https://doi.org/10.7592/MT2021.81.pragmaatika

Ei saa me läbi “Pragmaatika” korpuseta. Korpuspragmaatika ja pragmaatikakorpus

Külli Prillop
Tartu Ülikooli eesti ja üldkeeleteaduse instituudi teadur
kulli.prillop@ut.ee

Tiit Hennoste
Tartu Ülikooli eesti ja üldkeeleteaduse instituudi kaasprofessor
tiit.hennoste@ut.ee

Külli Habicht
Tartu Ülikooli eesti ja üldkeeleteaduse instituudi kaasprofessor
kulli.habicht@ut.ee

Helle Metslang
Tartu Ülikooli eesti ja üldkeeleteaduse instituudi emeriitprofessor
helle.metslang@ut.ee

Teesid: Projektis “Pragmaatika grammatika kohal” uurime (inter)subjektiivsuse (S/IS) väljendumist tänapäeva eesti keeles. Projekti üks ülesanne on võrrelda S/IS markerite (nt *vist*, *ma arvan*, *võib juhtuda*) kasutust eri registrites ja tekstiliikides. Olemasolevad korpused ei sobi niisuguse uuringu aluseks, sest kõik vajalikud registrid või tekstiliigid pole kaetud, tekstiliikide määratlus ei ole täpne ja eri registrid või tekstiliigid on esindatud väga erinevas mahus, mis muudab võrdleva statistilise analüüsi keerukaks. Uurimist takistab ka see, et korpuspäringusüsteemid ei väljasta piisavalt pikka konteksti, mistõttu pole igal üksikjuhul võimalik otsustada, kas tegu on S/IS markeriga või mitte. Nimetatud probleemide vältimiseks koostame Tartu Ülikoolis uut, korpuspragmaatika põhimõtetele tuginevat “Pragmaatika” korpust, mis sisaldab võrdses mahus kümne tekstiliigi tekste, kokku vähemalt viis miljonit sõnet.

Märksõnad: diskursusemarker, eesti keel, keelestatistika korpuspragmaatika, register, (inter)subjektiivsus, tekstiliik

Sissejuhatus. Pragmaatikast korpuspragmaatikani

Pragmaatika on keeleteaduse haru, mis uurib, kuidas keelelist teadmist kontekstis (suhtluses) kasutatakse. Pragmaatikal puudub talle ainuomane metodoloogia, üldjuhul on ta laenanud meetodeid nt semantikast, diskursuseanalüüsist

ja vestlusanalüüsist. Pragmaatikat iseloomustavad kolm mõistet: funktsioon, kontekst, kvalitatiivne uurimine.

Pragmaatilise analüüsi eesmärgiks on ennekõike leida, millised keelelised vahendid mingeid funktsioone täidavad. Sellisena on see oma klassikalisel kujul funktsioonilt vormile liikuv uurimine (*function-to-form*). Analüüsi võtmemõisteks on kontekst. Keelekasutus toimub alati kindlas kontekstis ja see kontekst mõjutab keele kasutamist. Selle juures vajab pragmaatiline analüüs tihti lausest suuremat konteksti (nt lõik või mitmest voorust koosnev järjend, mis võimaldab avada üksuse pragmaatilise funktsiooni). (Clancy & O'Keeffe 2020)

Keele ja konteksti seoste uurimine saab olla teoreetiline spekulatsioon või toetuda tegelikule keelekasutusele. Nii on ka pragmaatikas olnud olulised teoreetilised arutlused (nt kõneaktide teooria, koostööprintsiip jms) ja ka tegeliku keelekasutuse uurimine. Kuna keel ei ole ühtne, siis on osa pragmaatikast keskendunud sellele, et võrrelda keelekasutust erinevates allkeeltes, ennekõike suulises ja kirjalikus keelekasutuses (nt Biber 1988). Viimasel ajal on kujunemas ka võrdlev pragmaatika, mis võrdleb samades funktsioonides kasutatavaid eri keelte vahendeid (nt Dingemanse & Roberts *et al*. 2015). Andmepõhine pragmaatika on olnud ennekõike kvalitatiivne distsipliin, milles korpused (ja lihtsalt andmekogud) on materjali otsimise allikad.

Korpuslingvistika on kvantitatiivne distsipliin, milles uurimine toimub üldjuhul vormiüksustest funktsioonide suunas (*form-to-function*). Vormilt funktsioonile liikumine võimaldab hästi uurida keelenähtusi, mis on korpuspäringusüsteemide abil kergesti leitavad. Tõsi, siin tekib probleeme nt suulise keelega, milles on palju ebasüsteemset varieeruvust. Nt küsisõna *või* võib suulises keelekasutuses esineda kujul *või*, *vä*, *võ*, *võe*, *ve*. See nõuab keerukamat päringusüsteemi, mis võimaldaks ilma variante eelnevalt märgendamata neid siiski otsida. Näiteks võib olla Tartu Ülikooli suulise korpuse SEKK päringusüsteem (Orasmaa & Käärik *et al*. 2010). Üldjuhul tegeleb korpuslingvistika konkreetsete üksustega, mitte nende ning konteksti suhetega. Lause on korpusuuringutes ja -otsingutes tavaline piir.

Viimase kümne-viieteistkümne aasta jooksul on arenema hakanud korpuslingvistikat ja pragmaatikat ühendav korpuspragmaatika (*corpus pragmatics*) (vt Taavitsainen & Jucker *et al.* 2014; Aijmer & Rühlemann 2015; Rühlemann 2019; Clancy & O'Keeffe 2020). Selles ühendatakse korpuslingvistiline nähtuste otsimise metodoloogia (vertikaalne lugemine, *vertical reading*) pragmaatika konkreetsete nähtuste interpreteerimisega kontekstis (horisontaalne lugemine, *horizontal reading*). Selles lähenemises seotakse omavahel vormilt funktsioonile ja funktsioonilt vormile suunad. Vormilt funktsioonile suund alustab kindlatest üksustest. Näiteks võib alustada sõnast *või*, heita morfoloogilise märgendamise ja ühestaja abil kõrvale *või* nimisõna ja verbivormina, otsida

ülejäänud kasutusjuhud koos kontekstiga ning seejärel analüüsida, millistes kontekstides ja rollides seda kasutatakse (sidend, kinnitusküsimuse küsisõna, voorualguline partikkel, küsijätk jne). Selline lähenemine on korpuspragmaatikas kõige tavalisem, kuid puuduseks on asjaolu, et see ei too välja teisi sama funktsiooniga üksusi, meid huvitaval juhul kinnitusküsimuse küsisõnu. Selleks analüüsiks on vajalik suund funktsioonilt vormile, mis on korpusuuringutele mõeldes keerukam. Selleks et alustada otsingut funktsioonidest, tuleb korpus eraldi märgendada. See aga omakorda tähendab, et need funktsioonid on kas korpuspõhiste uuringute või teooria abil ette ära määratud. Üks võimalus on luua märgendatud korpus konkreetse uuringu tarvis, märgendades vajalikud üksused. Teine variant on teha laiemalt ja pinnapealsemalt märgendatud korpus, mis toob välja rohkem olulisi funktsioone. See on töömahukas kas käsitsi või heal juhul poolautomaatselt tehtav tegevus ja korpusi, milles oleks märgendatud pragmaatilised üksused, on vähe. Ühe sellisena võib nimetada SPICE-Ireland korpust (Kallen & Kirk 2012). Kohalikuks näiteks sobib Tartu Ülikoolis tehtud suulise infosuhtluse korpus ja selle jaoks välja töötatud dialoogiaktide märgendus (Hennoste & Koit *et al*. 2004).

Pragmaatika grammatika kohal

Projektis "Pragmaatika grammatika kohal: subjektiivsus ja intersubjektiivsus eesti keele registrites ja tekstiliikides (PRG 341)" uurime subjektiivsuse ja intersubjektiivsuse väljendumist erinevates eesti keele kirjalikes ja suulistes variantides. Subjektiivsust ehk kõnelejale orienteeritust käsitame kui kõneleja/kirjutaja hoiakute ja veendumuste väljendumist tekstis/dialoogis, intersubjektiivsust ehk kuulajale orienteeritust kui tekstis väljenduvat kuulaja/lugeja arvestamist ja kaasamist (vt nt Narrog 2017; Verhagen 2005; Traugott 2003). Subjektiivsust/intersubjektiivsust (S/IS) võivad väljendada näiteks diskursusemarkerid, modaal-, kõnelemis- ja tajuverbid, kõneviisid, direktiivid, hinnanguline leksika.

S/IS markerid on üldjuhul muutumatud sõnavormid või fraasid, mida on sellistena hõlpus otsida isegi morfoloogiliselt märgendamata korpustest. Teisalt iseloomustab S/IS markeritena kasutatavaid vahendeid mitmetähenduslikkus ja polüfunktsionaalsus, mida ükski automaatne märgendaja ei erista. Mõnel vahendil on S/IS väljendamine põhiline, mõnel aga kaasnev funktsioon. Meie uuringu keskmes on kaks S/IS funktsioonidomeeni: edastatavale infole lisatavad tõenäosus- ja väärtushinnangud. Näiteks sõna *vaevalt* "Eesti keele seletava sõnaraamatu" (EKSS) näidetest väljendab ainult esimene (1a) kõneleja subjektiivset hinnangut situatsiooni tõepärasusele – kahtlust. Eesti Keele Instituudi

(EKI) Sõnaveebis[1] on sõnale *vaevalt* antud kolm tähendusseletust, millest S/IS markeri funktsiooni esindab samuti ainult üks: '(väljendab kahtlust mingi tegevuse või olukorra võimalikkuse suhtes:) vähe usutav'.

(1) a. *Vaevalt* see saal kõiki peolisi ära mahutab
b. Jõudsin *vaevalt* uinuda, kui mind juba äratati.
c. Koduni jäi *vaevalt* paarsada meetrit.
d. Haavatu püsis *vaevalt* sadulas.

Projekti üks ülesanne on uurida konkreetsete markerite kasutuskontekste, distributsiooni ja funktsioone. Meie lähenemine on ennekõike vormilt funktsioonile, st oleme enne määratlenud diskursusemarkerid, mida käsitleme: (1) modaalse ja hinnangulise tähendusega adverbid, nt *vist*, *ilmselt*, *tegelikult*, *vaevalt*, (2) tunnetusverbe sisaldavad pealaused, nt *(ma) arvan (et)*, *(mulle) tundub (et)*, (3) modaal- ja performatiivverbid, nt *võib juhtuda*, *peaks tulema*; *kinnitan* / *väidan olevat*.

Projekti teine ülesanne ja olulisim uus joon on markerite kasutuse süstemaatiline võrdlus eri keelevariantides. Analüüsime erinevaid registreid ja tekstiliike suulises suhtluses, netisuhtluses ja trükitekstides.[2] Eesmärk on välja selgitada neile omane, neid eristav S/IS vahendite kasutamine. Lisaks uutele lingvistilistele teadmistele pragmaatiliste markerite kasutuse kohta on projekti tulemuste üks võimalikke praktilise kasutuse valdkondi tekstiõpetus ja kirjutamiskoolitus (eri tüüpi tekstide analüüs ja eesmärgipärane loomine).

Meie analüüs on ühelt poolt kvalitatiivne ja teisalt kvantitatiivne. Kvantitatiivne analüüs tähendab selle väljaselgitamist, kui palju esineb uuritavaid markereid eri registrites ja tekstiliikides. Kvalitatiivne analüüs tähendab markerite kasutuse uurimist, et leida kontekstuaalsed tegurid, mis mõjutavad markerite kasutamist, nende hulka ja kombinatsioone eri registrites ja tekstiliikides.

Korpuspragmaatika ja korpuste probleemid

Analüüsimetoodika täpsustamiseks alustasime projekti pilootuuringuga (Hennoste & Habicht *et al*. 2020, 2021), mille materjali kogusime olemasolevatest tekstikorpustest, vt tabel 1.[3] Enamik vabalt kasutatavaid eesti keele korpusi on koondatud Keeleveebi (http://www.keeleveeb.ee/). Keeleressursse, sh korpusi, majutab ka Eesti Keeleressursside Keskus (https://korp.keeleressursid.ee/), mis tegeleb päringusüsteemi KORP arendamisega. Kõik meie pilootuuringusse valitud avalikult kättesaadavad korpused paiknevad Keeleveebis. Avalikud ei ole suulise keele ja netivestluste korpused, kuid need on kasutatavad Tartu Ülikooli suulise ja arvutisuhtluse labori kaudu, mille töötajad osalevad projektis.

Tabel 1. Pilootuuringus kasutatud korpused ja nende umbkaudsed mahud sõnedes

I. Trükitekstid	
Ilukirjanduslik proosa	5,8 miljonit
Ajakirjandus:	
ajakiri Kroonika	580 000
nädalaleht Eesti Ekspress	7,2 miljonit
päevaleht Õhtuleht	45,5 miljonit
päevaleht Eesti Päevaleht	87,9 miljonit
päevaleht Postimees	32,9 miljonit
Teadus:	
teadusartiklid	1,35 miljonit
doktoritööd	2,3 miljonit
Populaarteadus:	
ajakiri Horisont	260 000
II. Netitekstid	
Reaalajas argidialoog:	
netivestlused	95 000
jututoad	7,0 miljonit
Avalik mittereaalajas suhtlus:	
foorumid	8,7 miljonit
uudisgrupid	4,6 miljonit
kommentaarid	1,9 miljonit
III. Suulised tekstid	
Ametlik-avalik suhtlus	310 000
Argisuhtlus	240 000

Pilootuuring tõi välja mitu probleemi, mis on tingitud uuringu aluseks olnud korpuste omadustest. Esimese probleemi tekitas S/IS markerite määratlemiseks vajalik kontekst. Kuna samad keelelised üksused esinevad tekstides nii S/IS markerina kui ka muus rollis, tuli kõigepealt määratleda iga vaadeldava üksuse markerina esinemise tingimused ning seejärel analüüsida konkreetseid korpusenäiteid ja selekteerida välja need, mis vastavad markeri tingimustele.

Peamiseks probleemiks S/IS markerite eristamisel muudest kasutustest sai asjaolu, et olemasolevate korpuste avalikud päringusüsteemid võimaldavad küll otsida lauseid, kus uuritav sõna esineb, aga ei kuva lausest pikemat konteksti. Näiteks markerite puhul, mis on kujunenud pealausest (*(ma) arvan (et)*, *(ma) usun (et)* jt), on keskseks määramisaluseks see, millist lauseosa mööda liigub dialoogi või teksti sisuline pealiin. Markerina toimiva formaalse pealause korral

liigub suhtlus või tekst mööda grammatilist kõrvallauset (Thompson 2002). Selle määramiskriteeriumi rakendamiseks on vaja näha mitmelauselist konteksti. Näites (2) on H vastus V küsimusele formaalselt kõrvallause, *ma=arvan et* aga pealause. Samas pragmaatiliselt on *ma=arvan et* siin subjektiivsusmarker, mille saab ilma tekstiliini lõhkumata ära jätta. Näites (3) seevastu liigub teksti pealiin mööda pealauset, st vastaja *arvan, et* on tavaline pealause. (Hennoste & Habicht *et al.* 2020: 69–70.)

> (2) V: siis me `kuuendal ok`toobril sinna `sisse läheme tead=`noh? (0.8) aga noh seal=et `siis e (0.3) et ma=i `tea ütleme siis=et siis et et **`millal sa siis võtmed saad=ju**.
> H: ma=arvan et **siis kui=mul seal `korter `üle antakse ju**. (Argisuhtlus)

> (3) **- Mida arvate** tervetest inimestest kes liiguvad alla 10 tuhande sammu päevas?
> **- Arvan, et** neil pole aega iga päev 2 tundi kõndimise alla panna ja nad hoiavad end vormis näiteks trennis käimisega (ajakulu sellele on märksa väiksem). (Foorum)[4]

Lausest pikemat konteksti pole vaja mitte ainult markerite märgendamiseks, vaid ka analüüsimiseks. Näiteks *(ma) arvan (et)* esineb tekstides neljal kujul: *ma arvan et*, *arvan et*, *ma arvan*, *arvan*. Püstitasime hüpoteesi, et ilma subjektpronoomenita *arvan (et)* on rohkem grammatiseerunud kui subjektiga *ma arvan (et)*. Esialgu ei leidnud hüpotees kinnitust, aga kuna olemasolevad korpused ei võimalda jälgida terveid lõike, jäi kontrollimata, kuivõrd mõjutab subjekti ärajätmist teksti sidusus. On võimalik, et *ma* puudub tõenäolisemalt siis, kui esimene isik on tekstis varem mainitud, nagu näites (4).

> (4) "Reegel on, et õlut mu sõbrad tuua ei või. **Ma** ise koolitan neid välja, et missugune jook missuguse toidu kõrvale käib. Vabariigi aastapäevi oleme tähistanud aga eriti pidulikult."
>
> Seekord olid nad abikaasaga kokku pannud seitsmekäigulise õhtusöögi menüü, kus õllel polnudki kohta. Iga käigu kõrvale pakuti hoopis mõnd mullidega jooki… "**Arvan, et** vabariigi aastapäevi tuleb mitte ainult tähistada, vaid ka tähtsustada!" (Ajakirjandus)[5]

Teine suurem probleemide ring tuleneb vajadusest registreid/tekstiliike markerite kasutuse poolest omavahel võrrelda. Olemasolevate korpuste mahud

erinevad enam kui sajakordselt (vt tabel 1), mis tekitab küsitavusi peamiselt siis, kui soovime leida, kas ja mil määral eristuvad registrid/tekstiliigid üksteisest neis kasutatavate S/IS markerite koguhulga poolest: kui palju on eri registrites/tekstiliikides neile iseloomulikke S/IS markereid, millised need on ja kuidas koonduvad registrid/tekstiliigid sel alusel gruppidesse. Et korpustes ei ole diskursusemarkerid märgendatud, kasutame siinkohal probleemi näitlikustamiseks adverbe, ja proovime kindlaks teha, kui palju leidub ühes või teises registris just sellele registrile iseloomulikke adverbe, s.o niisuguseid määrsõnu, mille sagedus selles registris on küllalt suur. Väikese sagedusega adverbid võivad olla uuritavasse tekstivalimisse sattunud juhuslikult ja sellisena iseloomustavad need ainult konkreetse autori või teksti keelekasutust.

Näites kasutame oma kogutud foorumite korpust, ilukirjanduskorpust ja ajakirjanduskorpust, igaühes 500 000 sõnet. Esmalt otsime adverbe, mille sagedus on vähemalt 10 esinemust 500 000 sõne kohta. Seejärel jagame foorumite korpuse neljaks alamkorpuseks, säilitades kõik korpuse omadused peale suuruse. 200 000 sõne suurustest alamkorpustest otsime adverbe, mille sagedus on vähemalt 4 (st normaliseeritult 10 esinemust 500 000 sõne kohta), ja 100 000 sõne suurusest alamkorpusest adverbe, mille sagedus on vähemalt 2 (normaliseeritult samuti 10 esinemust 500 000 sõne kohta). Tulemused on esitatud tabelis 2.

Tabel 2. Adverbide hulk, mille sagedus on normaliseerituna vähemalt 10 esinemust 500 000 sõne kohta, erinevates korpustes

	Korpus	**Korpuse maht**	**Adverbe sagedusega vähemalt 10/500 000**
1	Ilukirjandus	500 000	593
2	Kommentaarid	500 000	461
3	Foorumid	500 000	491
4	Foorumite alamkorpus A	200 000	533
	Foorumite alamkorpus B	200 000	541
5	Foorumite alamkorpus C	100 000	586
	Foorumite alamkorpus D	100 000	547

Kasutades võrdluseks väiksemamahulisi foorumite korpusi, jõuaksime tulemusele, et suhteliselt sagedaste adverbide hulga poolest sarnanevad foorumipostitused rohkem ilukirjandustekstide kui veebikommentaaridega (tabel 2,

read 1, 4 ja 5 *vs.* rida 2). Otsustades võrdse mahuga korpuste põhjal, on tulemus vastupidine (tabel 2, rida 1 *vs.* read 2 ja 3).

Vastuoluliste järelduste põhjuseks pole mitte väiksemal valimil põhinevate otsustuste suurem statistiline viga, vaid seaduspära, et igas tekstis on väike hulk suure sagedusega sõnu ja suur hulk väikese sagedusega sõnu. Matemaatiliselt kirjeldab seda sõltuvust Zipfi seadus $f(r) = C / r^a$, kus $f(r)$ on sõna esinemissagedus antud tekstis, r on sõna astak (koht) antud teksti põhises sagedussõnastikus ning C ja a on keelest ja tekstikorpuse suurusest sõltuvad konstandid, kusjuures tavaliselt on a väärtus ligikaudu 1 (Zipf 1935: 40–48). Eestikeelsete ilukirjandustekstide autorikõne 100 000-sõnelise valimi põhjal arvutatult $C = 4122$ ja $a = 0{,}86$ (Tuldava 1977: 151). Zipfi seadust on hiljem täpsustatud (vt ülevaadet Piantadosi 2014).

Zipfi seaduse järgi on ootuspärane, et sõnu (sh adverbe), mille sagedus on vähemalt kaks, on tekstis protsentuaalselt rohkem kui sõnu, mille sagedus on 10. Tabelis 2 on väiksemate korpuste põhjal leitud sagedused seetõttu suuremad kui suuremate korpuste põhjal leitud sagedused. See tähendab, et sageduspiiri ei saa korpuste suurusega proportsionaalselt normaliseerida. Väiksema korpuse jaoks sobiv sageduspiir on võimalik määrata suurema korpuse põhjal binomiaalse interpoleerimise teel (Baayen 2001: 63–69), aga pole selge, kuidas toimida juhul, kui uuritakse nii üksiksõnu kui ka sõnajärjendeid (nagu meie projektis). Samuti pole teada, kuidas võtta mudelis arvesse lisakriteerium, mis määrab, kui paljudes tekstides peab sõna esinema, et kuuluda registrile omase sõnavara hulka (Bestgen 2020).

Sõna leviku hindamisel tuleb lisaks sõna esinemissagedusele arvesse võtta ka seda, kui mitmel autoril või kui mitmes tekstis sõna esineb, kuna sõnakasutus pole autorist ega tekstist sõltumatu. Keeleveebi teadustekstide korpuses esineb S/IS marker *küllap* 47 korral, kusjuures seda on korpuses esindatud 55 autorist kasutanud 15. Neist andmetest lähtuvalt võiks otsustada, et *küllap* kuulub teadustekstis üldkasutatavate S/IS markerite hulka, sagedus korpuses on oluliselt suurem kui 10/500 000. Siiski selgub, et 72% kõigist *küllap* kasutustest kuulub ainult kolmele autorile, seega on tegu pigem nende autorite idiolekti iseloomustava väljendusviisiga, mitte teadustekstile üldiselt omase S/IS markeriga.

S/IS markerite sageduste analüüsimisel tuleb arvesse võtta sedagi, et mitmed nendest markeritest on alles grammatiseerumas, mistõttu peaksid uuritavad tekstid võrreldavate tulemuste tagamiseks olema suhteliselt samaaegsed ja ajaliselt sarnase jaotusega. Praegused korpused sisaldavad eri aastakümnete tekste ja ka korpuste ajaline ulatus on erinev, nt Delfi kommentaarid pärinevad kahekuulisest perioodist veebruar kuni märts 2004, Postimehe tekstid viie-

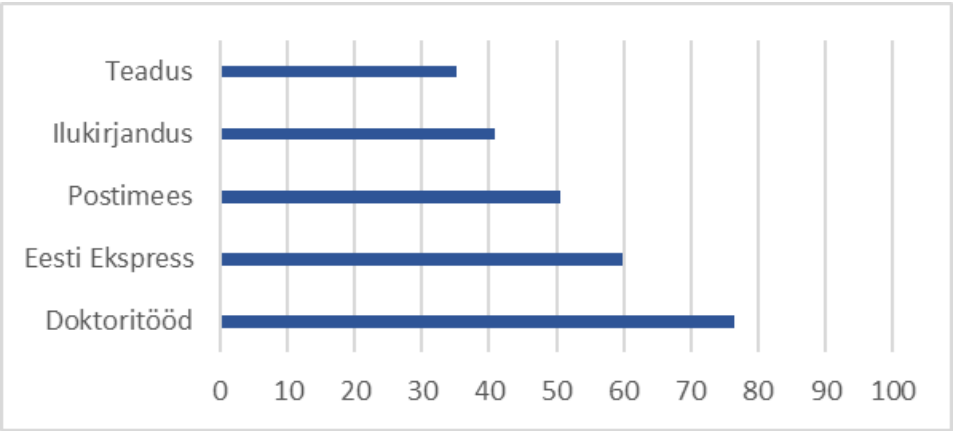
aastasest perioodist 1996 kuni 2000, foorumitekstid aastatest 2000 kuni 2008, kusjuures päringutulemus ei sisalda alati infot teksti kirjutamise aja kohta.

Viimane suurem probleem puudutab korpusetekstide liigitamist. Kuna meie eesmärk on võrrelda S/IS markerite kasutust eri keelevariantides, peavad tekstiliigid/registrid kasutatavates korpustes olema piisavalt rangelt eristatud ja kõik olulised keelevariandid peavad olema esindatud (vt Hennoste & Habicht *et al*. 2021). Praegustes korpustes on aga materjal organiseeritud väljaannete, mitte tekstiliikide ega registrite kaupa, nii et näiteks ajakirja Horisont korpus sisaldab lisaks populaarteaduslikele tekstidele ka toimetaja veergu, anekdoote jm. Korpustest puuduvad reklaami ja ajakirjanduse piirile jäävad sisuturundustekstid, samuti juhendid. Mõned korpused (nt eesti veeb 2013/2017/2019) eristavad temaatilisi tekstitüüpe (toit, tervis, religioon sport jms), mis aga samuti võivad jaguneda eri registrite vahel.

Tulemusi moonutab ka see, kui üht keelevarianti esindavas korpuses on suurel hulgal tekste, milles esitatud näitematerjal või tsitaadid kuuluvad hoopis teise keelevarianti. Isegi kui protsentuaalselt on tsitaate vähe, on neil potentsiaalselt oluline mõju meie analüüsile, juhul kui tsitaadid erinevad põhitekstist just S/IS markerite suhtelise kasutussageduse poolest. Näiteks selgus olemasolevaid korpusi kasutavast pilootuuringust, et teadustekstide *arvan*-lausedesinevad vaid tsiteeritud intervjuudes, küsitluste vastustes ja mõningates ülevaateartiklites, mis sellisena ei kuulu teadustekstide hulka. Sama kehtib ka teiste pealausest kujunenud tunnetusverbilähteste markerite, nt *(ma) usun (et)*, *(ma) loodan (et)* ja *(ma) kardan (et)* kohta, vt (5).

> (5) Eesti tuntumaid lavastajaid Voldemar Panso kirjutas artiklis "Näidend teatri poolelt vaadatuna": "**Usun, et** näitekirjanik peaks suutma kahestuda – ta peaks suutma ette kujutada, kuidas tema loodud elu hakkab lavalaudadel hingama, tema tegelased käituma ja rääkima; --" (Teadusartiklid)

Kui võrdlesime pealausest kujunenud markerite sagedusi omavahel, et teha kindlaks, millised korpustes esindatud tekstid eelistavad sisult neutraalseimat markerit *(ma) arvan (et)* teistele struktuurilt sarnastele, aga rohkem markeeritud tähendusega markeritele *(ma) usun (et)*, *(ma) loodan (et)*, *(ma) kardan (et)* jm, saime ebaootuspärase pildi, kus teadusartiklid on ilukirjandustekstidega sarnasemad kui doktoritöödega (vt joonis 1).



Joonis 1. (ma) arvan (et) *osakaal pealausest kujunenud S/IS markerite hulgas.*

Kokkuvõte ja lahendus: uurijate endi kogutud "Pragmaatika" korpus

Olemasolevad korpused on väga mahukad ja sisaldavad erinevaid tekste, kuid ei sobi korpuspragmaatiliseks registreid ja tekstiliike võrdlevaks uurimiseks peamiselt kolmel põhjusel:

a) markerite määramiseks vajalikku lausest pikemat konteksti pole võimalik näha;

b) markerite üldise kasutussageduse usaldusväärseks võrdlemiseks registriti või tekstiliigiti on korpuste mahud ja ajapiirid liiga erinevad;

c) registrite/tekstiliikide omavaheliseks võrdlemiseks ei ole need piisavalt rangelt piiritletud ja osa uurimiseks olulisi registreid puudub.

Niisiis koostame oma projekti vajadustele vastava pragmaatikakorpuse, kus tekstid on suhteliselt ühevanused (pärit aastatest 2010–2020) ja kus registrid on rangelt eristatud ning võrdses mahus esindatud. Korpuse tööplatvormiks valisime SketchEngine'i (Kilgarriff & Baisa *et al.* 2014), mis võimaldab vaadata mitmelauselist konteksti. Lisaks sellele on SketchEngine'is muidki meie uuringutes kasulikke funktsioone, nt sõnavisandid, N-grammid, päringuvastuste märgendamine ja CQL-päringute võimalus.

Kogutav korpus koosneb kümnest alamkorpusest, millest igaüks esindab eri registrit (tabel 3). Korpus ei ole avalik, sest sisaldab eravestlusi (netisuhtlus) ja autoriõigustega kaitstud tekste (ajakirjandus, ilukirjandus).

Me ei lisa korpusesse ajalehtede ning ajakirjade terviknumbreid, mis sisaldavad eri liiki tekste, vaid valime näiteks Horisondist ainult populaarteaduslikud artiklid, päevalehtedest ainult uudised, arvamused, intervjuud ja juhtkirjad. See tähendab, et meil pole võimalik koguda korpust täiesti automaatselt veebi kraapimise teel, nagu tehakse tänapäevaste suuremahuliste kolmanda põlvkonna korpuste puhul. Automaatne kogumine oleks mõttekas, kui saaksime tekstiliike ja registreid automaatselt määrata. Paraku suudab eesti keele jaoks välja töötatud tekstide liigitaja (Vaik & Muischnek 2018) eristada vaid kirjakeele normi järgivaid ja mittejärgivaid tekste, kusjuures täpsusega 74%, mis ei ole meile piisav. Dimensionaalne tekstimudel (Vaik & Sirts *et al.* 2020), mis võimaldaks täpsemat liigitust, on esialgu veel vaid teoreetiline.

Tabel 3. Pragmaatikakorpuse liigendus

I. Toimetatud trüki- ja veebitekstid
1. Ilukirjandus
2. Ajakirjandus (Kroonika, Eesti Päevaleht, Postimees)
3. Sisuturundus
4. Teadus
5. Populaarteadus (Horisont, Eesti Mets, Oma Keel, Eesti Loodus, Novaator)
II. Netisuhtlus
6. Reaalajas netivestlused
7. Foorumid
8. Kommentaarid
III. Suuline suhtlus
9. Ametlik-avalik suhtlus
10. Argisuhtlus

Iga alamkorpus meie korpuses sisaldab 500 000 tekstisõna. Mahu planeerimisel sai lisaks tekstide kogumise meetodile määravaks asjaolu, et praeguste teadmiste juures puudub võimalus pragmaatilisi üksusi automaatselt märgendada. Käsitsi ja ka poolautomaatselt märgendamine on aeganõudev, isegi kui piirduda ainult teatud hulga S/IS markerite tuvastamisega tekstist. Seega pidime korpuse mahu kavandamisel arvestama ka analüüsimisele kuluva ajaga.

Korpuse abil uurime S/IS markerite kasutuse erinevusi registriti/tekstiliigiti ega plaani luua automaatanalüüsi vahendeid. Korpuslingvistika arenguid arvestades võib tunduda, et liigiliselt määratud ja piiratud mahuga korpust luues liigume ajas tagasi, ent registritele/tekstiliikidele omaste keelevahendite

mõtestatud korpuspragmaatilise analüüsiga edasi liikumiseks vajame sellise usaldusväärse baasuuringu tuge.

Tänuavaldused

Artikli valmimist on toetanud Euroopa Regionaalarengu Fond (Eesti-uuringute Tippkeskus) ja Eesti Teadusagentuur (projekt PRG341). Täname retsensente kasulike tähelepanekute eest.

Kommentaarid

[1] https://sonaveeb.ee/ (18.11.2021).

[2] Register on allkeel, mis seostub teatud kasutussituatsiooni ja selle funktsioonidega (Hennoste 2001; Biber & Conrad 2009). Registrile on omased konsituatsioonist mõjutatud ja seda teistest registritest eristavad keeleliste tunnuste kimbud. Tekstiliik on kultuuris väljakujunenud ja teadvustatud tekstirühm, millele on omased konventsionaalselt temaga seotud tunnused. Tekstiliikideks oleme arvanud traditsioonilised žanrid (nt romaan, ajaleheuudis, teadusartikkel jm) ning lisaks tekstirühmad, mida ühiskonnas žanridena ei tõlgendata (nt doktoritöö, veebikommentaar jms).

[3] Tabelis on esitatud kasutatud korpused. Nende koostamise alused on eklektilised ega ole alati määratletud registrite ega tekstiliikide terminites.

[4] https://foorum.perekool.ee/teema/mida-arvate-tervetest-inimestest-kes-liiguvad-alla-10-tuhande-sammu-paevas/ (18.11.2021).

[5] https://arileht.delfi.ee/artikkel/73736213/ollemeistri-suur-ulevaade-milline-marjuke-toidu-korvale-koige-paremini-sobib (11.09.2021).

Kirjandus

Aijmer, Karin & Rühlemann, Christoph (toim) 2015. *Corpus Pragmatics: A Handbook*. Cambridge University Press.

Baayen, Harald R. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Bestgen, Yves 2020. Comparing lexical bundles across corpora of different sizes: The Zipfian problem. *Journal of Quantitative Linguistics* 27 (3), lk 272–290 (DOI: 10.1080/09296174.2019.1566975).

Biber, Douglas 1988. *Variation across speech and writing*. New York: Cambridge University Press.

Biber, Douglas & Conrad, Susan 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

EKSS = *Eesti keele seletav sõnaraamat* (http://www.eki.ee/dict/ekss/ – 19.11.2021).

Clancy, Brian & O'Keeffe, Anne 2020. Corpus Pragmatics. O'Keeffe, Anne & Clancy, Brian & Adolphs, Svenja (toim). *Introducing Pragmatics in Use*. 2nd edition. Routledge, lk 47–68.

Dingemanse, Mark & Roberts, Seán G. & Baranova, Julija & Blythe, Joe & Drew, Paul & Floyd, Simeon & Gisladottir, Rosa S. & Kendrick, Kobin H. & Levinson, Stephen C. & Manrique, Elizabeth & Rossi, Giovanni & Enfield, Nick J. 2015. Universal Principles in the Repair of Communication Problems. *Plos One* 10 (9) (DOI: 10.1371/journal.pone.0136100).

Hennoste, Tiit 2001. Allkeeled. Hennoste, Tiit (toim). *Eesti keele allkeeled*. Tartu Ülikooli eesti keele õppetooli toimetised 16. Tartu, lk 9–56.

Hennoste, Tiit & Habicht, Külli & Metslang, Helle & Prillop, Külli & Laanesoo, Kirsi & Ogren, David & Pärismaa, Liina & Pärt, Elen & Rumm, Andra & Rääbis, Andriela & Simmul, Carl Eric 2020. Diskursusemarker *(ma) arvan (et)*. Erelt, Mati & Reinsalu, Riina (toim). *Emakeele Seltsi aastaraamat* 65. Tallinn: Teaduste Akadeemia Kirjastus, lk 63–90 (DOI: 10.3176/esa65.03).

Hennoste, Tiit & Habicht, Külli & Metslang, Helle & Prillop, Külli 2021. Kuue (inter) subjektiivsuspartikli kasutus eesti keele registrites. Erelt, Mati & Reinsalu, Riina (toim). *Emakeele Seltsi aastaraamat* 66. Tallinn: Teaduste Akadeemia Kirjastus, lk 91–123 (DOI: 10.3176/esa66.04).

Hennoste, Tiit & Koit, Mare & Rääbis, Andriela & Valdisoo, Maret 2004. Developing a dialogue act coding scheme: An experience of annotating the Estonian Dialogue Corpus. Oostdijk, Nelleke & Kristoffersen, Gjert & Sampson, Geoffrey (toim). *LREC 2004. IV International Conference On Language Resources and Evaluation. Workshop: Compiling and Processing Spoken Language Corpora.* 24th May 2004. Lisboa, Portugal. Lisboa, lk 40–47.

Kallen, Jeffrey & Kirk, John 2012. *SPICE-Ireland: A User's Guide*. Documentation to accompany the SPICE-Ireland Corpus: Systems of Pragmatic annotation in ICE-Ireland. Belfast: Cló Ollscoil na Banríona.

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, lk 7–36 (DOI: 10.1007/s40607-014-0009-9).

Narrog, Heiko 2017. Three types of subjectivity, three types of intersubjectivity, their dynamicization and a synthesis. Van Olmen, Daniel & Cuyckens, Hubert & Ghesquière, Lobke (toim). *Aspects of Grammaticalization: (Inter)subjectification and Directionality*. Berlin/Boston: De Gruyter Mouton, lk 19–46 (DOI: 10.1515/9783110492347-002).

Orasmaa, Siim & Käärik, Reina & Vilo, Jaak & Hennoste, Tiit 2010. Information Retrieval of Word Form Variants in Spoken Language Corpora Using Generalized Edit Distance. Calzolari, Nicoletta & Choukri, Khalid & Maegaard, Bente & Mariani, Joseph & Odjik, Jan (toim). *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. ELRA, lk 623–629.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21, lk 1112–1130 (DOI: 10.3758/s13423-014-0585-6).

Rühlemann, Christoph 2019. *Corpus linguistics for Pragmatics.* A Guide fot Research. Routledge Corpus Linguistics Guides. Taylor and Francis. Kindle Edition.

Zipf, George K. 1935. *The psycho-biology of language: an introduction to dynamic philology*. Boston: Houghton-Mifflin.

Taavitsainen, Irma & Jucker, Andreas H. & Tuominen, Jukka (toim) 2014. *Diachronic corpus pragmatics.* Pragmatics & beyond new series. Amsterdam: John Benjamins (DOI: 10.1075/pbns.243).

Thompson, Sandra 2002. "Object complements" and conversation: towards a realistic account. *Studies in Language* 26, lk 125–164 (DOI: 10.1075/sl.26.1.05tho).

Traugott, Elizabeth Closs 2003. From subjectification to intersubjectification. Hickey, Raymond (toim). *Motives for language change*. Cambridge: Cambridge University Press, lk 124–139 (DOI: 10.1017/CBO9780511486937.009).

Tuldava, Juhan 1977. Sagedussõnastik leksikostatistilise uurimise objektina. *Töid keelestatistika alalt* 2. Keelestatistika. TRÜ toimetised. Tartu, lk 141–171.

Vaik, Kristiina & Muischnek, Kadri 2018. Eestikeelsete veebitekstide automaatne liigitamine. *Eesti Rakenduslingvistika Ühingu aastaraamat* 14, lk 215–229 (DOI: 10.5128/ERYa14.13).

Vaik, Kristiina & Sirts, Kairit & Muischnek, Kadri 2020. Dimensionaalne tekstimudel. Teoreetiline ülevaade. *Keel ja Kirjandus* 10, lk 875–898.

Verhagen, Arie 2005. *Constructions of Intersubjectivity: Discourse, Syntax, and Cognition*. Oxford: Oxford University Press (DOI: 10.1093/acprof:oso/9780199226702.001.0001).

Summary

We can't get by without the pragmatics corpus. Corpus pragmatics and the pragmatics corpus

Külli Prillop
Research Fellow
Institute of Estonian and General Linguistics
University of Tartu
kulli.prillop@ut.ee

Tiit Hennoste
Associate Professor
Institute of Estonian and General Linguistics
University of Tartu
tiit.hennoste@ut.ee

Külli Habicht
Associate Professor
Institute of Estonian and General Linguistics
University of Tartu
kulli.habicht@ut.ee

Helle Metslang
Professor Emerita
Institute of Estonian and General Linguistics
University of Tartu
helle.metslang@ut.ee

Keywords: discourse marker, Estonian language, language statistics corpus pragmatics, register, (inter)subjectivity, text type

Within the project "Pragmatics above grammar: Subjectivity and intersubjectivity in Estonian registers and text types" (PRG341) we are studying the expression of subjectivity and intersubjectivity in different written and spoken registers of modern Estonian. We focus on adverbs that function as discourse markers (e.g. *vist* 'maybe, probably', *ilmselt* 'apparently, obviously', *tegelikult* 'actually'), markers that develop from main clauses containing cognition verbs that take sentence complements (e.g. *(ma) arvan* 'I think'*, usun* 'I believe'*, (mulle) tundub* 'it seems (to me), it appears (that)') as well as modal and performative verbs (e.g. *võib (juhtuda)* 'can (happen)', *peaks (tulema)* 'should (come)'; *kinnitan*/*väidan (olevat)* 'I affirm/claim'). The analysis combines quantitative corpus-linguistic and qualitative pragmatic approaches, thus belonging to the field of corpus pragmatics. Unlike previous studies of related topics, the project systematically compares the usage of markers in different registers (spoken, online communication, print texts) and text types.

The pilot studies performed thus far have revealed several problems with the existing Estonian corpora, important in the study of pragmatics. Firstly, some text types are underrepresented or not represented at all, the text types cannot always be distinguished, and the particular text may not always correspond to the nominal text type (e.g. an academic text may contain quotes from texts of other types). All of this makes it difficult to do comparative statistical analysis of different text types. Secondly, the markers under examination are multifunctional and identifying their (inter)subjective function requires consideration of context broader than a single sentence. However, the public search systems for the existing corpora do not provide this context. For instance, the discourse marker function of cognition verbs is indicated primarily by the fact that the topic of the conversation or text follows through the subordinate clause, not the main clause. Since the available search systems do not provide context larger than a single sentence, the identification of the topic of the discourse, and therefore of the potential discourse-marker function of the verb, is made more difficult.

To avoid these problems, the project working group is developing a new "Pragmatics" corpus, being created in the SketchEngine environment. The corpus is made up of 10 subcorpora representing different text types and registers. Each subcorpus contains roughly 500,000 words.

Külli Prillop (PhD) on Tartu Ülikooli eesti fonoloogia teadur. Lisaks fonoloogia uurimisele osaleb ta pragmaatika projektirühmas, on vana kirjakeele korpuse üks koostajatest ja on loonud korpuste märgendusprogramme.

Külli Prillop (PhD) is Research Fellow of Estonian phonology at the Institute of Estonian and General Linguistics, University of Tartu. She is also a member of the project team of pragmatics, one of the compilers of the corpus of old literary Estonian, and has created tagging solutions for text corpora.

kulli.prillop@ut.ee

Tiit Hennoste (PhD) on Tartu Ülikooli eesti ja üldkeeleteaduse instituudi kaasprofessor. Tema peamine uurimisvaldkond on suuline eesti keel ja suhtlus, lisaks eesti keele allkeeled ja spontaanse netisuhtluse keel.

Tiit Hennoste (PhD) is Associate Professor at the Institute of Estonian and General Linguistics of the University of Tartu. His main fields of research are spoken Estonian and spoken interaction, varieties of Estonian, and the language of spontaneous online communication.

tiit.hennoste@ut.ee

Külli Habicht (PhD) on Tartu Ülikooli eesti ja üldkeeleteaduse instituudi eesti keele kaasprofessor. Tema põhilised uurimisvaldkonnad on eesti vanem kirjakeel, kirjakeele morfosüntaktiline varieerumine ja muutumine ning pragmaatikaüksuste kujunemine.

Külli Habicht (PhD) is Associate Professor of the Estonian language at the Institute of Estonian and General Linguistics, University of Tartu. Her main research areas are older literary Estonian, morphosyntactic variations and development of the literary language, and the development of pragmatic units.

kulli.habicht@ut.ee

Helle Metslang (PhD) on Tartu Ülikooli emeriitprofessor. Tema uurimisvaldkondadeks on morfosüntaks, pragmaatika, keele dünaamika, keele varieerumine, ajalooline sotsiolingvistika, kontrastiivlingvistika ja keeletüpoloogia.

Helle Metslang (PhD) is Professor Emerita at the University of Tartu. Her research interests include morphosyntax, pragmatics, language dynamics, language variation, historical sociolinguistics, contrastive linguistics, and language typology.

helle.metslang@ut.ee