnates, kas partitsiipvormist on kujunenud või kujunemas adjektiiv).

2. Prototüüpse adjektiivi tunnused

1970. aastatel tekkis kognitiivses psühholoogias prototüübiteooria (Rosch 1978) näol alternatiiv klassikalisele selgepiiriliste kategooriatega opereerivale lähenemisviisile. Kui klassikalist kategooriat iseloomustavad oma olemuselt binaarsed (+/–) olulised (ingl *necessary*) ja piisavad (*sufficient*) tunnused, saab prototüübipõhist kategooriat kirjeldada kontiinumina. Prototüübiteooriat saab rakendada ka

[2] Tänapäeva eesti keelega tegelevad leksikograafid töötavad eesti keele ühendkorpusega (artikli kirjutamise ajal on värskeim versioon aastast 2019, Kallas ja Koppel 2020). Korpuse on morfoanalüsaatori EstNLTK 1.6 abil lemmatiseerinud, märgendanud ja ühestanud Lexical Computing Ltd. Siinkohal on oluline märkida, et korpuseliideses on kõik oleviku partitsiibid (*v*-kesksõnad) automaatselt märgendatud adjektiivideks ning leksikograafil tuleb (sageli massiliselt) liigendamata kasutusnäited omal käel läbi analüüsida.

keeleteaduses. Sellisel juhul ei ole näiteks kõik samasse sõnaliiki kuuluvad sõnad oma klassi samavõrd head esindajad – on tüüpilisemaid ehk tsentrisse kuuluvaid ning ebatüüpilisemaid, mida saab teatud tingimustel liigitada ka mõne teise sõnaliigi alla (sõnaliigipiiride ebamäärasusest vt nt Erelt 1977). Prototüüpseteks adjektiivideks võime pidada näiteks sõnu *suur*, *sinine*, *kuri* ja vähem prototüüpseteks käändumatuid (*lontis*, *eht*) või partitsiipseid adjektiive (*huvitav* ja *täissöönud*). Juba toona tõdeti siiski, et keelekirjeldus ei saa läbi tunnusteta, mille põhjal sõna kas vastab tüüpilise sõnaliigi esindaja tunnustele või mitte. Tunnuste osas oletati mingit laadi hierarhia kehtimist: teatud tunnus või tunnuste kimp on olemuslikult olulisem kui teine. (Viks 1977)

Tänapäeval on kättesaadavad suuremahulised tekstikorpused ning võimalikuks on saanud keelendi käitumise jälgimine tegelikes kasutuskontekstides paljude kasutuskordade lõikes. Tunnuste esinemist binaarsetena (ühe või teise variandi olemasolu tuvastades) või kontiinumina (ühe või teise variandi osakaaludena) saab hinnata automaatanalüüsi abil loodavate statistiliste ülevaadete kaudu. Selles uurimuses järgime prototüübipõhise kategoriseerimise põhimõtteid ning oletame, et adjektiiviklassi prototüüpsed esindajad peaks jagama kõiki või enamikku adjektiividele iseloomulikuks peetavaid tunnuseid, ebatüüpilised adjektiiviklassi liikmed vaid mõnda neist. Järgnevas lahkamegi adjektiivile omaseid tunnusjooni, millest osa kasutame korpusmustrite uuringus.

Nii nagu teisigi sõnaliike, iseloomustab adjektiivi lihtsustatult[3] kolme tüüpi tunnuste kimp: morfoloogilised, süntaktilised ja semantilised omadused, millele võib lisanduda iseloomulik pragmaatiline käitumine (Erelt 2017: 58). Adjektiivi morfoloogilised tunnused on muutumine käändes ja arvus ning võimalus moodustada võrdlusastmeid. Siia kuuluvad ka adjektiive moodustavad tuletusliited (nt *-ne*, *-line*, *-lik*, *-kas*, *-jas*, *-tu*/*-matu*, *-s*/*-as*/*-as*/*-us*/*-es*/*-is*, *-mine* jne[4] (Kasik 2015: 348–367)).

Adjektiivi põhiline semantiline ülesanne on viidata omadusele (nt *punane vihmavari*). Omadused on paljudel juhtudel skalaarsed – neid saab hindaja meelest olla tavalisest rohkem –, sestap alluvad omadussõnad tüüpiliselt võrdlemisele, mis vormi osas realiseerub komparatiivi ja superlatiivi vormide võimalikkusena. Suhtelise (relatiivse) või absoluutse (mitteskalaarse) omaduse väljendusvõimalusest sõltub omakorda adjektiivifraasi struktuur – skalaarseid ehk gradeeritavaid omadussõnu võivad laiendada intensiivsusastet väljendavad määrsõnad (*kohutavalt ilus*, vrd **kohutavalt tõeline*) ning nagu ülal mainitud, skalaarsetest omadussõnadest saab moodustada võrdlusastmeid (*ilusam*, vrd ?*maavillasem*). Ka mitteskalaarsed adjektiivid võivad saada määrsõnalaiendi omaduse olemasolu või seisundis oleku täielikkuse astet väljendavate adverbide näol (*absoluutselt*/*peaaegu tõeline*). (Erelt 2017: 406–408)

Suhtelist omadust väljendavate adjektiivide skalaarsusest tuleneb nende "hüüupotentsiaal" ehk osalemine lausungites pragmaatilisi funktsioone kandvate hüüufraasidena (*Õudne! Oivaline!*) (vt Erelt 1986: 113, Erelt 2017: 533, Paradis 2001). Sellistes kõneleja emotsiooni ja hinnangut vahendavates hüüdlausetes esinevad adjektiivid väljendavad eeskätt vastandust omaduse keskmise normiga, millest kõrvalekaldumine on ebaootuspärane ning kutsub seetõttu kõnelejas esile

[3] Lihtsustatult selles mõttes, et reaalselt esinevad eri keeletasanditel kirjeldatavad protsessid koostoimes.
[4] Reet Kasik (2015: 367–368) käsitleb partitsiipvormilisi omadussõnu tuletistena, eristades *v*-omadussõnaliidet, mis hõlmab põhiliselt *v*-kesksõnu; *v*-lõpulised võivad olla ka nt nimisõnast tuletatud sõnad, nagu *verev*, *terav*.

üllatusreaktsiooni (Tarabarova 2014: 17–18). Adjektiivsed hüüdlaused on üks peamisi vaegseid predikaadita hüüdlausekonstruktsioone (*Nii vahva!*) (*ibid*. 99).

Adjektiivifraasi võib moodustada omadussõna üksi või koos oma laienditega. Lauses esineb adjektiivifraas täiendi (*hea laps*), öeldistäite (*laps on hea*) või predikatiivadverbiaali (*nali mõjus kohatuna*) funktsioonis. Täiendi rollis ühildub adjektiiv substantiiviga arvus ja käändes. Omadussõnafraasi moodustab ka järgarvsõna (*viies laps*) või aseomadussõna (*selline laps*), kuid neis fraasides omadussõnana käituvatel sõnadel on juba teistsugune või pleekunum tähendus. Adjektiivi võib laiendada substantiivifraas (*kokakunstis vilunud*), kaassõnafraas (*lapse vastu tõre*), infinitiivifraas (*kiire lugema*), adverbi funktsioonis adjektiivifraas (*hirmus pisike*), kvantorifraas (*kahe kilo raskune*), adverbifraas (*eriti raske*) ja kõrvallause (*see oli raskem, kui ma arvanud olin*). (Erelt 2017: 63, 405)

3. Uurimuse valim ja meetod

Testsõnade ja parameetrite valikul pidasime silmas oma eesmärki – katsetada adjektiivile omaseks peetavaid morfosüntaktilisi jooni võrdluses teiste sõnaklassidega ning testida korpuseandmetele tuginedes, kas ja kuivõrd eristub adjektiivi käitumine tegelikkuses. Seetõttu sisaldavad testsõnad nii prototüüpseid adjektiive, adjektiiviklassiga piirnevaid rühmi kui ka teisi sõnaliike. Võrdlus teiste sõnaliikidega teenib ka meie kaugemat eesmärki luua digitaalne tööriist, mis kasutades ettemääratud parameetreid näitab sõna kaugust sõnaliigi prototüüpsest esindajast. Uurimuses kasutatud parameetrid on arendatud välja adjektiivi prototüüpsete tunnuste põhjal.

3.1. Testsõnade valik

Adjektiivi morfosüntaktiliste korpusmustrite kindlakstegemiseks analüüsisime 12 rühma, millest igaühte kuulus 10 sõna (vt lisa 1). Rühmad ja sõnad valisime välja käsitsi, võttes arvesse peamisi leksikograafide murekohti (vt Paulsen jt 2020: 187–188): adjektiivi tõlgendusvõimalust a) partitsiibina, b) substantiivina ja c) adverbina. Kõik testsõnad kuuluvad ka EKI ühendsõnastiku (2021) märksõnastikku (välja arvatud rühm "verbi partitsiibid", mis kuuluvad sõnastikus vastava verbi paradigmasse).

Testsõnad võib jagada suures plaanis kaheks: adjektiiviga piirnevate sõnade rühm (sihtrühm) ning kontrollrühm. Sihtrühma valimis on esindatud kuus alarühma: kõik on sõnad, mida võib suurema või vähema mööndusega nimetada adjektiiviks. Tunnuste eristusjõu kontrolliks testisime ka esindajad muudest sõnaliikidest (samuti kuus rühma). Sihtrühmad ja kontrollrühmad on järgmised.

Sihtrühm:

1. adjektiivid I
2. adjektiivid II
3. adjektiivid/partitsiibid
4. adjektiivid/substanstiivid
5. adjektiivid/adverbid
6. käändumatud adjektiivid

Kontrollrühm:

1. partitsiibid
2. substantiivid
3. adverbid
4. verbid
5. proadjektiivid
6. ordinaalid

Sihtrühm püüab välja selgitada adjektiivseid omadusi ehk adjektiiviklassi enda statistilist varieerumist, kontrollrühm kaardistab sõnaliikidevahelisi erisusi. Adjektiivide I rühma sõnad on korpuses eeldatavasti peaaegu täielikult kasutuses adjektiividena (nt *tõeline*, *rahvusvaheline*, *harv*). Adjektiivide II rühma sõnad on peamiselt kasutusel adjektiivina, ent leidub ka pisut kasutust teistes sõnaliikides, näiteks verbipartitsiibina (nt *toetav*, *särav*) või substantiivina (nt *külm*, *uudishimulik*, *hea*), mis on adjektiividele väga iseloomulik[5]. Omaette rühma moodustavad siiski adjektiivid/substantiivid, mille puhul eeldame substantiivset kasutust juba suuremal määral (nt *teismeline*, *kodutu*, *võõras*).

Kuna adjektiivide käitumine korpuses võib varieeruda vastavalt sellele, kas tegemist on adjektiivide/substantiivide, käändumatute adjektiivide (nt *katoliku*, *eht*, *lugupeetud*), adjektiivide/adverbide (nt *krussis*, *tohutu*, *alasti*) või adjektiivide/partitsiipidega (nt *tulev*, *kehtiv*, *kurnatud*), siis analüüsime nimetatud rühmi eraldi. Erilise tähelepanu all on tulemuste analüüsis rühmad adjektiivid/partitsiibid ja (adjektiviseerumata) partitsiibid, mille eristamises leksikograafid abi vajavad. Täpsemalt soovivad leksikograafid teada, kui suurel määral on partitsiip korpuses adjektiivsena kasutusel, nt *algaval aastal on* (adjektiivne kasutus) *vs*. *varakult algav* (verbiline kasutus). Võrdleme partitsiipide ja adjektiivide näite igas tulemuste osas eksplitsiitselt. (Edaspidi kasutame terminit *verbipartitsiip* adjektiviseerumata partitsiipide sünonüümina.)

Et tabada erinevusi sõnaliikide vahel, kaasasime valimisse järgmised rühmad: partitsiibid (nt *hakanud*, *pandud*, *tallav*), substantiivid (nt *kass*, *padi*, *sülearvuti*), adverbid (*selili*, *võõrsil*, *sõbralikult*), verbid (nt *jooksma*, *vestlema*, *itsitama*), proadjektiivid (nt *niisugune*, *sama*, *milline*) ja ordinaalid (nt *esimene*). Arvestasime rühma sõnavalikus ka sõnasagedusi – esindatud on nii suure kui ka väikese sagedusega sõnad.

3.2. Parameetrid

Testimiseks valitud parameetrid tuginevad nii adjektiivi prototüüpsetele tunnustele kui ka praktilistele tähelepanekutele leksikograafilises töös. Lisaks pidime arvesse võtma, et parameetrid oleksid olemasolevate märgenduste toel korpusest ekstraheeritavad ja eritleksid just sõnastikutöös kriitiliseks osutunud kohti, nagu adjektiivi piir verbipartitsiibiga. Esimesed neli parameetrit põhinevad teaduskirjanduses adjektiivile omaseks peetavatel tunnustel, viies parameeter on saanud tõuke leksikograafilises töös otstarbekaks osutunud analüüsimeetodist ning kuues on selgitatud välja korpusmaterjali analüüsil. Parameetrid on järgmised:

[5] Adjektiividele omast süsteemset nimisõnadegaga toimuvat sõnaliigivahetust analüüsib Vare (2006: 199) sõnaliigilise transpositsiooni ehk sõnaliigimuutusena, mille puhul sõna liigub teise sõnaliiki nii, et lähtesõna säilib endisel kujul (*haige laps – kaks haiget*).

1) **Atribuudiparameeter** (testsõna+substantiiv ehk test_S) tugineb adjektiivi kalduvusele esineda lauses substantiivi atribuudina. Eeldame, et üks sagedamaid adjektiivi kollokatsioonimalle on adjektiiv+substantiiv (nt *kohevad juuksed*, *kulla lapsele*), seega selgitab atribuudiparameeter, kas ja mil määral järgneb testsõnale substantiiv.
2) **Ühildumisparameeter** lähtub adjektiivide omadusest ühilduda substantiiviga arvus ja käändes ning kontrollib testsõna ühildumist, kui sellele järgneb substantiiv (test_S_ÜHILD). Ühildumiseks lugesime juhud, kus testsõna ja sellele järgnev substantiiv olid korpuses märgendatud sama arvu ja käändega (nt *headele lastele*).
3) **Võrdeparameeter.** Et adjektiivi iseloomustab võrdluskategooria, siis selle parameetriga kontrollisime testsõnast tekitatud võrdevormi olemasolu ja osakaalu testsõna enda sagedusega võrreldes. Valisime võrdluskategooria esindajaks keskvõrde, kuna ülivõrde moodustamine ja korpusest leidmine on kahe erineva kuju (*kõige* + keskvõrre ja lühike ülivõrre) tõttu keerulisem. Lisaks ei esine ülivõrre üksi: kui sõnal on olemas ülivõrre, siis esineb ka keskvõrre. Keskvõrdekatse läbiviimiseks tekitasime kunstlikult keskvõrde vormid kõigist valimi sõnadest, k.a nimisõnad, arvsõnad jm (**kassim*, **viiendam*, **selilim*) ning ka käändumatutest adjektiividest (**venem*, **katolikum* jne), millel võrdlusastmed üldiselt puuduvad. Sealjuures järgisime (kus võimalik) keskvõrde reeglit: ainsuse omastava vorm + *m*. Kontrollisime konstrueeritud keskvõrrete esinemist ühendkorpuses korpuspäringusüsteemi Sketch Engine'i abil (Kilgarriff jt 2014).
4) **Adverbiparameeter** (adverb+testsõna ehk D_test) mõõdab, kas ja kui sageli esineb korpuses testsõna ees adverb. Kuna nii skalaarsed kui ka mitteskalaarsed adjektiivid võivad saada määrsõnalaiendeid (vt ptk 2), on adverb+adjektiiv (nt *väga külm*, *äärmiselt kahetsusväärne*) oletatavasti samuti adjektiividele iseloomulik korpussuhe.
5) **Lausealguseparameeter** (testsõna+substantiiv (LA_test_S) selgitab, kas ja kui sageli alustab testsõna lauset täiendi positsioonis substantiivi ees. Parameetri üks eesmärk on eristada adjektiviseerunud partitsiipe adjektiviseerumata verbivormidest (näiteks lause algus "Tuleval suvel…" oleks igati loomulik ja normipärane, kuid "Oleval suvel…" mitte). Parameeter kasvas välja sõnaraamatu "Eesti keele naabersõnad 2019" koostamisel kasutusel olnud võttest, mis aitas automaatselt loodud andmebaasis puhastada adjektiivi-substantiivi kollokatsioonipaaris adjektiivide seast välja verbivorme (Kallas jt 2015).
6) **Neljakohaline parameeter**, mis selgitab, kui sageli esineb testsõna mustris verb+määramata sõna+testsõna+substantiiv (V_?_test_S). See neljakohaline järjend pärineb käesolevale uuringule eelnenud pilootuuringust, kus testisime erinevaid parameetreid morfoloogiliselt ühestatud korpuses[6] ning just see parameeter osutus adjektiviseerunud ja adjektiviseerumata partitsiipide parimaks eristajaks. Adjektiviseerunud partitsiipe esines mustris V_?_test_S adjektiviseerumata partitsiipidest sagedamini (nt *on lihtsalt <u>hõivatud</u> mees*, *leidus palju <u>tuntud</u> hitte*, *jäägu punased <u>punutud</u> pastlapaelad*). Kui teised parameetrid põhinevad adjektiivi keeleteaduslikult

[6] Morfoloogiliselt ühestatud korpus koosneb käsitsi ühestatud tekstidest ja on kättesaadav aadressil https://www.cl.ut.ee/korpused/morfliides/ (28.4.2022).

määratletud prototüüpsetel omadustel, siis neljakohaline parameeter on n-ö alt-üles (*bottom-up*) lähenemisviisi tulemus, aluseks korpusmaterjali analüüs ja kontekstimustrite sõelumine.

Üks adjektiivi iseloomulikke lausefunktsioone on ka võimalus esineda öeldistäitena. Praegusesse uurimusse me öeldistäite morfosüntaktilise mustri testimist ei kaasanud, sest korpuseotsing *olema*-verbi vormide kaudu tooks kaasa verbi liitajavormide tulemused (*on käidud*) ning hägustaks adjektiviseerunud ja adjektiviseerumata verbivormide üldpilti. Öeldistäite morfosüntaktilise mustri lisamist adjektiivi tunnusparameetrite hulka tasub järgmistes uurimustes kindlasti kaaluda, kui õnnestub leida viise eristada adjektiviseerunud partitsiibid verbi liitaegade vormidest.

3.3. Testmustrite pärimine korpusest

Uurimuses testisime kuut parameetrit eesti keele ühendkorpuses 2019, mis sisaldab 1,5 miljardit tekstisõna ning on suurim eestikeelsete tekstide korpus. Kõikide parameetrite testimiseks koostasime programmeerimiskeeles Python loogilised avaldised (vt tabel 1) ning eraldasime nende abil korpusest testsõna esinemisjuhud, mil konkreetne parameeter (muster) kehtis. Testsõna enda sõnaliik oli korpuspäringul alati määramata, vaba.

Tabel 1. Parameetrite korpusest eraldamise avaldised

Jrk	Nimetus	Tähis	Loogiline avaldis
1	atribuudiparameeter	test_S	lemmas[i].lower() in test_words and postags[i+1] == 'S'
2	ühildumisparameeter	test_S_ÜHILD	lemmas[i].lower() in test_words and postags[i+1] == 'S' and forms[i] == forms[i+1]
3	keskvõrdeparameeter	keskvõrre	lemmas[i] in test_comparatives
4	adverbiparameeter	D_test	postags[i] == 'D' and lemmas[i+1].lower() in test_words
5	lause alguse parameeter	LA_test_S	i == 0 and lemmas[i].lower() in test_words and postags[i+1] == 'S'
6	neljakohaline parameeter	V_?_test_S	i < sent_len - 3 and postags[i] == "V" and lemmas[i+2] in test_words and postags[i+3] == "S"

Testmustritele vastavate isendite[7] ekstraheerimisel rakendasime EstNLTK korpuse töötlemise vahendeid[8]. Seejuures arvestasime lausete piiridega, st ükski testmustri isend ei ületanud lause ega osalause piire. Iga lause analüüsis kasutasime EstNLTK teavet selle tekstisõnade, tekstisõnade lemmade ning vormide kohta. Lemmasid läks vaja üksnes testsõnade tuvastamiseks ja sõnavormide teavet testmustri "test_S_ÜHILD" isendite kindlakstegemiseks. Tekstisõnad olid olulised

[7] *Isend* on objektorienteeritud lähenemisviisis (infotehnoloogias) klassi esindaja või klassi kuuluv eksemplar, meie uurimuses mustrile vastav konkreetne esinemisjuhtum.

[8] Vaata täpsemalt: https://nbviewer.org/github/estnltk/estnltk/tree/version_1.6/tutorials/corpus_processing/ (28.4.2022).

testmustritele vastavate isendite teksti fraaside eraldamisel. Kõik testmustrite isendid on ekstraheeritud üksteisest sõltumatutena: eraldi ei peetud arvestust juhtudel, mil üks testmustri isend sisaldab teist. Näiteks test_S sisaldub nii mustris test_S_ÜHILD, LA_test_S kui ka V_?_test_S, ometi loendatatakse neid kõiki üksteisest sõltumatult[9].

Pärast seda, kui korpusest olid päritud testsõnade üldsagedused (lemma sagedused) ning esinemisjuhtumite arv parameetritega määratud tingimustel, arvutasime parameetrile vastamise osakaalud iga testsõna jaoks, mis ongi meie põhiline statistilise analüüsi üksus.

Automaatanalüüsi tulemused sõltuvad korpuse märgendamise kvaliteedist ja arvestada tuleb vigadega nii sõnaliigi määramisel (millele eelneb morfoloogiline analüüs) kui ka lemmatiseerimisel. Sagedusandmete tulemusi mõjutab morfoloogilise ühestamise kvaliteet. EstNLTK 1.6 kombineerib reeglipäraseid ja statistilisi meetodeid ning toetub oma analüüsis Vabamorfi leksikonile ja morfoanalüsaatorile, mille täpsus on (vähemalt) 97% (Kaalep, Vaino 2001).

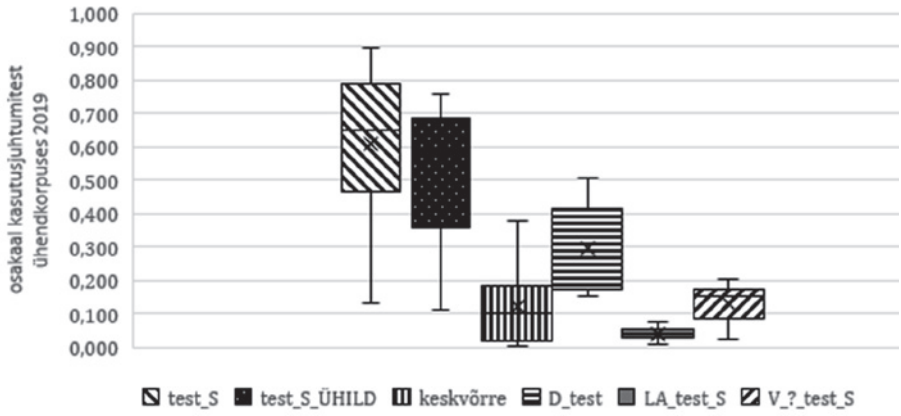
4. Tulemused

Esitame analüüsitulemused vastavalt kasutatud meetodile. Esiteks moodustasime parameetrite põhjal prototüüpse adjektiivi korpuskäitumise profiili. Et saada teada erinevate sõnarühmade üldisi parameetrite esinemise proportsioone, kasutasime meetodina keskmistamist ning analüüsisime väärtuste varieerumist parameetrite lõikes. Selgitamaks välja, kuivõrd parameetrid suudavad adjektiive teistest sõnaliikidest eristada ja seda, mil moel erinevad omavahel eri tüüpi adjektiivid, viisime läbi kõrvalekaldeanalüüsi, mille käigus tegime kindlaks, kas ja mil määral hälbisid teised testrühmad prototüüpse adjektiivi iseloomulikest väärtustest. Mõõtmaks, kuidas suhestub iga uuritav üksiksõna prototüüpse adjektiivi tervikprofiiliga (st kokkuvõtlikult näidud kõigil parameetritel), rakendasime eukleidiliste kauguste arvestamise metoodikat.

4.1. Prototüüpse adjektiivi korpuskäitumise profiil

Esmalt selgitasime välja, kuidas kirjeldavad parameetrid prototüüpseid omadussõnu (rühm adjektiivid I). Kümne sõna andmeid (keskmisi, mediaane ning varieeruvuse piire) näitab joonis 1, arvandmed on esitatud tabelis 3. Jooniselt on näha, et osa parameetreid on kõrgema keskmise väärtusega ja osa madalamaga, samuti on näha, kui suurtes piirides väärtused parameetri lõikes varieeruvad.

[9] Täpne programmikood on allalaetav: https://github.com/ahtilohk/PSG227/blob/main/Adjective_patterns_occurrences_in_ENC2019.py (28.4.2022).



Joonis 1. Parameetrite väärtused prototüüpsete adjektiivide rühma (adjektiivid I) puhul

Tabel 3. Prototüüpset adjektiivirühma kirjeldav statistika

Näitaja	Test_S	Test_S_ühild	Keskvõrre	D_test	LA_test_S	V_?_test_S
keskväärtus	0,608	0,519	0,120	0,296	0,041	0,131
standardhälve	0,222	0,198	0,112	0,120	0,020	0,055
mediaan	0,651	0,585	0,102	0,269	0,039	0,151
miinimumväärtus	0,133	0,113	0,003	0,154	0,008	0,025
maksimumväärtus	0,896	0,760	0,377	0,507	0,078	0,205

Atribuudiparameeter (test_S) näitab, et keskmiselt 60% kasutusjuhtumitest esineb prototüüpne adjektiiv vahetult substantiivi ees, st klassikalise eestäiendina. Ülejäänud kasutuskordadel paigutub ta kuidagi teisiti. Ühildumisparameeter (test_S_ÜHILD) osutab, et prototüüpsele adjektiivile on omane, et keskmiselt pooltel kasutusjuhtudest on ta kasutatud nimisõna käändes (ja arvus) ühilduva eestäiendina. Teine pool adjektiivi kasutusjuhtudest esineb järelikult kas teistsugustes konstruktsioonides (nt öeldistäitena) või ei ühildu mõõdetavalt (nt *suure jutuga* – ühildumine on ainult arvus ja mitte käändes). Madalamate näitudega otsas on mõlema parameetri osas kaks sõna: *napisõnaline* ja *kahetsusväärne*, mille puhul on osakaalud vastavalt 0,3 ja 0,1.

Adverbiparameeter (D_test) näitab, et prototüüpsetel adjektiividel on esinemist koos vahetult eelneva adverbiga (nt *väga*, *eriti*) keskmiselt kolmandikul sõna kasutusjuhtudest.

Madalamaid väärtusi joonisel 2 näitavad esiteks komplekssed parameetrid, mis kombineerivad kolme ja nelja muutuja koosesinemist (LA_test_S ja V_?_test_S) ning madal keskmine väärtus on ka keskvõrdetunnusel. Nimelt osutub, et kümnest prototüüpseks peetud omadussõnast vaid paarist kasutatakse komparatiivivorme arvestataval määral (üle 20% kasutuskordadest), nt *pikk*, *tõsine*; kahel sõnal – *rahvusvaheline* ning *tõeline* – on võrrete osakaal väga madal, mis on vägagi mõistetav nende semantikat arvestades.

Prototüüpsete adjektiivide rühma parameetrite osakaalud (varieerumise vahemikud ja mediaan, mis esitatud tabelis 3) on orientiiriks, millega edasises analüüsis võrreldakse teiste sõnarühmade käitumist.

4.2. Koondanalüüs

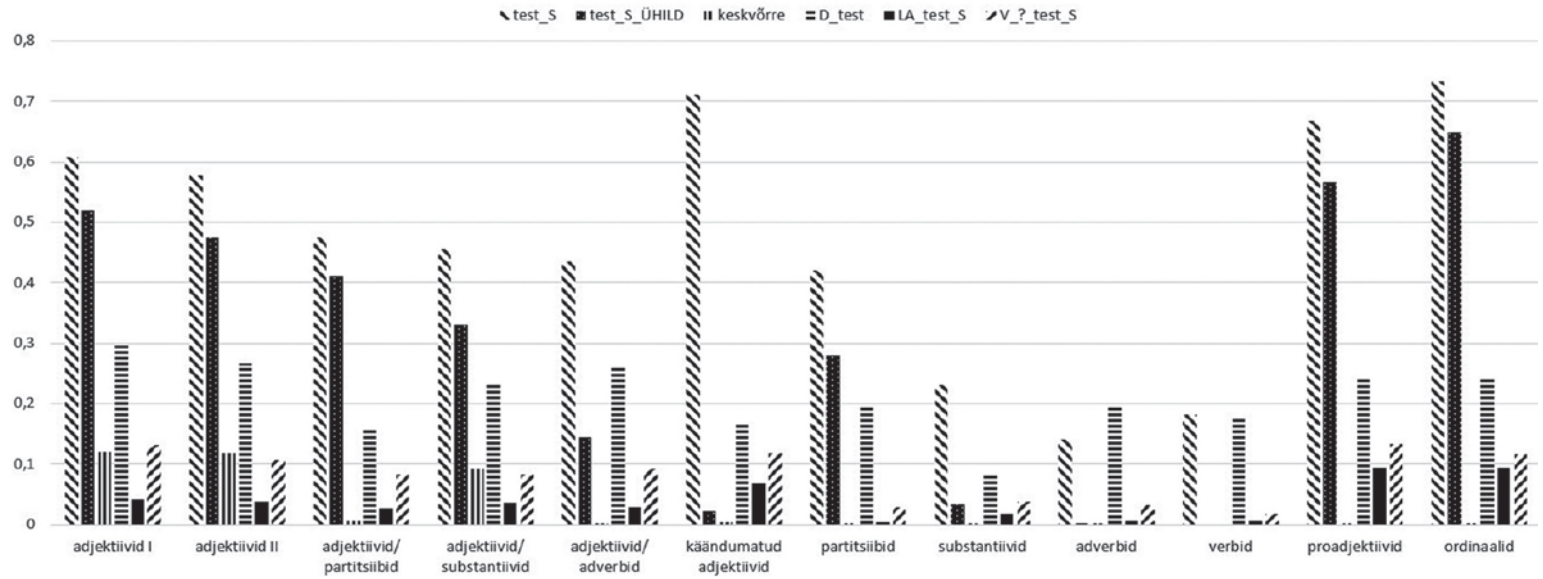
Joonis 2 võtab kokku kõigi parameetrite testimise keskmised tulemused testrühmade kaupa. Analüüsime iga testitud parameetri tulemusi eraldi ning lahkame nii sõnarühmade kui ka mõningate üksiksõnade juures väärtuste varieerumist, mis mõjutab rühma aritmeetilist keskmist.

Atribuudiparameeter (test_S) mõõdab kõige kõrgemaid tulemusi ordinaalide rühmas (nt sõnadel *kaheksas* ja *üheksas*) ja käändumatute adjektiivide rühmas (nt sõnadel *soome-ugri*, *luteri* ja *eri*, mis esinevad kollokatsioonides nagu *soome-ugri rahvad*, *luteri kirik* ja *eri riikides*). Kõrgem on see näitaja ka proadjektiivide ja prototüüpsete adjektiivide rühmas, samas madalama tulemusega eristuvad selgelt substantiivide, adverbide ja verbide rühm. Sellegipoolest ei võimalda atribuudiparameeter adjektiive kõigist teistest sõnaliikidest täpselt eristada, kuna on selleks liiga üldiselt esinev tunnus. Partitsiipide rühma tulemus sarnaneb adjektiivide/adverbide ja adjektiivide/substantiivide tulemusega ning ka adjektiivirühmade enda vahel on varieeruvust. Prototüüpsetest adjektiividest kõrgemal paiknevad käändumatud adjektiivid ja ordinaalide rühm, mis käitubki selles aspektis adjektiividega sarnaselt.

Ühildumisparameeter (test_S_ÜHILD) näitab selgeid erinevusi sõnarühmade vahel. Parameeter piiritleb uuritud sõnarühmadest ordinaalide ja proadjektiivide rühma, mis näitavad kõrget keskmist väärtust (∼0,6) ja väikest varieeruvust. Keskmiselt kõrgema väärtusega on ka rühmad adjektiivid I, adjektiivid II, adjektiivid/verbivormid ja adjektiivid/substantiivid. Selgelt nullväärtusega on adverbide ja verbide rühmad. Madal keskmine näit iseloomustab substantiive (kuna otsisime ühildumist testsõnale järgneva substantiiviga), käändumatuid adjektiive ja adjektiive/adverbe.

Ühildumisparameeter eristab sõnade käitumist järgmistes rühmades: adjektiivid/verbivormid, adjektiivid/adverbid ja verbi partitsiibid. Rühmas adjektiivid/verbivormid on kolm sõna – *armunud*, *kurnatud* ja *läinud* –, mis seda tunnust ei jaga; rühmas adjektiivid/adverbid on sõnad *meeletu* ja *tohutu*, mis seda tunnust jagavad prototüüpsete adjektiividega samal määral; sõna *täis* mõnevõrra vähem. Ülejäänuid iseloomustab adverbile omane 0-väärtus sellel parameetril. Kaheks jaguneb ka verbi partitsiipide rühm: adjektiivilaadseid väärtusi omavad *koosnev*, *olev*, *tallav*, *toimuv* ja veidi vähem *nakatuv*; samas ükski *nud*- ja *tud*-partitsiipidest ühildumist üles ei näidanud.

Kokkuvõttes eristab ühildumisparameeter hästi kaheks sõnad (ja rühmad), mis jagavad seda tunnust adjektiividega samal määral ning need, mis ei jaga üldse, ning jagab kahte lehte sõnad heterogeensetes rühmades. Parameetri esinemine vaatlusaluste verbivormide ja partitsiipide rühmas näitas täielikku morfoloogilise tunnuse (kas *-v*, *-tav* või *-nud*, *-tud*) järgimist, mis lubab järeldada, et kui partitsiipe kasutatakse adjektiividega sarnastes positsioonides/funktsioonides, siis *v*- ja *tav*-vormid järgivad prototüüpse adjektiivi eeskuju ning *nud*- ja *tud*-kesksõnad atüüpilise nn käändumatu omadussõna eeskuju.



Joonis 2. Testrühmade keskmised tulemused parameetrite lõikes

Keskvõrdeparameetri tulemus oli kõrgeim kolmes rühmas (adjektiivid I, adjektiivid II ja adjektiivid/substantiivid). Võrdevormid puuduvad proadjektiividel, ordinaalidel, verbidel, adverbidel, substantiividel, verbi partitsiipidel ning adjektiividel/adverbidel. Ootuspäraselt esines kõige enam keskvõrdeid kahes esimeses adjektiivirühmas (adjektiivid I, adjektiivid II), kuid parameetri keskmine tulemus oli kõrge ka testrühmas adjektiivid/substantiivid, kus omakorda eristus sõna *hull*, mille puhul keskvõrde esinemisarv ulatus ligikaudu 80%-ni *hullu* enda üldsagedusest. Kuna valim ei olnud suur, siis tõstis ühe sõna väga kõrge osakaal terve rühma osakaalude keskmise uuritud rühmade kõige kõrgemaks.

Adjektiivirühmades mõjutab konkreetsete testsõnade semantika samuti rühma keskmist: vähe võrdevorme esineb sõnadel *tõeline*, *rahvusvaheline*, *kahetsusväärne*, *napisõnaline*, *toasoe* ja *kaitsealune*. Need sõnad tähistavad mitteskalaarset omadust, mille võrdlemine ei ole pragmaatiliselt mõttekas.

Kokkuvõttes eristab keskvõrde parameeter hoolimata oma tagasihoidlikest väärtustest hästi rühmi ja sõnu, mida kasutatakse korpuses adjektiiviga sarnaselt.

Adverbiparameetri (D_test) väärtus on kõrgeim prototüüpsete adjektiivide rühmas keskmise tulemusega 0,29, vahetult järgnevad rühmad adjektiivid II tulemusega 0,27 ja adjektiivid/adverbid tulemusega 0,26 (nt *väga purjus*). Madalate adverbile järgnemise näitajatega eristub substantiivide rühm, mille keskmine parameetri tulemus on 0,08. Adjektiivirühmade lähedusse paiknevad ka proadjektiivide (0,24) ja ordinaalide rühm (0,24). Adverbiparameetri tulemuste varieeruvus on võrreldes teiste parameetritega väiksem ja väärtused jagunevad ühtlasemalt. Partitsiipide rühma keskmine sagedus (0,193) peaaegu kattus adverbirühma (0,194) mõõtmistulemusega. Kokkuvõtvalt on adverbiparameetri väärtused kõrged adjektiivirühmades ning proadjektiivide ja ordinaalide rühmas ning madala tulemusega eristub teistest substantiivide rühm.

Lause alguse parameetri (LA_test_S) väärtused on üldiselt madalamad kui eelnevatel, kuna tegu on mitut tingimust kombineeriva päringuga. Kõige kõrgemad näidud mõõdab lause alguse parameeter proadjektiivi (0,093) ja ordinaali rühmades (0,094) (erandiks *mingisugune*, mis ei alusta lauset). Kõrge keskmine näit (0,068) on ka käändumatute adjektiivide rühmas, kuid selle taga on ühe sõna – *lugupeetud* – ülisage esinemine selles positsioonis (nt *lugupeetud kuulajad*). Oluline on, et sel parameetril on äärmiselt madalad näidud verbi partitsiipide, verbide ja adverbide rühmas. Seetõttu on ta hea kriteerium, kui on vaja eristada adjektiveerunud verbivorme regulaarsetest verbi partitsiipidest. Adjektiive iseloomustavad vahepealsed väärtused sellel parameetril (võrreldes ühelt poolt proadjektiivide ja ordinaalidega ning teiselt poolt verbide, verbi partitsiipide ja adverbidega).

Neljakohaline parameeter (V_?_test_S) eristas adjektiividest kõige paremini verbi partitsiipvorme, verbe, adverbe ja substantiive. Nende rühmade keskmised näitajad (verbi partitsiibid 0,028, verbid 0,016, adverbid 0,031, substantiivid 0,037) jäid oluliselt adjektiivirühmadele alla (adjektiivid I 0,131, adjektiivid II 0,107). Kõige rohkem esines mustrit verb + vaba sõna + testsõna + substantiiv proadjektiivide rühmas (0,133, nt *sisaldab täpselt sama toimeainet*) ja prototüüpsete adjektiivide rühmas (nt *näeme noori edukaid inimes*i).

4.3. Kõrvalekaldeanalüüs

Uurimaks, kuivõrd suudavad parameetrid adjektiive teistest sõnaliikidest eristada ja seda, mil moel erinevad omavahel eri tüüpi adjektiivid, viisime läbi kõrvalekaldeanalüüsi. Kõrvalekaldeanalüüs selgitas välja, kas ja mil määral hälbisid teised testrühmad prototüüpsete adjektiivide järgi sätitud piirmääradest. Adjektiivi piirmäärad määrasime kindlaks prototüüpsete adjektiivide rühma põhjal (adjektiivid I). Määrasime igale parameetrile I rühma alusel adjektiivide miinimum- ja maksimumväärtuse (tabel 3) ja uurisime, kuidas teiste testrühmade sõnad piiridesse mahutuvad. Arvutasime välja nende kõrvalekalde adjektiivirühma väärtustest.

Tabelis 4 on tulemused esitatud testrühmade kaupa. Tabel näitab, kui palju konkreetse testrühma liikmetest mahtus iga parameetri puhul adjektiivi piirmääradesse (100% tähistab täielikku kattumist ja 0% näitab, et ühegi testrühma liikme tulemus ei paigutunud konkreetse parameetri põhjal adjektiivide miinimum- ja maksimumväärtuse vahele).

Tabel 4. Kõrvalekaldeanalüüsi tulemused

	test_S	test_S_ÜHILD	keskvõrre	D_test	LA_test_S	V_?_test_S
adjektiivid II	100%	100%	70%	80%	100%	100%
adjektiivid/partitsiibid	60%	50%	30%	50%	50%	60%
adjektiivid/substantiivid	100%	100%	50%	70%	90%	100%
adjektiivid/adverbid	100%	30%	20%	90%	70%	100%
käändumatud	70%	10%	20%	70%	70%	80%
partitsiibid	70%	50%	0%	20%	10%	50%
substantiivid	100%	0%	0%	0%	100%	80%
adverbid	70%	0%	0%	70%	10%	80%
verbid	50%	0%	0%	70%	10%	10%
proadjektiivid	100%	100%	0%	90%	20%	90%
ordinaalid	100%	100%	0%	100%	20%	100%

Kõige sarnasemad rühmad prototüüpsete adjektiividega (adjektiivid I) on: adjektiivid II, adjektiivid/substantiivid, proadjektiivid ja ordinaalid. Esimesed koosnevadki adjektiividest (kuigi mitte täiesti prototüüpsetest, kuna korpuses esineb nende rühmade sõnadel kasutust ka nt substantiivina), nii et nende kokkulangemine alusrühmaga on ootuspärane. II rühma adjektiivid mahuvad nelja parameetri põhjal 100%-liselt prototüüpsete adjektiivide piiridesse, kuid ka nende hulgas esineb sõnu, mis esimese rühma põhjal määratud piiridest irduvad, nt *kallis*, mille keskvõrret *kallim* esineb korpuses rohkem, ning *toasoe* ja *kaitsealune*, mille keskvõrde parameetri väärtused on semantikast tingitult madalamad.

Teiste sõnaliikide hulgast paistavad silma proadjektiivid ja ordinaalid, mille käitumine sarnaneb korpuses adjektiividega. Nende rühmade puhul osutusid aga headeks eristajateks keskvõrde parameeter ja lause alguse parameeter (LA_test_S); lause alguse parameetri piirmääradest jäävad välja 80% proadjektiividest ja 80% ordinaalidest ning keskvõrde parameetri põhjal erinevad prototüüpsetest adjektiividest kõik uuritud proadjektiivid ja ordinaalid.

Kokku 88% eksperimendisõnadest hälbivad prototüüpsete adjektiivide piirmääradest vähemalt ühe parameetri põhjal. Üldse ei hälbi kuus sõna adjektiivide

II rühmast (*erinev*, *hea*, *külm*, *soe*, *särav*, *uudishimulik*), rühmast adjektiivid/verbivormid kaks sõna (*elav*, *karjuv*). Rühmast adjektiivid/substantiivid viis sõna (*noor*, *haige*, *rumal*, *tuttav*, *võõras*) ja rühmast adjektiivid/adverbid üks sõna (*meeletu*). Suurem osa mittehälbivatest sõnadest kuulub rühmadesse adjektiivid II ja adjektiivid/substantiivid. Prototüüpsetest adjektiividest kõige erinevamalt käituvad teistest rühmadest verbipartitsiibid – *osanud*, *tahtnud* ning adverb *võõrsil* ja verb *oskama*, mis irdusid adjektiividest kõigi kuue testitud parameetri põhjal.

Problemaatilisi rühmi (adjektiivid *vs*. partitsiibid) eristavad kõige paremini kolm parameetrit: lausealguseparameeter 90% täpsusega, adverbiparameeter 80% täpsusega ja keskvõrde parameeter 100% täpsusega. See tähendab, et nende parameetrite põhjal ei mahu partitsiipide tulemused adjektiivide piirmääradesse. Teiste sõnaliikide lõikes (rühmad 6–12) on parimad eristajad lause alguse parameeter (eristab ~78% sõnu), keskvõrde parameeter (eristab 100% sõnu!) ja ühildumisparameeter (eristab ~59% sõnu, sealhulgas substantiive, adverbe ja verbe 100%-liselt).

Korpuspõhise sõnaliigieristaja väljatöötamise seisukohast on oluline välja selgitada, millised testitud parameetritest on kõige täpsemad (millele saaks tulevikus automaatanalüüsis ehk ka suurema kaalu anda). Niisiis vaatasime, milliste tunnuste suhtarvud toovad kõige paremini esile adjektiivide sõnaliigipärast käitumist. Kõigi testrühmade lõikes osutusid parimateks eristajateks keskvõrde parameeter (83%), ühildumisparameeter (~51%) ja lause alguse parameeter (~50%). Võttes arvesse, et pooled testsõnadest moodustavad eri tüüpi adjektiivid, siis täielikku eristust ei saakski eeldada.

Kõige nõrgemad eristajad kõikide testrühmade lõikes olid atribuudiparameeter, mis suutis prototüüpsetest adjektiividest eristada kõigest 16% testsõnadest, ja neljakohaline parameeter, mis eristas prototüüpsetest adjektiividest 23% testsõnadest. Atribuudiparameetri tulemus oli seega kahjuks loodetust vähem ilmekas, kuna substantiiv kõige sagedama ning mitmesuguste süntaktiliste rollidega sõnaliigina sattus korpuses testsõnale vahetult järgnevaks naabriks hoolimata testsõna sõnaliigist. Adverbiparameeter, 35% eristusvõimega, paigutus parameetrite tõhususes keskmisele kohale. Kuigi iga testitud parameeter suutis eristada mõnd sõnaklassi prototüüpsetest adjektiividest, võib tulevikus arvestada nende üldist eristusvõimet ja kaaluda test_S ja V_?_test_S parameetri väljajättu.

Testitud parameetrite komplekt tervikuna eristas kõiki testrühmi arvestades adjektiividest kõige paremini verbi partitsiipe ja verbe (st partitsiipide ja verbide väärtused irdusid testrühmadest enim adjektiivirühma põhjal määratud vahemikest). Kas ja kuivõrd käituvad konkreetsed verbi partitsiibid korpuses juba nagu adjektiivid, oligi leksikograafide põhiküsimus, millele lootsime uuringuga jälile saada.

4.4. Üksiksõnade võrdlus adjektiivi korpuskäitumisega

Kuivõrd leksikograafil on huvi eeskätt konkreetse sõna korpuskäitumise hindamise vastu – kas ja kuivõrd sarnane on see tüüpilise adjektiivi käitumisele – siis peaks huvipakkuva sõna profiili (6 parameetri arvväärtused) saama võrrelda otse prototüüpse adjektiivi profiiliga. Leidsime, et selliseks võrdluseks võiks sobida eukleidilise kauguse mõõdik, millesse saab kaasata võrreldavate nähtuste mitmeid

parameetreid korraga. Võtsime analüüsitulemustest kõigi katses osalenud sõnade individuaalsed profiilid kuue parameetri lõikes ning valisime tüüpilist adjektiivi esindama profiili, mis moodustub prototüüpsete adjektiivide rühma mediaanväärtustest (vt tabel 3, ptk 4.1).

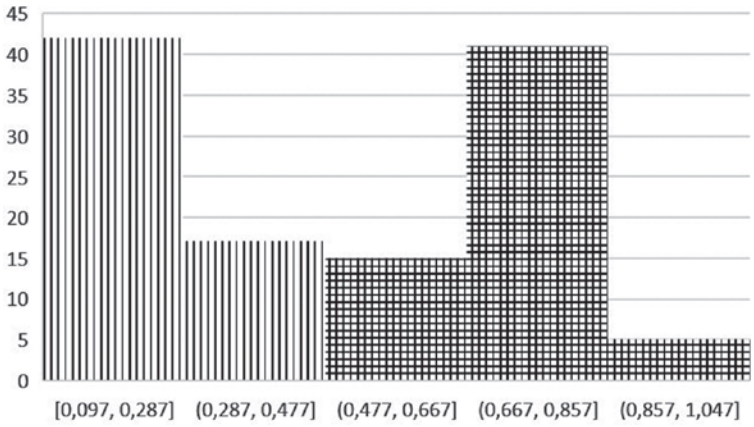
Eukleidline kaugus on (optimeeritult) ruutjuur kuuedimensioonilise parameetrite ruumi parameetrite kauguste ruutude summast.

$$Eukl_{dist} = \sqrt{\sum_{i=1}^{n=6}(V_i - O_i)^2}$$

V_i – võrdlusaluse i-nda parameetri väärtus
O_i – võrreldava omadussõna i-nda parameetri väärtus

Arvutasime kõigile 120 sõnale eukleidilise kauguse võrdlusaluseks võetud profiilist. Mõõdetud kaugused on positiivsed arvud, kusjuures mida suurem arv, seda suurem on mõõdetud profiili koguerinevus võrdlusalusest ja mida väiksem on eukleidiline kaugus, seda suurem on sarnasus võrdlusalusega. Seda, kummas suunas erinevus esineb (st kas mingit omadust on rohkem või vähem, kui standard eeldab), kaugus arvesse ei võta.

Sarnasuse mõõtmine ühe sõna kaupa laseb meie katseandmestikus esile tulla üleüldisel sarnasuse skaalal. Ka prototüüpseks peetud adjektiivide rühma sees on mõned, mis on võrdlusalusele lähedasemad, st sarnasemad prototüübile (nt *terviklik*, *edukas* – väärtused 0,13 ja 0,16), ja teised, mis sellest tublisti kaugemal (*kahetsusväärne* – 0,48 ja *napisõnaline* – 0,72). Eukleidiliste kauguste arvutus toob välja mõningad lähedased sõnad teistest rühmadest, nt *elav* (adjektiiv/partitsiip, väärtus 0,1), *võõras* (adjektiiv/substantiiv, väärtus 0,166) ning *meeletu* ja *tohutu* (adjektiivid/adverbid, mõlemal väärtus 0,2), *olev* (verbi partitsiibid, väärtus 0,2), *nisuke* (proadjektiiv, 0,12).

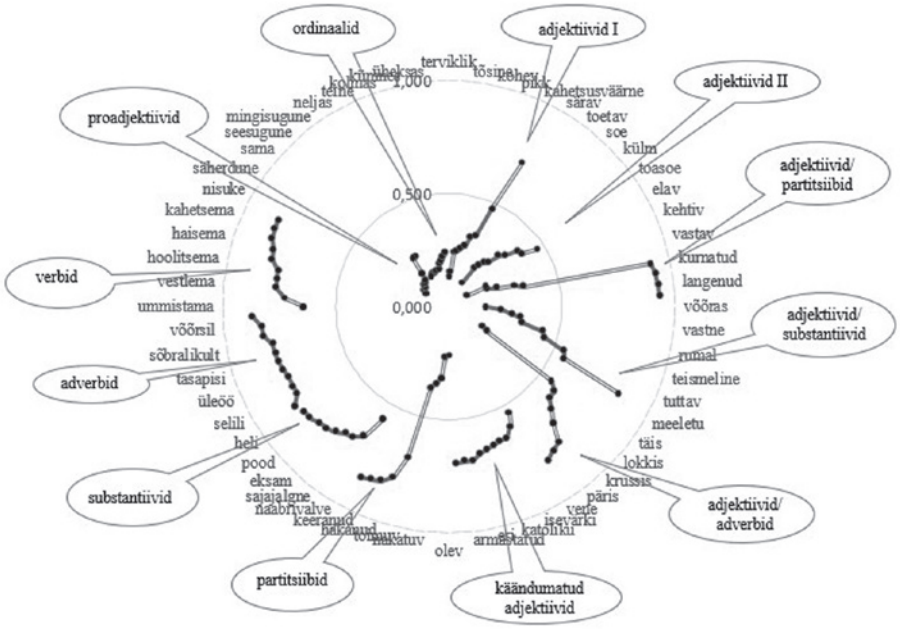


Joonis 3. Eukleidiliste kauguste histogramm

Eukleidiliste kauguste histogramm joonisel 3 näitab, et väärtused kuhjuvad kahte piirkonda: vahemikku 0,097–0,477, mis on väike kuni mõõdukas kaugus

võrdlusalusest, ning vahemikku 0,477–1,047, mis on suurem kaugus võrdlusalusest. Selline tulemus läheb hästi kokku meie katseandmete valikuprintsiibiga: siia kuulusid rühmad, millelt eeldati adjektiivile sarnanemist, ning kontrastiks rühmad, mis pidanuksid näitama üles erinevat käitumist. Tähistasime joonisel 3 adjektiivile sarnasemad andmed triibulise ja sellest erinevamad andmed ruudulise markeeringuga.

Järgmisel diagrammil (joonis 4) on kõik katsesõnad andmepunktidena paigutatud rühmakuuluvuse ja rühma sees kasvava eukleidilise kauguse järgi. Välisäärel on kuvatud valik rühmadesse kuuluvatest sõnadest. Eukleidilise kauguse väärtusi osutavad vahejooned diagrammil on paigutatud tsentraalselt sammuga 0,5. Väärtus 0,5 piirab "siseringi", kuhu paigutuvad sõnad, mille korpuskäitumine vaadeldud parameetrite lõikes meenutab kõige enam prototüüpsete adjektiivide profiili. Nagu näha, jääb siseringist välja üks protüüpse adjektiivirühma liige – *napisõnaline* –, mille näitajad on kõikidel parameetritel hästi madalad. Pilk korpusesse näitab, et see madala esinemissagedusega omadussõna tuleb esile peamiselt öeldistäite positsioonis (nt *on napisõnaline*, *jääb napisõnaliseks*). Täielikult osutuvad prototüüpse adjektiivi sarnaseks rühm adjektiivid II ning nii proadjektiivide kui ka ordinaalide rühmad, mille korpuskäitumine osutub selle metoodikaga mõõtes vägagi adjektiivi sarnaseks. Neli rühma – verbid, adverbid, nimisõnad ja käändumatud omadussõnad – jäävad sellest siseringist väljapoole. Järelikult on nende korpuskäitumine meie parameetritega mõõdetult prototüüpse adjektiivi omast üsnagi erinev. Joonisel 4 on näha ka rühmad, mille liikmetest mõned käituvad adjektiivisarnaselt ning osa liikmetest sellest erinevalt. Sellised



Joonis 4. Sõnade eukleidilised kaugused prototüüpse adjektiivi korpuskäitumise profiilist katserühmade kaupa

rühmad on need, mis juba oma nimetuses sisaldavad kahte sõnaliiki: adjektiivid/partitsiibid, adjektiivid/adverbid, adjektiivid/nimisõnad ning lisaks rühm partitsiibid. Seega osutub, et eukleidilise kauguse mõõdik suutis teha eristusi heterogeensete rühmade sees.

Rühm adjektiivid/partitsiibid jaguneb kaheks selle alusel, kas tegemist on oleviku- või minevikupartistiipidega. *v*-lõpulisi teeb adjektiiviga sarnasemaks nende käänatavus ning seega ka täiendi positsioonis peasõnaga ühildumine. Sama põhimõte jagab kaheks ka partitsiipide rühma. Rühm adjektiivid/substantiivid jaotub hajusamalt; kõige kaugemale prototüüpse adjektiivi käitumisest jääb sõna *hull*. Rühmast adjektiiv/adverb ilmutavad kaks liiget (*meeletu*, *tohutu*) silmanähtavalt siseringi kuulumist, samas kui suurem osa selle rühma sõnu käituvad pigem ebatüüpilise adjektiivina.

Kõigis rühmades paigutuvad liikmed skaalale, st liikmed erinevad mingil määral üksteisest kauguse poolest prototüübist. Kõige vähem skalaarsust ja kõige rohkem väärtuste kuhjumist on ordinaalide rühmas, mis suhteliselt ühtsena sarnaneb väga tugevalt prototüüpse adjektiivi korpuskäitumise profiiliga.

Eukleidilise kauguse metoodika laseb hinnata sarnasust tüüpilise adjektiivi profiiliga, samas ei ütle see midagi selle kohta, kas mõõdetavad erinevused on samas suunas või sootuks erinevas. Diagrammi ringikujulisus viib meie andmestikus olnud sõnarühmade otsad kokku, st paigutab kõrvuti alguses olnud prototüüpsed adjektiivid ning järjekorras viimastena ordinaalid ja proadjektiivid. See, et need rühmad osutuvad oma mõõdetud käitumisprofiili poolest nii sarnasteks, ei olnud küll ette läbi mõeldud, kuid on lõppkokkuvõttes üsna loogiline ja loomulik tulemus.

5. Kokkuvõte ja arutlus

Uurimuse eesmärk on testida prototüüpse adjektiivi tunnuseid korpuseandmestikul. Tunnused formaliseerisime kuueks parameetriks, mis mõõdavad testsõna esinemist nimisõna eestäiendina, ühildumist järgneva sõnaga, keskvõrrete olemasolu, laiendamist eelneva adverbiga; kaks kitsamalt määratletud parameetrit kombineerivad testsõna paiknemist nimisõna ees lause alguses ning järjendis, kuhu kuulub verb ning üks määratlemata sõna. Uurimuse sihtrühma moodustavad kuus adjektiivide ja nendega piirnevate ja segunevate sõnade rühma (prototüüpsed adjektiivid, vähesel määral teistes sõnaliikides esinevad adjektiivid ning partitsiipsed, substantiivsed, adverbilised ja käändumatud adjektiivid). Võrdluseks on kaasatud kuus kontrollrühma (verbipartitsiibid, substantiivid, adverbid, verbid, proadjektiivid ja ordinaalid).

Tulemused näitavad, et prototüüpsed adjektiivid ületavad kõiki teisi uuritud sõnarühmi keskvõrdeparameetri ja adverbiparameetriga mõõtes. Näeme ka varieeruvust, mis kinnitab, et tunnuste reaalne olemasolu tekstimassiivis on olemuselt pigem skalaarne kui binaarne. Jah/ei-otsustused saab asendada tunnuse esinemise suhtelise sageduse ehk tõenäosusnäitajatega.

Kõige adjektiivsemaks tunnuseks võib testitulemuste põhjal pidada (kesk)võrdevorme, mida praktiliselt ei tule esile teistes sihtrühma ega ka kontrollrühma testsõnades. Samas, ka adjektiivide endi hulgas esilduvad võrdevormid korpuses suhteliselt madala osakaaluga.

Mitme parameetri (nt ühildumis-, atribuudi- ja adverbiparameetri) suhtes näitasid kontrollrühma kuuluvad proadjektiivid ja ordinaalid ühtsemalt (st vähema varieeruvusega) adjektiividelt eeldatavat käitumist. Adjektiividest eristab neid otsustavalt semantika (abstraktsus, järjestuse väljendamine), mistõttu võrdevorme proadjektiivide ja ordinaalide seast ei leia. Hästi eristas proadjektiive ja ordinaale kõrvalekaldeanalüüsis kaks testitud parameetrit: lause alguse ja keskvõrde parameeter. Prototüüpsete adjektiivide põhjal seadistatud piirmääradest jäid lause alguse parameetriga mõõtes välja 80% proadjektiividest ja 80% ordinaalidest ning keskvõrdeparameetri põhjal erinesid prototüüpsetest adjektiividest kõik uuritud proadjektiivid ja ordinaalid.

Käändumatut omadussõna ja ordinaali iseloomustab mõnevõrra kõrgem esiletulek atribuudiparameetri põhjal, kuid kuna see tunnus iseloomustab kõiki rühmi, oleks eestäiendi suhet hea kitsendada süntaktilise fraasipiiranguga juba otsingul.

Kuivõrd uurimust motiveerivad leksikograafidele enim peavalu valmistavad adjektiividega seotud kitsaskohad, siis pöörasime erilist tähelepanu sellele, kuidas suutsid testitud parameetrid eristada adjektiividest partitsiipe. Partitsiipide rühma eristas prototüüpsete adjektiivide rühmast kõige paremini kolm parameetrit: lausealguseparameeter 90% täpsusega, adverbiparameeter 80% täpsusega ja keskvõrde parameeter 100% täpsusega. (100% täpsus tähendab, et kõigi uuritud partitsiipide tulemused jäid kõrvalekaldeanalüüsis väljapoole adjektiivide järgi sätitud piirmäära.) Lause alguse parameeter eristab niisiis edukalt partitsiipvormide adjektiivset käitumist: verbiühendi osana partitsiip ei alusta lauset, kuid adjektiivilaadne partitsiip võib seda teha.

Arvestades kõiki testrühmi, eristaski parameetrite komplekt tervikuna kõrvalekaldeanalüüsis adjektiividest kõige täpsemalt partitsiipe ja verbe. Nende väärtused irdusid testrühmadest enim adjektiivirühma miinimum- ja maksimumväärtuse põhjal määratud vahemikest. Tulemused annavad lootust, et suudame pakkuda leksikograafidele sõnaliigiotsustustes tuge ja esitada testitud parameetreid kasutades sõna korpuskäitumisest ülevaate, mis iseloomustab seda, kuivõrd käitub konkreetne partitsiip korpuses adjektiivilaadselt.

Kuna tulevane korpuspõhine sõnaliigieristaja peaks eristama adjektiividest ka teisi sõnarühmi peale partitsiipide, siis selgitasime, millised testitud parameetritest eristavad enim teiste sõnaliikide sõnu adjektiividest. Nendele parameetritele on tööriista seadistamisel võimalik suurem kaal anda. Teiste sõnaliikide lõikes olid parimateks eristajateks lause alguse parameeter (eristas ~78% sõnadest), keskvõrde parameeter (eristas 100% teistest sõnaliikidest sõnu) ja ühildumisparameeter (eristas ~59% sõnadest, sealhulgas substantiive, adverbe ja verbe 100%-liselt). Võrreldes kõigi testrühmade tulemusi, k.a adjektiiviga piirnevad rühmad, osutusid parimateks eristajateks keskvõrde parameeter (üle 83%), ühildumisparameeter (~51%) ja lause alguse parameeter (~50%). Siin tuleb silmas pidada, et pooled testrühmadest olid eri tüüpi adjektiivid, mistõttu on osaline kattumine prototüüpse adjektiivirühma tulemustega ootuspärane. Iga testitud parameeter suutis kõrvalekaldeanalüüsis eristada vähemalt üht testrühmadest tüüpilistest adjektiividest, niisiis väärivad need kõik edasi katsetamist.

Uurimuse viimases osas mõõtsime katseliselt kõikide katsesõnade sarnasust prototüüpse adjektiivi korpuskäitumise profiiliga eukleidilise kaugusena sellest. See metoodika paigutas kõik katsesõnad sarnasuse skaalale, lasi mõõta prototüüpsust

adjektiiviklassi sees ning näitas ära terved rühmad, mis adjektiivile sarnanevad (ordinaalid, proadjektiivid). Adjektiivist kõige erinevamaks osutusid kontrollina kaasatud sõnarühmad (verbid, adverbid ja substantiivid), aga ka sihtrühmast käändumatud adjektiivid, mis osutusid tervikuna ebaprototüüpseks. Eukleidilised kaugused lasid jagada liikmed adjektiiviga sarnanevateks ning erinevateks nende rühmade sees: adjektiiv/partitsiip, partitsiip ja adjektiiv/adverb, adjektiiv/substantiiv. Seega peaks taoline mõõdik sobima lahendama leksikograafide poolt tõstatatud peamisi sõnaliigi määratlemise probleeme.

Uurimuse tulemusi võis mõjutada asjaolu, et sõnad olid käsitsi valitud ning valim ei olnud suur. Seega võis üksikutel sõnadel olla tugev mõju rühma keskmistele väärtustele iga parameetri puhul. Eriti ilmne oli see juhtudel, kus üks testrühma sõna irdus selgelt teiste oma rühma sõnade tulemustest (nt *lugupeetud* käändumatute adjektiivide rühmas atribuudiparameetri tulemustes või *hull* adjektiivi/substantiivi rühmas keskvõrde parameetri puhul). Üldisema ülevaate sõnarühmast (kas sõnaliigipõhiselt või adjektiiviklassi alaliikide kaupa) saaks, kui vaadata suuremat sõnahulka korraga. See võimaldaks piirmäärade kehtestamisel paremini arvestada rühma ühisosa ja jätta erandlikud tulemused kõrvale. Siiski oli uurimuse väiksem valim ka eelis, kuna võimaldas meil käsitsi tulemusi kontrollida. Samuti saime valida välja just sobivad sõnad, mis asuvad adjektiivirühma piirialadel ja vajavad täpsemat korpusanalüüsi sõnaliigilise otsuse tegemiseks (adjektiivid/substantiivid, adjektiivid/adverbid, adjektiivid/partitsiibid).

Kuna ei selgunud tunnuseid, mille poolest adjektiivide väärtused ületaksid uuritud korpuses selgelt kõiki teisi sõnaliike, saaks leksikograafilise tööriista häälestamisel parema tulemuse, kui määrata iga parameetri puhul vahemik, millest üleminekul, samamoodi kui allajäämisel, hakkaks adjektiivse käitumise tõenäosus kahanema. Kasu võiks olla ka "karistuste" süsteemist, mis negatiivsete tunnuste olemasolul hindab sõna adjektiivsust madalamaks (nt *ei* otsisõna ees viitab verbilisele kasutusele: *ei armastatud* vs. *armastatud laulja*). Lõplike normide kehtestamiseks peaks aga analüüsima suuremat rühma prototüüpseid adjektiive.

Testitud parameetrid on kõik peale keskvõrde parameetri võimalik praegusel kujul kasutusele võtta. Keskvõrde parameetri testimiseks korpuses on esmalt vaja luua automaatne keskvõrde generaator. Praegune korpusmärgendus adjektiivi tema võrdevormidega ei seo ega võimalda ühiselt välja otsida. Keskvõrde genereerimisel saaks alusena kasutada näiteks Eesti Keele Instituudi sõnastikubaasi Ekilex (Hein jt 2020) noomeni paradigmat ja pärida sealt välja omastava vormid. Kuigi võrdevormide puhul võib korpuses esineda ka juhuslikku kokkulangemist (nt *auto-autom*, mispuhul esineb vorm *autom* sõna *automaat* lühendina), siis aitaks seda välistada või vähemalt vähendada võrdevormi käändevormide kontollimine (nt **automat*, **automasse*). Homonüümid jäävad paratamatult tulemusi veidi hägustama.

Sõnaliigist lähtuva morfosüntaktilise korpusanalüüsi vajadust motiveerib leksikograafide soov näha täpsemat ülevaadet korpuse toorandmetest – need tuleks töödelda sellisele kujule, et leksikograaf saaks langetada kõige õigema ja ka kiirema otsuse sõna omaduste kohta (Paulsen jt 2019). Et teha järeldusi tegelikust keelekasutusest, on tarvis kokkuvõtvat analüüsi uuritava sõna käitumismustritest mahult ja tekstiliigiliselt koostiselt võimalikult esinduslikus korpuses, alahindamata seejuures leksikograafi rolli tulemuste tõlgendamisel (automaatanalüüsitud tekstikogudest leksikograafi töös vt Langemets jt 2021: 760). Teisest küljest on automaatse eesti

keele morfoloogilise analüüsi loojad algusest peale nentinud raskusi sõnaliigiliste määrangute osas (nt Muischnek, Vider 2005, Habicht jt 2000). Nende raskustega oleme kokku puutunud ka enda üritustes luua rakendust, mis laseks statistilise esiletuleku põhjal hinnata sõnavormi iseseisvumise astet (Paulsen jt 2021, Vainik jt 2021). Kahtlemata mõjutab märgenduse täpsus ka käesoleva uurimuse tulemusi. Kuivõrd aga kõik korpuspõhiselt mõõdetud parameetrid siiski andsid mõttekaid ja mõtestatavaid tulemusi, siis oleme optimistlikud selles suhtes, et neid saab kaasata rakendusse, mis aitaks leksikograafidel hinnata sõnavormi käitumist tekstis ning seeläbi langetada paremini põhjendatud sõnaliigiotsustusi.

Viidatud kirjandus

Allemann, Elo 2002. Kesksõnad atribuudi ja predikatiivina ajakirjanduses ja õiguses ['Participle as an attribute and predicative in journalism and law']. Magistritöö. Tallinn: Tallinna Ülikool.

EKI ühendsõnastik 2021. Eesti Keele Instituut, Sõnaveeb.

Erelt, Mati 1977. Ebamäärasusest sõnade liigitamisel ['About uncertainty in the classification of words']. – Keel ja Kirjandus, 9, 525–528.

Erelt, Mati 1986. Eesti adjektiivisüntaks ['The Syntax of Estonian Adjectives']. Eesti NSV Teaduste Akadeemia Emakeele Seltsi toimetised nr 19. Tallinn: Valgus.

Erelt, Mati 2017. Omadussõnafraas ['Adjective phrase']. – Mati Erelt, Helle Metslang (Toim.), Eesti keele süntaks. Eesti keele varamu 3. Tartu: Tartu Ülikooli Kirjastus, 405–415.

Habicht, Külli; Kaalep, Heiki-Jaan; Muischnek, Kadri; Müürisep, Kaili; Rääbis, Andriela 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest ['Do the available morphological descriptions of Estonian work on a real text?']. – Keel ja Kirjandus, 9, 623–633.

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

Hein, Indrek; Männiko, Kaur; Kallas, Jelena; Koppel, Kristina; Langemets, Margit; Nurk, Tõnis; Plado, Merily; Vaus, Mari; Viks, Ülle; Tavast, Arvi; Laubre, Martin; Sharma, Yogesh; Niilo, Hardi 2020. Ekilex 2020. Eesti Keele Instituudi sõnastiku- ja terminibaas. Eesti Keele Instituut.

Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias ['Syntagmatic Relations of Estonian Content Words in Corpus and Pedagogical Lexicography']. Tallinna Ülikooli humanitaarteaduste dissertatsioonid 32. Tallinn: Tallinna Ülikool.

Kallas, Jelena; Koppel, Kristina; Tuulik, Maria 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel ['New possibilities in corpus lexicography based on the example of the Estonian Collocations Dictionary']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 11, 75–94. https://doi.org/10.5128/ERYa11.05

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

Kaalep, Heiki-Jaan; Vaino, Tarmo 2001. Complete morphological analysis in the linguist's toolbox. – Congressus Nonus Internationalis Fenno-Ugristarum. Pars V. Dissertationes sectionum: linguistica. II. Tartu, 9–16.

Karelson, Rudolf 2005. Taas probleemidest sõnaliigi määramisel [‘Once again about the problems in word class determination’]. – Eesti Rakenduslingvistika Ühingu aastaraamat, 1, 53–70. https://doi.org/10.5128/ERYa1.03
Kasik, Reet 2015. Sõnamoodustus [‘Word Formation’]. Tartu: Tartu Ülikooli Kirjastus.
Kerge, Krista 1998. Vormimoodustus, sõnamoodustus ja leksikon: oleviku kesksõna võrdluse all [‘Form Formation, Word Formation and Lexicon’]. Tallinn: TPÜ Kirjastus.
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. https://dspace.ut.ee/handle/10062/67138
Langemets, Margit; Koppel, Kristina; Kallas, Jelena; Tavast, Arvi 2021. Sõnastikukogust keeleportaaliks [‘From a collection of dictionaries to a language portal’]. – Keel ja Kirjandus, 8–9, 754–769. https://doi.org/10.54013/kk764a6
Lindström, Liina; Bakhoff, Liisi; Kalvik, Mari-Liis; Klaus, Anneliis; Läänemets, Rutt; Mets, Mari; Niit, Ellen; Pajusalu, Karl; Teras, Pire; Uiboaed, Kristel; Veismann, Ann; Velsker, Eva 2006. Sõnaliigituse küsimusi eesti murrete korpuse põhjal [‘Word classes in Estonian Dialect Corpus’]. – Ellen Niit (Toim.), Keele ehe. Tartu Ülikooli eesti keele õppetooli toimetised 30. Tartu: Tartu Ülikool, 154–167.
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
Paradis, Carita 2001. Adjectives and boundedness. – Cognitive Linguistics, 12 (1), 47–64. https://doi.org/10.1515/cogl.12.1.47
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; Tuulik, Maria; Lohk, Ahti 2019. The lexicographer’s voice: Word classes in the digital era. – I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (Eds.), Electronic Lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 319–337.
Paulsen, Geda; Vainik, Ene; Lohk, Ahti; Tuulik, Maria 2021. Catching lexemes: The case of Estonian noun-based ambiforms. – I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, C. Tiberius (Eds.), Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. Brno: Lexical Computing CZ, s.r.o, 288–311.
Rosch, Eleanor 1978. Principles of Categorization. Cognition and Categorization. Hillsdale–New York: Lawrence Erlbaum, 27–48.
Tarabarova, Olga 2014. Eesti keele täiskujuline hüüdlause kui konstruktsioon 20. sajandi kirjakeeles [‘Estonian exclamative sentence as a construction in 20th century written Estonian’]. Magistritöö. Tartu Ülikool. http://hdl.handle.net/10062/43718
Tiits, Mai 1982. Seisundiadverbidest [‘About adverbs of state’]. – Keel ja Kirjandus, 1, 17–21.
Vainik, Ene; Lohk, Ahti; Paulsen, Geda 2021. The Distribution Index Calculator for Estonian. – I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, C. Tiberius (Eds.), Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 Conference. Brno: Lexical Computing CZ, s.r.o, 121–138.
Vainik, Ene; Paulsen, Geda; Lohk, Ahti 2020. A typology of lexical ambiforms in Estonian. – Z. Gavriilidou, M. Mitsiaki, A. Fliatouras (Eds.), Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. 1. Alexandroupolis: Democritus University of Thrace, 119–130. https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p119-130.pdf (28.4.2022).

Vare, Silvi 2006. Adjektiivide substantivatsioonist ühe tähendusrühma näitel ['On the substantivation of adjectives in Estonian']. – Ellen Niit (Toim.), Keele ehe. Tartu Ülikooli eesti keele õppetooli toimetised 30. Tartu: Tartu Ülikool, 205–222.
Viks, Ülle 1977. Sõnaliik kui niisugune ['Part of speech as such']. – Keel ja Kirjandus, 9, 521–525.

Võrguviited

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
Morfoloogiliselt ühestatud korpus. https://doi.org/10.15155/1-00-0000-0000-0000-00085L
Sketch Engine. https://www.sketchengine.eu/ (20.9.2021).

Maria Tuulik (Eesti Keele Instituut, Tartu Ülikool). Uurimisvaldkonnad on adjektiivide semantika, leksikograafia ja sõnaliigipiirid.
Roosikrantsi 6, 10119 Tallinn, Estonia
maria.tuulik@eki.ee

Ene Vainik (Eesti Keele Instituut). Uurimisvaldkond on semantika ja keele psühholoogiaga piirnevad aspektid (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). Uurimisvaldkonnad on leksikaalne semantika, leksikograafia ja korpuslingvistika.
Roosikrantsi 6, 10119 Tallinn, Estonia
geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se

Ahti Lohk (Tallinna Tehnikaülikool, Eesti Keele Instituut). Uurimisvaldkonnad *onwordnet*'i semantiliste hierarhiate valideerimine graafipõhiste meetoditega ja tekstikaeve algoritmid kasuliku, uue ja rakendatava teabe eraldamiseks struktureerimata tekstist.
Akadeemia tee 15A, 12618 Tallinn, Estonia
ahti.lohk@eki.ee, ahti.lohk@taltech.ee

Lisa 1. Uurimuses kasutatud testsõnade rühmad koos sagedusandmetega eesti keele ühendkorpuses 2019

Rühm	Sõna	Sagedus ÜK 2019-s
adjektiivid I	edukas	198033
	harv	29656
	kahetsusväärne	13281
	kohev	11083
	napisõnaline	2067
	pikk	1160497
	rahvusvaheline	856030
	terviklik	55205
	tõeline	374280
	tõsine	373905
adjektiivid II	erinev	2551550
	hea	4436736
	kaitsealune	17471
	kallis	288627
	külm	285799
	soe	394299
	särav	57786
	toasoe	7060
	toetav	68868
	uudishimulik	14360
adjektiivid/verbivormid	armunud	235
	elav	283058
	karjuv	9939
	kaunistav	1549
	kehtiv	223701
	kurnatud	382
	langenud	234
	läinud	146
	tulev	215938
	vastav	691647

Rühm	Sõna	Sagedus ÜK 2019-s
verbi partitsiibid	hakanud	40320
	keeranud	6090
	koosnev	67628
	nakatuv	104
	olev	920039
	osanud	13083
	pandud	141157
	tahtnud	26038
	tallav	237
	toimuv	193030
nimisõnad	eksam	24489
	elund	4005
	heli	30469
	kass	61422
	klots	5502
	naabrivalve	3665
	padi	9330
	pood	108046
	sajajalgne	289
	sülearvuti	9339
adverbid	kolmekesi	3847
	kõvemini	3078
	mürinal	280
	seaduslikult	2940
	selili	3760
	sõbralikult	4535
	tasapisi	12937
	võõrsil	7874
	üleöö	5403
	üsna	121304

Rühm	Sõna	Sagedus ÜK 2019-s
adjektiivid/substantiivid	haige	154376
	hull	126312
	kodutu	20479
	noor	934772
	rumal	92956
	teismeline	33366
	tuttav	175133
	tööealine	14881
	vastne	19893
	võõras	217971
adjektiivid/adverbid	alasti	13078
	korras	3765
	krussis	451
	lokkis	2283
	meeletu	68900
	purjus	32803
	päris	308801
	tohutu	135426
	täis	357997
	valmis	201257
käändumatud adjektiivid	armastatud	29446
	eht	679
	eri	264541
	indoeuroopa	1638
	isevärki	2224
	katoliku	19898
	lugupeetud	128289
	luteri	7606
	soome-ugri	16742
	vene	624958

Rühm	Sõna	Sagedus ÜK 2019-s
verbid	arvama	412011
	haisema	4573
	hoolitsema	29716
	itsitama	2312
	jooksma	145873
	kahetsema	12193
	oskama	139978
	sätendama	522
	ummistama	4325
	vestlema	16484
proadjektiivid	milline	1103060
	mingisugune	215827
	nihuke	1191
	niisugune	257266
	nisuke	1061
	sama	2828434
	seesugune	29039
	selline	4690920
	säherdune	4442
	taoline	186320
ordinaalid	esimene	3573597
	kaheksas	49516
	kolmas	905988
	kuues	101481
	kümnes	44710
	neljas	278150
	seitsmes	79926
	teine	5862469
	viies	174721
	üheksas	37092

HOW TO RECOGNIZE ADJECTIVES? AN ANALYSIS OF CORPUS PATTERNS

Maria Tuulik, Ene Vainik, Geda Paulsen, Ahti Lohk
Institute of the Estonian Language

This study was inspired by a survey of Estonian lexicographers (Paulsen, Vainik and Tuulik 2019), where the lexicographers expressed the need for a new digital tool that would facilitate word class identification for ambiguous cases. In the case of adjectives, the lexicographers emphasized the difficulty of determining if a verb participle has sufficient adjectival use to be included in dictionaries as an adjective.

In the article, we examine the morphosyntactic features characteristic of the adjective class and test different parameters in the corpora to differentiate adjectives from other word classes. We provide an overview of the test results of six parameters. In the study we analysed 12 groups of 10 words each. The test groups and test words were chosen manually, with consideration given to the problematic cases outlined by the lexicographers. We compared different types of adjectives or near to adjectives (the test groups) as well as different word classes (the control groups).

To analyse the parameters' capability to set adjectives apart, a deviation study was conducted. We determined a normative range for prototypical adjectives and set the minimum and maximum value for every parameter. In addition, we calculated the deviation of other test groups from the prototypical adjective range.

The groups of particular focus (regular verb participles vs. adjectives) were best differentiated by three parameters. The sentence beginning testword+noun parameter (which determined if and how often a test word starts a sentence in the complement position) sets participles apart with 90% accuracy. Also, the parameter that measured the existence of comparative forms for test words was 100% accurate. The adverb parameter (which measured how often a test word is preceded by an adverb) was able to distinguish adjectives from verb participles with 80% accuracy. Among all groups, the comparative form parameter was the most accurate in the deviation study at setting prototypical adjectives apart from other test groups.

A Euclidean distance analysis was able to differentiate adjective-like test words and test groups from others that do not behave similarly to prototypical adjectives.

As all tested parameters produced meaningful results and were able to differentiate some word classes from adjectives, they can be input for a new digital tool which would show a word's deviation from prototypical word class representatives to help lexicographers with word-class-related decisions.

Keywords: parts of speech, morphosyntax, lexicography, language technology, Estonian