Dimensionaalne tekstimudel

Teoreetiline ülevaade

KRISTIINA VAIK, KAIRIT SIRT, KADRI MUISCHNEK

Interneti võidukäiguga on kasvanud elektrooniliste tekstide hulk, mis ühest küljest on muutnud info kättesaadavaks, kuid teisalt on tekitanud nõudluse tarkvaralahenduste järele, mis oleksid suutelised automaatselt eraldama tekstikogudes peituvat vajalikku infot. Lihtsaim info eraldamise viis on tekstide klassifitseerimine, mis aga lihtsusest hoolimata kätkeb mitmesuguseid probleeme. Varem loodud tekstikorpustesse valiti tekstid esindamiseks eelnevalt kindlaksmääratud tekstiliigilist (žanrilist) jaotust, st tekstide tüüp või liik oli juba korpuse loomise käigus märgendatud. Praegused kolmanda põlvkonna tekstikorpused kogutakse aga automaatselt veebist ning nende jaoks puudub üldlevinud taksonoomia. Isegi kui luua mõni oletatavasti sobilik taksonoomia (nt artiklid, blogi-, foorumitekstid jmt), siis peale andmemahu kasvu on suurenenud ka tekstide variatiivsus: „klassikaliste” kirjalike tekstide kõrvale on tekkinud suur hulk kasutajate loodud sisu, mistõttu olemasolevate väga üldiste taksonoomiate alusel on automaatne märgendamine raske ning see ei pruugi tekstide kohta piisavalt infot anda. Lihtsustatult: pole täpselt teada, millised žanrid ja žanride hübriidid olemas on ja millised on neid eristavad tunnused. Käesolev artikkel üritabki välja pakkuda ühe võimaliku teoreetilise mudeli ja selle rakendamise meetodi nende küsimustega tegelemiseks.

Tekstide liigitamise teoreetiliste käsitluste analüüsil on probleemiks üheselt aktsepteeritud terminite ning definitsioonide puudumine (teemat on lähemalt käsitlenud nt Lee 2001; Mehler jt 2010; Crowston jt 2011). Erinevate lähenemiste ja terminoloogia rohkus on andnud põhjuse nimetada seda olukorda metafoorselt džungliks (Lee 2001; Sharoff 2010). Loomuliku keele töötluses, tekstitüpoloogias ja diskursuse analüüsis kasutatakse tekstide liigitamisel peamiselt mõisteid *stiil*, *tüüp*, *liik* ehk *žanr* või *register*, kuid puudub konsensus, kuidas neid mõisteid defineerida; mille alusel otsustada, kas konkreetne keelevariant on tüüp, register, žanr või hoopis midagi neljandat. Terminoloogiline segadus keerleb peamiselt mõistete *žanr* ja *register* ümber, mis võib olla tingitud koolkondlikest erinevustest: mõned uurijad eristavad rangelt neid kaht (nt Lee 2001; Eggins, Martin 1997; Ferguson 1994) ning on uurijaid, kes on kasutanud neid mõisteid läbisegi (nt Biber 1994, 1995). Eri mõistete kasutamine ja (mitte)eristamine on põhjustanud olukorra, kus heal lapsel on mitu nime, st sarnase nähtuse kirjeldamiseks on kasutatud erinevaid termineid või definitsioone. Seetõttu on oht neid mõisteid väärti mõista või isegi omavahel segi ajada, kuna definitsioonide piirid on abstraktsed ja hägused.

Korpusetekstide liigitamisel on võimalik lähtuda ka keelevälise ja -sisese kriteeriumi mõistetest (Sinclair, Ball 1996). Nende vahele on võimalik tõmmata rangelt eristav joon, kus keeleväline kriteerium on defineeritud teksti funktsiooni ning

keelesisene kriteerium keeleliste tunnuste kaudu. Keelesisene kriteerium on keeleline nähtus, nt otsustamaks, kas konkreetne tekst on formaalne või mitteformaalne, tuginetakse teksti keelelistele tunnustele. Keeleväline kriteerium peegeldab teksti funktsiooni või eesmärki, nt ilma arvamuseritust lugemata võime teha eeldusi selle artikli eesmärkide kohta. Kuigi keelevälised ja -sisesed kriteeriumid sõltuvad üksteisest, on keeleväliste kriteeriumide alusel tehtud taksonoomiaid aja jooksul tekkinud mitmeid (nt BNC, Brown, MASC jm), kuid sisemiste kriteeriumide alusel loodud taksonoomiaid laialdaselt aktsepteeritud ei ole. Suletud korpusse ajastul loodi korpusi keelevälise kriteeriumi alusel, kuna puudus arvutuslik võimekus tekste sisemistest kriteeriumidest lähtuvalt analüüsida ning väidetavalt ei suuda sisemised kriteeriumid üksi anda edasi infot keele ja selle kasutamise konteksti kohta (Atkins jt 1992). Nüüdisajal koostatakse suuri keelekorpusi peamiselt veebi kraapimise teel (keeleväline kriteerium). Kuid kuna on kasvanud keeletehnoloogiliste tarkvaralahenduste hulk, mille abil tekste keelelistest tunnustest lähtuvalt analüüsida, siis tegelikkuses puuduvad takistused liigitada tekste sisemiste kriteeriumide alusel.

Käesolevas artiklis pakume välja tekstide liigitamise dimensionaalse mudeli teoreetilised alused, lähtudes sisemiste ja väliste kriteeriumide eristusest. Dimensionaalse tekstimudeli põhiidee seisneb tekstide kategoriseerimises sisemiste kriteeriumide alusel, nii et tekstid, mis jagavad samasuguseid sisemisi kriteeriume, on ka funktsiooni (ehk väliste kriteeriumi) järgi sarnased. See on süntees Douglas Biberi multidimensionaalsest analüüsist (Biber 1986, 1988) ning Serge Sharoffi funktsionaalsetest tekstidimensioonidest (Sharoff 2018).

Biberi ja Sharoffi lähenemised jagavad sarnast ideed: keelelistel tunnustel on funktsioon ning funktsiooni väljendatakse keeleliste tunnuste kaudu. Kuid nende eesmärgid on erinevad: Biberit huvitavad žanre eristavad keelelistes tunnused ja see, kuidas need tunnused kogunevad dimensioonideks, Sharoffi siht on eelkõige rakenduslik – luua veebitekstide automaatseks klassifitseerimiseks sobiv tüpoloogia. Kuna siinses artiklis loodame need kaks eesmärki ühitada, st luua tekstide žanrilise klassifitseerimise raamistik ja ühtlasi saada uusi teadmisi žanre eristavate tunnuste kohta, siis on nii Biber kui ka Sharoff meile olulised. Artiklis välja pakutav dimensionaalne tekstimudel püüab neid lähenemisi lõimida, võttes Biberi lähenemisest kaasa tekstisisese kriteeriumi koos dimensiooni mõistega – dimensioon on keeleliste tunnuste komplektide alusel mõõdetav parameeter – ning Sharoffi lähenemisest tekstide hübriidsuse idee, mille järgi tekste ei klassifitseerita binaarselt (kuulub või ei kuulu), vaid need võivad samaaegselt kuuluda mitmesse žanri. Eesmärk on pakkuda välja universaalne mudel, mis otseselt ei sõltuks žanrilisest kategoriseerimisest ega defineerimisest, vaid annaks raamistiku tekstide liigitamiseks keeleliste tunnuste abil kirjeldatavate omaduste ehk dimensioonide kaudu. Selline universaalsus tagab võimaluse lisada uusi tunnuseid ning rakendada mudelit uutele andmekogudele ilma vajaduseta muuta all olevat teoreetilist mudelit.

1. Tekstilistest dimensioonidest lähtuvad varasemad tööd

Dimensionaalne tekstimudel lähtub Biberi (1986, 1988) multidimensionaalsest analüüsist ning Sharoffi (2018) funktsionaalsete tekstidimensioonide teooriast. Järgnevalt anname ülevaate mõlema lähenemise eesmärkidest, raamistike ülesehitusest ning olulistest mõistetest.

1.1. Multidimensionaalne analüüs

Multidimensionaalne analüüs (MDA) on korpuspõhine metodoloogiline lähenemine, mis kasutab faktoranalüüsi selleks, et võrrelda erinevaid tekste keeleliste parameetrite komplektide ehk dimensioonide kaudu. Oluline on mainida, et MDA raamistikus on tekstide žanriline kuuluvus ette teada ning eesmärk pole tekstide liigitamine, vaid žanre eristavate dimensioonide kui suhtlusfunktsioonide keelelise väljenduse komplektide uurimine. Douglas Biberi (1985, 1986) 1980. aastate keskpaigas rajatud MDA kulmineerus mõned aastad hiljem suuremahulise metodoloogilise uurimusega varieerumisest kõnes ja kirjas (1988).

MDA põhirõhk oli algselt kirjaliku ja suulise keelekasutuse erinevuste kvantitatiivsel uurimisel, mille kohta Biberi (1986: 384–387, 1988: 52–53) arvates varasemad tööd ei suutnud pakkuda adekvaatseid üldistusi, kuna toetusid vaid vähestele tekstidele, suulise ja kirjaliku keele mitmekesisest žanrihulgast kasutati võrdluseks vaid paari väljavalitut, tüüpiliselt üht suulist ja üht kirjalikku. Samuti põhinesid Biberi kritiseeritud kvantitatiivsed uuringud liiga vähestel keelelistel tunnustel ning eba-proportsionaalne oli ka kirjutajate ja kõnelejate valik: üldjuhul esindas inglise keelt keskklassi kuuluva akadeemilises maailmas tegutseja keelekasutus. Lisaks põhines enamik kritiseeritud uuringuid ainult ingliskeelsetel tekstidel. Biberi eesmärk oligi vältida varasemate uurimuste kitsaskohti, rakendades kvantitatiivset analüüsi, kasutades heterogeensemata ja suuremahulisemat korpus, suuremat hulka keelelisi tunnuseid ning rakendada oma uudset lähenemist ka teistele keeltele.

Multidimensionaalses analüüsis on tähelepanu kahel mõistel: *dimensioon* ja *keeleliste tunnuste koosinemine*. Dimensioon on keelelist variatsiooni mõõtev parameeter, mis põhineb keelesisestest kriteeriumide ehk leksikaalsete ja grammatiliste tunnuste koosinemise mustritel. Dimensioonidel on samaaegselt nii keeleline kui ka funktsionaalne sisu: keeleline sisu avaldub esilduvamate keeleliste tunnuste tekstides koosinesinamisena, kuid tunnuste koosinemine pole juhuslik, vaid need täidavad ühist kommunikatiivset funktsiooni või funktsioone: näiteks otsekõne, personaalpronoomenite ja imperatiivivormis verbide koosinemine viitab vahetule suhtlusele, noomenifraaside rohkus infotihedale tekstile jm. Biberi väitel ei piisa dimensioonide omavaheliste sarnasuste ja erisuste kirjeldamiseks ühest või paarist keelelistest tunnusest. MDA eesmärgiks ongi kaasata võimalikult suur hulk potentsiaalselt olulisi tunnuseid. Nii tagatakse see, et dimensioonid põhinevad sellistel keelelistel tunnustel, mis tegelikes tekstides koos esinevad, mitte neil, mida uurija oletab mingi suhtlusfunktsiooni täitmisel kasutatavat. Dimensioon näitab mingi tekstilise omaduse määra kindla žanri tekstides: näiteks dimensiooni *abstraktne* skaala ühes otsas

on teaduskirjandus, ametlikud dokumendid ning teises otsas telefonivestlused, romantiline kirjandus jm. Faktoranalüüsi rakendamisel jagunevad tekstid faktoritesse ning seejärel toimub faktorite ehk dimensioonide induktiivne tõlgendamine nende suhtluseesmärgi põhjal. Olgugi et koosinemise mustrid on saadud kvantitatiivselt, on dimensiooni tõlgendamine faktori põhjal siiski eksploraatiivne ehk otsinguline. (Biber 1988: 12–13, 55)

Biber (1988) analüüsis Lancasteri-Oslo/Bergeni kirjaliku inglise keele korpust (vt LOB) ja Londoni-Lundi suulise inglise keele (vt LLC) korpust 67 keelelise tunnuse jaotuse alusel. Faktoranalüüsi tulemuste interpreteerimisel saadi inglise keeles seitse dimensiooni¹ (vt lisaks Biber 1988, 1995; eesti keeles on dimensioone kirjeldanud ja nende üle pikemalt arutlenud Tiit Hennoste (2000)): D1 informatiivne vs. kaasav tekst; D2 narratiivne vs. mittenarratiivne tekst; D3 eksplitsiitne, kontekstist sõltumatu viitamine vs. mittespetsiifiline, situatsioonist sõltuv viitamine; D4 (avalik, teksti pinnal väljenduv) veenmine; D5 keelekasutuse abstraktsuse määr; D6 info edastamine reaalselt; D7 akadeemiline argumentatsioon koos nn pehmenusstrateegiatega (ingl *hedging*). Neist esimesed viis on kesksed ning viimased kaks on perifeersed dimensioonid ega pruugi igas korpuses esineda. Peale inglise keele on multidimensionaalset analüüsi rakendatud ka paljudele teistele keeltele, nt somaali (Biber, Hared 1992), Nukulaelae tuvalu (Besnier 1988), Brasiilia portugali (Sardinha jt 2014), hispaania (Biber jt 2006; Parodi 2007), Ameerika inglise (Passonneau jt 2014; Grieve jt 2011; Grieve 2014), mandariini (Jang 1998), korea (Kim, Biber 1994), vene (Katinskaia, Sharoff 2015), gaeli (Lamb 2002), dagbani (Purvis 2008) ja pakistani keelele (Shakir, Deuber 2019).

Inglise keeles esildub kesksedest dimensioonidest peamiselt D1, mille üht poolust esindab informatsioonitihedus ning teist afektiivne, suhtluslik ja üldine sisu. Selles dimensioonis on kaks peamist parameetrit: a) kõneleja või kirjutaja eesmärk on informatsiooni edastamine või suhtlus kui niisugune; b) loodavat teksti on võimalik toimetada (tüüpiliselt kirjalikku teksti on) või toimub tekstiloomes reaalselt (tüüpiliselt suuline suhtlus). Ka kõigis teistes MDA meetodil analüüsitud keeltes esilduvad väga tugevalt dimensioonid, kus tekib suulise vs. kirjaliku registri vastandus. Teine tugevalt esilduv keskne dimensioon nii inglise kui ka teistes keeltes on narratiivsus. Narratiivsuse tunnused seostuvad minevikus toimunu kirjeldamisega ning toetavad narratiivse teksti eesmärki: edastada sündmusi nende toimumise järjekorras.

1.2. Funktsionaalsed tekstidimensioonid

Sarnaselt Biberiga on žanrilise klassifitseerimisega tekstide dimensionaalsuse abil tegelenud ka Serge Sharoff (2018). Tema raamistik tugineb funktsionaalsetele tekstidimensioonidele (FTD), kus iga dimensioon esindab funktsionaalset kategooriat², mis kirjeldab teksti suhtluseesmärki. FTD sai alguse veebikorpuste liigitamisest, kus

¹ Mõistet *dimensioon* tähistatakse nüüd ja edaspidi sünonüümselt tähega D.

² Dimensioon ehk funktsionaalne kategooria on väga sarnane tüüpiliste žanriliste taksonoomiate kategooriatega. Sharoff (2018: 9) kasutab küll mõistet *dimensioon*, kuid peab selle all siiski silmas žanri kui keelekasutusviisi, millel on oma kommunikatiivne eesmärk või funktsioon.

puudub eelnev teadmine teksti žanrilise või registrilise kuuluvuse kohta ning on vaja arvestada ka tekstide hübriidsusega. Tüüpilised automaatsed klassifitseerimiskatsed viimasega ei arvesta, kuigi nt *artiklid* ja *blogid* võivad olla väga mitmenäolised ning seega ei saa olla kindlapiirilised žanrid.

Sharoffi lähenemises pannakse žanriline taksonoomia paika tekstiväliste kriteeriumide alusel. Taksonoomia luuakse testküsimuste abil, kus iga küsimus tähistab n-ö prototüüpset žanri ning inimmärgendaja annab igale küsimusele vastuse 4-punktilisel Likerti skaalal (dimensioon ei esildu, esildub vähesel määral, esildub osaliselt, esildub tugevalt). Testküsimused on esitatud ühetaolisel kujul ja paluvad hinnata erinevate omaduste määra tekstis, nt *Mil määral sisaldab tekst ilmset argumentatsiooni lugeja veenmiseks?* või *Mil määral tekstis paistab osalevat mitu osapoolt?* jne. Sellise taksonoomia loomisega püütaksegi arvesse võtta tekstide hübriidsust, sest testküsimused hindavad, kui tugevalt iga dimensioon tekstis esineb.

Taksonoomia loomisega tekkinud tekstikorpust kasutatakse hiljem masinõppe mudeli treenimiseks, mille abil saab klassifitseerida uusi tekste eelnevalt loodud taksonoomia kategooriate alusel, mh eesmärgiga välja selgitada igale hübriidžanrile iseloomulikud tekstisisesed kriteeriumid. Sharoffi väitel on FTD universaalne, st rakendatav kõikidele tekstidele või tekstikorpustele; lihtne, st võrreldes enamiku žanriliste taksonoomiatega iseloomustab FTD-d vähene parameetrite ja kategooriate hulk; ning üldistatav, st erinevad märgendajad suudavad selle taksonoomia alusel tekste sarnaselt kategoriseerida (ehk hindajatevaheline suur kooskõla). Olulise uuendusena võtab FTD kasutusele distantseerimise mõõdiku: tekste esitatakse funktsionaalsete dimensioonide ruumis ja nii on võimalik mõõta mis tahes teksti kaugust mingit žanri n-ö prototüübina esindavast tekstist; kui tekst paikneb mitmemõõtmelises ruumis prototüüpsetest tekstidest kaugemal, siis on võimalik, et tegu on hübriidse tekstiga.

Kuidas aga tekkis žanride loend? Sharoff võttis alustuseks John Sinclairi ja Jackie Balli (1996) kommunikatiivsete eesmärkide tüpoloogiast üle kuus tekstikategooriat ning laiendas need kaheksaks funktsionaalseks žanriklassiks (mida ta nimetaski *Functional Genre Classes*, Sharoff 2010). Neist aga ei piisanud tekstide märgendamiseks, nii et mitme märgendamiskatse (peamiselt Sharoff 2010; Forsyth, Sharoff 2013; Sorokin jt 2014; Katinskaia, Sharoff 2015) tulemusena jõuti lõpliku 18 dimensioonist koosneva taksonoomiani. Need jagunevad omakorda esmasteks ja teisteks dimensioonideks. Esmased dimensioonid on funktsionaalselt motiveeritud ja jaotatud nelja suuremasse rühma:

- I. informatiivsed dimensioonid, mille prototüüpsed esindajad on *instruktsioonid* (nt käsiraamatud, retseptid), *uudised*, *juriidika* (nt lepingud, seadused), *infotekstid* (nt CV, entsüklopeedia);
- II. dimensioonid, mille põhisisu on diskussioon: *argumentatiivsed tekstid* (nt juhtkirjad, poliitilised debadid), *teadustekstid* (nt esseed, artiklid) ja *hinnangulised tekstid* (nt arvustused);

III. narratiivsed dimensioonid: *ilukirjandus, isikliku sisuga tekstid* (nt päevikulaadsed blogid, kirjad) ja *poetiliselt tekstid*;

IV. dimensioonid, mille eesmärk on edastada reklaamilaadset informatsiooni: *reklaamtekstid, propaganda, apellatiivsed tekstid (adressaadile suunatud)*.

Vähemalt ühes esmases dimensioonis peab iga tekst saama nullist suurema väärtuse. Teised dimensioonid on selleks, et esitada esmaste dimensioonide sisest variatiivsust. Võib juhtuda, et kaks teksti on funktsioonilt samasugused (ehk saavad esmastes dimensioonides samasugused väärtused), kuid need võivad olla stiililt erinevad. Sekundaarsete dimensioonide kaasamisega proovitaksegi tabada need väikesed erisused. Sekundaarsed dimensioonid on *emotsionaalsus, meelelahutuslikkus, mitteformaalsus* (keelega hõlbimine standardsest keelekasutusest), *erialalisus* (teksti mõistmine nõuab erialalist taustainformatsiooni) ja *dialoogilisus*.

1.3. MDA ja FTD kitsaskohad

MDA on oluline teoreetiline lähenemine, kuna enne seda ei olnud süstemaatilist viisi žanride omavahelise ja sisemise varieeruvuse võrdlemiseks keeleliste parameetrite kaudu. Erinevalt MDA-st, mis uurib teadaolevate žanride keelekasutust ja suhtlusfunktsioone, sai FTD alguse vajadusest klassifitseerida veebikorpust, mille kohta pole teada, missugusesse žanri või registrisse mingi tekst kuuluda võiks. FTD-d ja MDA-d eristab see, et MDA tugineb tekstisisestele kriteeriumidele ehk keeleliste tunnustele, kuid FTD tekstivälisele kriteeriumile ehk kommunikatiivsele funktsioonile.

MDA puhul kritiseeritakse (nt McEnery, Hardie 2012: 112–114) põhiliselt uuringu aluseks olevat korpust ning uuringu mittekorduvust. Korpuse osas on ette heidetud selle väikest mahtu ning terviklike tekstide asemel lühemate tekstisegmentide kasutamist. Hilisemad püüdlused tulemusi korrata pole andnud samasuguseid tulemusi, mis näitab seda, et MDA-ga ei ole võimalik teha tekstide variatiivsuse kohta universaalseid järeldusi, vaid dimensioonide tõlgendamine on otsinguline ning korpusepõhine. Tunnuste valiku osas on kritiseeritud seda, et toetudes varasemate uurijate töödele ja/või grammatikakirjeldustele, võivad mõned olulised tunnused märkamata jääda.

Ühe kindla korpuse analüüsist alguse saanud MDA pole üdini universaalne teooria, mida näitab ka see, et eri keelte MDA tulemused on saanud erinev arv dimensioone, mis on omakorda tingitud kasutatavatest tunnustest ja korpusest. Näiteks Ameerika inglise keele analüüsiks kasutasid Rebecca Passonneau jt (2014) ning Jack Grieve (2014) enam-vähem samasuguseid tunnuseid (erinevusega, et esimeses võeti kasutusse ka nimeüksused), kuid erinevaid korpust, ning nende MDA käigus tekkinud dimensioonid erinevad mõnel määral. Passonneau jt (2014) võtsid aluseks MASC-i korpust,³ milles MDA tuvastas neli dimensiooni;⁴ Grieve (2014) kasutas

³ 1990. aastate Ameerika inglise keele korpust, mis koosneb poolest miljonist sõnast.

⁴ D1 esindab suulist vs. kirjalikku diskursust (ingl *involved* vs. *informational*), D2 ühel pool on tugevaks tunnuseks nimeüksused ning teisel pool üldnimed (ingl *name entities* vs. *common*)

ajakirjatoimetusele saadetud kirjade korpus⁵, millest MDA tuvastas kolm dimensiooni.⁶ Keeled ja tunnused (v.a nimeüksused) on küll samad, kuid korpused ja seega ka tulemused on erinevad: saadi ainult kaks sarnast dimensiooni, millest üks sarnaneb Biberi dimensiooniga *informatiivne* vs. *kaasav* (ühel pool on suur informatiivne suhtluslik ja üldine sisu) ning teine Biberi dimensiooniga *narratiivne* vs. *mittenarratiivne*. Lisaks, Passonneau'lt (2014) esildus andmestikust dimensioon *nimeüksused* vs. *üldnimed*, mida kusagil varem nähtud pole ja mis oli otseselt tingitud tunnuste valikust.

Ühe probleemina tooksime omalt poolt välja dimensioonide hägususe või hübriidsuse: paljud dimensioonid on segu pisematest „alldimensioonidest”. See hägusus on osaliselt faktoranalüüsi tõlgendamise tulemus, kuna faktoranalüüsil tekib tunnuste arvust väiksem hulk faktoreid, mis peegeldavad ühist nähtust või aspekti. Kuid selle nähtuse või aspekti nägemine on oskus omaette ning sellisena küllaltki subjektiivne. Näiteks on igas keeles leitud tunnuste komplekt, mis väljendab kirjutaja isiklikke veendumusi, hoiakuid ja tundeid, kuid paraku mõnes (nt Ameerika inglise (Grieve 2014); soomaali (Biber, Hared 1992)) keeles ei esildu see iseseisva dimensioonina, vaid on suurema konstrukti osa või on jagatud mitme dimensiooni vahel, nt Brasiilia portugali keeles (Sardinha jt 2014) kajastub kirjutaja meelsus isegi kolmes erinevas dimensioonis.

FTD eesmärk oli MDA-s kasutatavaid väliseid kriteeriume (ehk žanre) destruktureerida ning pakkuda stabiilsemat korpusete (ka kogu veebi võib vaadelda korpusena) tekstivälise kriteeriumide alusel tehtud taksonoomiat. Kui tavaliselt žanrilisel klassifitseerimisel eiratakse tekstide hübriidsust (vt Santini 2007; Crowston jt 2011), siis FTD-s arvestatakse sellega. Selline tekstide ruumis modelleerimine funktsiooni alusel on leidlik lahendus, ent siiski ei vaadata kõiki tekste samamõõtmelises dimensionaalses ruumis. Nimelt ei pruugi FTD-s tekstid saada igas dimensioonis nullist suuremat väärtust. Need dimensioonid, mis saavad nullväärtuse, lülitatakse ruumist välja. See tähendab, et kui nt *tekst1* saab nullist suurema väärtuse dimensioonides D1 ja D2, aga *tekst2* ainult dimensioonis D1, siis *tekst1* ja *tekst2* dimensionaalne ruum on erinev ehk terviklikku 18-dimensionaalset ruumi ahendatakse vastavalt sellele, kui mitmes dimensioonis sai tekst nullväärtuse. Ja ikkagi jääb kahtlus, kas FTD erinevate žanriliste taksonoomiate põhjal koostatud dimensioonide (= žanride) loend suudab katta kogu teadmata sisuga veebikorpuset.

nouns), D3 põhineb narratiivsel diskursusel (ingl *narrative*), D4 põhineb abstraktsel ja informatiivsel diskursusel (ingl *abstract*).

⁵ Aastatel 2000–2012 saadetud kirjad, kokku üle 26 miljoni sõna.

⁶ D1 on ühelt poolt struktuurilt lihtne ja mitteformaalne ning teisalt üksikasjalik ja formaalne (ingl *simple* vs. *elaborate writing style* / *informal* vs. *formal style*), D2 põhineb narratiivsel vs. kirjeldaval diskursusel (ingl *narrative* vs. *expository*), D3 on seotud argumentatsiooni ja veenmisega (ingl *opinionated* vs. *informational style*).

2. Dimensionaalne tekstimudel

Siinses artiklis väljapakutav dimensionaalne tekstimudel (DTM) kombineerib MDA ja FTD ideid, hoidudes samal ajal mõlema kohta välja toodud kitsaskohtadest. MDA-st mugandatakse tekstisisene kriteerium koos dimensiooni mõistega, st tekstide kirjeldamiseks olulised dimensioonid avalduvad keeleliste tunnuste kaudu. FTD-st võetakse üle tekstide hübriidsuse idee, st iga tekst võib omada väärtust mitmes dimensioonis korraga, kuid DTM-is laiendame seda ideed nii, et tekst peab omama väärtust kõikides dimensioonides ehk nii tekib terviklik tekstidimensioonaalne ruum. Kui MDA ja FTD järgi on dimensioonidel samaaegselt nii keeleline kui ka funktsionaalne sisu ehk teksti funktsioon avaldub keeleliste tunnuste kaudu, siis DTM-i järgi dimensioonil endal funktsiooni ega eesmärki ei ole, vaid see kirjeldab teatud omaduse määra ehk skaalat. DTM-is tuleneb keeleväline kriteerium dimensioonide koosesinemise mustritest – tekstid, mis paiknevad mitmemõõtmelises ruumis üksteisele lähedal, jagavad ka sarnast funktsiooni. Seega, DTM on üldine raamistik, mille alusel saab tekste keelesiseste ja -väliste kriteeriumide kaudu uurida ja liigitada.

DTM on hierarhiline mudel (vt ka joonist 1), mille kolmel tasandil paiknevad selle kolm põhikomponenti.

I. Mudeli alumisel tasandil asuvad sisemist kriteeriumi kajastavad tunnused, mis on otseselt tekstist mõõdetavad, nt nimisõnade arv, sõnavara suurus, relatiivlausete hulk, abstraktsete sõnade hulk jm. Tunnused võivad olla keelteülesed või ka mingile konkreetsele keelele eriomased.

II. Mudeli keskmisel tasandil asuvad latentsed dimensioonid, mis on kirjeldatavad tunnuste komplektide kaudu. Kõik tekstid saavad väärtuse igas dimensioonis, aga erinevus tuleneb omaduse määra, st mõnes tekstis on mingi omadus tugevalt väljendunud, teises keskmiselt ja kolmandas üldsegi mitte. Dimensioonid ise on keelteülesed, kuigi tunnustekomplektid, mille kaudu dimensioonid väljenduvad, võivad olla keelespetsiifilised.

III. Kõige ülemisel tasandil asuvad žanrid,⁷ mille funktsioone või eesmarke dikteerib „käesolev” situatsioon. Žanre defineerivad dimensioonide väärtusmustrid. Tekstid, mille dimensioonide avaldumismuster on sarnane, jagavad samasuguseid funktsioone ning seega võiksid kuuluda ühte žanri.

DTM-i dimensioonid vastavad järgmistele tingimustele:

- I. dimensiooni aluseks on tekstides esinev tuumiknähtus või -omadus, mida saab kirjeldada keelespetsiifiliste tunnuste kaudu;
- II. dimensiooni väärtus on pidev, see pole dihhotoomne;

⁷ Kasutame mõistet *žanr* tinglikult, sest ei erista omavahel žanri, registrit ega tüüpi, vaid laseme teistel uurijatel vastavalt vajadusele mõisteid valida ja defineerida.

III. dimensioon on iseseisev konstrukt, seega dimensioonid on üksteisest sõltumatu.⁸ Dimensioon üksi ei suuda kirjeldada ühtegi žanri, žanrid koosnevad dimensioonide esilduvusmustritest. Küll aga on dimensioon võimeline kirjeldama üht žanri omadust (nt teadusteksti keelekasutust iseloomustab keerukus).

Joonis 1. Dimensionaalne tekstimudel ja selle komponendid.

DTM-i dimensioonid pakkusime MDA ja FTD analüüsi ja kombineerimise tulemusena. Kõigi kirjanduses mainitud keelte MDA meetodiga saadud dimensioonid rühmitasime nii, et sõelale jäid need, mis on neis keeltes ühised, ja välja jätsime keele- ja kultuurispetsiifilised dimensioonid (nt korea keelele omase lugupidamise väljendamine (ingl *honorification*) dimensiooni). Enamik MDA dimensioone on suuremad omaduste konstruktid, seega MDA dimensioonid on destruktureeritud selliselt, et iga MDA dimensioon on kirjeldatav ühe või mitme DTM-i dimensiooni kaudu. Näiteks esildub kõigis MDA meetodil analüüsitud keeltes dimensioon, mis eristab tugevalt kirjalikku ja suulist diskursust, kus ühel pool on rõhk informatsiooni edastamisel, teisel suhtlusel kui niisugusel. Seega olulisteks tunnusoonteks on informatsiooni edastamine, interaktiivsus (üks vs. mitu osalejat), afektiivsus ja spontaansus. Lisaks, näiteks dimensiooni *narratiivne–mittenarratiivne* pooluseid tugevalt eristavateks joonteks on (üldjuhul) tegevuse ajaline lineaarsus ning tegevussubjekti eksplitsiitne väljendamine.

FTD dimensioonid kõrvutasime MDA dimensioonidega selliselt, et iga MDA dimensioon oleks kirjeldatav FTD dimensioonide kaudu. Selle tulemusena eristusid sellised FTD dimensioonid, mis kattusid üksüheselt MDA dimensioonidega (nt *emotsionaalsus*, *instruktsioonid*, *infotekstid*, *mitteformaalsus*), ja sellised, mis olid

⁸ Kuigi osa dimensioone võib teatud olukordades omavahel korreleeruda.

suuremad konstruktid väiksematest omadustest (nt *reklaamtekstid*, *apellatiivsed tekstid*, *meelelahutuslikkus*, *ilukirjandus* jm). Kuna MDA dimensioonide omadused on hägused või hübriidsed ning FTD dimensioonid väga žanrikesked ja mitte üheselt mõõdetavad, siis MDA ja FTD dimensioonide sellisel viisil kombineerimisel lõime dimensioonide komplekti, kus iga dimensioon on iseseisev konstrukt ning kirjeldab vaid üht mõõdetavat omadust. Järgnevas alapeatükis kirjeldame iga DTM-i dimensiooni detailsemalt.

2.1. DTM-i dimensioonid

Dimensionaalne tekstimudel võimaldab vaadelda tekstide hübriidsust ja variatiivsust, selmet tekste paigutada rangepiirilistesse kategooriatesse, nagu seda tavaliselt tekstide klassifitseerimisel tehakse. Pakume järgnevalt välja komplekti dimensioone,⁹ mis põhinevad MDA ja FTD dimensioonide kombineerimisel:

D1 abstraktsus	D7 impersonaalsus
D2 afektiivsus	D8 ajalise olulisus
D3 instrueerivus	D9 interaktiivsus
D4 informatsioonitihedus	D10 subjektiivsus
D5 spontaansus	D11 keerukus
D6 formaalsus	D12 argumentatiivsus

D1 abstraktsuse skaala ühes otsas on tekstid, mis esitavad entiteete, mida pole võimalik meeltega vahetult kogeda (nt *demokraatia*, *armastus*), ja teises otsas konkreetsed, meeltega vahetult tajutavad esemed (nt *laud*), mille vahele paigutuvad meeltega tajutavad nähtused (nt *tuul*) ja tegevused (nt *ehitama*). Abstraktsus avaldub tekstides peamiselt leksikaalsel tasandil. Dimensiooni näitlikustamiseks võrdleme kaht tekstilõiku:

(1) Gödel kodeeris mittetäielikkuse tõestamiseks formaalse aksiomaatika aritmeetikasse. Nimelt saab kogu nimetatud formaalse süsteemi ja kõik väited esitada aritmeetika enda teoreemidena, s.t teoreemidena täisarvude kohta. Seega õnnestub kirja panna aritmeetikateoreem A, mille sisuline tähendus formaalses süsteemis on, et seesama teoreem A ei ole aritmeetika aksiomaatikast tõestatav. Sellest näiliselt ainult aritmeetikasse puutuvast spetsiifilisest teoreemist järeldub, et ühtegi piisavalt keerulist matemaatilist süsteemi ei saa lõpliku hulga aksiomide abil täielikult aksiomatiseerida. (etTenTen kalah.zzz.ee, doc id = 133615)

(2) Elas kord kapsauss, kellele üle kõige maailmas meeldis hapukapsas. Kõik oma kaks elupäeva oli ta unistanud hapukapsast ja öösel sellest undki näinud. Lõpuks ei pidanud ta enam vastu ja hakkas läheduses oleva maja poole roomama. (etTenTen www.poogen.ee, doc id = 679414)

⁹ Dimensioonide nimekiri on esialgne, kuna võib juhtuda, et teatud dimensioonid osutuvad liiasteks, kuid see selgub alles edasise empiirilise töö käigus.

Näide 1 on lõik matemaatilist loogikat käsitlevast tekstist, näide 2 jutustuse katkend. Näide 1 sisaldab abstraktseid sõnu ning väljendeid (nt *mittetäielikkus*, *aksiomaatika*, *matemaatiline süsteem*, *lõplik hulk*), sh ka abstraktse tähendusega verbe (nt *kodeerima*, *esitama*, *tõestama*). Näide 2 on konkreetsem, lugejale hoomatavam, esitades hästi tuntud objektide (nt *kapsauss*) konkreetseid tegevusi (nt *roo-mama*).

D2 afektiivsus mõõdab tundeid ning emotsionaalseid reaktsioone tekstis. Tekstide afektiivsuse taset määratakse peamiselt leksikaalsete üksuste afektiivsuse alusel, kusjuures ka esmapilgul neutraalsena tunduvad sõnad võivad esile kutsuda nii häid kui ka halbu emotsioone, nt *suvi* ja *päikesepaiste* seostuvad pigem positiivse emotsionaalsusega. Emotsionaalset suhtumist väljendatakse tekstides ka kirjavahe-märkidega (?) ning süntaktilisel tasandil hüüd- ja küsilauseetega (nn retoorilised küsimused). Afektiivsuse dimensiooni näitlikustamiseks vaatame kaht järgnevat näidet:

(3) Olles näinud igasugust rõvedust, ületab see foto oma sündsuse piirid ikka mitmekordsest.. mul ei jätku sõnu kirjeldamiseks seda õõvastavat vaatepilti.. see on midagi enneolematut, seni nägematut ja vaevast, et kunagi enam midagi sellist juhtuda lastakse.. vot nii sündsusetu on see pilt! kuidas ansipi valitsus sellel juhtuda lasi? (etTenTen publik.delfi.ee, doc id = 28979014)

(4) Politoloogia magistriõppe tudengitel on võimalik teoreetiline seminar läbida endistel alustel, ehk saada 1AP semestri jooksul kohalkäimise eest. Kõikide teiste erialade tudengid (kellel teoreetiline seminar on valikaine) on kohustatud ainet valides täitma allpool nimetatud tingimused. (etTenTen riigiteadused.tlu.ee, doc id = 4383412)

Näide 3 pärineb Delfi kommentaariumist, näide 4 on katkend Tallinna Ülikooli riigiteaduste instituudi kodulehelt. Näide 3 on selgelt afektiivsem: afektiivsus avaldub leksikaalsel tasandil, nt sõnades *rõvedus*, *õõvastav*, *enneolematu*, *seninägematu*, *sündsusetu*, morfoloogilisel tasandil määra- või viisimäärsõnade kasutamisenä, hüüu- ja küsimärkide kasutamisenä jm. Näites 4 on afektiivsus väga madal, selles ei peegeldu kirjutaja emotsioonaalne või tundmuslik reaktsioon.

D3 instrueerivus annab teksti lugejale juhiseid teatud toimingute tegemiseks või kirjeldab mingi protsessi toimumise etappe, vt näidet 5. Instrueerivus avaldub peamiselt morfoloogilisel ja süntaktilisel tasandil (nt käskivas kõneviisis verbid, ka *da*-infinitiivsed ja impersonaalsed öeldised, käsklused).

(5) võta toorjuust ja või ca tund varem kapist välja pehmenema. pane ahi soojenema 180 kraadi peale. sega või suhkruga, lisa toorjuust ja klopi ühtlaseks (ei pea vahtu kloppima). lisa ükshaaval munad, segades iga muna enne järgmise lisamist korralikult taighasse. püreeri banaanid ja lisa eelnenule. sega omavahel jahu, küpsetuspulber ja sooda ning sõelu taighasse. (etTenTen www.pilleriin.ee, doc id = 64846)

See on lõik retseptist, kus lugejale antakse juhendeid, missuguses järjekorras eesmärgi saavutamiseks konkreetseid tegevusi peab tegema. Peale retseptide võivad instrueerivad tekstid olla nt käsiraamatud, juhendid, õppematerjalid, ka ravimite infolehed.

D4 informatsioonitihedus mõõdab tekstis esineva informatsiooni hulka. Informatsioon koosneb semantiliste üksuste võrgustikust, mida tüüpiliselt esitatakse väitlauses esinevate propositsioonide abil. Informatsioonitihedus mõõdabki tekstis sisalduvate propositsioonide ning propositsiooni predikaadiga seotud argumentide ehk osaliste hulka, nt *poiss jookseb poodi* on informatiivsem kui *poiss jookseb*. Informatsioonitihedus avaldub peamiselt leksikaalsel ja süntaktilisel tasandil, väljendudes leksikaalse rikkuse ning atribuudiühendite nimisõnafraasidena, nimisõnade ning arvsõnade või arvude suure osakaaluna tekstis. Peale entsüklopeediliste kirjade võib informatsioonilisust suurel määral esineda avalik-õiguslikes, teadus- ja ajakirjandustekstides, infolehtedes (nt ravimiinfoleht, dokumendid jm). Näide 6 on lõik tüüpilisest informatsioonitihedast tekstist.

(6) Hiidpandasid esineb looduslikult vaid Hiina edelaosa metsades. [...] Pikka aega arvati, et maailmas on vähem kui 1000 pandat, kuid 2006. aastal tehtud uuem uurimus selgitas väljaheidetest saadud DNA põhjal, et neid võib olla 2000–3000. Hiidpandad on looduskaitse all. Varasem pandade levila hõlmas Hiina lõuna- ja idaosa, Myanmar ja Vietnami põhjaosa. 1980. aastatel vähenes pandade levila bambuse loodusliku hävinemise tõttu. Hiina valitsuse tellitud 2014. aasta uuringuga leiti vabast loodusest 1864 hiidpandat. (Vikipeedia: hiidpanda)

D5 spontaansus iseloomustab peamiselt sellist teksti, mille loomisel on reaaliajalised piirangud. Seega tuleb spontaansus kõige rohkem esile suulisel diskursuses, kus öeldut ei saa otseselt tagasi võtta, kuid siiski on säilinud eneseparanduste võimalus. Spontaanne tekst võib olla nii suuline (vestlused, monoloogid) kui ka kirjalik (kommentaarium, foorumid, netivestlused). Spontaanus avaldub nii leksikaalsel kui ka süntaktilisel tasandil. Leksikaalsel tasandil on iseloomulikud ajalisele survet tingitud trükivead (nt *segle* vs. *selge*), kirjavead (nt kirjavahemärkide puudumine, kirjavahemärkide sõnade külge kleepumine, täpitähtede asemel numbrite kasutamine) ja keelevead (nt *kellelgi* vs. *kellelgi*, *autosi* vs. *autosid*), samuti mitmesugused lühendamisvõtted (nt *ple*, *naq*, *tahax*). Süntaktilisel tasandil tuleb esile süntaktiliste üksuste omavaheline lõtv seotus: lauseid luuakse tükkahaaval nii, et vajaduse korral on võimalik juppe aina juurde lisada, liitlausetel struktuur on lihtne, kasutatakse rindlauseid ning lihtsaid põimlauseid.

(7) igastahes hinded on tähtsamad, sõbrad tulevad ja lähevad, aga koolis käiakse vaid 1 kord, ning lõputunnistust on vaja. yhes6naga ma olen t2hele pannud et siis kui kellegil un naq head hinded siis nad tundide ajal eriti v2lja teistest ei tee ja pyyavad 6ppida ning seetõttu ple h2id s6pru (mitte alati ofc.) aga neil kellel on head sõbrad siis naq kipuvad tunnis jutustama liiga palju ja mitte kuulama ning siis hinded lange-

vad...et ma naq tahax teada kumb on teiearust t2htsam... (etTenTen www.lapsemure.ee, doc id = 138925)

(8) Vaikne soe päikesepaisteline pärestlõuna, maja oli enamjaolt tühi ja mõned üksikud õpilased rändasid mööda koridoridest moodustunud labürinti. Üks neist, Aadam, jalutas mõtliku ilmega ühe klassiruumi poole. Ta oli oma garderoobikapist leidnud kirja, armukirja, kui täpsustada. Sõnum oli lihtne: Tule peale kaheksandat tundi klassi 313, sa ei kahetse seda. Lõpus olid veel mõned südamekesed. Alguses rebis poiss selle tükkideks ja viskas prügikasti, arvates, et see on sõprade poolt mingi tobe nali. (etTenTen www.poogen.ee, doc id = 380465)

Näide 7 on katkend foorumist. Võrreldes näitega 8, mis on pärit kirjandusportaalist, on näide 7 oluliselt spontaansem ning sarnaneb suulise keelega, kus lausete (lausungite) piirid on hägustunud. Näide 8 on viimistletud kirjalik tekst, mille spontaansus on madal.

D6 f o r m a a l s u s kirjeldab tekstis kasutatava keele ametlikkuse laadi, avaldu- des süntaktilisel tasandil keeruka lauseehituse ning nominalisatsioonide rohkusena ning leksikaalsel tasandil viisakusväljendite (nt pöördumised) ning kõnekeelsuste vältimise, nn korrektse sõnakasutusena.

(9) Palume Teil võimalusel 4. maiks teatada, kas Teie asutus oleks huvitatud töömessil osalemisest, et saaksime paremini planeerida ruume, stende jmt. Ootame kindlasti osalema ka mittetulundusühinguid, sest tudengid vajavad ka praktikakohti ja/või töökogemust tähtajaliste projektide näol. (etTenTen www.ngo.ee, doc id = 112983)

(10) Tegelt tõesti, su õde on sinust 2 aastat noorem. Pane talle nn 'blokk' peale, tõrju teda.. ning sa peaksid talle ütleva, et sulle käib see pinda võimagi, ta ei tee sulle midagi selle eest ju? kui sa teda kardad.. hmm räägi emale sellest. (etTenTen www.lapsemure.ee, doc id = 377167)

Näide 9 on näitest 10 selgelt formaalsem, mis avaldub peamiselt leksikaalselt (nt *Teie* kasutamine), üldise neutraalse sõnakasutusena. Näide 10 näib kahe sõbra vahelise vestlusena: vestlejaid sinatakse ning vestluse toon on familiaarne. Näite 9 tekst on suunatud tundmatule lugejale, näites 10 vestluskaaslased näiliselt tunnevad üksteist, kuigi tõenäoliselt päriselus mitte. Seega ei suurenda formaalsuse taset auditooriumi anonüümsus, vaid see sõltub sellest, milline on suhtluskanal ja kuidas auditooriumi tajutakse.

D7 i m p e r s o n a a l s u s e fookuses on tegevus ja/või tegevusobjekt ehk toime- alis, tegevussubjekt on tagaplaanil või varjatud. Impersonaalsuse dimensiooni ühes otsas ongi sellised tekstid, kus agent puudub, ja teises otsas sellised tekstid, kus aktiivne tegevuse sooritaja on selgelt väljendatud. Impersonaalsus avaldub nii leksikaalsel (pärisnimede ja isikuliste asesõnade väike hulk) kui ka morfoloogilisel (impersonaalivormis verbid) tasandil. Prototüüpseteks impersonaalse teksti näide- teks on erinevad kasutusjuhendid, infolehed ja ka teaduskirjandus. Võrdluseks võib vaadata kaht järgnevat näidet: üks on lõik Kärstna kooli kodulehelt ja teine on kat- kend ühest muinasjutust.

(11) Suurem kapitaalremont viidi koolihoones läbi 1957.a ning 1964.a sai möisa-hoone juurdeehituse (4 klassiruumi). Enam kui kümnel viimasel aastal on Kärstna kooli keskkonda parendatud mitmete ümberehituste ja remontidega: 2009.a suvel - renoveeritud juurdeehituse I korrus, kool viidud maaküttele; kooli ümbrus saanud kivatendid ja välisvalgustuse. (etTenTen www.karst.vil.ee, doc id = 645762)

(12) Elas kord üks väike valge draakon, kellele meeldis õudselt lennata. Ta lendas pikki tunde kõrgel õhus, imetledes rohelisi metsi, kuldkollaseid põlde ja sinavaid mägesid. Talle meeldis ka lennata järvede kohal, vaadates siledalt veepinnalt enda peegeldust. Väike valge draakon armastas maad ning selle kohal lennates oli ta väga õnnelik. Kuid väikesel valgel draakonil oli ka üks suur mure. Nimelt ei olnud ükski teine draakon valge. Kõik teised olid mustad ja rohelised ja punased ja kollased ja kuldsed ja hõbedased ning purskasid ka musta ja rohelist ja punast ja kollast ja kuldset ja hõbedast tuld. Väike valge draakon aga oli valge ja purskas valget tuld. (<http://kodu.ut.ee/~kerstit/jutud/draakon.html>)

Näiteid 11 ja 12 eristab see, et esimesel puudub otsene agent (*keegi tegi kapitaalremonti ja parendas kooli keskkonda*), kuid teisel on konkreetset tegevust kontrolliv agent (*valge draakon*).

D8 *ajalisuse olulisus* näitab, kas ja kuivõrd on tekstis oluline ajalise järgnevuse väljendamine. Tüüpiliselt on ajalisus oluline narratiivsetes tekstides (nt uudised, ilukirjandus jm). Dimensiooni teise otsa paigutuvad sellised tekstid, kus sündmustel puudub ajaline mõõde, tegevus on seotud käesoleva kõnehetkega või on tegu geneerilise ehk üldajaga (vrd näiteid 13 ja 14). Ajalisuse olulisuse dimensioon avaldub eelkõige morfoloogilisel ja leksikaalsel tasandil, peamised tunnused on verbide ajavormid ning ajamäärused, tuleviku väljendamisel ka verbid *saama* ja *hakkama*.

(13) Õlithas oli püsti juba tänavu kevadel, suvel pidi selle sisseseade saama viimase lihvi, et septembris hakkaks põlevkiviõli voolama ning selle aasta lõpuks lootis Eesti Energia toota juba 50 000 tonni põlevkiviõli. (etTenTen www.juhtimine.ee, doc id = 532382)

(14) On teada ütlus, et eestlase jaoks on parim magustoit teine eestlane. Eestlasele on omane pidev naabrimehe või sõbra ületrumpamise fenomen. (etTenTen www.rajaleidja.ee, doc id = 510304)

D9 *interaktiivsus* on suhtluslik ja reageeriv, oluline on mitme „vestleja” olemasolu tekstis. Interaktiivne võib olla ühe autori tekst (ilukirjandus, ajakirjanduse intervjuud jm), samuti saavad dia- või polüloogi moodustada erinevate autorite tekstid (nt internetifoorumites jm). Interaktiivsus avaldub leksikaalsel tasandil esimese ja teise isiku pronoomenite, samuti pöördumisvormelite (nt *kuule*, *vaata*) kasutamisenä, süntaktilisel tasandil nt küsilausestena. Keerulisem on leida interaktiivsuse tunnuseid tekstidest, mis on küll üksteisest mõjutatud, kuid otseses dialoogis üksteisega ei ole, nt allolev näide 16. Vaatame järgnevat kolme näidet.

(15) „Kuule, kas sina tead, kes on tänapäeval need nn grillkanad,” küsin temalt. „Ega vist, kes need siis on?” tõstab teine imestunult silmad praetaldriku kohalt. „Ma ausalt öelda panen imeks, et sa seda terminit ei tea,” on minu kord hämmelduda. „Grillkanad on need solaariumiga liialdavad ja ülepleegitatud juustega tütarlapsed suuremate linnade ööklubides, kes oma seltskonda napside, lusti ja mugavuse eest müüvad.” (etTenTen arvamus.postimees.ee, doc id = 210366)

(16) maali 11-02-11 13:45 Valimislubadused on nagu kasukahunnikud laskmata karu nahast. Igakord olen mõelnud, et kust ja kelle arvelt see raha võetakse, lõpuks ometi on lubaduste oletatava hinna vastu huvi tuntud ja arvestusi tehtud! Oluline polegi see, kui sendipealt öiged need on, vaid see, et teema kui selline üles võetud ja ehk osa rahvast hakkab ka mõtlema, et parim valik polegi see mees, kes suurema kotitäie raha lubab. Üllar 11-02-11 13:32 Valimistel on iga hää tähtis ja lubadused ei maksa midagi! Mati 11-02-11 13:20 Raha rahaks aga tegelikult tekib küsimus, Reformierakond ja IRL on ju võimul olnud, mida siis nende lubaduste suunas tehtud on? Sellised lubaduste kalkulasioonid on sisuliselt valimiskampaania meetod opositsiooni mahategemiseks! (etTenTen vikerraadio.err.ee, doc id = 239307)

(17) Samas rääkis üks kaitse tunnistaja, Magerovi sõber ja alluv, et tema arvates Magerov midagi talle süüks pandavat ei teinud, hoopis politseinikud tarvitasid tema kallal vägivalda. (etTenTen www.pohjarannik.ee, doc id = 140357)

Näited 15 ja 16 on selgelt interaktiivsed, kuid see avaldub neis erinevalt. Näites 15 avaldub interaktiivsus süntaktilisel tasandil otsekõnena ning leksikaalsel tasandil asesõnade, täpsemalt 1. ja 2. isiku pronoomenitena. Näide 16 on pärit foorumist, kus vestlus toimub küll samal teemal, autorid otseselt omavahel ei suhtle, aga on siiski üksteise tekstist mõjutatud (nt kasutajad *Üllar* ja *Mati* reageerivad kaudselt kasutaja *maali* jutule). Kui näide 16 on selgelt vahetu ja kiire kirjalik vestlus, siis näide 17 paigutub interaktiivsuse dimensiooni teise otsa, kus vestluses osalejate vahel „reaalajaliselt” mingisugust vastastiksuhtlust ei toimu, vaid iseloomulik on hoopis narratiivi esitamine.

D10 s u b j e k t i i v s u s väljendab peale primaarse sisu ka autori hinnangut sisu kohta, viimase väljendamiseks kasutatakse eelkõige hinnangusõnu, st leksikaalseid vahendeid, samuti nn arvamisverbe.

(18) Ja mina arvan, et meie valijat alaväärtustakse sellega. Valija, isegi keskmine valija, ei ole nii loll. Kas see näitab, et poliitik lähtub sellisest valijakuvandist, keda tegelikult ei ole olemas? Viik: Ma arvan et poliitilised kampaaniad on disainitud madalama laubalistele, kui on keskmine inimene. Miks see nii on? Sest kumbki pool ei võta demokraatiat tõsiselt. (etTenTen www.ngo.ee, doc id = 82920)

(19) Suunaõppes süvendatakse üldteadmisi majandusest ja õpitakse tundma selle konkreetsete valdkondade seaduspärasusi. Baasteadmised saadakse majandusteooriast (mikro- ja makroökonomika), rahvamajandusest (majanduspoliitika, avaliku sektori ökonomika) ja ettevõtetmajandusest (juhtimine, turundus, majandusarvestus ning finantsjuhtimine). Erialaõppes läbitakse peamiselt rakendusliku

majandusteaduse aineid, mis valmistavad ette praktiliste majandusotsuste tegemiseks tulevikus. (etTenTen www.ut.ee, doc id = 460269)

Näide 18 on katkend Vabaühenduste Liidu veebilehelt ning näide 19 on pärit Tartu Ülikooli kodulehelt, kus kirjeldatakse majanduse eriala õpinguid bakalaureuseõppes. Näites 18 väljendub subjektiivsus väga tugevalt, avaldades nii leksikaalsel kui ka morfoloogilisel tasandil (nt isikulised ja enesekohased asesõnad, verb *arvama*).

D11 *keerukus* nõuab lugejalt lisapingutust, sest teksti on vaja kognitiivselt töödelda, et see muutuks arusaadavaks. Siin me defineerime keerukust keelesüsteemi omadusena, mis avaldub peamiselt leksikaalsel (võõrsõnad, erialasõnavara) ja süntaktilisel (lausepikkus, mitmetasandilised põimlaused, fraasipikkus, fraaside arv osaluses, sisestatud konstruktsioonid, nominalisatsioonid jm) tasandil. Keerukuse dimensiooni iseloomustab näide 20.

(20) Õigusökonoomiliste uurimismeetodite kritiseerijad on aga leidnud, et tegelikult see nii ei ole ja sellest järeldub, et regulatsioon, mis võib olla majanduslikult efektiivne, võib olla täielikult vastuvõetamatu õigluse aspektist vaadatuna. (Lahe 2005: 36–37)

D12 *argumentatiivsus* kirjeldab tekste, kus autor esitab teema või nähtuse kohta oma seisukohti, kuid erinevalt subjektiivsest hinnangust on need seisukohad põhjendatud objektiivsete väidetega. Tüüpilisteks argumentatiivseteks tekstideks võib pidada teaduskirjandust, vastandatuna nt kirjeldavale või narratiivsele tekstile. Argumentatiivse teksti distinktiivseks tunnuseks on põhjuse-tagajärje seoste esitamine, mis avaldub süntaktilisel tasandil tingimus- ja põhjuslausetena ning leksikaalsel tasandil kausaalsust või tingimust väljendavate sidenditena (nt *sellepärast et; sel juhul*). Argumentatiivne tekst on objektiivne, aga igasugune isiklike hinnangute või arvamuste lisamine võib muuta objektiivse teksti subjektiivsemaks. Argumentatiivsust iseloomustab näide 21.

(21) „Jaanuarist märtsini ja aprillist juunini on kõige pikemad veerandid koolis, seega oleks mõistlik, eriti arvestades võimalike lõputöid ja eksameid, pikendada kevadist koolivaheaega,” sõnas EÕEL aseesimees Kristo Peterson. (etTenTen www.ngo.ee, doc id = 464475)

2.2. Dimensioone kirjeldavad tunnused

Dimensionaalses tekstimudelil on dimensioonid keeleliste tunnuste komplekside alusel mõõdetavad parameetrid, mis kirjeldavad teatud omaduse määra tekstis. Teksti paigutumine dimensiooni skaala ühte otsa näitab selle omaduse tugevat esinemist, paigutumine teise otsa näitab aga, et see omadus avaldub tekstis nõrgalt või puudub üldse. Lähtudes eri keelte MDA tulemustest, saab väita, et tunnused on

keele- ja kultuurispetsiifilised, kuigi teatud keeled võivad tunnuste osas jagada ühisosa. Järgnevalt anname põgusa ülevaate võimalikest tekstilistest ja keeelistest tunnustest, millest võib lähtuda dimensioonide uurimisel. Samuti kirjeldame, milliseid keeletehnoloogilisi vahendeid saab nende tunnuste automaatseks ekstraheerimiseks kasutada.

Märgendamata korpuse korral on võimalik kasutada vaid leksikaalseid või tekstilisi tunnused. Leksikaalsed tunnused on sõnaloendid. Tüüpiliselt on need sõnapõhised, näiteks Zhang (2016) kasutas uurimiseks kirjalike registrite erinevaid metadiskursuse markerite¹⁰ loendeid. Sõnaloendid võivad koosneda peale terviklike sõnade ka väiksematest või suurematest üksustest, nt Scott Crossley ja Max Louwerse (2007) kasutasid sõnade bigramme eristamaks suulisi ja kirjalikke registreid. Tekstilised tunnused annavad täiendava ülevaate tekstis esinevatest nähtustest ning neid on tüüpiliselt üsna kerge arvutada. Näiteks tekstisõnade ja teksti sõnavara suhet (*type/token ratio*) kasutatakse leksikaalse mitmekesisuse mõõdikuna. Ka inglise keele multidimensionaalses analüüsis (Biber 1988) kasutatakse leksikaalseid ja tekstilisi tunnuseid, mille eristamisel koosinemine tekstis on osutunud üheks olulisemaks kirja ja kõne eristamise meetodiks. Tekstilisi tunnuseid nimetab Biber leksikaalseks spetsiifilisuseks. Sõnavara mitmekesisus on Biberi järgi omane just formaalsetele akadeemilistele tekstidele, kuid mitte reaalajalisele kõnele ajalise ja mälu surve tõttu. Teise leksikaalse spetsiifilisuse tunnuseks on kasutatud sõnapikkust, mis on osutunud samuti oluliseks registrite eristamisel, sest pikematel sõnadel on lühematest sõnadest spetsiifilisemad tähendused. George Zipf (1949) on näidanud, et sagedasti kasutatavad üksused lühenevad ja omavad üldisemat tähendust. Inglise keele MDA järgi on pikkade sõnade ja rikka sõnavara koosinemine omane just informatiivsetele tekstidele, mille eesmärk on pikemate ja rohkemate sõnade abil olla võimalikult arusaadav võimalikult paljudele inimestele.

Leksikaalsete ja tekstiliste tunnuste eraldamine või märgendamine korpuses ei nõua erilisi arvutuslikke või tarkvaralisi ressursse. Olulisteks tööriistadeks on sõnestaja ja lausestaja. Mitmete MDA meetodil analüüsitud keelte märgendamiseks on kasutatud 1980. aastate lõpus Biberi (1988) loodud mittevabavaraalist tööriista *POS tagger*. See tööriist pole enam kuigi aktuaalne, sest praegu eksisteerib mitmeid erinevaid lahendusi leksikaalse ja tekstilise info ekstraheerimiseks. Biberi tööriista teevad aga unikaalseks erinevad sõnaloendid (näiteks spetsiifilised verbiklassid, abstraktsete sõnade loetelu jm), mille koostamine on aja- ja töömahukas. Lisaks kaasneb taoliste sõnaloendite koostamisega vajadus kontrollida nende valiidust.

Peale leksikaalsete ja tekstiliste tunnuste on võimalik kasutada morfoloogilisi ja süntaktilisi tunnuseid. Morfoloogilisteks tunnusteks on lemmad ning info sõnaliigi ja grammatiliste kategooriate kohta. Süntaktiliste tunnustena saab tekstist ekstraheerida lausete ja osalausete tüüpe (käsk- või hüüdlause; tingimus- või põhjuslause), osalausete omavahelisi suhteid, liitlause struktuuritüüpi, fraasiliike, lausete ja fraaside pikkust ja sügavust, süntaktilisi funktsioone, sõltuvussuhteid jms. Kui varasemad MDA uurimused pidid sisendit rohkemal või vähemal määral käsitsi

¹⁰ Keelevahend, mida kasutatakse diskursuse struktureerimiseks või autori kohalolu väljendamiseks, nt personaalsed markerid *I, you*; teksti teema markerid *examine, explore* jm (Zhang 2016).

märgendama, siis nüüdisajal pole käsitsi märgendamine aktuaalne, kuna paljude keelte jaoks on olemas morfoloogiliselt ja süntaktiliselt märgendatud korpused (vt Universal Dependencies), mille alusel on võimalik treenida mudelipõhiseid analüsaatoreid.

Tunnuste defineerimisel saab osaliselt tugineda varasematele töödele, mis on uurinud DTM-i dimensioonidega ühtivate omaduste keelelisi tunnuseid. Näiteks teksti formaalsuse määramisel on Krista Kerge oma kolleegidega (Kerge jt 2007; Kerge 2010; Kerge, Pajupuu 2010; Puksand, Kerge 2012) lähtunud sõnaliikidest ning Heyligheni ja Dewaele formaalsusindeksist, mis näitab, et pronoomenite, verbide, adverbide ja interjektsioonide esinemissagedus muudab teksti kontekstuaalsemaks ja mitmemõttelisemaks, samal ajal kui noomenite, adjektiivide ja adpositsioonide esinemissagedus langetab kontekstuaalsust ehk muudab teksti formaalsemaks. Keerukust on põhjalikult käsitlenud Kerge väitekiri (2003; vt ka Puksand, Kerge 2012), kus keerukust on hinnatud nominalisatsiooni kaudu (nimisõna vs. tegusõna osatähtsuse võrdlemine). Keskmisest nominaalsemat teksti on raskem mõista, sest rohkete nimisõnade kohalolu muudab viidatavate mõistete ahela tihedamaks ja raskemini mõistetavaks. Katseliste meetoditega on teksti emotsionaalset valentsi püüdnud tuvastada Hille Pajupuu koos kolleegidega, kelle loodud emotsioonidetektor põhineb emotsionaalse valentsi järgi eelnevalt märgendatud sõnastikul (Pajupuu jt 2012, 2016). Heidi Meier (2002, 2003) on uurinud erinevaid tekstitüüpe (peamiselt esseed, teadus- ja ilukirjandustekstid) ning teinud keeleliste tunnuste korrelatsioonanalüüsi. Meier toob muuhulgas välja, et nt narratiivsetes tekstides, kus oluline on süžee ja selle jutustajad, kasutatakse rohkem isikulisi asesõnu, isikulisust ja minevikulisust näitavaid *nud*-kesksõnu, asesõnu *see* ja *keegi*. Argumenteerivates tekstides esineb rohkem umbisikulist tegumoodi ning alistavaid sidesõnu. Meier täheldas ka seda, et narratiivsetes tekstides seob jutustavaid lauseid aeg, mida kas väljendatakse konjunktiivide (nt *ja*, *siis*) või teiste ajalisele järgnevusele viitavate keeleliste tunnuste kaudu. Liina Lindström (2000) on uurinud narratiivsete lausete sõnajärge.

3. Kokkuvõte ja edasine töö

Dimensionaalne tekstimudel võimaldab vaadelda tekstide hübriidsust ja variatiivsust, selmet neid rangelt erinevatesse kategooriatesse paigutada, nagu tekstide klassifitseerimisel tavaliselt tehakse. DTM ei sõltu žanrilisest kategoriseerimisest ega defineerimisest, vaid olulisteks lähtepunktideks on mõisted *tunnus*, mis on otsest mõõdetav teksti karakteristik (nt umbisikuline tegumood, nimisõnade hulk), ja *dimensioon*, mis on tunnuste komplektiga kirjeldatav teatud omaduse määr (nt dimensioon *impersonaalsus* võiks olla kirjeldatav umbisikulise tegumoe suure hulga ning verbiga seotud pärisnimede ja isikuliste asesõnade vähesuse hulga kaudu). DTM-i eesmärk on pakkuda raamistik, millega kirjeldada tekste suhteliselt väikese hulga dimensioonide abil, mis omakorda avalduvad keeleliste tunnuste kaudu. Kui dimensioonide hulk on fikseeritud ning dimensioonid ise on keelest sõltumatud,

siis tunnuste hulk ei ole kuidagi piiratud ning need on suuremal või vähemal määral keelespetsiifilised.

Siinne artikkel esitas dimensionaalse tekstimudeli teoreetilised alused, mis annab raamid ja suunad edasiseks empiiriliseks tööks eestikeelsete tekstidega. Üks võimalik lähenemine on alustada dimensioonide ja tunnuste seoseid kajastavate hüpoteeside sõnastamisest ning nende hüpoteeside testimisest sobivate keelekorpusete peal. Iga dimensiooni jaoks tuleb defineerida potentsiaalselt olulised tunnused ning seejärel katsetada teatud tekstihulga peal nende tunnuste statistilist olulisust. Teine võimalus on anda tekstidele dimensioonide skaalal hinnanguid, nt kui informatiivne või afektiivne on mingi kindel tekst. Seejärel saab analüüsida, kuivõrd mingite tunnuste esinemine korreleerub dimensioonide hinnangutega. Kui dimensioone kajastavate tunnuste komplektid on leitud, saab tulemusi omakorda kinnitada. Näiteks saab püstitada hüpoteese, et mingi tekst on kõrge või madala väärtusega mingis dimensioonis sõltuvalt sellest, millised tunnused selles tekstis avalduvad, ning seejärel testida neid hüpoteese automaatse klassifitseerimise ja inimhinnangute kaudu. Dimensioonide ja tunnuste märgendamiskatsete puhul on oluline hinnata hindajatevahelist kooskõla, sest see määrab selle, kas erinevad inimesed tajuvad neid sarnaselt. Hindajatevahelisest kooskõlast sõltub, kas formaalselt defineeritud dimensioonid ja nende tunnused kehtivad ning on ka inimese jaoks eristatavad. Hindajatevaheliste suurte erinevuste korral on tarvis dimensioonid ja/või nende tunnused ümber formuleerida nii, et hindajatevahelised erinevused oleksid võimalikult väikesed.

Tekstide liigitamine teatud nähtuse alusel on tingitud mitmest praktilisest vajadusest. Info keelenäidete žanrilise kuuluvuse kohta on keeleuurijatele väga oluline, kuna see annab võimaluse uurida keeleliste väljendusvahendite (mitte)kasutamist funktsioonilt erinevates tekstides. Loomuliku keele automaattöötuse (nt kõnetehnoloogia (vt nt Kerge jt 2008), masintõlge, korrektsem morfoloogiline ja süntaktiline analüüs jm) seisukohalt on vajalik, et tekstid oleksid mingisuguse nähtuse järgi eristatavad, sest sellest sõltub sobivate töötlusvahendite ja rakendatavate mudelite valik. Loomuliku keele töötluses pole harv selline juhus, kus ühe andmekogu peal treenitud keelemudel või meetod annab sama andmekogu raames häid tulemusi, kuid mõnele teisele andmekogule rakendatuna ei pruugi nii edukas olla. Seetõttu vajavad keeletehnoloogilised rakendused valdkonna spetsiifilisemaid andmekogusid. Tekstilisi ja keelelisi tunnuseid on aga võimalik tekstist eraldada ka siis, kui info teksti eesmärgi või funktsiooni kohta pole saadaval. Selline tekstisisestest kriteeriumidest lähtuv analüüs võimaldaks luua domeenispetsiifilisemaid andmekogusid, mille abil saaks arendada erineva profiiliga teksti automaattöötlusvahendeid.

Kirjutise valmimist on toetanud Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus).

VEEBIVARAD

LLC = The London-Lund Corpus of Spoken English. <http://www.helsinki.fi/varieng/CoRD/corpora/LLC>

LOB = The Lancaster-Oslo/Bergen Corpus. <http://www.helsinki.fi/varieng/CoRD/corpora/LOB>

etTenTen. <https://doi.org/10.1515/1-00-0000-0000-0000-0011FL>

Universal Dependencies. <https://universaldependencies.org>

Wikipedia. <https://et.wikipedia.org>

KIRJANDUS

Atkins, Sue; Clear, Jeremy; Ostler, Nicholas 1992. Corpus design criteria. – *Literary and Linguistic Computing*, kd 7, nr 1, lk 1–30.

Besnier, Niko 1988. The linguistic relationships of spoken and written Nukulaelae registers. – *Language*, kd 64, nr 4, lk 707–736.

Biber, Douglas 1985. Investigating macroscopic textual variation through multifeature/multidimensional analyses. – *Linguistics*, kd 23, nr 2, lk 337–360.

Biber, Douglas 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. – *Language*, kd 62, nr 2, lk 384–414.

Biber, Douglas 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas 1994. An analytical framework for register studies. – *Sociolinguistic Perspectives on Register*. (Oxford Studies in Sociolinguistics.) Toim D. Biber, Edward Finegan. Oxford: Oxford University Press, lk 31–56.

Biber, Douglas 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas; Davies, Mark; Jones, James K.; Tracy-Ventura, Nicole 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. – *Corpora*, kd 1, nr 1, lk 1–37.

Biber, Douglas; Hared, Mohamed 1992. Dimensions of register variation in Somali. – *Language Variation and Change*, kd 4, nr 1, lk 41–75.

Crossley, Scott A.; Louwerse, Max M. 2007. Multi-dimensional register classification using bigrams. – *International Journal of Corpus Linguistics*, kd 12, nr 4, lk 453–478.

Crowston, Kevin; Kwaśnik, Barbara; Rubleski, Joseph 2011. Problems in the use-centered development of a taxonomy of web genres. – *Genres on the Web: Computational Models and Empirical Studies*. (Text, Speech and Language Technology 42.) Toim Alexander Mehler, Serge Sharoff, Marina Santini. Dordrecht: Springer Publishing Company, lk 69–84.

Eggins, Suzanne; Martin, James R. 1997. Genres and registers of discourse. – *Discourse as Structure and Process: Discourse Studies: A Multidisciplinary Introduction*. Toim Teun A. van Dijk. London: Sage, lk 230–256.

- Ferguson, Charles 1994.** Dialect, register and genre: Working assumptions about conventionalization. – Sociolinguistic Perspectives on Register. (Oxford Studies in Sociolinguistics.) Toim Douglas Biber, Edward Finegan. Oxford: Oxford University Press, lk 15–30.
- Forsyth, Richard S.; Sharoff, Serge 2013.** Document dissimilarity within and across languages: A benchmarking study. – Literary and Linguistic Computing, kd 29, nr 1, lk 6–22.
- Grieve, Jack 2014.** A multi-dimensional analysis of regional variation in American English. – Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber. Toim Tony Berber Sardinha, Marcia Veirano Pinto. Amsterdam–Philadelphia: John Benjamins Publishing Company, lk 3–34.
- Grieve, Jack; Biber, Douglas; Foriginal, Eric; Nekrasova, Tatiana 2011.** Variation among blogs: A multi-dimensional analysis. – Genres on the Web: Computational Models and Empirical Studies. (Text, Speech and Language Technology 42.) Toim Alexander Mehler, Serge Sharoff, Marina Santini. Dordrecht: Springer Publishing Company, lk 303–322.
- Hennoste, Tiit 2000.** Allkeeled. – Eesti keele allkeeled. (Tartu Ülikooli eesti keele õppetooli toimetised 16.) Tartu: Tartu Ülikooli Kirjastus, lk 9–56.
- Jang, Shyue-Chian 1998.** Dimensions of Spoken and Written Taiwanese: A Corpus Based Study. PhD thesis. University of Hawaii at Manoa.
- Katinskaia, Anisia; Sharoff, Serge 2015.** Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. – The 5th Workshop on Balto-Slavic Natural Language Processing associated with the 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015), Hissar, Bulgaria 10–11 September 2015: Proceedings. Shoumen, Bulgaria: Incoma Ltd., lk 65–74.
- Kerge, Krista 2003.** Keele variatiivsus ja *mine*-tuletus allkeelte süntaktilise keerukuse tegurina. (Tallinna Pedagoogikaülikooli humanitaarteaduste dissertatsioonid 10.) Tallinn: Tallinna Pedagoogikaülikooli Kirjastus.
- Kerge, Krista 2010.** Kirjanduse keeleparameetrid mitme tekstiliigi taustal. – Emakeele Seltsi aastaraamat 55 (2009). Tallinn: Emakeele Selts, lk 32–62.
- Kerge, Krista; Pajupuu, Hille 2010.** Text-types in speech technology and language teaching. – Analizar datos > Describir variación / Analysing data > Describing variation. Toim Jorge L. Bueno Alonso jt. Vigo: Universida de Vigo, Servizo de Publicacións, lk 380–390.
- Kerge, Krista; Pajupuu, Hille; Altrov, Rene 2007.** Tekst, kontekstuaalsus ja kultuur. – Keel ja Kirjandus, nr 8, lk 624–637.
- Kerge, Krista; Pajupuu, Hille; Tamuri, Kairi; Meier, Heidi 2008.** Kõnetehnoloogia vajab žanrlist lähenemist. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 4, lk 53–65.
- Kim, Yong-Jin; Biber, Douglas 1994.** A corpus-based analysis of register variation in Korean. – Sociolinguistic Perspectives on Register. (Oxford Studies in Sociolinguistics.) Toim D. Biber, Edward Finegan. Oxford: Oxford University Press, lk 157–182.
- Lahe, Janno 2005.** Süü deliktioõiguses. (Dissertationes iuridicae Universitatis Tartuensis 16.) Tartu: Tartu Ülikooli Kirjastus.
- Lamb, William E. 2002.** Scottish Gaelic Speech and Writing: Register Variation in an Endangered Language. PhD thesis. Belfast: Cló Ollscoil na Banríona.
- Lee, David 2001.** Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. – Language Learning and Technology, kd 5, nr 3, lk 37–72.

- Lindström, Liina 2000.** Narratiiv ja selle sõnajärg. – Keel ja Kirjandus, nr 3, lk 190–200.
- McEnery, Tony; Hardie, Andrew 2012.** Corpus Linguistics: Method, Theory and Practice. (Cambridge Textbooks in Linguistics.) Cambridge: Cambridge University Press.
- Mehler, Alexander; Sharoff, Serge; Santini, Marina 2010.** Riding the rough waves of genre on the web. – Genres on the Web: Computational Models and Empirical Studies. (Text, Speech and Language Technology 42.) Toim A. Mehler, S. Sharoff, Marina Santini. Dordrecht: Springer Publishing Company, lk 3–33.
- Meier, Heidi 2002.** Olulisi aspekte tekstitüübi võrdluses. – Tekstid ja taustad: artikleid tekstianalüüsist. (Tartu Ülikooli eesti keele õppetooli toimetised 23.) Toim Reet Kasik. Tartu, lk 101–114.
- Meier, Heidi 2003.** Essee allkeelte võrdluses. – Tekstid ja taustad II: tekstianalüüsi vaatepunkte. (Tartu Ülikooli eesti keele õppetooli toimetised 26.) Toim Reet Kasik. Tartu, lk 116–135.
- Pajupuu, Hille; Altrov, Rene; Pajupuu, Jaan 2016.** Identifying polarity in different text types. – Folklore. Electronic Journal of Folklore, nr 64, lk 25–42.
- Pajupuu, Hille; Kerge, Krista; Altrov, Rene 2012.** Lexicon-based detection of emotion in different types of texts: Preliminary remarks. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 8, lk 171–184.
- Parodi, Giovanni 2007.** Variation across registers in Spanish: Exploring the El-Grial Pucv Corpus. – Working with Spanish Corpora. Toim G. Parodi. London: Continuum, lk 11–53.
- Passonneau, Rebecca J.; Ide, Nancy; Su, Songqiao; Stuart, Jesse 2014.** Biber redux: Reconsidering dimensions of variation in American English. – Proceedings of COLING 2014: The 25th International Conference on Computational Linguistics: Technical Papers. Dublin: Dublin City University and Association for Computational Linguistics, lk 565–576.
- Puksand, Helin; Kerge, Krista 2012.** Õpiteksti analüüs kirjaoskuse omandamise kontekstis. – Emakeele Seltsi aastaraamat 57 (2011). Tallinn: Emakeele Selts, lk 162–217.
- Purvis, Tristan M. 2008.** A Linguistic and Discursive Analysis of Register Variation in Dagbani. PhD thesis. Bloomington: Indiana University.
- Santini, Marina 2007.** Automatic Identification of Genre in the Web Pages. PhD thesis. Brighton: University of Brighton.
- Sardinha, Tony Berber; Kauffmann, Carlos; Acunzo, Cristina Mayer 2014.** A multi-dimensional analysis of register variation in Brazilian Portuguese. – Corpora, kd 9, nr 2, lk 239–271.
- Shakir, Muhammad; Deuber, Dagmar 2019.** A multidimensional analysis of Pakistani and U.S. English blogs and columns. – English World-Wide, kd 40, nr 1, lk 1–23.
- Sharoff, Serge 2010.** In the garden and in the jungle. – Genres on the Web: Computational Models and Empirical Studies. (Text, Speech and Language Technology 42.) Toim Alexander Mehler, S. Sharoff, Marina Santini. Dordrecht: Springer Publishing Company, lk 149–166.
- Sharoff, Serge 2018.** Functional text dimensions for the annotation of web corpora. – Corpora, kd 13, nr 1, lk 65–95.
- Sinclair, John; Ball, Jackie 1996.** Preliminary Recommendations on Text Typology. EAGLES Document EAG-TCWG-TTYP.

- Sorokin, Alexey; Katinskaia, Anisia; Sharoff, Serge 2014.** Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus. – Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference „Dialogue” (Bekasovo, June 4 – 8, 2014). Moscow: RGGU, lk 646–658.
- Zhang, Man 2016.** A multidimensional analysis of metadiscourse markers across written registers. – Discourse Studies, kd 18, nr 2, lk 204–222.
- Zipf, George K. 1949.** Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, Massachusetts: Addison Wesley Press.

Kristiina Vaik (snd 1990), MA, Tartu Ülikooli eesti ja üldkeeleteaduse instituudi doktorant (Jakobi 2, 51005 Tartu), kristiina.vaik@hotmail.com

Kairit Sirts (snd 1980), PhD, Tartu Ülikooli arvutiteaduse instituudi keeletehnoloogia teadur (Narva maantee 18, 51009 Tartu), kairit.sirts@ut.ee

Kadri Muischnek (snd 1965), PhD, Tartu Ülikooli eesti ja üldkeeleteaduse instituudi arvutilingvistika dotsent (Jakobi 2–426, 51005 Tartu), kadri.muischnek@ut.ee

The dimensional text model: A theoretical overview

Keywords: corpus linguistics, text classification, text typology, functional text dimensions, multidimensional analysis

Corpus linguists and language technologists are increasingly turning to the Web as a source of language data. However, automatically crawled corpora have some shortcomings: lots of data but the content is unknown. This has created a need for software which is able to extract all the necessary information from the raw corpus. One such information extraction task in natural language processing is automatic text classification, which in practice imposes several challenges, such as confusion around the terminology, the absence of a generally accepted taxonomy, etc. Even if the latter existed, Web corpora include noisy user-generated content with lots of variation, meaning that all this variety may not fit well into generally accepted taxonomies. In this article we propose a novel theoretical framework for text classification – the Dimensional Text Model (DTM). This approach does not depend on existing genres or genre taxonomies but rather relies on some *text-external* (function) and *text-internal* (linguistic features) criteria according to which the texts that express a similar set of linguistic features share a similar function. DTM is a combination of the Multidimensional Analysis (MDA, Biber 1988) and Functional Text Dimensions (FTD, Sharoff 2018). From MDA we adapt the concepts and definitions for *text-internal criteria* and *dimension* – dimension is a quantifiable measure of a set of co-occurring linguistic features. From FTD we adapt the notion of *hybridism* where instead of classifying a text belonging to a class (genre) or not, it characterizes a

text through a combination of several parameters, i.e dimensions, and describes it in the space of dimensions. The aim of DTM is to propose a cross-linguistic universal model which does not depend on defining or classifying genres, instead it offers a framework for classifying texts based on their characteristic linguistic features, describing them in a single space of dimensions and interpret the function of these texts based on their location in the DTM space.

Kristiina Vaik (b. 1990), MA, University of Tartu, Institute of Estonian and General Linguistics, PhD Student (Jakobi 2, 51005 Tartu), kristiina.vaik@hotmail.com

Kairit Sirts (b. 1980), PhD, University of Tartu, Institute of Computer Science, Research Fellow in Language Technology (Narva maantee 18, 51009 Tartu), kairit.sirts@ut.ee

Kadri Muischnek (b. 1965), PhD, University of Tartu, Institute of Estonian and General Linguistics, Associate Professor in Computer Linguistics (Jakobi 2-426, 51005 Tartu), kadri.muischnek@ut.ee