Radioloogiauuringute vastuste lühendite ja lühendamise korpuslingvistiline analüüs

Eola Valdre[1], Peeter Ross[2,4], Katrin Tsepelina[1], Kaarel Veskis[1], Tarmo Vaino[1], Heiki-Jaan Kaalep[3]

Eesti Arst 2014; 93(9):502–512

Saabunud toimetusse: 23.12.2013
Avaldamiseks vastu võetud: 01.04.2014
Avaldatud internetis: 31.10.2014

[1] TÜ eesti ja üldkeeleteaduse instituut,
[2] TTÜ tehnomeedikum,
[3] TÜ arvutiteaduse instituut,
[4] Ida-Tallinna Keskhaigla

Kirjavahetajaautor:
Peeter Ross
peeter.ross@ttu.ee

Võtmesõnad:
terviseandmed, radioloogiauuringu vastus, vabatekst, lühendamine, korpuslingvistika

Taust ja eesmärgid. Elektrooniline haiguslugu nõuab selle täitjatelt võimalikult ühetaolist keelekasutust. Samas erineb Eesti meditsiini akadeemiliste ja igapäevatöö tekstide kirjakeel märgatavalt. Haigusloo vabateksti sõnavara (sh terminikasutuse) ühtlustamiseks ja/või standardimiseks on seetõttu oluline uurida igapäevatöös kasutatavat keelt. Uuringu eesmärk on rakendada korpuslingvistilist analüüsi, et uurida radioloogiauuringute vastustes kasutatud lühendamisstrateegiaid ja hinnata lühendite variatiivsust.

Metoodika. Radioloogiauuringute vastuste lühendite uurimiseks kasutati korpuslingvistilist analüüsi. Analüüsiks koostati radioloogiauuringute isikustamata vastustest keelekorpus, mis hõlmas 207 534 vastust ning sisaldas üle 11,8 miljoni sõne. Lühendite tuvastamiseks korpusest rakendati selleks koostatud erireegleid, mis arvestasid lühendamise, vastuste ja eesti meditsiinikeele eripära. Uuringuvastuste lühendite sagedust ja tähendust analüüsisid eraldi kaks arsti.

Tulemused ja järeldused. Reeglipõhise otsingu tulemusel arvati lühenditeks ja analüüsiti 10 606 sõnet, mis esinesid korpuses 446 158 korda. Lühendikasutus oli suhteliselt meelevaldne ja ebajärjepidev: analüüsitud 2453 ühetähenduslikule lühendivariandile vastas 811 mõistet. Lühendite rohkus ja varieeruvus võib olla seletatav ajasurve ja tööpingega, suhtumisega kirjakeele normi, mida ei peeta kliinilises töös esmatähtsaks, ning keelestandardi puudumisega. Lühendite sünonüümia tuleneb osaliselt ka nende keelelisest algupärast ning peegeldab muutuva keelekeskkonna mõju: toimub nihe ladina keelelt inglise keelele. Ühtlustamata ja standardimata keelekasutus võimaldab teksti mitmeti tõlgendada ning pärsib infotehnoloogilisi rakendusi. Korpuslingvistiline analüüs on meditsiinikeele uurimise efektiivne vahend. Meditsiinilingvistika on multidistsiplinaarne valdkond.

Elektroonne haiguslugu on tinginud meditsiinikeele ühtlustamise ja terminoloogia korrastamise vajaduse. Digiandmebaaside kasulikkus on selles, et andmeid saab kasutada eri eesmärkidel ja korduvalt, kuid seda ainult siis, kui andmed on järjepidevad ja omavahel võrreldavad, teisisõnu – andmebaaside kasulikkus ja kasutatavus sõltuvad oluliselt andmekvaliteedist.

E-haigusloo tekstandmed muutuksid oluliselt väärtuslikumaks, kui neis sisalduv teave oleks hõlpsasti kättesaadav, töödeldav ja analüüsitav (1). Kahjuks võib tekstandmete automaatne analüüs nii konkreetse haigusjuhu, konkreetse patsiendi kui ka konkreetse valimi puhul osutuda ebapiisava andmeesituskvaliteedi tõttu tüsilikuks või isegi võimatuks. Tervise infosüsteem on kiiresti arenenud, kuid tekstandmete koostamine on võrreldes paberil haigusluguде ajaga jäänud üsna muutumatuks. Sageli kasutatakse arvutit vaid andmete talletamiseks, nende edaspidist kasutamist silmas pidamata. Automaattöötluse ja -analüüsi jaoks peab arvuti kirjutatut üheselt mõistma. Arvutil puudub n-ö vaikeseade teksti mõistmiseks seni, kuni vajalikku teavet pole talle ette antud või õpetatud. Selleks et mõista, mis teavet võiks arvuti tekstist üheselt arusaamiseks vajada, tuleb eelnevalt uurida ja kirjeldada ka vabateksti keelekasutust.

Arstiteaduse omakeelsus arenes Eestis alles XIX sajandil, s.o oluliselt hiljem kui Euroopa suurriikides. Nii on esimesed omakeelsed meditsiinitekstid Inglismaal juba IX–XII sajandist ja esimesed prantsuskeelsed XIII sajandist (2). Kuigi ladina keel oli meditsiini *lingua franca*, tingisid valdkonna vahetu seotus praktikaga ning otsene vajadus suhelda ladina keelt mittevaldavate patsientidega omakeelse meditsiini tekke. Iga eriala keele arengut mõjutab see, mis keeles eriala õpetatakse. Eesti meditsiinikeelt on peamiselt mõjutanud ladina, saksa, vene ja inglise keel. Veelgi enam – nii saksa, vene kui ka inglise keel on olnud või on teaduskeelena eesti keele konkurendid. Praegu kasvab eriti kiiresti inglise keele mõju (3). Eestis võib täheldada meditsiinikeele kahestumist: 1) igapäevane töökeel ning 2) akadeemilise meditsiini, s.o arstiteaduse, meditsiiniõpetuse ja erialase enesetäienduse keel. Viimase puhul kasutatakse lisaks eesti keelele väga sageli ka inglise keelt, kuid igapäevatöö tekstid on peaaegu alati eestikeelsed. Seega mõjutavad Eesti meditsiinikeele arengut ühtviisi oluliselt selle kaks eri kasutusvaldkonda: akadeemiline ja töine kasutus. Akadeemiliste eestikeelsete meditsiinitekstide varieeruvus on palju väiksem kui igapäevatöös tekkinud tekstandmetel. Osaliselt seletab seda asjaolu, et akadeemilised tekstid on toimetatud tekstid. Samuti mõjutavad akadeemilise teksti loomet vähem kiirustamine, tööstress, kollegiaalne žargoon ja/või suvaline lühendamine. Arsti igapäevatöös tekkivad tekstandmed sõltuvad konkreetse autori töö- ja kirjutamiskogemusest ning on keeleliselt toimetamata.

Radioloogiauuringu vastus on haigusloo tekstandmete liik, milles radioloog kirjeldab uuringu leiu ning üldjuhul esitab patsiendi uuringule suunanud kolleegile asjakohase arvamuse. Tehnoloogia arengule vaatamata pole radioloogiauuringu vastuse ülesehitus ega sisukontseptsioon eriti muutunud (4). Vastuse kliinilist sisu piiravad uuringu liik, konkreetsed ruumisuhted (nt kehapool), visualiseeritavad anatoomilised struktuurid ning seotus teatud tüüpi patoloogiaga. Vastus kirjutatakse etteantud andmeväljas vabas vormis (nn vabatekst) ning selle stiil arvestab taotluslikult ka leiu teatavat määramatuse komponenti („ei saa välistada", „võimalik…", „kahtlus…" jts) (5). Piltlikult võib radioloogi nimetada visuaalse kujutise kirjakeelde tõlkijaks (6). Nagu iga tõlget, mõjutab radioloogi vastuse teksti väga oluliselt sihtrühm: vastuse keelekasutus sõltub sellest, kui põhjalikku valdkonna tundmist radioloog suunava arsti puhul eeldab. Suhtlus radioloogi vastuse kaudu on kollegiaalne ja toimib vaikimisi eeldusel, et kumbki osapool valdab nii erialakeelt kui ka kogu seotud mõistevälja, seetõttu ei pruugi kõik üksikasjad olla alati eraldi kirjeldatud, kehtib n-ö teadmise ja mõistmise vaikeseade (7). Uuringu vastustes kasutatakse palju lühendamist, mis eeldab samuti konkreetse (vahel väga spetsialiseeritud) eriala mõistevälja väga head tundmist.

Uuringu vastuses on kirjas vähemalt patsiendi identifikaator, kliiniline küsimus ja uuringu näidustus, uuringu nimetus ja metoodika, leiu kirjeldus ning kokkuvõtteks radioloogi arvamus (8). Kliinilisest aspektist iseloomustavad head vastust selle selgus, õigsus, usaldusväärsus, kompaktsus, täielikkus, järjepidevus, teabevahetuse sujuvus, piisav nõustamine, õigeaegsus ja andmeesituse standarditus (4). Kuigi tavaliselt on vastuse tekst piiranguteta vabatekst, võib see olla rohkem või vähem struktureeritud, järgides kokkuleppelist liigendust ja/või kasutades standarditud keelt, sh klassifikaatoreid ja sõnastikke (RadLex, SNOMED jt) (8). Struktureerimisel võidakse kasutada vastuse kirjutaja ja lugeja jaoks vastuse erinevat ülesehitust (*reporting into structure and reading structure*) (9). Leiu kirjeldamisel kasutatakse sageli lühendeid, tüüplauseid või -fraase. Publikatsioonid, kus on käsitletud vastuse loetavust sõltuvalt sellest, kas vastuses kasutati struktureeritud või vabateksti, on mõneti vastuolulised (10, 11). Vastuse struktureerimise vajadust põhjendatakse ühetaolisema andmeesitusega, mis võimaldab tekstandmete palju ulatuslikumat kasutamist (sh ajakriitilist, diakroonilist ja populatsiooniülest), muu hulgas tõhusat andmekaevet, vajaduse korral kohandamist eri huvirühmadele (raviarstid, administratsioon ja IT-spetsialistid, patsient jt) ning ka muid rakendusi (nt masintõlget). Vabateksti pooldajad osutavad vajadusele kirjeldada eripäraseid kliinilisi olukordi. Teabe kohandamisel kliendikeskseks ning andmekaeve võimaldamiseks on andmeesituse standardimine vältimatu eeldus (12).

Radioloogiauuringu vastuse vabateksti on ka meil üritatud struktureerida. Esimene radioloogiauuringu vastuse sõnavara uuriv

projekt Eestis oli *Baltic eHealth*, milles koostati vastuse tõlkemudel piiriüleseks radioloogiateenuseks. Radioloogiauuringud tehti Taanis, kuid kujutist kirjeldati Eestis või Leedus. Projekti tulemusena loodi algoritm põlve röntgenuuringute struktureeritud kirjeldamiseks ning vastuse automaatseks tõlkeks. Eestikeelse sõnavara valik põhines mitme radioloogi ühisarvamusel (13, 14).

Praeguseni on vabateksti struktureerimise üks põhitakistusi meditsiinipersonali äärmine hõivatus igapäevatööga. Sellest tuleneb ressursikasutuse küsitavus: olukorras, kus on kriitiline tagada kõigi radioloogiliste kujutiste õigeaegne ja asjatundlik kirjeldamine, ei ole võimalik kasutada kvalifitseeritud meditsiinipersonali otseste töökohustuste väliste või vaid kaugtulemusi võimaldavate küsimustega tegelemiseks. Võimalik kasu terviseandmete täielikumast kasutamisest tundub suhteliselt vähekonkreetne ning ebamäärases tulevikus.

Meditsiinikeele uurimist on siiski võimalik ratsionaliseerida. Keele eri kasutusviiside (nt suulise ja kirjaliku keele) ning eri kasutajarühmade iseloomulike allkeelte (nt jututubade ja foorumite keele) uurimiseks kasutatakse spetsiaalseid keeleressursse – lingvistilist analüüsi võimaldavaid tekstikogusid, mis koosnevad sihipäraselt valitud ning märgendatud tekstidest. Selliseid tekstikogusid nimetatakse korpusteks. Tekstikorpuste suureks eeliseks on võimalus hinnata mis tahes keeleelementide või -struktuuride kasutussagedusi ning kirjeldada konkreetsele kontekstile iseloomulikke kasutusmustreid. Eesti korpuslingvistika ja keeletehnoloogia on väga arenenud, juurdepääs olemasolevatele keeleressurssidele ja keeletehnoloogia vahenditele on võimalik TÜ arvutilingvistika töörühma kodulehe (http://www.cl.ut.ee/korpused/) ja Eesti Keele Instituudi kodulehe (http://portaal.eki.ee/keeletehnoloogia.html) kaudu.

Digitaalsete terviseandmete tõhusamaks kasutamiseks on mitmes riigis uuritud meditsiinikeelt ja koostatud meditsiinikeele korpusi. Üks probleem on, et enamik olemasolevatest korpustest kajastab vaid akadeemilist (peamiselt teaduskirjanduse) keelt, igapäevatööd kajastavaid tekstikorpusi on seni väga vähe. Andmekaitsenõuete tõttu on autentsed tekstid lingvistile kättesaamatud. Teine probleem on autentse töise meditsiinikeele äärmiselt suur spetsiifika, sh kollegiaalne žargoon ja vaikeseaded, ning tihe seotus konkreetsete kliiniliste olukordadega: sellise keele uurimine eeldab lisaks (arvuti-) lingvistika teadmistele ka meditsiinivaldkonna süvatundmist, teisisõnu – pädevaid uurijaid on vähe.

Ida-Tallinna Keskhaiglas (ITK) – ja ka mujal Eestis – kasutatakse radioloogiauuringu vastuses vabateksti, mis on jagatud kliiniliste andmete, uuringu parameetrite, leiu kirjelduse ja radioloogi arvamuse või soovitusega kokkuvõtte väljadeks. Radioloogi töö lihtsustamiseks kasutatakse ka tüüpvastuseid. Vastuse terminivaliku, tüüpfraaside ning lühendite kasutamise otsustab iga radioloog ise.

Uuringu eesmärk on korpuslingvistilise analüüsi abil uurida radioloogiauuringute vastuste lühendeid ja lühendamisviise, et nende põhjal hinnata vastuste keelekasutuse ühtlustamise ja standardimise vajadust.

METOODIKA

Radioloogiauuringute vastuste teksti uurimiseks kasutati korpuslingvistilist analüüsi. Radioloogiauuringute vastustest koosneva tekstikorpuse jaoks materjali saamiseks tegime isikustamata päringu ITK infosüsteemist. Päring hõlmas ajavahemikku 01.07.2009–01.07.2011 ja sisaldas järgmisi andmeid: uuringu liik ja nimetus (sh haigekassa kood ja uuritav piirkond), uuringu unikaalne identifitseerimisnumber, patsiendi vanus, patsiendi sugu, uuringu tellija (arst, raviasutus või osakond), tellimise aeg, vastamise aeg, vastaja, kliiniline küsimus ja kaasnevad kliinilised andmed, radioloogilise leiu kirjeldus ja kokkuvõtteks radioloogi arvamus. Lühendite korpuslingvistiliseks analüüsiks kasutasime järgmisi andmevälju: uuringu identifitseerimisnumber ja radioloogiauuringu vastus (kliinilised andmed, leiu kirjeldus ja kokkuvõte). Tekstides esinenud meditsiinitöötajate nimed kodeerisime eelkõige analüüsi võimaliku kallutatuse vältimiseks. Samuti eemaldasime algmaterjalist kaks juhtu, kus kliiniliste andmete unikaalsuse tõttu (nt vigastuse tekke aeg ja koht) oli teoreetiliselt võimalik seostada andmeid konkreetse haigusjuhuga. Korpuse materjal kohandati UNIX-süsteemis töötamiseks, korpus teisendati XMLi (http://www.w3.org/2001/XInclude), märgendati ja valideeriti TEI P5 järgi (http://www.tei-c.org/). Korpust ei lausestatud, sest seoses

eri moel tähistatud kuupäevade ja lühendite rohkusega tekstis ning osaliselt sellest tingitud lauselõpumärgi ebaharilikult sageda kasutuse ja lausealguse näpuvigade (väiketäht suurtähe asemel) tõttu oli lausete ja osalausete piire täisautomaatselt usaldusväärselt võimatu määrata. Selleks tuleb eelnevalt teksti lühendid, kuupäevad, muutelõpu ebakonventsionaalsed lisamised jms kirjeldada ning seejärel võimaluse korral märgendada olukorrad, kui lauselõpumärki on kasutatud mujal kui lause lõpus.

Korpuseotsinguks kirjeldasime otsitava tekstiosa (nt lühendid) reeglipõhiselt üldistatud kujul, s.o regulaaravaldisena, mille leidmiseks korpusest, ja esinemissageduse määramiseks kasutasime UNIXi-põhiseid käsuridu ja -jadasid. Korpusest lühendite tuvastamiseks koostasime reeglid nii, et tekstist saaks välja sõeluda võimalikult palju potentsiaalseid lühendikandidaate, seetõttu on reeglid kohati osaliselt kattuvad (vt tabel 1). Sõneks lugesime tühikutega eraldatud märgijada (stringi), kus oli vähemalt üks tähemärk või number. Käsujadade (skriptide) jaoks määratleti kõik tähemärgid eesti ja inglise keele tähestiku ning võimalikud (kuigi mitte kõik) muutelõpud eesti keele järgi. Iga reeglit rakendati kogu korpusele eraldi, tulemused koondati ühisloendisse, korduvad read eemaldati ning tulemus järjestati sageduste (esinemiskordade arv korpuses) järgi järgmise käsureaga:

cat $1/*.txt | LC_COLLATE=C sort -V | uniq -c | sort -nr > $1/_uniq_sorted_list.

Reeglite põhjal saadud lühendite sagedusloendid vaatas eraldi läbi kaks arsti, kes täpsustasid, mis leitud sõnedest olid lühendid ja mida need tähendasid, ning esitasid lühendite eestikeelsed vasted, teadaolevad sünonüümid ning võimaluse korral ka ladina- ja/või ingliskeelsed vasted.

Ühetäheliste lühendite puhul oli mitmetähenduslikkus algselt eeldatav ning tähenduste tegeliku esinemissageduse paremaks tõlgendamiseks koostasime ka sageduspõhise loendi trigrammidest, milles sõnekolmiku keskel oli uuritav ühetäheline lühend. Analüüsisime ükshaaval kõik ühetähelised lühendid ja määrasime nende tähendused, samuti analüüsime kõik vähemalt viis korda korpuses esinenud mitmemärgilised ühetähenduslikud lühendid ja määrasime nende tähendused. Mitmemärgiliste mitmetähenduslike lühendite ja ebakonventsionaalsete lühendamisviiside kirjeldamise jätsime käesoleva uuringu raames kõrvale.

Lühendid analüüsisime kahes etapis. Kõigepealt vaatasime läbi kõik lühendikandidaadid ning rühmitasime need lühenditeks ja lühendamisviisideks. Viimasteks nimetasime mis tahes sõnade (sh üldkeele

Tabel 1. Vastuste tekstidest lühendite otsingu kriteeriumid

	Nr	Sõne pikkus	Mall	Näide
Tüüp 1 (väikese tähemärkide arvuga või lühikesed sõned)	1	1–2 sümbolit	Sõne alguses on tähemärk, millele võib järgneda üks kirjavahemärk või erisümbol.	a., a, A., A, a), a-, A-
	2	3–5 sümbolit	Sõne sisaldab vähemalt üht tähemärki ja vähemalt üht side- või mõttekriipsu või punkti. Sõne lõpus tohib olla üks kirjavahemärk või erisümbol.	pt., kr., CT., VAS., KATE., mts?., ves., man., AXIN., PCN-s, XII-L, Dgn.:, palp., cerv., ESV., hebd., IIst.
	3	piiramata	Sõne alguses on 1–3 tähemärki, millele järgneb mis tahes arv numbreid.	Th11, S1, L1, L5, L3
Tüüp 2 (keelele mitteomased sõned)	4	2–3 sümbolit	Sõne koosneb 2–3 vokaalist või 2 suvalisest tähemärgist.	ii, aa, MR
	5	3–6 sümbolit	Sõne alguses on 2 konsonanti, millele järgneb 1–3 tähemärki. Sõne lõpus tohib olla üks kirjavahemärk või erisümbol.	vrdl, gr., pt., kroon., FLAIR, tse, prk, Dm-ga, fract jne. (Haarab ka peamiselt võõrkeele algupära sõnu, mis ei ole lühendid, nagu klass, tsüst, grade.)
Tüüp 3 (lühemad sõned, millele on käändelõpp liidetud)	6	3–11 sümbolit	Sõne alguses on 1–5 tähemärki, millele järgneb side- või mõttekriips, millega on liidetud muutelõpp.	EKG-ga, KT-l, pt-l, KATE-t, Rö-lt, VEM-il, UH-ga, cm-ni, AVM-le, Op-tud, ACI-de, prk-s
Tüüp 4 (laiendavad reeglid)	7	piiramata	Sõne koosneb tähemärkidest, mis on eraldatud kaheks osaks side- või mõttekriipsuga. Sõnes võib olla ka numbreid. Sõne lõpus tohib olla üks kirjavahemärk või erisümbol.	l-sõlmi, KT-uuring., L4-L5, ThI-ThXII, MT-luu, SI-liidus
	8	piiramata	Kõik sõned, mis lõppesid punktiga nii, et järgnev sõne algab väiketähega. Viimane reegel haarab osaliselt ka lauset lõpetavad suvalised sõnad, kui järgnev lause ei alga näpuvea tõttu suurtähega, kuid mall võimaldab leida tekstist täiendavalt kui tahes pikki punktiga lõppevaid lühendeid.	bilat., patol., isel., Bronch., Pneum., Postop., intervert., degenerat., auskult., talocr., hypert.

sõnade) harva esinevaid ning ootamatuid sõnakatkeid, mis tegelikult ei lühendanud sõna oluliselt ja mille kasutus üldkasutatava lühendina on kaheldav (nt *k-nääre, lumbosakraal., degenerat., tõenäol.*). Kõik lühendid rühmitasime läbivaatamisel omakorda ühe- ja mitmetähenduslikeks lühenditeks. Ühetähenduslike lühendite rühmast analüüsisime kõik kuni viis korda korpuses esinenud lühendid nii, et alguses määras nende tähenduse üks arst ning seejärel vaatas kogu töö üle teine arst (radioloog). Ühetähenduslikud lühendid jagasime ülevaatlikkuse huvides järgmiselt: üldkeele, anatoomia (sh lülisamba anatoomia lühendid, mida vaatlesime lühendite eripärase numbrikasutuse tõttu eraldi), patoloogia ja kliinilise meditsiini lühendid (vt tabel 2).

Korpuslingvistiliseks analüüsiks ei ole kasutatud isikustatud andmeid, isikustamata andmete uuringuks on saadud luba Tallinna meditsiiniuuringute eetikakomiteelt (14.10.2010. a otsus nr 2169).

TULEMUSED

Korpuse kirjeldus

Radioloogiauuringute vastuste korpuse koostamiseks kasutati kokku 207 534 vastuse teksti. Algmaterjalis oli kokku 11 865 356 sõnet. Materjal jagunes uuringu liikide kaupa järgmiselt (uuringute arv; sõnede hulk kokku): röntgenuuringud (139 998; 4 663 958), ultraheliuuringud (34 020; 2 970 399), KT-uuringud (20 725; 2 751 990), MRT-uuringud (11 037; 1 293 070), nukleaarmeditsiini uuringud (1754; 185 939). Keskmiselt koosnes vastus 57,2 sõnest, keskmine sõnede arv oli kõige väiksem (33,1) röntgenuuringute vastustes, suurim keskmine sõnede arv (132,8) oli KT-uuringute vastustes (vt joonised 1–5). Lühimad leiu kirjeldused koosnesid kõigest ühest sõnest (nt *Iileus, Normis, Normileid*), pikimas vastuses oli 4882 sõnet.

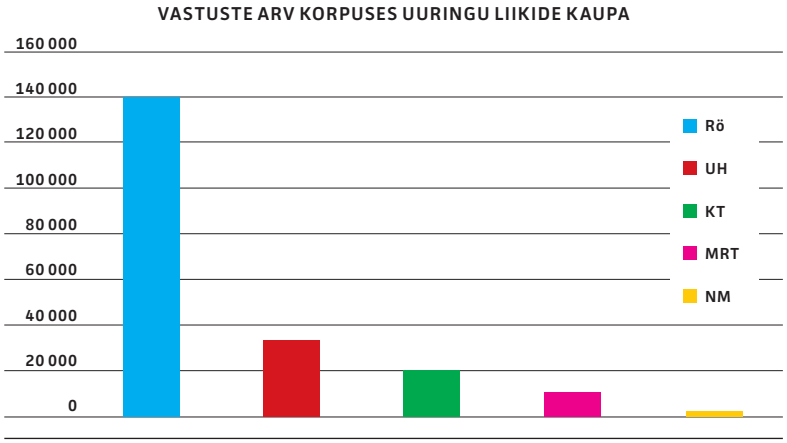
Kuigi korpuse materjali ei analüüsitud uuringu vastuse koostajate kaupa, jäid silma teatud radioloogiomased erisused (nt vastuste kirjutamine ainult suurtähtedega, ladina keeles, osa teksti rõhutamine suurtähtedega). Vahel oli kokkuvõtte väli, kus pidanuks olema radioloogi arvamus, tühi, kuid arvamus oli ikkagi esitatud uuringu leiu kirjelduse lõpus. Sageli esines tekstides tühikuvigu, ülearuseid tühje ridu, lauselõpumärgita lauseid ning sõnede n-ö kokkukleepumist teiste sõnede või mittetähemärkidega. Kõik see raskendas teksti automaatset analüüsi oluliselt.

Lühendamise analüüs

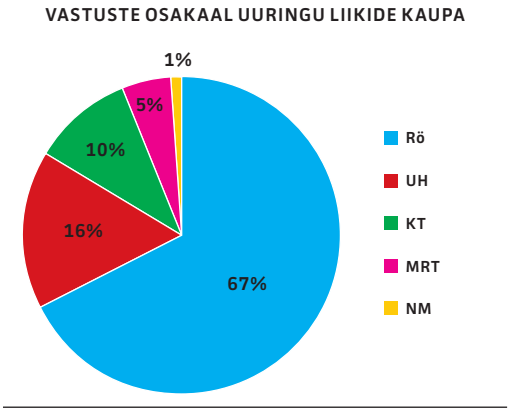
Lühendite analüüsiks õnnestus reeglite alusel korpusest välja sõeluda 14 961 lühendikandidaati, mida oli kokku kasutatud 1 250 260 korda (s.o 10,5% korpuse kõigi sõnede suhtes). Läbivaatamisel osutus neist tegelikeks lühenditeks 10 606 ja nad esinesid 446 158 korda (s.o 3,8%). Ülejäänud olid kas lühikesed tavasõnad (nt *ja, ei*), näpuvead, lauselõpu sõnad, millele järgnes punkt, eesvõi järelliited, muutelõpud, sidekriipsuga paarissõnad (nt *aeg-ajalt, enam-vähem*), RHK-koodid, kodeeritud vastajate nimed, tekstis tähemärkidest eraldi esinenud Rooma numbrid (s.o I kuni XII või sõned, milles lisaks Rooma numbritele oli mõni mittetähemärk (nt *I-II, III/X, 5XII*)). Rooma numbrite puhul oli sageli võimatu üheselt määrata nende võimalikku tähendust, s.t jäi selgusetuks, kas nad tähistavad numbrit, kuud, segmenti, lüli või veel midagi.

Tabel 2. Mitmemärgiliste ühetähenduslike lühendite analüüs: radioloogiauuringute vastuste ühetähenduslike lühendite arv kõigi uuritud (kuni viis korda esinenud) mitmemärgiliste lühendite seas

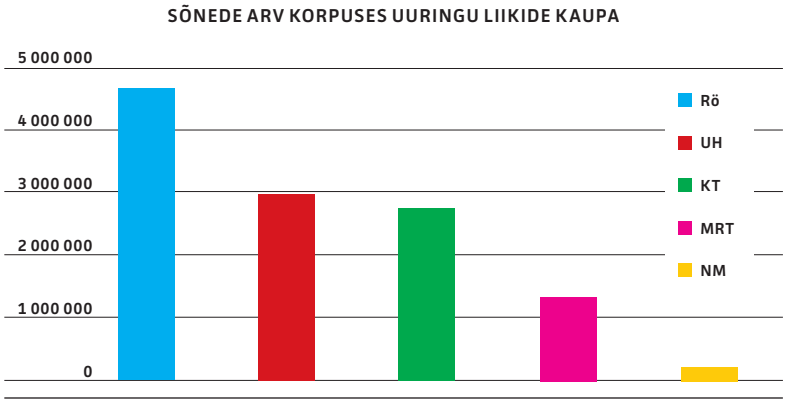
VALDKOND	LÜHENDID		LÜHENDAMISVIISID		KOKKU	
	Esinemiskordade arv	Mõistete arv	Esinemiskordade arv	Mõistete arv	Esinemiskordade arv	Mõistete arv
üldkeel	106 902	47	1720	17	108 622	73
radioloogia	80 748	45	778	11	81 526	62
anatoomia	74 666	154	11 497	45	86 163	199
lülisamba anatoomia	25 041	89	0	0	25 041	89
patoloogia	13 212	123	3564	18	16 191	141
kliiniline	15 521	100	1441	8	16 962	108
KOKKU	316 090	558	19 000	99	334505	672



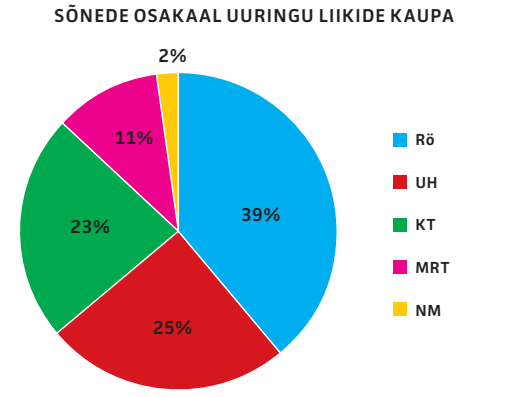
Joonis 1. Korpuse kirjeldus: radioloogiauuringute vastuste jaotus korpuses uuringuliikide kaupa.



Joonis 2. Vastuste osakaal uuringu liikide kaupa.



Joonis 3. Korpuse kirjeldus: radioloogiauuringute vastuste sõnede arv korpuses uuringuliikide kaupa.

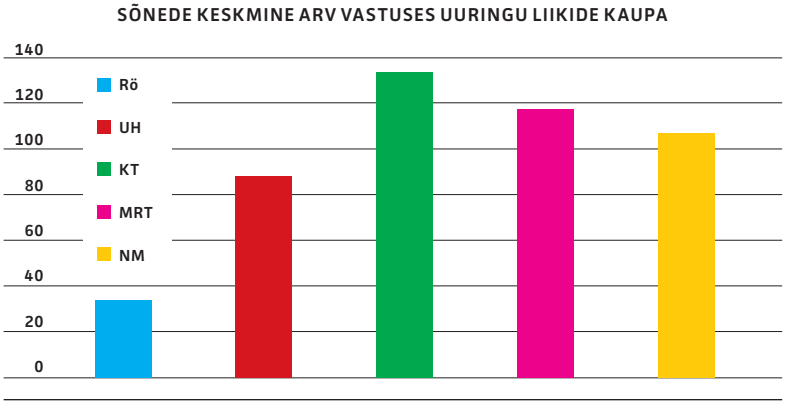


Joonis 4. Korpuse kirjeldus: radioloogiauuringute vastuste sõnede osakaal uuringuliikide kaupa.

Korpuses kuni viis korda esinenud mitmemärgiliste sõnedega oli ühetähenduslikult lühendatud 672 mõistet, kokku kasutati ühetähenduslikke lühendeid 334 505 korda, mis moodustab kogu korpuse sõnedest umbes 2,8% (vt tabel 2).

Mitmetähenduslikest lühenditest analüüsiti mittetõstutundlikult ja lisatud kirjavahemärkidest sõltumatult kõigi ühetäheliste lühendite kõik tähendused (vt tabel 3). Üksiktähti esines korpuses kokku 80 571 korda. Lühenditena kasutati neid 44 034 korda, tähisena 34 741 korda, üksiktähed osutusid näpuvigadeks 305 korda ning nende tähendus jäi teadmata 906 korral (ligikaudu 1%). Üksiktähed tekstis võisid olla ka muutelõpud (402) või mitmetäheliste lühendite osad jms (183), mis jäid tekstis tühiku või kirjavahemärkide vale kasutuse tõttu eraldi.

Kolm sagedaimat lühendit olid üksiktäht X (x) tähenduses „korda“ (31 177), *P-A* tähenduses „posterio-anterioorne“ (26 123) ja sh. tähenduses „sealhulgas“ (25 316).



Joonis 5. Korpuse kirjeldus: radioloogiauuringute vastuste sõnede keskmine arv vastuses uuringuliikide kaupa.

Joonistel on kasutatud järgmised lühendid: Rö – röntgenuuringud, UH – ultraheliuuringud, KT –kompuutertomograafilised uuringud, MRT – magnetresonantstomograafilised uuringud, NM – nukleaarmeditsiini uuringud.

Kõige sagedamini esinenud lühend X (x) oli mitmetähenduslik, rohkem mitmetähenduslikke lühendeid kümne sagedaima lühendi hulgas ei olnud (vt tabel 4).

Kogu lühendite analüüs võimaldas tekstist ühetähenduslikult kindlaks teha

811 mõistet, neid kirjeldavaid termineid oli kokku lühendatud 2453 korda (keskmiselt 3 lühendit mõiste kohta). Mõisteid, mida kirjeldavaid termineid lühendati alati vaid ühtmoodi, oli 46,7% (nt uuringute metoodika spetsiifilised lühendid nagu *MRA, TRUS, PNB, MRCP, FLAIR, MTF*) ja mõisteid, mille terminite lühendamisel kasutati palju erinevaid lühendivariante, oli 53,3% (sh eriti anatoomiliste struktuuride nimetused, nt *viies metatarsaalluu* oli lühendatud 25, *ühissapijuha* 24 ja *õndlaarter* 27 eri lühendivariandiga). Ühe mõiste kohta esinevate lühendite rohkust ja varieeruvust mõjutab oluliselt lühendi algupära keel: üldiselt on eesti- ja ingliskeelsete lühendite varieeruvus oluliselt väiksem kui ladinakeelsetel lühenditel. Anatoomiliste struktuuride kohta kasutati eelistatult ladinakeelsete terminite põhiseid lühendeid. Kirjeldatud mõistete kohta oli kõige rohkem ladina algupära lühendeid (69,8%), eestikeelseid lühendeid oli 20,6%, inglise algupära lühendeid 7,5% ja rahvusvahelisi lühendipõhiseid tähiseid 2,1% (nt mõõtühikud, keemilised elemendid).

Tabel 3. Ühetäheliste lühendite analüüs: ühetäheliste lühendite tähendused ja tähenduste esinemissagedus korpuses

Lühend või tähis	Tähendus ja esinemiskordade arv tekstis
A	*arteria* 7100, aasta 2054, *articulatio* 1123, ligikaudu 278, A-tüüp või staadium 24, aort 11, *truncus arteriosus* 2, anamnees 2, anteroposterioorne 2, Weber A 2, hemofiilia A 2, Apgar 1, Kerley A jooned 1, A H1N1-tekkene pneumoonia 1, A-ösofagiit 1, A kaksik 1
B	B FFE 1421, Kerley B jooned 22, B-staadium 19, *bulbus* 18, B-hepatiit 15, B-staadium 12, Brickeri juha 8, Weber B 7, B_{12}-vitamiin 5, B-tüüp 5, teine 2, *bursa* 1, Balthazar B 1, B-ösofagiit 1, HLA B27 1
C	*cervicalis* 305, *cavum/cavitas* 157, C-hepatiit 104, *colon* 33, C-staadium 14, Celsius 13, RHK-kood 9, *corpus* 3, *carotis* 3, Weber C 3, *columna* 2, *cauda* 1, *carpometacarpalis* 1, *circulus* 1, *conus* 1, *(sectio) caesara* 1, C-tsirroos 1, C-ösofagiit 1, loetelu punkti tähis 1
D	*ductus* 438, diameeter 144, *dexter* 31, D-dimeerid 27, *digitus* 13, diagnoos 9, distants 7, Douglase õõs 6, 3D 6, *discus* 4, *diabetes mellitus* 2, RHK-kood 1, Balthazar D 1
E	ehk 401, *e* 11, RHK-kood 8, esmaspäev 2, *E. coli* 1, epilepsia 1
F	frontaal- 319, F (dreeni suuruse tähis) 15, funktsioon 2, *fissura* 1, *femoris* 1, *fascia* 1, film 1, *fossa* 1
G	*gauge* 281, *genu* 12, gramm 11, *grade* 4, gastrointestinaalne 2, Gram-positiivne 1, *gyrus* 1
H	*in hebdominis* 245, *hiatus* 50, tund 27, kõrgus 17, *haemispherium* 5, Hunteri kanal 4, hingamiskahin 3, *haematoma* 1, haigla 1, *helicobacter* 1, *humeroscapularis* 2
I	esimene 7231, *intima media* 22, jaanuar 7, RHK-kood 3, intra- 2, jood-131 2
J	jood-131 3, RHK-kood 3, järgi 2, järgne 2
K	kuu 54, kontrastaine 8, kell 6, K-varras 5, kord 4, kaalium 3, RHK-kood 1
L	*lumbalis* 1387, esimene 95, *lobus* 80, liiter 43, *loco* 3, ligament 1, *linea* 1, L-kujuline 1
M	*musculus* 3404, *morbus* 665, *membrum* 92, meeter 73, *manus* 17, *membrana* 2, mees 1, *medulla* 1, RHK-kood 4
N	*nervus* 903, norm 404, *nucleus* 43, nädal 38, nina-CPAP 2
O	oktsipitaal- 52, *os* 17, *oculus* 6, objektiivselt 1, *ostium* 1, O-seis 2
P	päev 179, parietaal- 177, parem 105, pärast (sh post) 25, *pulmonis* 13, *papilla* 3, *phalangea* 3, *pupilla* 1, prostata 1, põsk-(koobas) 1, pool 1
Q	Qmax 110, Q-sakk 2, RHK-kood 2
R	rasedusnädal 8, *ramus* 7, *right* 1
S	suund 1061 (2 s – kahes suunas), *solutio* 127, segment (kopsu-, maksa-) 117, S1 87, S-kujuline 65, *substantia (nigra)* 39, S-skolioos 32, *sinister* 19, sakrum 6, sümptom 3, *status* 3, S niverdus 2, suur (kõverik, varvas) 2, sündroom 2, *sectio* (*caesarea*) 2, S deviatsioon 1, sümptomaatika 2, *sinus* (*cavernosus*) 1, *Staphylococcus* 1
T	temperatuur 150, tund 95, temporaal- 61, tüüp 60, lülisamba torakaalosa 56, täpsustamata 14, tuumor 11, T-sakk 10, *truncus coeliacus* 7, T (tesla) 3, T2 2, aeg 1, tablett 1, tõus 1, tänav 1, Tallinn 1, T (Spair) 1
U	umbes 1504, *ulcus* 1, uuring 1, *unit* 1, U-kiud 2, U-kujuline 1
V	*vena* 8886, viies 3050, *volume* 168, *vesica* 99, väike 82, vasak 52, või 16, *vertebra* 14, mai 6, *valva* 5, venoosne 1, vaba 1
W	vatt 310, Willisi ring 3
X	korda 31842, HK-kood korda (haigekassa kood x n) 269, kümnes 155, oktoober 8, X-seis 2
Y	Y-protees 2, Y-kujuline 1
Z	RHK-kood 1

*Ühetähelised lühendid esinesid nii suur- kui ka väiketähtlühenditena.

ARUTELU

Kättesaadavad eesti keeleressursid ei võimalda igapäevatöö meditsiinikeele uurimist. Ajakirja Eesti Arst korpus (15) on representatiivne akadeemilise eesti meditsiinikeele, mitte töise keelekasutuse uurimiseks. Seda, et akadeemiline meditsiinikeel ei sobi töise keelekasutuse uurimiseks, on tõdetud juba varasemates uuringutes (1, 16, 17). Valisime lühendite uurimiseks teadlikult vaid ühe vabatekstiliste terviseandmete liigi – radioloogiauuringute vastuse, sest selle tekst on suhteliselt lühike, kuid uurimiseks piisava mahuga ning kliinilisest seisukohast piiritletud. Eeldasime, et radioloogiauuringute vastustes võib esineda kordusi ja tüüpfraase, võõrkeele mõjusid, kollegiaalset žargooni ja rohkelt lühendeid.

Heas keelekorpuses on vajalik teave kirjeldatud ja asjakohaselt märgendatud. Keele seisukohalt võib eesmärgist sõltuvalt olla vaja näiteks lausepiiride, kuupäevade, lühendite, võõrkeele märgendamist. Olemasolevas korpuses oli automaatselt usaldusväärselt võimalik märgendada ainult tekstide struktuuri, usaldusväärseks automaatseks lausestamiseks ja morfoloogiliseks märgendamiseks ning ühestamiseks on vaja lisateavet. Analüüsisime lühendeid, sest neid esineb terviseandmete vabatekstis palju ning nende märgendamine võimaldaks ehk paremini korpust automaatselt lausestada.

2012. aastal avaldasid Rootsi kolleegid Kvisti ja Velupillai artikli, milles kirjeldasid radioloogiauuringute vastuste korpust, mille abil uuriti vastuste sõnavara ja keelekasutust eesmärgiga seda lihtsustada nii, et vastus oleks patsiendile arusaadavas keeles (1). Selles korpuses on vastuseid umbes kaks korda rohkem kui meie omas (434 427 vastust), kuid sisu on võrreldav, sest uuriti samuti ainult radioloogide vastuseid. Autorid sedastasid, et 100 korpuses enim esinenud sõna seas oli lühendeid 18 ehk ligi viiendik, mis ei ole iseenesest kuigi üllatav tulemus, kui arvestada, et mitmed neist (kokku 7) olid üldkeele lühendid või tähised (*nt* ja telefon – *tel*). Reisbergi jt Eesti Arstis

Tabel 4. Suurima esinemissagedusega lühendid korpuses

Olemus	Valdkond	Mõiste	Mõiste lühendamiskordade arv	Sagedaim variant	Sagedaima lühendivariandi esinemiskordade arv
ühetäheline, X	üldkeel	korda	**31 177**		31 177
mitmemärgiline	anatoomia	posterio-anterioorne	**26 123**	P-A	26 007
mitmemärgiline	üldkeel	sealhulgas	**25 316**	sh.	24 762
mitmemärgiline	radioloogia	vaibeaeg T2 (relaksatsiooniaeg T2)	**19 904**	T2	19 890
mitmemärgiline	radioloogia	vaibeaeg T1 (relaksatsiooniaeg T1)	**19 448**	T1	19 442
mitmemärgiline	anatoomia	parem; paremal	**16 225**	par.	10 079
mitmemärgiline	anatoomia	vasak; vasakul	**16 181**	vas.	10 195
mitmemärgiline	üldkeel	sentimeeter	**16 155**	cm.	13 016
mitmemärgiline	üldkeel	milliliiter	**14 918**	ml.	14 683
mitmemärgiline	radioloogia	lahus (*Solutio*)	**12 359**	Sol.	12 359
ühetäheline, V	anatoomia	*vena*	**8886**		
mitmemärgiline	üldkeel	millimeeter	**8404**	mm.	5301
mitmemärgiline	radioloogia	magnetresonantstomograafia-	**8194**	MRT-	7160
mitmemärgiline	radioloogia	Rö	**7932**		
ühetäheline, I	üldkeel	esimene	**7231**		
ühetäheline, A	anatoomia	arter	**7100**		
mitmemärgiline	üldkeel	milliliitrit sekundis	**5069**	ml/s.	5069
mitmemärgiline	radioloogia	magnetresonants-	**4284**	MR	4029
ühetäheline, M	anatoomia	lihas (*musculus*)	**3404**		
mitmemärgiline	kliiniline	kliiniline	**3126**	kl.	2922
ühetäheline, V	üldkeel	viies	**3050**		
mitmemärgiline	kliiniline	patsient	**2600**	pt.	1767
mitmemärgiline	patoloogia	metastaas	**2363**	mts	2363
mitmemärgiline	radioloogia	kompuutertomograafia-	**2275**	KT	1131
ühetäheline, A	üldkeel	aasta	**2054**		
mitmemärgiline	kliiniline	operatsioon	**1984**	op.	1418
ühetäheline, U	üldkeel	umbes	**1504**		
ühetäheline, B	radioloogia	B FFE	**1421**		
ühetäheline, L	anatoomia	*lumbalis*	**1387**		
mitmemärgiline	anatoomia	metatarsaal-	**1381**	MT	1353
mitmemärgiline	anatoomia	side (*ligamentum*)	**1357**	Lig.	1181
mitmemärgiline	patoloogia	luumurd (*fractura*)	**1275**	fr.	986
mitmemärgiline	anatoomia	liiges (*articulatio*)	**1123**		
mitmemärgiline	kliiniline	diagnoos	**1080**	dgn.	924
ühetäheline, S	radioloogia	suund (kahes suunas)	**1061**		
mitmemärgiline	patoloogia	kopsuarteri trombemboolia	**1039**	KATE	1039
mitmemärgiline	radioloogia	ultraheli-	**1025**	UH	750
mitmemärgiline	patoloogia	patoloogia; patoloogiline	**1021**	patol.	434

avaldatud artiklis, kus uuriti Tartu perearstide infosüsteemi tekstide näitel muu hulgas lühendamist, oli 29 enim kasutatud sõna seas lühendeid 17 ehk 58,6% (17). Meie korpuse maht on võrreldav Reisbergi jt materjalis kasutatud radioloogiauuringute vastuste mahuga (neid oli kogu materjalis kokku 156 000), kuid mitte uuringute jaotusega. Uuringuliigiti oli Reisbergi jt kirjeldatud korpuses 113 000 radioloogilise (vastuse pikkusest lähtudes mõeldi tõenäoliselt röntgenuuringut) ja 43 000 ultraheliuuringu vastust, meie korpuses on representatiivses mahus esindatud kõik radioloogiauuringute liigid. Lühendeid oli Reisbergi jt kogu vabatekstilises materjalis 14%. Arvestades, et lühendite puhul kattuvad sõna ja sõne mõiste sageli (s.t mingeid muutevorme ei kasutata), on viimane suurusjärk võrreldav meie lühendikandidaatide arvuga (10,5%), kuid see on märkimisväärselt suurem tegelike lühendite esinemisest meie materjalis (3,8%). Meie ja Reisbergi jt uuringu tulemuste teatav erinevus on eelkõige tingitud erinevast uuringumaterjalist ja tõenäoliselt eri käsitlusviisist. Reisbergi jt artiklis ei täpsustatud, mille alusel lühendid muust tekstist eraldati, meie analüüsisime kogu korpuse lühendikasutust. Selleks määrasime endi koostatud kriteeriumide alusel tuvastatud lühendikandidaatidest materjali läbivaatamise käigus tegelikud lühendid ja nende tähendused. Seejärel koostasime lühendite sagedusloendid, et mõista, kas ja kuidas on võimalik lühendikasutust standardida. Analüüsi tulemusel võib väita, et sõnade lühendamise puhul on lisaks üldkeele ja erialakeele lühenditele (vt tabel 3 ja 4) võimalik vaadelda ka eri lühendamisviise. Viimaste puhul arvestatakse pigem keeleliste kui erialaste eripäradega (nt lühendatud sõnale sidekriipsuga keksõna või mitmuse tunnuse või käändelõpu lisamine: *-tud, -d, -sse*). Seda võtet kasutati palju ning need mustrid tuleks samuti täpsemalt kirjeldada, siis saaks paremini aru, miks sõnu pealtnäha suvalistest kohtadest katkestatakse. Katkestuskohti saaks tekstides hiljem automaatselt tuvastada.

Kahetsusväärne on, et mittetähemärke (punkte, side-, mõtte- ja kaldkriipse) ja suurtähti kasutatakse lühendite eristamisel muutelõppudest või teistest sõnedest suvaliselt. Elundite anatoomia (nt segmentide) kirjeldamisel kasutatakse nii Rooma kui ka araabia numbritega lühendamisviise ning nende eraldamiseks kasutatakse nii kald- kui ka sidekriipse (nt *TH1-TH2, Th1/Th2, Th1-2, Th1/Th2, Th1/2, Th I, T12, S I S1*). Probleemiks on ka tühikute kasutamine, sest automaatse analüüsi seisukohast tekitab see tekstis eraldi olevaid ebaselge tähendusega numbri- või täherühmi (nt *I-II* või *1-2* või *X* või *V*). Rooma numbrite põhjendamatu kasutamine on automaatanalüüsi jaoks eriti tüsilik, sest Rooma numbrid võivad tähendada ka palju muud peale järgarvu (vt tabel 3).

Eesti meditsiinikeeles on juurdunud ladinakeelsete terminite kasutamise tava, eriti anatoomiliste struktuuride või haiguste kirjeldamisel. Ladina keele põhiseid lühendivariante kasutati ka uuritud materjalis ülekaalukalt rohkem kui eestikeelseid lühendivariante. Kuid nii nagu teiste meditsiinikeelte puhul, on ka eesti meditsiinikeeles järjest enam märgata inglise keele mõju. Huvitav on märkida, et osa lühenditest, mis on algselt inglise keeles ning sellistena ka üle võetud (nt *FFE, DWI, SPIR, PACS, FNB, CNB, COPD, IUD*), on mõni siiski ka juba eestistatud (nt *PAKS, JNB, PNB*, aga ka *KOK, ESV*). Väga oluline nihe toimub anatoomialühenditega, kus enamik põhistruktuuride lühendeid on siiani olnud ladina algupära (nt *a., m., o., n.*). Samas on radioloogiauuringutega seotud pildi- ja virtuaalatlastes kasutusel ingliskeelsed suurtähtlühendid ja nii on näiteks *a. carotis interna* (mitmetes variantides kasutussagedus kokku 171) ja ACI (kasutatud 333 korda) kõrvale tekkinud üsna elujõuline lühend *internal carotid artery* ehk ICA (kasutatud 52 korda). Kui põhistruktuuri lühend oli ladina algupära, oli põhistruktuuri nimetusest ja täiendist koosnev lühend samuti ladina algupära, kuigi täiendi kirjapilt võis varieeruda: näiteks *M. hypertonicus, M. hypert., M. hyper., M. hyp.* ja ka *M. hüpertonicus, M. hüper* ja *M. hüp.* (kõik täiendid esinesid ka suure algustähe variandis, nt *Hypertonicus* ja nii punktidega kui ka ilma) või ka näiteks *a. mesenteerika, a. mesenterika, a. mesenterica, a. mesent.* ja *a. mes.* või *l/n., l/s., l-s, l-sõlm*. Ühetäheliste lühendite puhul trigrammiga piiratud kontekstis oli teadmata tähendusega 906 lühendit. Oluliselt pärssis lühendite tuvastamist asjaolu, et osa üksiktähti ei osutunud ei lühendiks ega tähiseks, isegi mitte näpuveaks, vaid olid mingis muus funktsioonis, näiteks muutelõpud või liitsõna lühendatud osa

lühendid. Kui soovime näiteks uurida, kui mitmel teatud kaebustega pöördunud kuni 30aastasest patsiendist olid teatud ajal lümfisõlmed mingil põhjusel teatud piirkonnas suurenenud või valulikud, tuleb arvuti jaoks iga päringuga seotud mõiste ning selle kõik võimalikud tähistamisviisid (s.o sünonüümid ja lühendid) eraldi kirjeldada või kodeerida ning tekib küsimus, kas väga suur ja reglementeerimata lühendamisviiside rohkus ja sünonüümia on terviseandmeid kirjeldavates vabatekstides mõistlik ja otstarbekas. Vähemalt lühendikasutust oleks võimalik väikese vaevaga standardida ning autentsel keelekasutusel põhinevad lühendite sagedusloendid aitaksid teha vajalikke valikuid (vt tabel 4).

Huvitav on, et lühendamisel ei lähtuta kaugeltki alati kõige lühemast veel arusaadavast variandist, vaid kasutussageduse järgi eelistatakse sageli mingil põhjusel pikemat varianti. Näiteks kasutati lüliarteri (*arteria vertebralis*) lühendit (siinkohal punkti ja suurtähe kasutuse kõikvõimalikes variantides) *a. vertebralis* 2297 korda, lühendit *a. vert.* kõigest 100 korda, lisaks kasutati vähemal määral veel ka muid lühendivariante. Üks pikema variandi eelistamise põhjus võib peituda tüüptekstide kasutamises vastuse põhjana. Tüüptekstides võivad eeldatavalt kirjeldamist vajavad struktuurid olla juba ette kirjutatud, lisatakse vaid neid iseloomustavad andmed: verevoolu kiirused, veresoonte läbimõõdud, lubinaastude olemasolu jms.

Oluline aspekt on sageli ka laiem kontekst, kuhu lühend valdkonnasiseselt paigutub. Nii võib nukleaarmeditsiini uuringutega seotult lühend *fr* tähistada frekventsi, ülajäseme röntgeni korral pigem fraktuuri. Ehk lühendi tähenduse automaatsel määramisel ei tuleks kasutada mitte ainult lühendi vahetut konteksti (nt otsustada lühendi tähenduse üle tema vahetute naabersõnede põhjal), vaid arvestama peab laiemat konteksti. Sageli võib uuringu modaliteet või piirkond lühendi osa võimalikke tähendusi lihtsalt välistada.

Lõpuks tuleks märkida veel üht aspekti. Nimelt on osa lühendeid kollegiaalses žargoonis sisuliselt muutunud sõnaks, näiteks operatsiooni asemel mugandus *opp* (*opp – opi – oppi – opile – opini – opiks*). Kui sõna rakendus kõrvale on keelde tekkinud lühem mugandus äpp ning 2013. aasta õigekeelsussõnaraamat lubab argikeeles kasutada ka sõna *opp* (18), siis võib-olla peaks uurima, missugused lühendid võiksidki olla kasutuses sõnana, see vähendaks oluliselt erisugust käändelõppude lisamist (ülakoma, rõhumärgi, side- või mõttekriipsu abil ning lühendile vahetult liidetud).

JÄRELDUSED

Uurimistöö võimaldas radioloogiauuringu vastuste lühendikasutuse uurimise näitel veenduda, et korpuslingvistilise analüüsiga saab meditsiinikeelt ratsionaalselt uurida: olemasolevatest vabatekstilistest andmetest saab tööjõu suhtes ressursisäästlikult määrata sõnavara ja sõnede kasutussagedust ning otsustada andmeesituskvaliteedi üle. Terminikasutuse, kirjakeele ja kirjutatu struktuuri ühtlustamine on parema andme(esitus)kvaliteedi saavutamise eelduseks. See võimaldab üheselt mõista teksti kliinilist infot ja eri eesmärkidel paremini tõlgendada ning analüüsida terviseandmeid. Korpuslingvistilise analüüsi eesmärgipärane kasutamine töise meditsiinikeele uurimiseks eeldab asutustevahelist ja erialaülest koostööd, sest meditsiinilingvistika on multidistsiplinaarne valdkond, mis eeldab meditsiinitöötajate ja lingvistide ühist tööd.

TÄNUAVALDUS

Uuringut on rahastatud Eesti keeletehnoloogia 2011–2017 programmi raames (projekt EKT6).

VÕIMALIKU HUVIKONFLIKTI DEKLARATSIOON

Autoritel puudub seoses uurimusega mis tahes huvikonflikt.

SUMMARY

Corpus-based analysis of abbreviations and abbreviating in Estonian radiology reports

Eola Valdre[1], Peeter Ross[2], Katrin Tsepelina[1], Kaarel Veskis[1], Tarmo Vaino[1], Heiki-Jaan Kaalep[3]

[1] Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia
[2] Technomedicum, Tallinn University of Technology, Tallinn, Estonia
[3] Institute of Computer Science, University of Tartu, Tartu, Estonia

Corresponding author:
Peeter Ross
peeter.ross@ttu.ee

Keywords:
health data, radiology report, free text, abbreviation, corpus linguistics

Background and aim. Electronic patient records set new requirements to medical language. The Estonian professional medical language is relatively young: it emerged in the 19th century and is still influenced by its predecessors and contemporaries Latin, German and Russian, as well as by the modern *lingua franca* of science, i.e. English. Medical texts in Estonian are written in either of the two sublanguages:

academic language or health data recording language. The latter reflects everyday work and is affected by fatigue, stress and tight schedules. It does not fully conform to any particular standard and is neither revised nor edited. Data standardisation is an absolute prerequisite for any E-health application meant to enable data mining, analysis and/or customisation. However, sustaining adequate data quality poses a real challenge, because it is impaired by inconsistency and ambiguity of free text. Hence, it is important to improve the language of free text to assure correctness and consistency of the content, and to reduce ambiguity and vagueness. In Estonia, the professional medical language has not been systematically studied, partly because there exist no suitable language resources. The aim of this study was to compile a relevant text corpus and to assess the overall suitability of the linguistic approach to study the language of radiology reports. We confined the study to analysis of abbreviations.

Methods. We compiled a corpus of depersonalised radiology reports. The corpus was converted to XML, annotated and validated against the TEI P5 encoding scheme. We established a specific set of rules and, by using UNIX commands based scripts, applied them to retrieve abbreviations from the corpus. Because of inherent ambiguity, all one-letter abbreviations were analysed as trigrams consisting of the abbreviation and the neighbouring tokens. The frequencies and meaning of the abbreviations were reviewed separately by two doctors.

Results. We compiled a corpus consisting of 207,534 depersonalised radiology reports with more than 11.8 million tokens. We retrieved 10,606 abbreviations (446,158 tokens, 3.8% of the corpus). Abbreviating appeared to be rather arbitrary and inconsistent. Mistyping was not an issue compared to ambiguity and/or inconsistent use of punctuation, space, and numbering; unconventional merging or breaking sentence structures and word boundaries, in particular, when adding case endings. The use of abbreviations of Estonian, Latin and English origin was often overlapping. This synonymy revealed an emerging shift from Latin- to English-based abbreviations.

Conclusions. The study of abbreviations in Estonian radiology reports showed an urgent need for standardisation of the medical language and confirmed an enormous potential of the linguistic approach for analysing the free text of health data. It is a feasible and resource-effective tool for analysing huge data sets and provides much needed insight into the existing problems of the medical language. Medical linguistics is an interdisciplinary field, meaning that inputs by linguists and medical professionals are equally important.

KIRJANDUS/REFERENCES

1. Kvist M, Velupillai S. Professional language in Swedish radiology reports – characterization of patient-adapted text simplification. Scandinavian Conference on Health Informatics, Copenhagen, 2013;55–9. http://www.ep.liu.se/ecp/091/012/ecp13091012.pdf, 16.12.2013.
2. Taavitsainen I, Pahta P. Medical and scientific writing in late medieval English. New York: Cambridge University Press; 2004.
3. Põlluste J. Eesti meditsiiniterminoloogia – kuidas edasi? Eesti Arst 2011;90:61–3.
4. Reiner BI. The challenges, opportunities, and imperative of structured reporting in medical imaging. J Digit Imaging 2009;22:562–8.
5. Reiner BI. Uncovering and improving upon the inherent deficiencies of radiology reporting through data mining. J Digit Imaging 2010;23:109–18.
6. Valdre E. Ükskeelse spetsialiseeritud korpuse rakendusvõimalustest kirjalikus tõlkes radioloogiliste kirjelduste näitel. Magistritöö. TÜ 2010. http://dspace.utlib.ee/dspace/bitstream/handle/10062/17232/Valdre_Eola.pdf, 16.12.2013.
7. Taira RK, Soderland SG, Jakubovits RM. Automatic structuring of radiology free-text reports. Radiology 2001;21:237–45.
8. Bozkurt S, Kahn CE. An open-standards grammar for outline-style radiology templates. J Digit Imaging 2012;25:359–64.
9. Sistrom CL, Honeyman-Buck J. Free text versus structured format: information transfer efficiency of radiology reports. Am J Roentgenol 2005;185:804–12.
10. Krupinski EA, Hall ET, Jaw S, Reiner B, Siegel E. Influence of radiology report format on reading time and comprehension. J Digit Imaging 2012;25:63–9.
11. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. Radiology 2011;260:174–81.
12. Reiner BI. Customization of medical report data. J Digit Imaging 2010;23:363–73.
13. Ross P, Sepper R, Pohjonen H. Cross-border teleradiology – experience from two international teleradiology projects. Eur J Radiology 2010;73:20–5.
14. Innovative Healthcare Services Exchange for Baltic Sea Region Improves Efficiency and Flexibility. Cisco Systems, Inc., IBSG Copyright © 2007 Cisco Systems, Inc. All rights reserved. http://www.cisco.com/web/DK/assets/docs/Baltic_CS_0514a.pdf, 04.08.2014.
15. Ajakirja Eesti Arst korpus: http://www.cl.ut.ee/korpused/segakorpus/eestiarst/, 16.12.2013.
16. Marciniak M, Mykowiecka A. Towards morphologically annotated corpus of hospital discharge reports in Polish. Proceedings of the 2011 Workshop on biomedical natural language processing, ACL-HLT 2011;92–100.
17. Reisberg S, Sirel R, Kalda R, Merzin M, Pruulmann J, Vilo J. Elektrooniliste terviselugude analüüsimise võimalused Tartu perearstide infosüsteemi näitel. Eesti Arst 2013;92:452–9.
18. Eesti õigekeelsussõnaraamat. ÕS 2013. Koost. Erelt, T jt. Toim. Raadik, M. Tallinn: Eesti Keele Instituut; 2013. http://www.keeleveeb.ee/, 29.07.2014.