EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 5, 187–196

SÕNAVARA LOOMULIK RIKKUS HARITUD KEELEOSKAJA TEKSTIDES

Hille Pajupuu,
Krista Kerge, Pilvi Alp

Ülevaade. Keeleoskuse üheks näitajaks on sõnavara rikkus ja selle ulatus. Uurimuses võrdlesime kõrgtaseme eesti keele eksami edukalt sooritanud kohalike venelaste sõnavara eesti keelt emakeelena rääkivate kõrgharidusega mittefiloloogide sõnavaraga kolme tüüpi tekstis: suuline dialoog, suuline esinemine ja kirjalik essee. Sõnavara rikkuse mõõtmiseks kasutasime Uber'i indeksit, sõnavara ulatuse määramiseks võrdlesime L1 ja L2 sõnavara sagedussõnastiku sõnavaraga (10 000 sagedasemat sõna).

Sõnavara rikkus erines kahel rühmal oluliselt: L1 oli L2-st rikkam dialoogis ja monoloogis, eriti aga essees. Sõnavara ulatus näitas seevastu sarnast mustrit: elementaarsõnavara e sagedussõnastiku esimese 3000 sõna hulka kuulus nii suulises kui ka kirjalikus keelevormis u 65% L1 sõnavarast ja 70% L2 omast. Harvaesinevaid sõnu oli nii L1 kui ka L2 suulistes ja kirjalikes tekstides u 20%. Võrreldes tulemusi samade L1 ja L2 tekstide formaalsusindeksitega, mis on L1 ja L2 puhul küllaltki sarnased, jõudsime järeldusele, et vaesem sõnavara ei ole vabal rääkimisel ja kirjutamisel takistuseks, juhul kui sõnakasutus on registri- ja žanrikohane.*

Võtmesõnad: sõnavara rikkus, sõnavara ulatus, tekstitüüp, žanr, Uber'i indeks, formaalsusindeks, L1, L2, eesti keel

Sissejuhatus

Uurimisprojekti "Rääkimise loomulikkus ja hindamine" raames üritame kirjeldada seda keelt, mida eesti ühiskond peab aktsepteeritavaks ehk loomulikuks.

Loomulik keel avaldub (situatsiooni ja žanrit arvesse võttes) hästistruktureeritud tekstina, mida iseloomustab spontaansus ja ladusus; oskus kasutada

* See artikkel on valminud tänu Eesti Teadusfondi grandile nr 6742.

keelt paindlikult ja tulemuslikult nii isiklikes kui ka avalikes oludes väljendeid eriti otsimata (vrd Raamdokument 2007: 39). Selline keel võimaldab suhtlejatel keskenduda sõnumi sisule, sõnumist arusaamist ei sega kõne kõla, keeleüksuste valik, vorm, järjestamine ega sidumine jm. Loomulikku keelekasutust ei samasta me standardiseeritud ehk normatiivse keelekasutusega, s.t sellise keelekasutusega, milleni isegi filoloogist L1-kõneleja ei pruugi jõuda ja mis seetõttu ei saa olla aluseks ka L2 keeleoskuse hindamisel (vt ka Ratcliff jt 2002).

Eeldame, et L1 loomulikkuse etaloniks on kõrgharidust nõudval ametikohal töötava mittefiloloogi spontaanne kõne ja enesekontrolli all kirjutamine (ingl *self-controlled writing*). Oleme loomulikku L1-kasutust kirjeldanud mitmest aspektist ja võrrelnud tulemusi sama haridustasemega L2-kõnelejate keelekasutuse tunnustega: aktsent ja selle taju (L. Meister, E. Meister 2007), lauseintonatsioon (Asu, ilmumas), erinevate tekstitüüpide pauseerimine (Pajupuu, Kerge 2006, Kerge jt 2008a, 2008b) ning nende kontekstuaalsus-formaalsus (Kerge jt 2007). Valdavalt tulevad loomuliku keelekasutuse tahkude juures tugevasti esile žanrilised erinevused.

Käesolev uurimus keskendub sõnavarale. Euroopa keeleõppe raamdokument kirjeldab vilunud keelekasutaja (C1) sõnavarakompetentsi kahest küljest: 1) sõnakasutus: "Tuleb ette väiksemaid keelevääratusi, kuid märkimisväärseid sõnakasutusvigu pole"; 2) sõnavara ulatus: "Valdab rikkalikku sõnavara ja oskab sõnavaralünkadest üle saada kaudse väljenduse abil; sõnade otsimist või mõne väljendi vältimist tuleb ette harva. Kasutab ka idioome ja argikeeleväljendeid" (Raamdokument 2007: 130).

Nii sõnakasutus kui ka sõnavara ulatus on lingvistilise kompetentsuse ja kõne voolavuse (ingl *fluent speech*) näitajaid (Little 2005, Read, Chapelle 2001). Hästi kirjeldab neid Eeva Tuokko (2007) doktoritöö, kust leiab ka terve rea asjakohase teooria vahendusi. Keeletestimises hinnatakse rääkimis- ja kirjutamisoskust subjektiivselt, toetudes hindamisskaaladele (Bachman 2001: 76). Oleme veendunud, et hindamisskaalade põhjal on võimalik hinnata, kuivõrd ladus on jutt ja kui sidus tekst (terminikasutuse kohta pikemalt Kerge 2008: 52–55), kuid hinnata sõnavara ulatust ja selle aspektina sõnavara rikkust (s.o jälgida iga teemaringi adekvaatset käsitlust lähtudes just sõnavarast) on – iseäranis suulise keelekasutuse juures – väga keeruline ülesanne, seda enam, et sõnavara rikkus on seejuures mitmel meetodil mõõdetav objektiivne näitaja ning et sõnavara ulatuse objektiivne mõõtmine nõuab sõnade keskmise kasutussageduse tundmist (vt tagapool).

Tekib küsimus, kas C1-taseme küllaltki nõudlike ülesannete muidu sujuva, kommunikatiivse ja asjakohase esituse hindamisel ongi mõtet L2 sõnavara ulatuse aspektidele eraldi tähelepanu pöörata.

Nii oleme seadnud eesmärgi kirjeldada sõnavara rikkust ja ulatust kõne loomulikkuse ühe tunnusena ja kaaluda selle kriteeriumi tähtsust L2 oskuse subjektiivse hindamise puhul. Meie uurimisküsimused on järgmised.

1) Kui rikas on haritud keeleoskaja L1 ja L2 sõnavara?
2) Kuidas iseloomustada haritud keeleoskaja L1 ja L2 sõnavara ulatust sõnade üldise sageduse aspektist eesti keeles?
3) Kas sõnavara rikkus ja ulatus erineb keelevormiti ja tekstitüübiti (suuline dialoog ja monoloog, kirjalik essee kui monoloog)?
4) Kas tulemustest lähtudes peaks keeleeksamitel eraldi keskenduma sõnavara rikkuse ja ulatuse hindamisele?

Sõnavara rikkust käibivate andmebaaside järgi Eestis uuritud ei ole. Ülle Rannuti doktoritöö küll viitab sõnavara rikkuse uurimisele kui oma eesmärgile (vt Rannut 2005: 11), kuid ei teosta seda läbipaistval meetodil ega võrdlemist lubaval viisil: jälgitakse intervjuude sõnavara liigilist koostist ja omandatud sõnade hulka ühe tekstiliigi (intervjuu) valitud lausetes, osutamata täpselt, kas ja kuidas on suhestatud sõnesid ja sõnu[1] (vt samas 19, 27–29).

Taust, materjal, meetod

Kuna materjal on kogutud eesti keele tasemeeksamil (täpsemalt vt allpool), siis valgustame pisut selle tausta. Eestis on 2000. aasta rahvaloenduse järgi veidi üle 1,37 mln elaniku ja see arv langeb (u 1,341 mln inimest jaanuaris 2008). Eesti riigikeelt räägib 2000. aasta andmetel emakeelena u 921 800 inimest; nende osatähtsus elanikkonnas kahaneb samuti (2000–2007 u 0,5% võrra).[2] Tulenevalt riigikeelse suhtlemise kohustusest testib Eesti riik eesti keele oskust eri tasemetel.[3] Kõrgeim keeleoskustase on nõutav peamiselt kõrgharidust vajavatel ametikohtadel (asutuste juhid, kõrgemad riigi- ja omavalitsusametnikud, juristid, arstid, psühholoogid, logopeedid, eesti keele või eestikeelsete ainete õpetajad, kõrgemad ohvitserid jne). Igal aastal on senisel kõrgtaseme eksamil osalenud u 1000 inimest, kelle emakeeleks on valdavalt vene keel[4] (vt REKK 2007).

Uurimuses kasutatava tekstimaterjali kogusime standardiseeritud situatsioonis: mitte-eestlaste oma kõrgtaseme (u B2+/C1) eesti keele eksamil, eestlaste oma selle eksamiga sarnastatud olukorras (sama eksamineerija, sama ajalimiit, samad ülesanded). Ülesandeid oli kolm: kirjalik essee (u 250 sõna, kirjutamise aeg 60 min), kahe testitava vestlus (suuline dialoog, kestus 5–7 min), lühiettekanne (suuline monoloog, kestus 1–2 min). Ette antud teemasid sidus valdkond: keskkond ja ühiskond. Materjal litereeriti arvutifailideks ja seda töödeldi programmiga WordSmith Tools 3.0 (Scott 1996).

Keelejuhid valisime põhimõttel, et tekiksid keelenõuete poolest hästi võrreldavad rühmad: 8 eestlast (4 naist, 4 meest, keskmine vanus 31,5, standardhälve 3,4) ja 8 eesti keelt kõrgtasemel oskavat mitte-eestlast (4 naist, 4 meest, keskmine vanus 32,5 standardhälve 14,2). Kõik nad töötavad kõrgharidust nõudvatel ametikohtadel, kasutavad töösuhtluses eesti keelt, kirjutavad ja räägivad eesti keeles vabalt.

Sõnavara rikkuse mõõtmiseks valisime Uber'i indeksi: $U = (_{\log}N)^2/(_{\log}N - _{\log}V)$, kus *N* on sõnede arv (ingl *tokens*) ja *V* eri sõnade arv (ingl *types*). Valem kujutab endast *V/N* matemaatilist transformatsiooni, mis vähendab mõnevõrra teksti pikkuse mõju sõnavara rikkuse hinnangule (vt Vermeer 2000). Sõnadeks tunnistasime

---

[1] Rannut (2005: 19) on oma materjali määratlenud järgmiselt: "Valitud intervjuudest selekteerisin välja keskmiselt 25–30 lauset intervjuu kohta vastavalt õpilaste jutukusele. Lausete arvu määras sõnavara hulk lauses, mis pidi andma mõlema rühma uuritavaks sõnavarahulgaks 2500 sõna ja sõnavormi (kokku 5000)". Nii on väga raske mõista ka tema analüüsi (vt samas: 27–29).

[2] Vt Eesti statistika andmebaas http://pub.stat.ee/px-web.2001/Dialog/statfile2.asp (30.09.2008).

[3] Kuni 1. juulini 2008 mõõdeti keeleseaduse alusel eesti keele oskuse alg-, kesk- ja kõrgtaset; alates 1.07.2008 kehtivad keeleseaduse muutmise seaduse tagajärjel Euroopa Nõukogu keeleoskustasemed A, B ja C, mida mõõdetakse A2-, B1-, B2- ja C1-taseme eksami vormis. Vt keeleseaduse § 5 lõige 4. Vastu võetud keeleseaduse muutmise seadusega 8. veebruaril 2007. a. Vt Elektrooniline Riigi Teataja. https://www.riigiteataja.ee/ert/act.jsp?id=12795872 (11.06.2008).

[4] Kõrgtaseme eksam on eesti keele oskuse riiklik standardeksam, mida Riiklik Eksami- ja Kvalifikatsioonikeskus on ALTE liikmena korraldanud 1999. aastast. C1-taseme standardeksameid hakati juurutama 1. juulist 2008. Siin kirjeldatav kõrgtase on juriidiliselt võrdsustatud tasemega C1, s.t kõrgtaseme tunnistus annab samasugused õigused kui C1-taseme oma. (Vt "Avalike teenistujate, töötajate ning füüsilisest isikust ettevõtjate eesti keele oskuse ja kasutamise nõuded". Vabariigi Valitsuse 26. juuni 2008. a määrus nr 105, § 16 lõige 3. Elektrooniline Riigi Teataja. https://www.riigiteataja.ee/ert/act.jsp?id=12983186 (30.09.2008).) Kogemuslikult võib oletada, et kõrgtase kaldub mõneti pigem B2-plusstaseme ehk väga hea B2-taseme suunas (pikemalt tasemesuhestuse kohta vt Kerge 2008: 17 jj).

ainult täielikult omandatud sõnad (need, mida ei ole kasutatud sobimatus kontekstis ega sellises vormis, mis raskendaks teksti mõistmist).[5]

Et standardiseeritud situatsioon tagas võrreldavate tekstide ligilähedaselt võrdse mahu, siis pidasime Uberi'i indeksit oma materjalile piisavalt sobivaks ega hakanud otsima teksti pikkuse mõju minimeerimise keerukamaid viise (vrd Duran jt 2004).

Mitmed autorid on sõnavara rikkuse hindamise kõrval pidanud otstarbekaks pöörata tähelepanu ka kasutatud sõnavara raskusele, lähtudes seisukohast, et sagedamini esinevaid sõnu teatakse paremini kui harvaesinevaid (vt diskussioon Vermeer 2000: 79, Witalisz 2007: 107). See viis meid mõttele võrrelda uuritavate keelekasutajarühmade sõnu eesti keele sagedussõnastikuga: 10 000 avalike tekstide sagedasima sõnaga (Kaalep, Muischnek 2002), mille jagasime sagedusjärkudeks (kuni 1000 sagedasimat sõna; 1001–3000 sõna, mis koos 1000 kõige sagedasema sõnaga moodustab elementaarsõnavara; 3001 kuni 10 000 sõna, mis koos 3000 sagedasima sõnaga moodustab tavalise sõnavara; väljapoole sagedussõnastiku 10 000 sõna piire jääv harvaesinev sõnavara). Selle võrdlusega üritasime saada pilti sõnavara ulatusest kitsamas tähenduses.

Sõnavara rikkust mõõtsime eraldi kõigis uuritavates tekstitüüpides: suulises esinemises, dialoogis ning kirjalikus essees, sõnavara (sagedus)ulatust suulises (monoloog + dialoog) ja kirjalikus keelevormis.

Tulemusi võrdlesime sama materjali peal saadud teksti formaalsusindeksitega (Kerge jt 2007), millest lähemalt tulemuste peatükis.

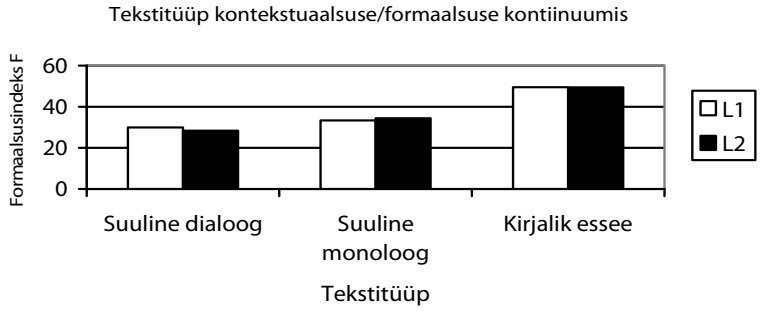
Tulemused

Keelekasutuse loomulikkuse sõnavaraga seotud parameetrite hulka kuulub nii siin kajastuv sõnavara rikkuse indeks ja sõnavara ulatus sagedusjärguti kui ka sõnaliigisuhteid kajastav formaalsusindeks *F*. Indeks näitab kontekstivaba ja kontekstisidusa sõnavara suhet kindlat liiki tekstis, iseloomustades seega teksti ühetimõistetavust ja jälgitavust: mida kõrgem on see indeks, seda ühemõttelisem on tekst. (Vt Heylighen, Dewaele 2002) Formaalsusparameeter on siinkohal oluline seepärast, et kõne loomulikkusele ei viita sõnakasutuse aspektist rikkus üksi, vaid kooskõlas loomuliku olukohase lausestusega, mida sõnaliigisuhe osutab.

Meie varasem eesti keele kontekstuaalsuse-formaalsuse uurimus[6] sellesama L1-L2 materjali võrdlusena viis kahe selge tulemuseni: 1) L1 ning L2 kõnelejad sellel dimensioonil ei erine ning 2) sõltumata uuritavate emakeelest on kõige ilmekam erinevus keelevormide ja tekstitüüpide vahel – kõige kontekstuaalsem on dialoog, kõige formaalsem kirjalik essee (vt joonis 1). Üldistatult: kontekstuaalsus kahaneb ja formaalsus kasvab suuliselt kirjalikule ja dialoogilt monoloogile. (Kerge jt 2007)

[5] Välja on arvatud näiteks *pulpulistlik* (*Pulpulistliku riigi juhtimine tõi sellist vilja*) jms ja paar täiesti arusaamatuks jäänud häälikukooslust.

[6] Teksti kontekstuaalsuse-formaalsuse andmed on saadud eesti keele jaoks kohandatud Heyligheni ja Dewaele (2002) valemiga, mis põhineb eeldusel, et mõnede sõnaliikide (pronoomenite, verbide, adverbide, interjektsioonide) sage esinemine tekstis muudab selle kontekstuaalsemaks ja seega mitmemõttelisemaks, samal ajal kui teiste (noomenid, adjektiivid, adpositsioonid, artiklid) sage esinemine langetab kontekstuaalsust, muutes teksti ühemõttelisemaks ehk formaalsemaks.



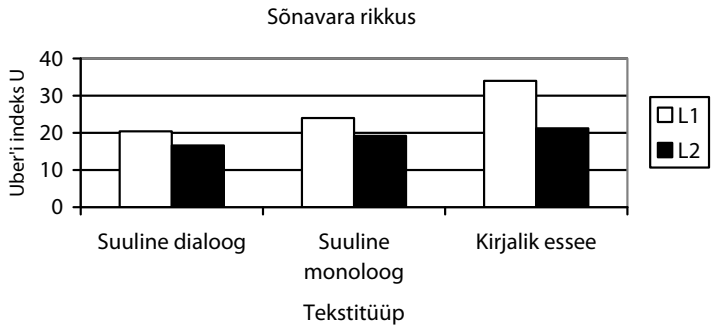
Joonis 1. L1 ja L2 kontekstuaalsuserinevused tekstitüübiti. Mida suurem *F*, seda formaalsem tekst (Kerge jt 2007)

Tekstitüüpide erinevusest lähtuvalt esitame ka sõnavara rikkuse mõõtmistulemused Uber'i indeksiga tekstitüüpide kaupa, tuues välja L1 ja L2 kasutajate indeksid (vt tabel 1).

Tabel 1. L1 ja L2 sõnavara rikkus erinevates tekstitüüpides. Esitatud on sõnade ja sõnede arv uuritavas tekstitüübis ning teksti sõnavara rikkust iseloomustav Uber'i indeks U. Mida suurem on U, seda rikkamaks võib pidada sõnavara

Tekstitüübid	**Eri sõnu (V)**		**Sõnesid (N)**		**Uber'i index (U)** $(_{\log}N)^2 /(_{\log}N-_{\log}V)$	
	L1	**L2**	**L1**	**L2**	**L1**	**L2**
Suuline dialoog	548	517	1752	2346	20,4	16,6
Suuline monoloog	477	402	1326	1348	24,0	19,2
Suuline kõne (dialoog + monoloog)	845	713	3078	3694	20,5	18,6
Kirjalik essee (monoloog)	736	666	1685	1824	34,0	21,2

Visuaalselt ilmekam on samade rühmade võrdlus tulpdiagrammil (vt joonis 2).

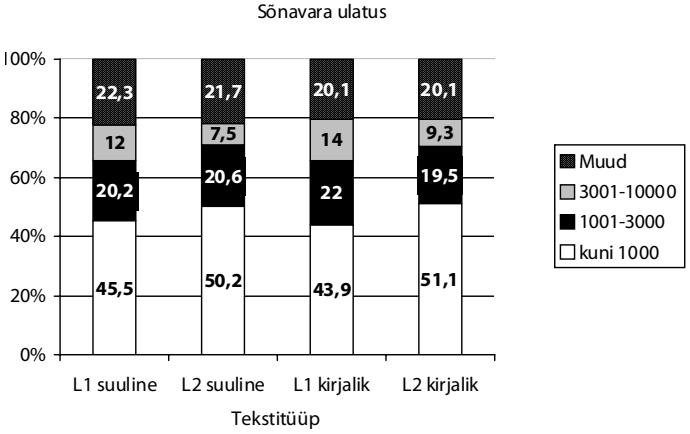


Joonis 2. L1 ja L2 sõnavara rikkus tekstitüübiti. Mida suurem on *U*, seda rikkamaks võib pidada sõnavara

Tulemustest nähtub, et tekstitüübid erinevad sõnavara rikkuselt. Loomulikuks võib pidada sõnavara rikkuse kasvamist dialoogilt monoloogi suunas ja suuliselt keelevormilt kirjaliku keelevormi suunas. Nii L1 kui ka L2 kasutajatel on sõnavara rikkuse indeks väikseim kõige kontekstuaalsemas uuritud tekstitüübis (dialoog) ja kõige suurem kõige formaalsemas tekstitüübis (kirjalik essee).

L2 sõnavara on siiski iga tekstitüübi piires märkimisväärselt vaesem kui L1 oma ja maksimaalne erinevus loomulikust ilmneb kirjalikus essees.

Uuritud rühmade sõnu eesti keele sagedussõnastikuga võrreldes (10 000 avalike tekstide sagedasimat sõna) jõudsime üsna üllatavale tulemusele: haritud emakeelekasutaja väga sagedase sõnavara osatähtsus oli kõigis tekstitüüpides märkimisväärselt suur – sageduselt esimese 3000 sõna järku jäi suulises tekstis 65,7% ja kirjalikus 65,9% sõnadest. L2 sõnavara ulatus on L1-ga väga sarnane: esimese 3000 sõna sagedusjärku jäi suulises tekstis 70,8% ja kirjalikus 70,6% sõnadest. Ka ülejäänud sõnade jaotus sagedussõnastiku taustal ei erine: sagedusrühma 3001–10 000 sõna jäi L1 puhul 12% ja L2 puhul 14% suulise teksti sõnadest ning vastavalt 7,5 ja 9,3% kirjaliku teksti sõnadest. Harvaesinevaid sõnu oli mõlemal kasutajarühmal sõltumata keelevormist veidi üle 20% (vt joonis 3).



Joonis 3. L1 ja L2 sõnavara võrdlus sagedussõnastikuga

Diskussioon ja järeldused

Sõnavara rikkuse kui osava (vilunud) suhtleja loomuliku keelekasutuse näitaja juures näib tähtsaim joon minevat kirjaliku ja suulise keelekasutuse vahelt ning teksti vormide vahelt (monoloog, dialoog). See tulemus toetab ja rikastab nn sõnavara rikkuse variatiivsuskeskset käsitlust (vt ingl *variationist view on lexical richness*: Gijsel jt 2005).

Suulisus-kirjalikkus ja monoloogilisus-dialoogilisus on ennekõike parameetrid, mis kirjeldavad žanrit (vt Chafe, Tannen 1987: 385): antud juhul kolme ühiskonna- ja keskkonnatemaatilise diskursuse kaudu seotud žanrit, s.o läbirääkimisi (arutelu ühise seisukoha leidmiseks), ettekannet ja esseed. Loomulikus keelekasutuses on kõige vaesem sõnavara interaktiivses arutelus, järgneb ettekanne, kõige rikkama

sõnavaraga on essee. See tulemus on kooskõlas näiteks Wallace Chafe'i tulemusega, et suulises tekstis on muuhulgas enam kordamist kui kirjalikus (vt osutust samas).

L2-kasutajate sõnavara on sõltumata žanrist märgatavalt vaesem kui L1 oma. Nii tekib oletus, et sõnavara rikkust oleks vaja testida eraldi. Seda tulemust interpreteerides tuleb aga arvestada, et kõigile uuritud L2-kasutajatele on juba antud eesti keele oskuse kõrgtaset kinnitav tunnistus (*de iure* C1) – tegemist on niisiis vilunud suhtlejatega. Siit võib järeldada, et keeleoskuse muud tahud kompenseerivad sõnavara suhtelist vaesust.

Kuivõrd suhteliselt vaene sõnavara ei ole seganud hindamast uuritavate keeleoskustaset suhteliselt kõrgeks, satub kahtluse alla ka eeltoodud oletus, et sõnavara rikkust oleks eraldi vaja testida. Rikkusparameetri reaalse tulemuse näivad korvavat teksti loomulik sõnaliigiline koostis, mis seisab sarnase formaalsusindeksi taga (joonis 1), ning sagedusjärkude L1-sarnane jaotus kui sõnavara ulatuse näitaja, mis tuvastati siinses uurimuses (joonis 3). Ka satuks sõnavara koostise detailne testimine rahuldava keelelise toimetuleku juures vastuollu raamdokumendi tegevus- ehk toimingupõhise ja õppijakeskse ideoloogiaga, millega seostub keelekasutaja vabadus valida suhtlusstrateegiaid (vt Raamdokument 2007: 24–28, 73, 143 jm; vt ka Kerge 2008: 25 jj, 55–56 jm).

Siiski väärib eritähelepanu L2 sõnavaene kirjalik keelekasutus (U-indeks 21,2): see sarnaneb pigem emakeelsete suulise keelekasutusega (20,5) kui kirjalikuga (34). Nii võib kaaluda sõnavara rikkuse mõõtmist kirjaliku teksti hindamise lisavahendina neil puhkudel, kui muud tunnused ei luba eksaminandi kirjaliku keelekasutuse taset kindlapiiriliselt määrata. (Suulise keelekasutuse juures, kus meetodi rakendamine oleks äärmiselt tüsilik, ei ole sellist lisavahendit meie andmetel vaja.)

Raske on interpreteerida tulemust, et elementaarsõnastiku osatähtsus läheneb loomulikus keelekasutuses kahele kolmandikule sõnadest. Edasises vajab sõnavara ulatuse see mõõde võrdlemist nii eri registrites kui ka suuremas tekstimassiivis. Siiski, vahendades paljude autorite tulemusi enam kui 30 aasta vahemikust, veenavad Wallace Chafe ja Deborah Tannen (1987: 385–86) lugejat, et suulise keelekasutuse puhul on sõnavara lihtsam ning verbide, asesõnade ja adverbide osatähtsus suurem. Seega iseloomustavad suulisust varasemate uurimuste järgi samad parameetrid, mis osutavad suurele kontekstuaalsusele Francis Heyligheni ja Jean-Marc Dewaele (2002) mõistes. W. Chafe'i ja D. Tannen'i seisukohad toetavad niisiis eespool viidatud kontekstuaalsusuurimuse tulemusi, millega kinnitasime L1- ja L2-kõnelejate sõnavara liigilise koostise suhtelist sarnasust, iseloomustamata täpsemalt, kuidas mõista sõnavara lihtsust (kas see haakub elementaarsõnavara leitud osatähtsusega).

Kokkuvõte

Tulemused lisavad loodavasse rääkimise loomulikkuse mudelisse vähemalt ühe, sõnavara loomuliku rikkuse kriteeriumi ja annavad uue mõõtme ka varem uuritud kontekstuaalsusele (F-indeksile). Uuringust ilmneb järgmine.

1) L1 kirjalikus kasutuses on sõnavara oluliselt rikkam kui suulises, samuti kasvab vähemsagedaste sõnade määr suuliselt kirjaliku suunas.
2) Sõnavara on monoloogis rikkam kui dialoogis.
3) Elementaarsõnavara osatähtsus L1 tekstides on ootamatult suur.

Funktsionaalne toimetulek L2 nõudlike tekstidega emakeelekõnelejast selgelt vaesema sõnavara juures suunab võrdlema keelekasutuse muid parameetreid. Sõnavara rikkuse mõõtmine võib seejuures osutuda kirjaliku keelekasutuse hindamise lisavahendiks.

Viidatud kirjandus

Asu, Eva Liina, ilmumas. Rising intonation in native and non-native Estonian. – Language History and Dialectology Issues. Vilnius: Lithuanian Language Institute.

Bachman, Lyle F. 2001 [1990]. Fundamental Considerations in Language Testing. Oxford Applied Linguistics. Oxford: Oxford University Press.

Chafe, Wallace; Tannen, Deborah 1987. The relation between written and spoken language. – Annual Review of Anthropology, 16, 383–407. doi:10.1146/annurev.an.16.100187.002123

Duran, Pilar; Marven, David D.; Richard, Brian J.; Chipere, Ngoni 2004. Developmental trends in lexical diversity. – Applied Linguistics, 25 (2), 220–242. doi:10.1093/applin/25.2.220

Gijsel van, Sophie; Speelman, Dirk; Geeraerts, Dirk 2005. A variationist, corpus linguistic analysis of lexical richness. – Proceedings from The Corpus Linguistics Conference Series 1 (1). Corpus Linguistics 2005, July 14-17 2005, Birmingham, UK. http://www.corpus.bham.ac.uk/pclc/index.shtml (30.09.2008).

Heylighen, Francis; Dewaele, Jean-Marc 2002. Variation in the contextuality of language: An empirical measure. – Foundations of Science, 7 (3), 293–340. doi:10.1023/A:1019661126744

Kaalep, Heikki-Jaan; Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik. Tartu: TÜ Kirjastus.

Kerge, Krista 2008. Vilunud keeleoskaja. C1-taseme eesti keele oskus. Tallinn: HTM, TLÜ, EKSA.

Kerge, Krista; Pajupuu, Hille; Altrov, Rene 2007. Tekst, kontekstuaalsus ja kultuur. – Keel ja Kirjandus, 8, 624–637.

Kerge, Krista; Pajupuu, Hille; Tamuri, Kairi 2008a. Where should TTS-synthesizer pause and breathe? – The Third Baltic Conference on Human Language Technologies. Vilnius: Vytauto Didžiojo Universitetas; Lietuviu kalbos institutas, 143–149.

Kerge, Krista; Pajupuu, Hille; Tamuri, Kairi; Meier, Heidi 2008b. Kõnetehnoloogia vajab žanrilist lähenemist. – Eesti Rakenduslingvistika Ühingu aastaraamat, 4, 53–65.

Little, David 2005. The Common European Framework and the European language portfolio: Involving learners and their judgements in the assessment process. – Language Testing, 22 (3), 321–336. doi:10.1191/0265532205lt3110a

Meister, Lya; Meister, Einar 2007. Perceptual assessment of Russian-accented Estonian. – ICPhS XVI: Proceedings of the 16th International Congress of Phonetic Sciences, 6–10 August 2007, Saarbrücken Germany. Saarbrücken: Universität des Saarlandes, 1717–1720.

Pajupuu, Hille; Kerge, Krista 2006. Hingav süntesaator ja pausid tekstiliigiti. – Keel ja Kirjandus, 3, 202–210.

Raamdokument 2007 = Euroopa keeleõppe raamdokument: õppimine, õpetamine ja hindamine. Tartu: Haridus- ja Teadusministeerium, 2007.

Rannut, Ülle 2005. Keelekeskkonna mõju vene õpilaste eesti keele omandamisele ja integratsioonile Eestis. Analüütiline ülevaade. TLÜ humanitaarteaduste dissertatsioonid 14. Tallinn: TLÜ Kirjastus.

Ratcliff, Ann; Coughlin, Sue; Lehman, Mark 2002. Factors influencing ratings of speech naturalness in augmentative and alternative communication. – AAC: Augmentative & Alternative Communication, 18 (1), 11–19. doi:10.1080/714043393

Read, John; Chapelle, Carol A. 2001. A framework for second language vocabulary assessment. – Language Testing, 18 (1), 1–32. doi:10.1177/026553220101800101

REKK 2007 = Riiklik Eksami- ja Kvalifikatsioonikeskus. Eesti keele tasemeeksamid. Statistika ja analüüsid. Vt http://www.ekk.edu.ee/eksaminandile/eesti-keele-tasemeeksamid/statistika-ja-analuusid (18.10.2008).

Scott, Michael 1996. WordSmith Tools. Oxford: Oxford University Press.

Tuokko, Eeva 2007. Mille tasolle perusopetuksen Englannin opiskelussa päästään? Perusopetuksen päästövaiheen kansallisen arvioinnin 1999 Eurooppalaisen viitekehyksen taitotasoihin linkitetyt tulokset. Jyväskylä Studies in Humanities 69. Jyväskylä: University of Jyväskylä.

Vermeer, Anne 2000. Coming to grips with lexical richness in spontaneous speech data. – Language Testing, 17 (1), 65–83. doi:10.1177/026553220001700103

Witalisz, Ewa 2007. Vocabulary assessment in writing: Lexical statistics. – Z. Lengyel, J. Navracsics (Eds.). Second Language Lexical Processes: Applied Linguistics and Psycholinguistic Perspectives. Second Language Acquisition. Clevedon: Multilingual Matters Ltd, 99–116.

Hille Pajupuu (Eesti Keele Instituut) uurimisvaldkonnad on kõneakustika, kultuuridevaheline kommunikatsioon, keeletestimine.
hille.pajupuu@eki.ee

Krista Kerge (Tallinna Ülikool) uurimisvaldkonnad on keele variatiivsus, tekstianalüüs, rakenduslingvistika (L1 ja L2 omandamine, õigus- ja haldussuhtlus, kõne paralingvistiliste komponentide ja süntaksi seosed).
krista.kerge@tlu.ee

Pilvi Alp (Riiklik Eksami- ja Kvalifikatsioonikeskus) uurimisvaldkond on keeletestimine.
pilvi.alp@ekk.edu.ee

NATURAL LEXICAL RICHNESS IN EDUCATED LANGUAGE USE

Hille Pajupuu, Krista Kerge, Pilvi Alp
Institute of the Estonian Language, Tallinn University,
The National Examinations and Qualifications Centre

Lexical richness/diversity and vocabulary range belong to measures of language competence. The vocabulary of local Russians with advanced Estonian proficiency (B2+/C1) was compared to the vocabulary of native Estonians with non-philological tertiary education. Three types of texts were used: oral dialogue, oral presentation and written essay. Lexical richness was measured by the Uber index. The vocabulary range was found by comparing the L1 and L2 vocabularies used by the subjects to a list of 10,000 most frequent words.

The two groups differed considerably on lexical richness: L1 results surpassed those of L2 in dialogue as well as in monologue, but most of all in essay. Vocabulary range, however, showed a similar pattern for the two groups: 65% of the L1 vocabulary used and 70% of that of L2 belonged, both in oral and written use, to the basic vocabulary, i.e. to the first 3000 words on the frequency list. The proportion of rare words (range over 10.000) was about 20%, in oral as well as written texts in both L1 and L2. Considering the above results together with the indices of text formality, which were rather similar for L1 and L2, we reached the conclusion that poorer vocabulary is no real hindrance to either free talking or writing, if the word use is adequate to register and genre.

Keywords: L1, L2 acquisition, C1, lexical richness, vocabulary range, frequent words, text type, Uber index, Estonian