doi:10.5128/ERYa9.17

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 9, 265–278

MEETODEID TEKSTIDE LEKSIKAALSETE JA GRAMMATILISTE ERINEVUSTE TUVASTAMISEKS MEDITSIINILISTE TARBETEKSTIDE NÄITEL

Raul Sirel

Ülevaade. Artiklis käsitletakse seni uurimata eestikeelset ressurssi: ravimipakendites sisalduvaid infolehti ja arstidele suunatud ravimeid tutvustavaid kokkuvõtteid. Nimetatud ainestiku analüüsimiseks kasutatakse mõnda läbipaistvat statistilist meetodit, mis võimaldavad kerge vaevaga tuvastada tekstide tüübist ja žanrist tulenevaid leksikaalseid ja grammatilisi erinevusi. Taolise analüüsi eesmärgiks on ühest küljest katsetada nimetatud meetodite efektiivsust tekste eristavate karakteristikute leidmisel, kuid ka koguda andmestikust lähtuvaid taustateadmisi keeletehnoloogiliste rakenduste efektiivsemaks loomiseks.*

Võtmesõnad: korpuslingvistika, tekstilingvistika, tekstikorpused, žanrianalüüs, keeletehnoloogia, eesti keel

1. Sissejuhatus

Infotehnoloogia võidukäiguga seoses on kasvanud kõiksugu dokumentide ja tekstide hulk ning kättesaadavus, mis omakorda on loonud nõudluse tehnoloogiate ja algoritmide järele, mis võimaldaksid seda ressurssi efektiivsemalt töödelda ja hallata. Nii näiteks on näha jätkuvalt kasvavat huvi erinevatest tekstidest informatsiooni kogumise (ingl *information extraction*), nende automaatse liigitamise (*document classification*), automaatse kokkuvõtmise (*automatic summarisation*) jms vastu.

Käsikäes saadaval olevate tekstide hulgaga on suurenenud ka nende varieeruvus, kuna tekste luuakse väga erinevatel eesmärkidel ning erinevatele sihtgruppidele. Keeletehnoloogias on paraku üsna tavapärane olukord, kus ühe keelekogu alusel loodud mudel või meetod toimib selle kogu raames suurepäraselt, kuid teisele korpusele rakendades annab meetod märksa tagasihoidlikemaid tulemusi (Kilgarriff 2001). Põhjus ongi enamasti erinevate tekstikogude keelekasutuse suures varieeruvuses, mistõttu on tarvis uurida funktsioonilt ja keelelistelt tunnustelt

* Artikli valmimist on toetanud Euroopa Regionaalarengu Fond Eesti Arvutiteaduse Tippkeskuse ja Tarkvara Tehnoloogia Arenduskeskuse kaudu, Eesti Teadusfond (grandiprojekt 9124) ning Haridus- ja Teadusministeerium (projektid SF0180078s08 ja EKT11005). Uurimistööks vajalikud andmed on taganud Ravimiamet. Suured tänud ka anonüümsetele retsensentidele sisulise tagasiside eest.

täiesti erinevaid tekstitüüpe ja -žanre, et luua universaalsemaid või vastupidi tekstispetsiifilisemaid mudeleid ja meetodeid.

Kuigi igapäevaelus on tekstižanrid tihtipeale määratletavad juba intuitiivselt või selle vormi analüüsides, siis Fairclough'i (2003: 66) järgi võib nõnda jõuda väära järelduseni, kuna teksti vorm ja sisu ei pea tingimata kokku langema. Seetõttu on funktsionaalses žanrianalüüsis teksti klassifitseerimisel tarvilik lähtuda eelkõige teksti sisust (Reinsalu 2011). Sisu funktsionaalne analüüs on paraku aeganõudev tegevus, mis nõuab lisaks ajale valdkondlikke teadmisi. Sel põhjusel käsitleb käesolev artikkel statistilisi meetodeid, mis võimaldavad üsna väikese vaevaga leida leksikaalsed ja grammatilised tunnused, mille suhtelised kasutussagedused on tekstiklassiti märkimisväärselt erinevad. Taoliste tunnuste analüüsimine võimaldab ühest küljest efektiivsemalt luua tekstispetsiifilisemaid mudeleid nende tekstide töötlemiseks, kuid samuti on need kasutatavad erinevate tekstitüüpide ja -žanride keelekasutuse profileerimiseks.

Tekstides sisalduvat keelekasutust on varasemalt uuritud näiteks nende formaalsuse ja kontekstuaalsuse alusel, milleks on analüüsitud tekstides esinevaid lausestruktuure (Pajupuu, Kerge 2010). Leksikaalsete ilmingute poolelt on eesti aja- ja ilukirjanduskeele erinevusi kirjeldanud Tiit Hennoste ja Kadri Muischnek (2000), kes on välja toonud märkimisväärse vahe (eriti isikuliste) pronoomenite kasutamises (ilukirjanduses selgelt enam; põhjuseks peetakse ilukirjanduse suuremat dialoogilisust).

Artiklis käsitletakse kahte seni uurimata eestikeelset ressurssi, milleks on ravimipakendites sisalduvad infolehed ning arstidele suunatud ravimeid tutvustavad kokkuvõtted. Nimetatud andmetest moodustati kaks tekstikorpust, analüüsiti neid leksikaalsest ja grammatilisest aspektist ning võrreldi tulemusi Tasakaalus korpusega[1], mis sisaldab võrdsel määral ilukirjanduse, ajakirjanduse ja teaduse keelt. Maailmas on ravimipakendi infolehti keeletehnoloogiliste ülesannete lahendamisel kohati rakendatud: nii näiteks on infolehtedest tekstikaeve meetoditega ekstraheeritud patsienti hoiatavaid fraase (Nabeta jt 2012) ning samuti on infolehtedest üritatud leida (Kuhn jt 2010) ning ennustada (Atias, Sharan 2011) ravimite kõrvaltoimeid. Eestikeelseid ravimipakendi infolehti ja ravimi omaduste kokkuvõtteid seni uuritud ei ole.

Artikli põhiosa on jagatud neljaks peatükiks, millest esimeses antakse ülevaade tekstide liigitamisega seotud problemaatikast, teises tutvustatakse uuritavat ainestikku ning kolmandas käsitletakse korpuste analüüsimiseks kasutatud meetodeid. Neljas peatükk on analüüsitulemuste tutvustamiseks.

2. Tekstide liigitamine teoorias ja praktikas

Tekstide liigitamine on mõneti keerulisem ülesanne, kui võiks tausta põhjalikumalt tundmata arvata. Seda eeskätt seetõttu, et puudub üheselt aktsepteeritav tüpoloogia tekstide klassifitseerimiseks ning liigitamine toimub väga erinevatel alustel: nii näiteks on võimalik tekste klassifitseerida kommunikatsioonikanali (nt suuline või kirjalik), auditooriumi ehk sihtgrupi, kasutusvaldkonna või ka suhtluseesmärgi vms järgi.

[1] Vt http://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=et (28.12.2012).

Üheks tuntumaks ning aktsepteeritumaks klassifikatsiooniks peetakse Werlich'i tüpoloogiat, mis jagab tekstid viide tüüpi: deskriptiivsed, narratiivsed, ekspositsioonilised, argumenteerivad ning instruktiivsed (Chilton, Schäffner 2002: 19). Taoline geneeriline jaotus põhineb Reet Kasiku (2007: 29) järgi inimese vajadusel kasutada keelt erinevatel eesmärkidel ning sellest lähtuvalt on tekstitüübid ka universaalsed – kultuurist ja harjumustest sõltumatud. Mõneti sarnaselt Werlich'ile on Kasik ise jaganud tekstid nende tüübi alusel kolme klassi: deskriptiivsed, narratiivsed ja argumenteerivad.

Lisaks tekstitüüpidele kasutatakse tekstide klassifitseerimisel tekstiliigi ehk žanri mõistet. Žanr on kultuurisidus keelekasutusviis, mis on aja jooksul välja kujunenud ning erinevalt tekstitüüpidest ei ole neis mingit universaalset alust, vaid need on pigem seotud harjumuste, kultuuri ning tavadega. Keelekasutusvaldkonna järgi eristatakse argikeelt, ilukirjanduskeelt ja tarbekeelt. Viimase alaliikidena räägitakse nt ajakirjanduskeelest, ametikeelest ja teaduskeelest, millest igaühel on veel hulgaliselt alaliike (Kasik 2007: 35).

Mõneti keeruliseks teeb tekstitüübi ja -žanri eristamise aga asjaolu, et inglise-keelses kirjanduses kasutatakse termineid *genre* ja *type* tihtipeale samatähenduslikuna, viidates enamasti tekstitüübile (Vardi 2000, Askehave, Swales 2001). Nii näiteks on James Robert Martin (1985) tekstitüübi (kasutades terminit *genre*) defineerinud kui keele sihipärase kasutamise mingi konkreetse eesmärgi saavutamiseks. Antud artiklis käsitletakse termineid *tekstitüüp* ja *tekstižanr* siiski sarnaselt Kasikule erinevate mõistetena.

Artiklis käsitletavad tekstid (ravimi infolehed ja omaduste kokkuvõtted) kuuluvad oma keelekasutusvaldkonna järgi tarbetekstide hulka: need on mõeldud suurele auditooriumile ning nende eesmärk on anda lugejale edasi mingi asja sisu ning teha seda võimalikult selgelt ja üheselt mõistetavalt (Kasik 2007: 43–44). Veel täpsemalt võib nimetatud tekste liigitada meditsiinivaldkonda kuuluvaks. Kuigi tekstid on käsitletavad tarbetekstidena ning kuuluvad samasse keelekasutusvaldkonda, siis tegelikkuses esineb neis ka märkimisväärseid erinevusi.

Tekstide erinevused on ühest küljest põhjendatavad erinevusega auditooriumis: neil on üsnagi erinevad sihtgrupid, kuna ravimipakendi infolehed on suunatud eelkõige ravimi tarvitajale, ravimi omaduste kokkuvõtted seevastu aga arstile. Teisest küljest ei saa kindlasti jätta tähelepanuta mõningaid erinevusi neid näiteks Werlich'i tüpoloogiasse paigutades. Nimelt kuuluvad ravimi omaduste kokkuvõtted kahtlemata deskriptiivsete tekstide hulka (neis kirjeldatakse ravimi (kõrval)toimeid, näidustusi, manustamisrežiime jne). Sama võiks esmapilgul arvata ka ravimipakendi infolehtede kohta (need sisaldavad samuti kirjeldusi kõrvalnähtude, näidustuste jms kohta), ent neis on ka teatavaid instruktiivseid jooni (juhendatakse ravimit õigesti tarbima ning probleemide ilmnemisel korrektselt käituma).

Seetõttu võiks eeldada, et ravimipakendi infolehed sisaldavad enam käskivat kõneviisi ning personaalpronoomeneid (tekstides pöördutakse ravimi kasutaja poole, kellel puudub igasugune valdkondlik kogemus). Võrreldes Tasakaalus korpuses sisalduvate tekstidega võib prognoosida mõnevõrra väiksemal hulgal tingiva kõneviisi kasutamist ravimitekstides. Seda peamiselt seetõttu, et tingivat kõneviisi kasutatakse väljendamaks viisakat käsku, kõneleja (täitumatut) soovi või hinnangut

tegevusele (EKK)[2]. Leksikaalsest küljest võiks aga eeldada, et ravimitekstide ja Tasakaalus korpuse leksikaalne ühisosa (vähemasti substantiivide lõikes) on üsna tagasihoidlik, kuna ravimitekstid sisaldavad suurel hulgal valdkonnaspetsiifilist terminoloogiat. Sarnaselt võib ka oletada, et ravimipakendi infolehtede ja ravimi omaduste kokkuvõtete leksikaalne ühisosa on märksa suurem kui ravimitekstidel ja Tasakaalus korpusel.

Käesoleval artiklil on kaks peamist eesmärki: katsetada eestikeelsete meditsiiniliste tarbetekstide peal statistilisi meetodeid, mis võimaldaksid automaatselt tuvastada tunnuseid, mille poolest kaks tekstitüüpi või -žanri erinevad, ning nimetatud meetodeid kasutades eelnevas lõigus esitatud hüpoteese kontrollida.

3. Andmestik

Uuritavaks ainestikuks on koostatud kaks tekstikorpust, millest esimene sisaldab 3977 Eestis registreeritud müügiloaga ravimipakendi infolehte ning teine samade ravimite omaduste kokkuvõtteid. Olemuslikult on esimeses informatsioon ravimi tarvitajale ning teises informatsioon arstile.

Käsitletav ainestik on saadud Ravimiametist ning on kättesaadav PDF-failidena Eesti ravimiregistri kodulehelt[3]. Korpuste koostamisel on järgitud põhimõtet, et korpustesse kuuluksid ainult selliste ravimite tekstid, millel on kättesaadavad nii pakendi infolehed kui ravimi omaduste kokkuvõtted. PDF-failidest ekstraheeritud andmed on analüüsitavates korpustes muutmata kujul, mis tähendab, et alles on jäetud näiteks peatükkide pealkirjad, tabelites sisaldunud arvandmed jms. Samuti ei ole korpustest eemaldatud suurel hulgal esinevaid korduvaid tekstiüksuseid, kuna selline korduvus on antud uurimistöö kontekstis tõlgendatav tekste iseloomustava parameetrina.

Selleks, et leida kvantitatiivsete meetoditega kaht tekstitüüpi või -žanri eristavad parameetrid, on tarvis võrrelda neid tekste sisaldavaid korpuseid. Võttes aga näiteks kaks meelevaldset tekstikorpust, võib juba enne nende analüüsimist täie kindlusega ütelda, et need on mingil määral erinevad. Nii võib näiteks ajalookirjandus tunduda majanduskirjandusest märksa narratiivsem, kuid võrreldes ilukirjandusega on mõlemad siiski äärmiselt mittenarratiivsed. Selleks, et näidata, kas erinevused on relevantsed või mitte, kasutatakse üldjuhul multifaktoriaalset analüüsi (Biber jt 1998: 169).

Artiklis on loodud tekstikorpuseid perspektiivi saavutamiseks võrreldud ka Tasakaalus korpusega, kuna see on ammendavalt kirjeldatud ning seda võib põhjendatult pidada kirjaliku eesti keele neutraalseks baastasemeks. Tasakaalus korpus sisaldab kirjaliku keelekasutuse kolme tähtsamat tekstiklassi: ilukirjanduse, ajakirjanduse ja teaduse keelt. Kasutatud tekstikorpuste mahud on näidatud tabelis 1.

Tabel 1. Analüüsitavate tekstikorpuste maht

	Ravimipakendi infolehed	Ravimi omaduste kokkuvõtted	Tasakaalus korpus
Sõnade arv	6 562 114	11 515 652	14 823 078

[2] Vt http://www.eki.ee/books/ekk09/index.php?id=198 (28.12.2012).
[3] Vt http://193.40.10.165/register/register.php?keel=est&inim_vet=inim (28.12.2012).

4. Analüüsimeetodid

4.1. Morfoloogiline märgendamine

Võrreldavad korpused on morfoloogiliselt märgendatud, kasutades eesti keele morfoloogilist analüsaatorit ja ühestajat. Morfoloogilise ühestamise käigus määratakse tekstisõna lokaalset konteksti (selle vahetus ümbruses asuvaid sõnu) analüüsides selle lemma ehk algvorm, sõnaliik ja sõnavormis kodeeritud grammatilised kategooriad (Kaalep 1997, Kaalep , Vaino 1998).

Morfoloogilise ühestaja väljundist on enne tekstikorpuste võrdlemist eemaldatud lausemärgi (nt –, /, ? jne) märgendi saanud sõned. Uurimistöös on morfoloogilist ühestajat kasutatud järgmiste grammatiliste tunnuste määramiseks:

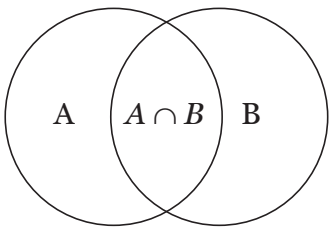
- sõna algvorm ehk lemma,
- sõnaliik,
- substantiivide kääne,
- verbide kõneviis,
- verbide aeg (olevik ja lihtminevik).

Suurest hulgast verbikategooriatest on valitud kõneviis ja aeg eeskätt sel põhjusel, et need avalduvad ühes sõnavormis (erinevalt näiteks verbi liitaegadest, mis avalduvad kahe sõnavormi ühendina), mistõttu ei ole tarvis rakendada automaatset süntaksianalüüsi.

4.2. Venni diagrammid leksikoni visualiseerimiseks

Venni diagrammiks nimetatakse diagrammi, millel hulgad esitatakse mingile pinnale joonistatud piirkondadena (Liikane, Kesa 2006). Tegemist on meetodiga hulkadevaheliste seoste tasapinnaliseks illustreerimiseks.

Käesolevas uurimistöös on Venni diagramme kasutatud võrreldavate korpuste leksikaalsete erinevuste illustreerimiseks. Täpsemalt on üritatud näidata, kui suur on erinevate korpuste leksikaalne ühisosa (kui leksikaalselt sarnased võrreldavad korpused üksteisele on) erinevate sõnaliikide lõikes. Korpuste leksikaalsete erinevuste visualiseerimisel ei ole kasutatud kõiki korpustes sisalduvaid sõnu, vaid ühes katses kümmet tuhandet kõige sagedamat lemmat ning teises katses kolme sõnaliigi (adjektiivid, substantiivid ja verbid) tuhandet kõige sagedamat lemmat.



Joonis 1. Näide Venni diagrammist, kus A ∩ B on hulkade A ja B ühisosa

Diagrammide loomisel on kasutatud veebipõhist tarkvara Area-Proportional Venn Diagram Plotter and Editor[4].

[4] Vt http://bioinforx.com/free/bxarrays/venndiagram.php (28.12.2012).

4.3. Sagedusprofileerimine

Sõnaliikide ja käänete kasutussagedusi on analüüsitud sagedusprofileerimist rakendades. Nimetatud meetod annab võimaluse leida kahe tekstikorpuse suurimad erinevused, analüüsides sõnavormide või märgendite sagedusi (Rayson, Garside 2000). Nimetatud meetodit on käesolevas uurimistöös kasutatud morfoloogilise ühestaja määratud sõnaliikide ja käänete märgendite esinemissageduste analüüsimiseks.

Sagedusprofiili koostamiseks on tarvis arvutada eeldatavad väärtused E_1 ja E_2 järgmiste valemite järgi, kus *a* on märgendi absoluutne sagedus esimeses korpuses, *b* on sama märgendi absoluutne sagedus teises korpuses, *c* on märgendite koguarv esimeses korpuses ning *d* on märgendite koguarv teises korpuses:

$$E_1 = \frac{c\,(a+b)}{c+d} \qquad E_2 = \frac{d\,(a+b)}{c+d}$$

Iga märgend saab logaritmilise tõepära (inglise keeles *log-likelihood*) väärtuse, asendades arvutatud E_1 ja E_2 väärtused järgmisesse valemisse:

$$\textit{Logaritmiline tõepära} = 2\left((a \times \log \frac{a}{E_1}) + b \times \log \frac{b}{E_2}\right)$$

Logaritmiline tõepära näitab antud märgendi sageduse erinevust kahes korpuses (mida suurem on logaritmiline tõepära, seda suurem on kahe tekstikorpuse vaheline erinevus antud parameetri suhtes). Logaritmilise tõepära alusel saab hinnata, kuivõrd sarnased või erinevad on kaks vaadeldavat teksti. Seda teadmist saab omakorda kasutada prognoosimaks, kui edukalt on võimalik muu tekstikogu jaoks väljatöötatud algoritmi või meetodit analüüsitavale tekstile rakendada (Kilgarriff 2001) ning milliseid muudatusi tuleks algoritmi või meetodi edukamaks kasutamiseks sisse viia.

4.4. Suhteline sagedus

Kõneviiside ja sõnavara kasutussageduse analüüsimiseks on kasutatud suhtelist sagedust, mis annab võimaluse analüüsida mingi omaduse esinemissagedust korpuse suurust arvestades. Suhtelist sagedust on kasutatud näiteks kõneviiside ja muude verbikategooriate ning käänete ja sõnaliikide kasutuse kvantitatiivseks iseloomustamiseks vaadeldavate tekstikorpuste lõikes. Suhtelist sagedust arvutatakse järgmist valemit kasutades:

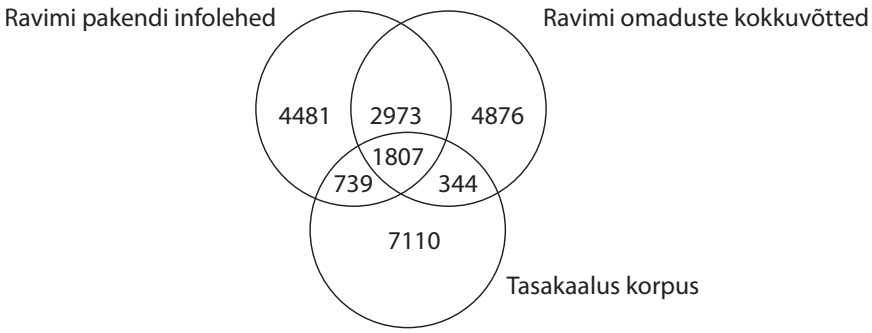
$$\textit{Suhteline sagedus} = \frac{\textit{sagedus}}{\textit{korpuse suurus}} \times 100$$

5. Analüüsitulemused

5.1. Leksikaalsed erinevused

Võrreldud on ravimipakendite infolehtede korpust ravimi omaduste kokkuvõtete korpusega ning mõlemat eelnevat Tasakaalus korpusega. Joonisel 2 on kujutatud Venni diagrammina kõigist kolmest korpusest leitud kümne tuhande kõige sagedama lemma kattuvus kõigi sõnaliikide lõikes.

Kõigi kolme korpuse kümne tuhande sagedama lemma ühisosaks osutus 1807 lemmat ehk ligikaudu 18% iga korpuse kümnest tuhandest kõige sagedamast lemmast, kusjuures ravimipakendi infolehtede ja ravimi omaduste kokkuvõtete korpuste sagedama leksikoni ühisosa oli 47,8%. Ravimipakendi infolehtede ja ravimi omaduste kokkuvõtete leksikoni Tasakaalus korpusega võrreldes selgus, et esimese ühisosa Tasakaalus korpusega oli 25,5% ning teise oma 21,5%, millest võib järeldada, et patsiendile suunatud tekstid on leksikaalselt mõneti tavakeelele sarnasemad kui arstile suunatud tekstid.



Joonis 2. Venni diagramm 10000 kõige sagedama lemmaga kolmes korpuses

Tabelis 2 on esitatud kõigis kolmes korpuses leiduvate unikaalsete lemmade arv kolme sõnaliigi lõikes, kusjuures lemmade arv on antud nii absoluutarvu kui suhtelise sagedusena. Lemmade koguarvu teades saab teha üldistusi, kuivõrd leksikaalselt piiratud analüüsitav korpus on, mida omakorda saab ära kasutada algoritmide ja meetodite kohandamiseks analüüsitavale tekstitüübile. Nii näiteks võib eeldada, et oluliselt piiratuma leksikoniga tekstides kasutatakse piiratumat terminoloogiat, mis tähendab, et sünonüümide hulk on piiratud.

Tabel 2. Korpustes sisalduvate unikaalsete lemmade hulk kolme sõnaliigi lõikes

Sõnaliik	Ravimipakendi infolehed		Ravimi omaduste kokkuvõtted		Tasakaalus korpus	
Adjektiivid	6 460	0,098%	10 339	0,090%	46 969	0,317%
Substantiivid	28 311	0,43%	42 770	0,371%	311 678	2,103%
Verbid	1 734	0,026%	1 920	0,017%	11 178	0,075%

Tabelis 3 on näidatud unikaalsete lemmade hulk, kusjuures välja on jäetud lemmad, mis on vaadeldavates korpustes esinenud kõigest üks kord. On selgelt näha, et Tasakaalus korpuses esines märksa enam lemmasid, mida oli kasutatud vaid

üks kord. See tähendab, et selles korpuses sisalduv keel on leksikaalselt märksa mitmekesisem kui ravimipakendite infolehtede ja ravimi omaduste kokkuvõtete korpustes sisalduvad. Samuti peegeldub tulemustes tõsiasi, et loodud korpused sisaldavad oma formaadi tõttu märkimisväärsel hulgal korduvaid tekstiüksuseid.

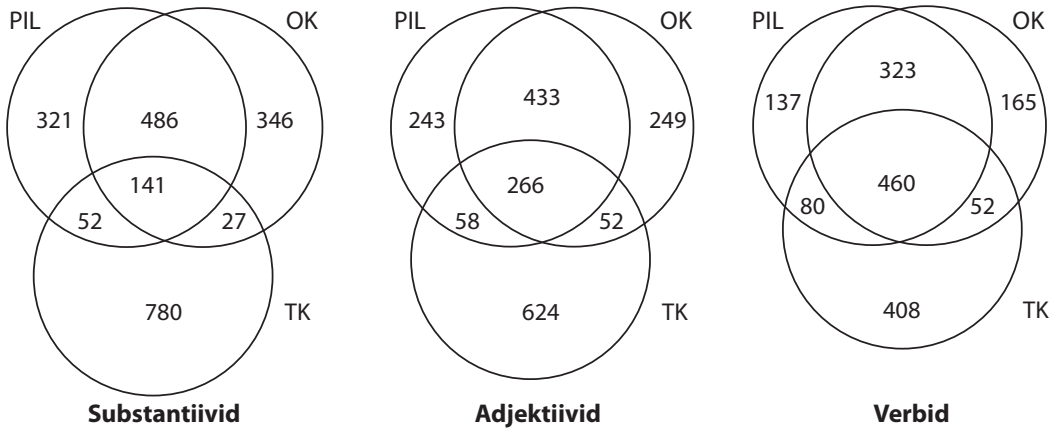
Tabel 3. Korpustes enam kui kord sisaldunud unikaalsete lemmade arv kolme sõnaliigi lõikes

Sõnaliik	Ravimipakendi infolehed		Ravimi omaduste kokkuvõtted		Tasakaalus korpus	
Adjektiivid	4 547	0,069%	7 670	0,067%	20 558	0,139%
Substantiivid	19 251	0,293%	30 407	0,264%	120 017	0,810%
Verbid	1 324	0,020%	1 539	0,013%	5 805	0,039%

Tabelites 2 ja 3 näidatud andmete suurusjärke hinnates on selgesti näha, et Tasakaalus korpuses sisalduv keel on oluliselt rikkalikum kui teistes vaadeldavates korpustes. Selle ning joonise 3 alusel saab järeldada, et ravimipakendi infolehtede ja ravimi omaduste kokkuvõtete korpustes sisalduvad tekstid on kõik küllaltki sarnase leksikoni ning ühtlase terminoloogiaga. Keeletehnoloogia seisukohalt on see teadmine oluline, kuna hõlbustab oluliselt näiteks informatsiooni ekstraheerimist tekstidest.

Loomulikult ei saa tabelite 2 ja 3 puhul teha lõplikke järeldusi lemmade täpse hulga kohta korpustes, kuna morfoloogilise märgendamise käigus on kahtlemata sisse sattunud vigu, mille ulatust on raske prognoosida.

Leksikoni sarnasuse hindamiseks sõnaliikide lõikes on joonisel 3 esitatud 1000 kõige sagedama lemmaga adjektiivide, substantiivide ja verbide Venni diagrammid. Kolme sõnaliigi lõikes osutus kolme korpuse tuhande kõige sagedama lemma ühisosa suurimaks verbidel (46%) ja adjektiividel (26,6%), väikseimaks aga substantiividel (14,1%). Kolme korpuse substantiivide küllaltki väike ühisosa on ilmselt põhjendatav küllaltki ravimispetsiifilise terminoloogiaga (toimeained, kõrvaltoimed, kaebused jms). Sellega on põhjendatav ka pakendi infolehtede ja ravimi omaduste kokkuvõtete üsna suur ühisosa substantiivide (62,7%) ja adjektiivide (69,9%) lõikes. Pakendi infolehtede ja ravimi omaduste kokkuvõtete tuhande kõige sagedama lemma ühisosa on 78,3%.



Joonis 3. Venni diagrammid 1000 kõige sagedama lemmaga ravimipakendi infolehtede korpuses (PIL), ravimi omaduste kokkuvõtete korpuses (OK) ja Tasakaalus korpuses (TK) kolme sõnaliigi lõikes

Kõigi kolme sõnaliigi puhul on sarnaselt kümne tuhande lemma katsega näha, et pakendi infolehed on leksikaalselt tavakeelele mõneti sarnasemad kui ravimi omaduste kokkuvõtted. Tekstikorpuste analüüsitulemused näitavad, et kuigi substantiivide ja adjektiivide ühisosa kolmes korpuses on üsna väike, siis märkimisväärselt suurem ühisosa pakendi infolehtede ja ravimi omaduste kokkuvõtete korpuses ning lemmade väike hulk nimetatud korpustes annavad üsna head eeldused edukaks tekstikaeveks.

5.2. Sõnaliigi- ja käändekasutuse sagedusprofiilid

Käände- ja sõnaliigikasutuse kirjeldamiseks on kasutatud peatükis 2.3 kirjeldatud sagedusprofileerimist. Joonistel 4 ja 5 kujutatud diagrammid on tõlgendatavad järgmiselt: vertikaalteljel on kujutatud logaritmiline tõepära, mis näitab kahe korpuse erinevust horisontaalteljel olevate atribuutide suhtes (mida suurem on logaritmiline tõepära, seda suurem on erinevus atribuudi esinemissageduste vahel kahes korpuses). Kuna diagrammidelt ei ole nähtav, kummas võrreldavas korpuses atribuut rohkem esindatud on, kasutatakse mõnel juhul erinevuse kirjeldamiseks suhtelist sagedust.

Joonisel 5 kujutatud käänete kasutussageduse erinevusest näeb selgesti, et enamiku atribuutide puhul on ravimipakendi infolehed ja Tasakaalus korpus üksteisele küllaltki sarnased (vastavad tulbad on diagrammil madalad). Joonisel 4 tulevad suuremad erinevused esile vaid kardinaalnumeraalide ja lühendite osas, kuna pakendi infolehed sisaldavad tarvitajale informatsiooni ravimite tarvitamisrežiimi kohta, mis on tihtipeale esitatud numbreid ja lühendeid kasutades (näited 1, 2, 3).

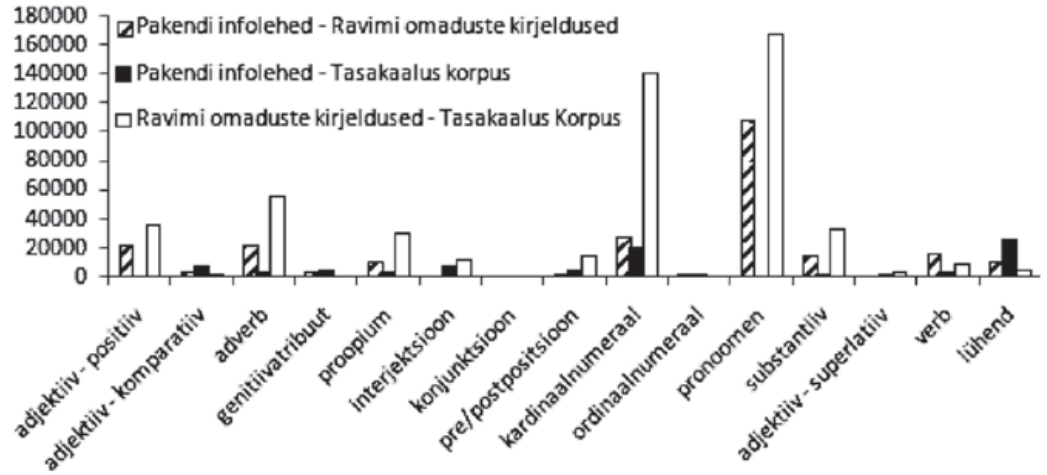
(1) 1 tbl. päevas.
(2) Kate: makrogool 400, makrogool 6000.
(3) Rytmonorm, 300 mg õhukese polümeerikattega tablettidel on märgistus “300”.

Oluline on ka täheldada suurt erinevust pronoomenite kasutuses, kusjuures ravimipakendi infolehtedel on väga suurel hulgal personaalpronoomeneid, kuna tekstides pöördutakse ravimi tarvitaja poole (näited 4 ja 5):

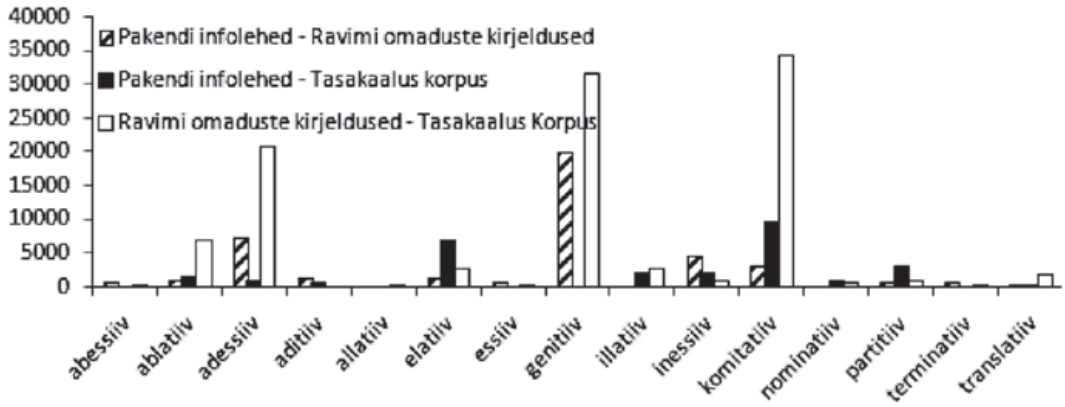
(4) .. kui teie veres on madal kaaliumisisaldus.
(5) Kui arst teile öelnud, et teil on teatud suhkrute talumatus ..

Võrreldes Tasakaalus korpuse ja ravimipakendite infolehtedega on suur pronoomenite kasutamise erinevus ka ravimi omaduste kokkuvõtete korpuses, kus need on tugevasti alaesindatud. Ravimi omaduste kokkuvõtete korpuses on seevastu suuresti üleesindatud kardinaalnumeraalid, kuna need tekstid sisaldavad hulganisti numbrilist informatsiooni ravimite toime ja manustamise kohta. Samuti on nii pakendi infolehtede kui ravimi omaduste kokkuvõtete korpuses tugevasti üleesindatud pärisnimed, kuna mõlemad sisaldavad suurel määral viiteid ravimi nimele, millest parasjagu räägitakse (näited 6 ja 7).

(6) Ärge võtke ravimit Rispolept ..
(7) AVAXIM 160 U kasutatakse ETTEVAATUSEGA.



Joonis 4. Korpuste sagedusprofiilide võrdlus (vertikaalteljel logaritmiline tõepära, horisontaalteljel sõnaliigid)



Joonis 5. Korpuste sagedusprofiilide võrdlus (vertikaalteljel logaritmiline tõepära, horisontaalteljel käänded)

Erinevused on ka näiteks interjektsiooni kasutuses, kuna pakendi infolehed ja ravimi omaduste kokkuvõtted neid ei sisalda. Oluline kvantitatiivne erinevus tuleb eri tekstide lõikes sisse ka adjektiivide kasutuses, kus Tasakaalus korpuses ja pakendi infolehtede korpuses olid suhtelised sagedused vastavalt 8,75% ja 8,73%, kuid ravimi omaduste kokkuvõtetes oli suhteline sagedus oluliselt väiksem: 3,31%.

Sõnaliigikasutuse sagedusanalüüs kinnitab leksikoni sagedusanalüüsist selgunut, et ravimipakendi infolehtedel sisalduv keel on Tasakaalus korpuses esindatud keelele oluliselt sarnasem kui ravimi omaduste kokkuvõtetes sisalduv. Samuti on huvitav tõdeda, et arstile ja patsiendile suunatud ravimitekstid on seniste tulemuste alusel selgesti statistiliselt eristatavad (näiteks on suured erinevused pronoomenite, pärisnimede, adverbide ja adjektiivide hulgas).

Käändekasutust kolmes võrreldavas tekstikorpuses on graafiliselt kujutatud joonisel 5, kust on näha, et suurimad erinevused tulevad esile taas Tasakaalus korpuse ja ravimi omaduste kokkuvõtete võrdluses, kus kõige enam on üleesindatud komitatiivi kasutus (suhteline sagedus 2,81%). Võrdluseks on Tasakaalus korpuses komitatiivi suhteline sagedus kõigest 1,25% ning sama väärtus pakendi infolehtede korpuses 2,09%. Komitatiivi on ravimi omaduste kokkuvõtetes kasutatud tihtipeale struktuurides nagu näidetes (8), (9) ja (10).

(8) Madopar 200 mg/50 mg ristuva poolitusvaoga tablett.
(9) Lisainformatsiooni saamiseks pidage nõu oma apteekriga.
(10) .. leevendab kudede pinget ja sellega kaasnevat valu.

Suuremad erinevused tulevad esile ka genitiivi ja adessiivi kasutuses, kusjuures suurimad lahknevused on täheldatavad taas ravimi omaduste kokkuvõtete ja Tasakaalus korpuse võrdluses ning mõlema parameetri järgi on pakendi infolehed oluliselt sarnasemad Tasakaalus korpusega kui ravimi omaduste kokkuvõtted.

Sarnaselt leksikoni ja sõnaliikide sagedusanalüüsile on ka käänete kasutussagedustest näha, et pakendi infolehed sarnanevad Tasakaalus korpuses sisalduvale keelele enam kui ravimi omaduste kokkuvõtetes sisalduv keel. Huvitavateks parameetriteks osutusid komitatiivi üleesindatus ravimi omaduste kokkuvõtetes ning elatiivi mõnetine alaesindatus pakendi infolehtede korpuses.

5.3. Verbikategooriate kasutus

Mõnede verbikategooriate suhtelist kasutussagedust kolmes tekstikorpuses iseloomustab tabel 4, kus esimeses veerus on kategooria, teises kuni kaks näidet antud kategooriast, kolmandas kategooria suhteline esinemissagedus pakendi infolehtede korpuses, neljandas kategooria esinemine ravimi omaduste kokkuvõtete korpuses ning viiendas kategooria esinemine Tasakaalus korpuses.

Tabel 4. Lihtaegade, käskiva ja tingiva kõneviisi esinemine analüüsitavates korpustes

Verbikategooria	Näide	Pakendi infolehed	Ravimi omaduste kokkuvõtted	Tasakaalus korpus
Olevik	*loeb*, *loeme*	10,99	8,70	9,49
Lihtminevik	*luges*, *lugesime*	1,03	2,31	5,48
Käskiva kõneviisi oleviku aktiivi jaatav kõne	*lugege*	1,73	0,04	0,03
Käskiva kõneviisi oleviku 2. isiku mitmuse aktiivi eitav kõne	*ärge*	0,350	0,004	0,003
Tingiva kõneviisi oleviku aktiivi jaatav kõne	*loeks*, *loeksid*	0,13	0,16	0,34

Tabelis 4 antud tulemustest võib olulise erinevusena välja tuua käskiva kõneviisi mitmuse aktiivi jaatava ning eitava kõne suurt üleesindatust pakendi infolehtede korpuses võrreldes teiste korpustega. See on põhjendatav pakendi infolehtedes sisalduvate pöördumistega ravimi tarvitaja poole, kus soovitatakse või keelatakse mõne toimingu läbiviimist (näited 11, 12 ja 13).

(11) Enne ravimi võtmist lugege hoolikalt infolehte.
(12) Ärge kasutage Madopar'i raseduse ajal.
(13) Kõrvaltoimete ilmnemisel konsulteerige arstiga.

Seevastu esineb ravimipakendi infolehtede ja ravimi omaduste kokkuvõtete korpustes märkimisväärselt vähem tingiva kõneviisi aktiivi jaatavat kõnet (vastavalt

0,13% ja 0,16%) kui Tasakaalus korpuses (0,344%). Aegade kasutusest võib välja tuua lihtmineviku vähese esindatuse pakendi infolehtede ja ravimi omaduste kokkuvõtete korpustes (vastavalt 1,03% ja 2,31%) võrreldes Tasakaalus korpusega (5,48%). Pakendi infolehtede puhul on see põhjendatav käskiva kõneviisi ulatusliku kasutusega.

Verbikategooriate suurima erinevusena saabki ilmselt välja tuua n-ö käskimise (või soovitamise) ja keelamise küllusliku esinemise pakendi infolehtedes, mis paistabki olevat selle tekstiliigi üheks olulisimaks karakteristikuks. Selline teadmine on oluline näiteks tekstidest kokkuvõtete tegemisel, kuna käsud-keelud on infolehtedes oluline informatsioon ja sisaldab juhtnööre ravimi doseerimiseks, tüsistuste vältimiseks jms.

6. Kokkuvõte

Artiklis on käsitletud mõnda läbinähtavat statistilist meetodit leksikaalsete ja grammatiliste erinevuste lihtsamaks tuvastamiseks tekstides sisalduvat keelekasutust analüüsides. Meetodeid on rakendatud ravimipakendi infolehtedest ja ravimi omaduste kokkuvõtetest loodud korpuste statistiliseks analüüsimiseks ja võrdlemiseks.

Loodud korpuseid analüüsides selgus, et ravimipakendi infolehed ja ravimi omaduste kokkuvõtted on võrreldes Tasakaalus korpuses sisalduvate tekstidega küllaltki piiratud leksikoniga, kusjuures pakendi infolehed sarnanevad Tasakaalus korpuses sisalduvale keelele mõneti enam kui ravimi omaduste kokkuvõtted. Samalaadne sarnasus tuli ilmsiks ka käände- ja sõnaliigikasutust analüüsides. Seevastu verbikategooriate kasutuses osutus oluliseks käskiva kõneviisi ning personaalpronoomenite laialdane esindatus ravimipakendi infolehtedes, mis tähendab, et patsientidele suunatakse oluliselt enam soovitusi-manitsusi kui arstidele.

Kõneviisi- ja sõnaliigikasutuse analüüs kinnitab eelnevalt püstitatud hüpoteese seoses teatavate erinevustega ravimipakendi infolehtede ja ravimi omaduste kokkuvõtete keelekasutuses. See tähendab, et kuigi mõlemad tekstid on žanrilt tarbetekstid ja kuuluvad samasse keelekasutusvaldkonda, siis tüübilt on tegemist siiski mõneti erinevate tekstidega, kuna nende auditoorium ja funktsioon on erinevad.

Kahtlemata on käsitletud tekstide näol tegemist andmestikuga, mis väärib ka edaspidist uurimist ning rakendamist. Viimast eriti tekstikaeve vallas, mis võimaldaks näiteks luua meditsiinitöötajatele kasutamiseks andmebaase ravimitest, nende kõrvaltoimetest, manustamisrežiimidest jms. Sellistest andmebaasidest oleks kliiniliste andmete kättesaamine oluliselt lihtsam ja efektiivsem kui infolehti ja omaduste kokkuvõtteid lugedes.

Viidatud kirjandus

Atias, Nir; Sharan, Roded 2011. An algorithmic framework for predicting side-effects of drugs. – Journal of Computational Biology, 18, 207–218.

Askehave, Inger; Swales, John M. 2001. Genre identification and communicative purpose: A problem and a possible solution. – Applied Linguistics, 22 (2), 195–212.

Biber, Douglas; Conrad, Susan; Reppen, Randi 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Chilton, Paul; Schäffner, Christina 2002. Politics as Text and Talk: Analytic Approaches to Political Discourse. Amsterdam: John Benjamins Publishing Co.
EKK = Erelt, Mati; Erelt, Tiiu; Ross, Kristiina 2007. Eesti keele käsiraamat. 3., täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.
Fairclough, Norman 2003. Analysing Discourse: Textual Analysis for Social Research. London, New York: Routledge.
Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. – Renate Pajusalu, Ilona Tragel, Tiit Hennoste, Haldur Õim (Toim.). Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: Tartu Ülikooli Kirjastus, 56–73.
Hennoste, Tiit; Muischnek, Kadri 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Tiit Hennoste (Toim.). Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: Tartu Ülikooli Kirjastus, 183–317.
Kaalep, Heiki-Jaan 1997. An Estonian morphological analyser and the impact of a corpus on its development. – Computers and the Humanities, 31 (2), 115–133. http://www.cl.ut.ee/yllitised/chum1997.pdf (28.12.2012). http://dx.doi.org/10.1023/A:1000668108369
Kaalep, Heiki-Jaan; Vaino, Tarmo 1998. Vale meetodiga õiged tulemused? Eesti keele morfoloogiline ühestamine statistika abil. http://www.cl.ut.ee/yllitised/kk_yhest_1998.pdf (28.12.2012).
Kasik, Reet 2007. Sissejuhatus tekstiõpetusse. Tartu: Tartu Ülikooli Kirjastus.
Kilgarriff, Adam 2001. Comparing Corpora. – International Journal of Corpus Linguistics, 6 (1), 1–37.
Kuhn, Michael; Campillos, Monica; Letunic, Ivica; Jensen, Lars J.; Bork, Peer 2010. A side effect resource to capture phenotypic effects of drugs. – Molecular Systems Biology, 6, artikkel nr 343. http://dx.doi.org/10.1038/msb.2009.98
Liikane, Lauri; Kesa, Marilin 2006. Arvutisõnastik. Elektrooniline versioon. www.keeleveeb.ee (28.12.2012).
Martin, James R. 1985. Process and Text: Two aspects of human semiosis. – J. D. Benson, W. S. Greaves (Eds.). Systemic Perspectives on Discourse, Vol 1: Selected Theoretical Papers from the 9th Int. Systemic Workshop. Norwood, N.J.: Ablex, 248–274.
Nabeta, Keita; Kimura, Masaomi; Ohkura, Michiko; Tsuchiya, Fumito 2012. Analysis on descriptions of precautionary statements in package inserts of medicines. – V. G. Duffy (Ed.). Advances in Human Factors and Ergonomics in Healthcare. USA: CRC Press, 257–266.
Pajupuu, Hille; Kerge, Krista 2010. Characteristics and assessment of educated L1 and L2 dialogue. – Rosario Caballero Rodrigues, Jesus Pinar Sanz (Eds.). Modos y formas de la communicacion humana / Ways and Modes of Human Communication. Cuenca: Universidad Castilla-La Mancha, 339–348.
Rayson, Paul; Garside, Roger 2000. Comparing corpora using frequency profiling. – WCC '00 Proceedings of the workshop on Comparing corpora, 9, 1–6. http://dx.doi.org/10.3115/1117729.1117730
Reinsalu, Riina 2011. Leping tekstiliigina: žanristruktuur. – Eesti Rakenduslingvistika Ühingu aastaraamat, 215–229.
Vardi, Iris 2000. Developing critical writers at the undergraduate level: some insights from critical thinking pedagogy and linguistics. – Cornerstones of higher education. Selected papers from the 1999 HERDSA Annual International Conference, Melbourne, Australia, 12-15 July 1999.

Raul Sirel (Tartu Ülikool) tegeleb peamiselt meditsiinilise tekstikaeve, korpuslingvistika ning dialoogsüsteemidega.
rsirel@ut.ee

METHODS FOR IDENTIFYING LEXICAL AND GRAMMATICAL DIFFERENCES IN MEDICAL APPLIED TEXTS

Raul Sirel
University of Tartu

This paper introduces some transparent statistical methods for identifying characteristics distinctive for patient information and specification leaflets for human medicines. Though the patient information leaflets and specifications for human medicines have been published by the Estonian State Agency of Medicines and been digitally available for some time, they have not been linguistically analysed nor used in the development of language technology applications.

It has been generally accepted that improving the quality of language technology applications often requires genre-specific approaches, for it is common that a model trained on one genre does not produce equally good results when applied to some other genre.

It is the aim of the present paper to identify the linguistic features that differentiate the patient information leaflets and specifications for human medicines from each other and from language represented in the Balanced Corpus of Estonian. In order to achieve that, two text corpora containing the texts from 3977 patient information leaflets and 3977 specifications for human medicines have been created and statistically compared with each other and the Balanced Corpus of Estonian.

The comparison of the corpora revealed that patient information leaflets and specifications for human medicines contain relatively limited lexicon compared with the Balanced Corpus. This knowledge is relevant, because confined lexicons tend to facilitate the tasks of information mining, automatic summarisation, etc. Furthermore, it appeared that the language in patient information leaflets was somewhat similar (compared to the language in specification leaflets) to the language represented in the Balanced Corpus.

Indubitably the collected corpora of patient information leaflets and specifications for human medicines are valuable resources and should be subjects for further research.

Keywords: corpus linguistics, text linguistics, text corpora, genre analysis, language technology