EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 3, 321–338

VENE-EESTI KOODIVAHETUSE KORPUS: KODEERIMISPÕHIMÕTETE VÄLJATÖÖTAMINE

Anastassia Zabrodskaja

Ülevaade. Artikli eesmärk on anda ülevaade Tallinna Ülikooli üld- ja rakenduslingvistika õppetooli juures valmivast vene-eesti koodivahetuse korpusest, mis pannakse tulevikus Tallinna Ülikooli eesti filoloogia osakonna kodulehele. Artiklis tutvustatakse lühendeid LIPPS ja LIDES ning nendega seonduvat, mh rahvusvahelise kakskeelse (*resp*. mitmekeelse) andmebaasi peamisi standardseid transkribeerimis- ja kodeerimispõhimõtteid[1] vene-eesti koodivahetuse korpuse näitel. Artikli viimases osas analüüsitakse sisestava (valdavalt ühesõnalise) vene-eesti koodivahetuse morfoloogia kodeerimisvõimalusi, sõnastatakse peamised üleskerkinud probleemid ning pakutakse ka võimalikke lahendusi.*

Võtmesõnad: korpuslingvistika, koodivahetus, eesti keel, vene keel

Sissejuhatus: mis on LIPPS ja LIDES

Viimastel kümnenditel on kakskeelsuse uurimine kujunenud keeleteaduse iseseisvaks valdkonnaks, selle uurijad on erinevatest keelekogukondadest kogunud märkimisväärse hulga andmeid. Ent erinevate uurimiseesmärkide ja -võimaluste tõttu polnud materjali üleskirjutamises ühtsust, mis raskendas keelelise materjali vahetamist ning tõstis esile vajaduse luua kakskeelse keeleainese transkribeerimisstandard. Kuna autorid kirjeldasid koodivahetuse juhtumeid eri termineid kasutades, väites, et just nende terminid on paremad, täpsemad ja sobivamad, rajas Euroopa Teadusfond (European Science Foundation) 1990. aastate esimesel poolel koodivahetuse ja keelekontaktide võrgustiku (Network on Code-Switching and Language Contact), mille eesmärk oli selgete ja üldaktsepteeritavate koodivahetuse põhimõistete arendamine (Milroy, Muysken 1995: 12).

Koodivahetuse ja keelekontaktide võrgustik soodustas keelte interaktsiooni andmevahetussüsteemi loomist. Keelte interaktsiooni ehk keelte vastasmõju (ingl *language interaction*) uurimine keskendub kakskeelsele (mitmemurdelisele)

[1] Artiklis kasutatakse terminit *kodeerimine*, lähtudes ingliskeelsest vastest *encoding*.

* Käesolev artikkel on kirjutatud ETF-i grandi nr 6151 "Koodivahetuse, eesti vahekeele ja lastekeele andmekorpuse koostamine ja üldkirjeldus" raames. Autor tänab prof Martin Ehalat, prof Anna Verschikut ja dr Ad Backust asjakohase terminoloogia arutamise ja kasulike nõuannete eest.

keelekasutusele erinevatest sotsiaalsetest ja lingvistilistest vaatenurkadest. Rohkem kasutatavate terminite *koodivahetus* (ingl *code-switching*) ja *keelekontakt* (ingl *language contact*) asemel valiti termin *keelte vastasmõju*, et hõlmata kõiki keelelisi nähtusi, mitte ainult neid, mis kuuluvad kahte eri süsteemi (LIPPS Group 2000: 133). Andmevahetussüsteem suurendas üldise transkribeerimis- ja kodeerimisjuhendi vajalikkust ning tõstis probleemi veel rohkem esile. See tähendas omakorda, et transkribeerimis- ja kodeerimisstandardite kehtestamisel pidi jõudma üksmeelele, võimaldamaks teiste keeleteaduse harudega (näiteks diskursusanalüüsi või süntaksiga) tegelevatel uurijatel keelte vastasmõju andmevahetussüsteemist samuti kasu saada.

Ljouwertis/Leeuwardenis 1994. aastal toimunud konverentsil tõdeti, et koodivahetuse uurijad saavad oma andmeid vahetada ainult eraviisiliselt, sest puudub formaliseeritud süsteem, standard jms (LIPPS Group 2000: 133). Samal ajal on lastekeele uurijatel olemas ühtne standardne viis andmete transkribeerimiseks ja kodeerimiseks – CHILDES (Children Language Data Exchange System, umbkaudu: lastekeele andmevahetussüsteem). Seetõttu on olemas ka andmebaasid, mida võib hõlpsasti omavahel vahetada. CHILDES võimaldab eri riikide keeleteadlastel jagada lastekeele materjali ja teha võrdlevaid rahvusvahelisi uurimusi ning analüüse.

Ülaltoodud asjaolud ajendasid LIPPS-i (Language Interaction in Plurilingual & Plurilectal Speakers, umbkaudu: mitmekeelsete kõnelejate keeleline vastasmõju) uurimisrühma loomist ja mitmekeelse interaktsiooni kodeerimisjuhendi LIDES (Language Interaction Data Exchange System) koostamist. Nii ilmuski mitme uurija pika koostöö tulemusel 2000. aastal ajakirja International Journal of Bilingualism erinumbrina "The LIDES Coding Manual". Sellel oli kaks peamist eesmärki:

1) pakkuda vilunud uurijatele kakskeelse ainese transkribeerimise ja kodeerimise juhtnööre olemasolevat arvutitarkvara kasutades;
2) julgustada algajaid uurijaid kakskeelseid andmeid kodeerima, toetudes üldstandardile, ning anda nende võimalikele küsimustele lihtsamas vormis vastused.

LIDES-i ettevalmistamisel kasutati hüppelauana Brian MacWhinney (1995) CHILDES-i projekti, mida tuli mugandada kakskeelse vestluse analüüsi jaoks sobivamaks (pidi tähistama selliseid ükskeelsest kõnest puuduvaid nähtusi nagu grammatiline ja semantiline transferents, leksikaalsed laenud, koodivahetus, koodisegu, diskursuse mallid jne).

Siinjuures tuleb mainida, et LIDES annab keeleliste nähtuste ja nende tähistuste loendi ning loomise põhimõtted. Igal uurijal tekivad aga oma täiendavad vajadused: näiteks ühte huvitab pigem mitmekeelse vestluse struktuur ja kõnevoorude ülesehitus, teist koodivahetuse grammatilised omadused jne. Seetõttu tuleb lisaks konventsionaalsetele põhimõtetele kodeerijal välja mõelda võtteid, mis lubaksid kajastada just teda huvitavaid seiku. Seda asjaolu näitlikustab ka vene-eesti koodivahetuse korpus. Järgmisena tutvustataksegi selle koostamisprintsiipe, -probleeme ja lahendusviise.

1. Vene-eesti koodivahetuse korpuse eellugu, sisu ja struktuur

1990. aastate poliitilised sündmused tõid eestlaste ja venelaste suhtlemismallidesse muutusi. Igapäevaelus hakkasid vene emakeelega kõnelejad kasutama üha enam eesti keelt. Kaasaegseid eesti-vene kontakte saadi hakata uurima kasutades või testides Lääne sotsiolingvistika teoreetilisi mudeleid ja uurimisparadigmasid. Eesti juhtivaks koodivahetusuurijaks on Anna Verschik ning tema juhendamisel on kaitstud hulk bakalaureuse- ja magistritöid, mis käsitlevad mõlema keele kontaktidest johtuvaid nähtusi.

Anastassia Zabrodskaja (2005) on vaadelnud Kohtla-Järve vene-eesti kooliõpilaste koodivahetust pragmaatilisest seisukohast. Jekaterina Ozernova (2005) on käsitlenud koodivahetust, laenamist ja konvergentsi[2] eesmärgiga uurida eesti leksikaalseid laene Eesti venelaste kõnekeeles. Tatjana Baškirova (2006) on näidanud koodivahetust teise keele omandamist soodustava vahendina. Indrek Konnapere (2006) on kirjeldanud ja analüüsinud kaitseväes aega teenivate vene keelt emakeelena rääkivate noormeeste vene-eesti koodivahetust ja selle funktsioone. Et nelja töö tulemusel saadud keelematerjal (lindistatud, osa käsitsi kirja pandud) ei läheks kaduma, nagu see juhtus omal ajal Nõukogude okupatsiooni eelse venelaste keelega (sellest on palju rääkinud Sergei Issakov, vt lähemalt Verschik 2001: 531), otsustati luua vene-eesti koodivahetuse korpus ning astuda LIPPS-i rühmaga ühendusse. Esimene töökohtumine leidis aset Limerickis 2006. aasta suvel 16. sotsiolingvistika sümpoosioni ajal. Kollokviumil "New Tools for Research in Bilingualism and Code-Switching: Studies using the Language Interaction Data Exchange System" (umbkaudu: kakskeelsuse ja koodivahetuse uued uurimisvahendid: LIDES-it kasutavad uurimistööd) tutvustasid A. Verschik ja A. Zabrodskaja valmivat vene-eesti koodivahetuse korpust laiale ringkonnale ning said asjakohast vastukaja (Verschik, Zabrodskaja 2006).

Kuna potentsiaalsed kasutajad on uurijad kogu maailmas, siis kõigi koodivahetuskorpuste metakeel on inglise keel (tekstide kommentaarid ja muu vajalik abiinfo, nagu situatsiooni kirjeldus, uurija kommentaarid jms). Et ka teised uurijad saaksid vene-eesti koodivahetuse korpuses orienteeruda, järgitakse sama konventsiooni. Andmaks artikli lugejatele parimat ülevaadet valmivast vene-eesti koodivahetuse korpusest, esitatakse näited samast põhimõttest lähtuvalt ning tõlgitakse eesti keelde erivajadusel.

Tavaliselt pole koodivahetuse korpused pelgad tekstikogumid, vaid sisaldavad ka mitmesugust taustainfot. Ühe näitena võib tuua türgi-hollandi andmebaasi, mille autor on Ad Backus. Oma korpuse juurde on ta teabefailis (ingl *readme file*) lisanud keelekogukonna ja informantide iseloomustuse ning kirjeldanud üksikasjalikult vastasmõju tüüpi (LIPPS Group 2000: 164–165).

Ka vene-eesti koodivahetuse korpusele on lisatud ingliskeelne Eesti venekeelsete kogukondade iseloomustus. Selline sissejuhatus on vene-eesti koodivahetuse korpuse puhul eriti vajalik. Selgitades Eesti etnolingvistilise situatsiooni kujunemist, tuleb rõhutada, et venekeelse elanikkonna kujunemine Eestis (vt Rannut 1994, Vseviov 2002) on küllaltki ainulaadne nähtus kogu sõjajärgse Euroopa ajaloos. Sarnast tüüpi keelekogukondi ei ole Läänes ega Idas, sest tegemist pole põlisvähemuse, venekeelsete immigrantide (selle sõna tavalises mõttes) ega vene diasporaaga. Kuigi

[2] Konvergentsi kohta vt Ozernova 2005, Zabrodskaja 2006.

venelaste immigratsioon Eestisse suurenes plahvatuslikult pärast II maailmasõda (vt Viikberg 1999), polnud need ümberasujad samalaadsed näiteks Hollandi türklastega (Backus 1996: 43–46) või muude immigrantlike keelekollektiividega Euroopas. Venelaste immigratsioon meenutab teatud mõttes koloniseerimist, sest venelased asusid ümber (keskvõimu heakskiidul ja soosimisel) territooriumile, mida peeti oma riigi osaks. Maailmas pole seda tüüpi kogukondade sotsiolingvistikat uuritud. Seega tutvustab valmiv korpus mitte ainult vene-eesti koodivahetuse nähtusi kaasaegse kontaktlingvistika metakeeles, vaid annab ka ülevaate Eesti eri paikade vene-eesti kontaktide ja venekeelsete keelekogukondade omapärast.

Kõnealune vene-eesti koodivahetuse korpus koosneb kolmest osast:

1. Eesti televisiooni kakskeelsed saated. Kokku 76 saadet, mis on litereeritud vaid osaliselt, kuna selles on palju ükskeelset materjali. Lisaks pole ükski saade kahe inimese dialoog, vaid koosnevad peamiselt väikestest, kuni kümnelauselistest vestlustest. Kõik vestlustes osalejate nimed on anonüümsuse huvides transliteratsiooni käigus asendatud lühenditega. Kui nimi on teadmata, siis on tähistatud näidetes eestlane tähega E(stonian), venelane tähega R(ussian). Kui ühes vestluses osaleb mitu inimest, on need tähistatud R1, R2, E1, E2 jne. Kui räägitakse kolmandatest inimestest, kes vestluses ei osale, siis nende nimed muudeti. Saatejuhi märgistus on H show_host, kui on kaks saatejuhti siis: H1 show_host_one, H2 show_host_two.
2. Tallinna venekeelsete elanike vene-eesti koodivahetus. Suur osa Tallinna koodivahetuse näiteid on kogutud kaubanduslikust suhtlusest (turult, poest jne) ning kirjalikest allikatest (poe- ja turuletid). Tekstid on litereerinud A. Verschik ja A. Zabrodskaja ning kodeerinud A. Zabrodskaja.
3. Ida-Virumaa vene emakeelega laste vene-eesti koodivahetus. Ida-Virumaal aastatel 2000–2006 käsitsi kirja pandud keeleaines on valdavalt pärit A. Zabrodskaja iseseisvast uurimusest, mis oli aluseks tema magistriväitekirjale (Zabrodskaja 2005). Translitereeringus on vestluses osalejate nimed kodeeritud korpuse 1. osas kirjeldatud meetodil. Õpilane (kui sugu on jäänud üleskirjutamise hetkel tähistamata) on näidetes tähistatud lühendiga S(tudent), õpetaja aga T(eacher).

Käsitsi kirja pandud kakskeelsetes vestlustes osalevad (vahel lihtsalt kõrvaltvaatajaks) kas A. Zabrodskaja või A. Verschik, neid tähistatakse vastavalt Researcher_one (RES_1) ja Researcher_two (RES_2). Lindistatud saated välja arvatud, on mõlemad uurijad olnud vaatlejad ja mõnikord ka vestluse algatajad.

Korpuse I alamosas on analüüsi aluseks aastatel 2000–2006 lindistatud 13 järgmist telesaadet:

1) “Dilemma” – 5 saadet 2005. aastast
2) “Kolmas sektor” – 5 saadet: 2 saadet 2001. a ja 3 saadet 2002. a
3) “Neli aastaaega” – 2 saadet: 1 saade 2000. a ja 1 saade 2001. a
4) “Pealtnägija” – 1 saade 2002. a
5) “Politseinädal” – 1 saade 2001. a
6) “Pressiklubi” – 3 saadet 2001. a
7) “Sputnik” – 13 saadet: 5 saadet 2001. a ja 8 saadet 2002. a
8) “Subboteja” – 1 saade 2005. a
9) “Subjektiiv” – 5 saadet: 4 saadet 2001. a ja 1 saade 2002. a
10) “Teadmiseks” – 2 saadet 2001

11) “Unetus” – 34 saadet: 6 saadet 2002. a, 3 saadet 2003. a, 5 saadet 2004. a, 18 saadet 2005. a ja 2 saadet 2006. a
12) “Uudistaja” – 3 saadet 2002. a
13) “Vaata mind” – 1 saade 2004. a

Korpuse II alamosas on analüüsi aluseks igapäevasuhtlus Ida-Virumaal:

1) Narva:
   – Narva koolid (koolide ametlikud nimed pole välja kirjutatud)
   – Tartu Ülikooli Narva Kolledž
2) Kohtla-Järve:
   – Kohtla-Järve 3. keskkool
   – Kohtla-Järve Pärna põhikool
   – Kohtla-Järve avaliku teeninduse punktid (kauplused, pangad, postkontor jms)

Korpuse III alamosas on analüüsi aluseks igapäevasuhtlus Tallinnas. Näiteid on kogutud järgmistest kohtadest:

1) Tallinna Ülikool
   – Tallinna Ülikooli Akadeemiline raamatukogu
   – Tallinna Ülikooli vene emakeelega üliõpilastega lindistatud intervjuud (20 tundi)
2) Tallinna bussijaam
3) kaubanduslik suhtlus:
   – Balti jaama turg (suurim alamkorpus – 36 näidet)
   – kauplused
   – avaliku teeninduse punktid (reisibürood jne)
4) õpetajate täienduskoolituskeskus (ametlik nimi on välja kirjutamata)

Tallinna Ülikooli kolmanda alamosana on kavas esitada 2005/2006. õppeaasta ainekursustel “Keeleteaduse alused”, “Üldkeeleteadus”, “Keeleteaduse uuemaid suundumusi” ning “Kultuuridevaheline kommunikatsioon” osalenud üliõpilaste kirjalikest töödest kogutud koodivahetusjuhtumid.

2. Transkribeerimise ja kodeerimise põhimõtetest

Konversatsioonianalüüs eeldab spontaanse kõne üleviimist kirjalikku formaati. Transkribeerimine ja kodeerimine on suulise kõne analüüsi olulisimad osad. Nii ühest kui teisest sõltub, mis tulemusteni uurija jõuab. Mary Bucholtz (2000: 1440) juhib tähelepanu mitmele mõlemat tegevust mõjutavale tegurile: nii keeleuurija enda ideoloogia kui ka poliitilised tingimused, tema uurimistöö eesmärgid ja potentsiaalne lugejaskond ning vahel isegi kodeerija tehnoloogiliste võimaluste piiratus. Peter Auer (1998: 8–10) on transkribeerinud enda loodud reeglite järgi Carol Myers-Scottoni (1993: 83) varem transkribeeritud mitmekeelset vestlust ja jõudnud vastupidistele järeldustele.[3]

Elinor Ochsi (1979) arvates on transkribeerimine subjektiivne tegevus. Kõigepealt peavad uurijad valima ja mugandama definitsiooni põhiüksuse kohta. Transkribeerija peab otsustama, kuidas erinevad üksteisest lause, lausung ja kõnevoor. Seejuures tuleb rõhutada, et lause defineerimine varieerub vastavalt iga

[3] Antud artikli raamid ei luba süveneda selle vestluse kahte analüüsi. Nendime vaid, et näide illustreerib lahknevusi erinevate kodeerimislähenemiste vahel.

keeleuurija isiklikele huvidele. Ühtset kokkulepet saavutada on raske. Tiit Hennoste (2000a) märkab õigustatult, et konversatsioonianalüüsi eri käsitlustes esile toodud üksused kattuvad suurel määral, ning samas, et üksuste piiritlemine on küllaltki udune. Kuigi põhilised kriteeriumid on süntaktilised, tuleb mõnikord tugineda ka intonatsioonile, nimelt siis, kui laused on elliptilised, neist puudub lauseliikmeid või nad on poolikud (Hennoste 2000a: 2227). Penelope Gardner-Chloros jt (1999: 412) osutavad lause, fraasi, väljendi defineerimise problemaatilisusele. Lause piiritlemiseks on mõningaid kriteeriume, näiteks üksiksõna, intonatsiooniline tervik, paus jms (Gardner-Chloros jt 1999: 408, 413). Fred Karlssoni (2002: 151) järgi on suulise keele põhiüksuseks lausung ehk kompaktne lõik, millel on füüsiliselt määratletavad piirid. T. Hennoste (2000a: 2223–2224) vaatleb eri autorite suulise kõne liigendusüksusi ning liigitab kesksed üksused kolme rühma:

a) semantilis-intonatsioonilised üksused: toonigrupp, süntagma, ideeüksus;
b) lausungid: prosoodiline lausung, vooruehitusüksus, makrosüntagma;
c) kõnevoor ja paratoon.

Mitmekeelset kõnet kodeeritakse kas lause-, fraasi- või üksiksõna tasandil (Gardner-Chloros jt 1999). Lundi ülikooli juures 1970. aastatel töötanud töörühm Talsyntax on pakkunud makrosüntagma mõiste, mis erinevalt lause mõistest hõlmab ka verbita väljendeid (vt tabel 1). Makrosüntagma ja lausungi ning nende piiridega seotud probleemidest on eesti keeles kirjutanud ka T. Hennoste (2000a, 2000b, 2000c).

Tabel 1. Suulise lause määratlemisviisid (Karlsson 2002: 152 ja seal kasutatud viited)

	Makrosüntagmad
1	täisstruktuuriga lausemakrosüntagmad ehk (normaal)laused
2	lausekatkendid: *tere päevast; tagasi ülikooli*
3	interjektsioonimakrosüntagmad: *hei; kurat; häh; nojah*
4	pöördumismakrosüntagmad: *poiss!; kallis sõber*
5	lõpetamata makrosüntagmad (katkendid): *mine sa*
6	mittegrammatilised ehk valesti moodustatud lausemakrosüntagmad

Tabelist 1 võib järeldada, et erinevatest sõnavormidest koosnev makrosüntagma on spontaanse kõne tüüpiline põhiüksus, mis on vene-eesti korpuse puhul väga vajalik.

Huvitav on seegi, et CHILDES-i koostamisreeglitest (MacWhinney 1995) ei leia keeleuurija lause definitsiooni, see jäetakse igaühe enda otsustada. Mõned kasutavad INTROS-i süsteemi (INformant's TRanscriptiOn String, umbkaudu: informandi transkribeeritud lint), milles pakutakse, et lähtuda tuleb süntaktilisest kriteeriumist ja defineerida lause nagu komponent (ingl *unit*), mis võib koosneda üksiksõnast, üksikfraasist, liht- või liitlausest (vt lähemalt LIPPS Group 2000: 173–174, 267).

Vene-eesti koodivahetuse korpuse suulise kõne lausete transkribeerimisel peetakse peamiseks lausungi piiriks pausi või vooru vahetumist. See viib aga kahjuks järgmise probleemini: kas vooru vahetumiseks pidada seda, kui kõneleja katkestab end ise või katkestatakse teda.

2.1. Tähestiku valiku problemaatika

Kui tegu on kahe keelega, siis tahaks iga uurija transkribeerimissüsteemi, mis võimaldaks neid omavahel eristada. Üks valik võib olla traditsiooniline kirjasüsteem, kui selline olemas. Kui uurija valib keeleliselt neutraalse (ingl *language-neutral*) tähestiku nagu IPA (International Phonetic Alphabet), siis peab näitama, mis keeles on konkreetsed transkribeerimise lõigud. Kuigi IPA on universaalne, kõigile keeleuurijatele ühtmoodi arusaadav transkribeerimisviis, kasutatakse ka mitmesuguseid muid transkribeerimis- ja transliteratsiooniviise. Näiteks LIDES-i autorid (LIPPS Group 2000: 214–246) on esitanud ladina tähestikku kasutavad keeled standardkujul. Ka vene-eesti koodivahetuse korpuses ei kasutata IPA-t. Eestikeelsed elemendid esitatakse eesti õigekirjas. Vene keele puhul on muidugi keerulisem, aga püütakse translitereerida põhimõttel "üks häälik – üks märk".

Keele tähistuse probleem aga siiski jääb, sest mõnedes lausetes esineb lausetäiteid (vene-eesti koodivahetuse korpuse puhul näiteks *hm*, *oh*, *o*, *a*, *aha(a)* jne), mida on raske omistada konkreetsele keelele. Seda enam, et võib olla ka sõnu/silpe, mis kuuluvad kolmandasse keelde. Vaatame näiteks järgmist kolmkeelset lauset:

(1)	Segodnja	*kes*	*on*	on	duty?
	täna	kes	on	PREP	korrapidaja
	vn	ee	ee	ingl	ingl

'Kes on täna korrapidaja?'

Mida sellistel puhkudel teha ja kuidas seda näidata, nõuab teoreetilisi otsuseid. Asi pole ainult meetodi valikus mõne sõna märkimisel (kas K1, K2 või hoopis K3). Suurt rolli mängivad ka süntaksi, fonoloogia ja pragmaatika aspektid. Kaks (või rohkem) keelt teevad transkribeerimist aina keerulisemaks, sest tuleb näidata, mis keelde iga keeleelement kuulub. Näite (1) juures võib tekkida küsimus, kuhu kuulub keeleüksus *on*. Ei saa sajaprotsendiliselt väita, et esimene *on* on just eestikeelse verbi *olema* 3. isiku ainsuse pöördeline vorm, aga teine *on* on ingliskeelne prepositsioon. Esimene *on* on vaieldav, sest seda võib interpreteerida ka ingliskeelse prepositsioonina. Teine *on* peab olema kindlasti ingliskeelne prepositsioon, vastasel korral ei omaks lause tähendust.

Korpuses saab esitada seda rida järgmiselt (näites (2) on transkribeeritud faili katkend).

(2)

@Languages:	Russian (1),		Estonian (2),		English (3)	
*G1:	segodnja@1		kes@2	on@2	on@3	duty@3 ?
%glo:	today		who	is	on	duty
%tra:	who is	on	duty	today		

On otsustatud, et venekeelseid elemente translitereeritakse, sest vene ortograafia nn joteeritud tähed vastavad tihti kahele häälikule, aga morfeemidevaheline piir jookseb sageli just nende kahe hääliku vahel (näiteks koodivahetatud *sai-ječk-a*, vt Zabrodskaja 2005: 41). Näeme, et vene tähestiku kasutamine ainult raskendaks glossimist. Inglise keele keskses CHILDES-i programmis on vene tähestiku kasutamine võimatu. Ladina tähestiku kasutamine ei loo visuaalset kontrasti vene ja eesti keele vahel ja raskendab lugemist ning koodivahetatud sõnast/väljendist arusaamist. Seetõttu tähistatakse iga elemendi keelelist kuuluvust, kasutades märke @1 jne.

2.2. Korpuse failide ülesehitusest

Korpuse valmiskogum peab sisaldama kolme või nelja faili: CHAT-andmefaili (ingl *datafile*), taustafaili (*depfile*), teabefaili ning võimalusel ka täiendfaili (*depadd*).

Korpus kui selline sisaldub just andmefailis. CHAT-andmefail koosneb tüüpiliselt kolmest allosast: faili päised (*file headers*), põhiread (*main tiers*) ja sõltread (*dependent tiers*). Kuigi kõik sõltread ei ole kohustuslikud, soovitatakse kasutada morfeemtõlke rida (%glo) ja lause vaba tõlke rida (%tra) (LIPPS Group 2000: 149).

CHAT-andmefaili koostamisel peab lähtuma järgmistest põhimõtetest:

1) faili iga sümbol peab olema ASCII (American Standard Code for Information Interchange – Informatsiooni Vahetamise Ameerika Standardne Kood) tähis;
2) iga rea peab lõpetama sisestusklahviga.

On ka teisi nõudeid, mis toimivad iga konkreetse punkti juures.

Tausta- ja täiendfailid teevad kindlaks koodivahetuskorpuse süntaksi ja struktuuri. Esimeses on kõik automaatsed kodeerimisvahended. Kõik uued koodid peavad sisalduma aga LIDES-i erilises alamkogumis – täiendfailis, nende seletus on teabefailis. Kodeerimisõpikus antakse isegi loetelu, mida täiendfail peab täpselt määratlema (LIPPS Group 2000: 162). Andmete sisestamis-, kodeerimis- ja selgitusviisi õigsust võib kontrollida programmi käskluse CHECK abil. See osutab ka vigadele valmiskogumi kõigis failides.

2.2.1. Faili päised

Faili päised on tekstiread, mis annavad informatsiooni faili kohta. Need päised asuvad iga transkribeeritud näite alguses ning algavad märgiga @, millele järgneb pealkiri. Mõned päised ei vaja mingit seletust, need on "tühjad" päised, näiteks @Begin või @End. Teistele päisetele järgneb informatsioon osalejate, situatsiooni jms kohta. Seda informatsiooni nimetatakse sissekandeks (ingl *entry*). Sissekannet nõudvatele päistele järgneb koolon ja tabeldusklahv, seejärel sissekanne ise. Näiteks:

(3)
@Participants: SHO shopkeeper, RES researcher, CHI child

Peale tühjade päiste ja informatsiooni sisaldavate päiste eristuse pakub CHAT ka kohustuslikke ja mittekohustuslikke ridu.

Järgmisi faili päiseid (vt tabel 2) on vaja programmi automaatse analüüsi käivitamiseks (MacWhinney 1995: 13–14; LIPPS Group 2000: 149).

Täiendavat infot sisaldavad mittekohustuslikud ehk fakultatiivsed read (LIPPS Group 2000: 149-150). Neid on kahte tüüpi: püsivad ja muutlikud.

Püsipäised (ingl *constant headers*) sisaldavad kasulikku informatsiooni, mis ei muutu terve faili jooksul. Samas võivad mõned päised olla kas püsivad või muutlikud, nt aeg (*date*), koht (*location*), situatsioon (*situation*) jne. Need read peab paigutama faili algusesse, s.t enne kõnet (LIPPS Group 2000: 150).

Kuna transkribeeritud vestlustes on kasutatud mitut keelt (antud korpuse puhul vähemalt kaht, vene ja eesti), soovitatakse kindlasti kasutada järgmist kaht rida (vt tabel 3).

Tabel 2. Transkribeeritud faili kohustuslikud read

@Begin	Iga transkribeeritud faili esimene rida. Tähistab terviku algust.
@Participants	Selles reas kirjutatakse lahti, kes on vestluse osalejaskond. See nõuab sissekannet, mis koosneb kahest või kolmest osast: a) kõnelejate ID (kohustuslik), mis koostatakse ühest kuni kolmest tunnusmärgist (suurtähtedega ja/või numbritega), nt STA või S81; b) kõnelejate nimed (mittekohustuslik); c) kõnelejate roll (kohustuslik), nt *Interviewer* 'intervjueerija'. Rolli, mida pole taustafailis, peab lisama täiendfaili. Viimased kaks osa (b ja c) võivad koosneda mitmetest elementidest, kuid neid tuleb ühendada allkriipsuga, nt *Stasja_Brodska* või *sister_in_law*.
@End	Iga transkribeeritud faili viimane rida. Tähistab terviku lõppu.

Tabel 3. Keelte tähistamine mitmekeelses korpuses

@Language(s):	Päis näitab transkribeeritud näidete peamist keelt või keeli.
@Language of XXX:	Päist võib kasutada, et kirjeldada konkreetse vestluspartneri peamist keelt (peamisi keeli). Sel juhul tuleb teha eraldi rida iga kõneleja kohta. Võib anda informatsiooni ka iga osaleja kakskeelsuse astmest, seda tuleb teha teabefailis.

Teisi näiteid püsipäistest (vastavalt vanus, haridus, sugu):

@Age of XXX:
@Education of XXX:
@Sex of XXX:

Muutpäised (ingl *changeable headers*) võivad esineda niihästi faili alguses koos püsipäistega kui ka faili põhiosas. Muutpäised sisaldavad informatsiooni, mis muutub faili jooksul. Nii näitavad need, kuidas saab vestlust etappideks jaotada tegevustest lähtuvalt. Seepärast sisestatakse need päised sinna, kus informatsioon muutub. Näite (4) neljas rida @Activities selgitab, mis juhtub vestluse erinevatel staadiumitel.

(4)
@Begin
@Participants: B1 boy_one, B2 boy_two
@Sex of Participants: male, male
@Age of Participants: 8, 8
@Mothertongue of Participants: Russian, Russian
@Languages: Russian (1), Estonian (2), Undecidable (0)
@Location: Kohtla-Järve Pärna School
@Date of boarding: JUN-02
@Date of coding: 04-APR-06
@Transcriber and Coder: Zabrodskaja, Anastassia
@Warning: this is only a small sample

*B1: xxx@o
@Activities: B1 dressed in hooded jacket stands in the classroom.
*B2: xxx@o

@Activities:	B2 is	talking	to	B1	behind	his	back.
@Activities:	B2 puts	the	hood	on	the	head	of B1.
*B1: ty@1	chto@1	loll@2		?			
%tra:	are	you	a	fool			
@Activities:	B1 is	turning		angrily		to	B2.
*B2: sam@1	takoj@1.						
%tra:	you	are					
@End							

Näitevestluses (4) esineb rida @Activities neli korda. See näitlikustab, mida osalejad teevad vestluse või situatsiooni jooksul ning aitab koodivahetuse põhjusest aru saada. Rida @Activities illustreerib vahel väga selgelt, miks konkreetne kõneleja läheb ühelt keelelt teisele, ehk kuidas koodivahetus on seotud tegevuse muutumisega.

2.2.2. Põhiread

Tegelikku kõnet transkribeeritakse andmefaili põhiridadel, mis reprodutseerivad kirjalikus vormis iga vestlusest osaleja ütluse. Nõudmised on järgmised (MacWhinney 1995: 8–9).

1. Põhiread osutavad väljaöeldule ning algavad tärniga (*). Iga põhirida peab sisaldama ühte ja ainult ühte lauset, kuid võib ulatuda mitmele tekstireale.
2. Pärast tärni * tuleb osaleja kood (ID), seejärel koolon ja tabulaator, nt

*RES_2:	kus	on	kolmas	klass	?

3. Põhirea jätkamine järgmisel tekstireal algab tabulaatoriga.
4. Laused lõpevad kas punkti (.), hüüumärgi (!) või küsimärgiga (?). Tavaliselt soovitatakse, et viimase sõna ja lauselõpumärgi vahele jäetaks tühik (LIPPS Group 2000: 151).
5. Suurtähtedega kirjutatakse välja ainult pärisnimed ja asesõna *I* 'mina'. Suurtähti ei kasutata ka lause esimese sõna algustähena.
6. Transkribeerimata keeleainest tuleb tähistada *www*-ga.
7. Arusaamatuks jäänud sõnu või väljendeid tähistab *xxx* või *xx* (kui ainult üks sõna on arusaamatu).
8. Ebaselge artikulatsiooniga sõnu saab märkida fonoloogilise vormina, võttes kasutusele ampersandi, nt *&a*.
9. Ebaselge artikulatsiooniga, kuid kontekstist aimatavaid sõnu saab täiendada, lisades sulgudes oletatava puuduvat osa, nt *pikapäeva(rühm)*.

2.2.3. Sõltread

Sõltread sisaldavad uurija kommentaare: tähelepanekuid, seletusi, kirjeldusi jms. Väga tähtis on see informatsioon välja tuua eraldi ridadele, põhiridadel osutub see mitteloetavaks. Nõudmised sõltridade kohta on järgmised (MacWhinney 1995: 8–9).

1. Sõltread kirjutatakse pärast transkribeeritud kõnevooru ning kommenteerivad või täpsustavad seda. Algavad alati sümboliga %.
2. Pärast sümbolit % kirjutatakse väikeste tähtedega kolmetäheline kood, siis koolon ja pärast tabulaatori vahet kirjutatakse juba informatsioon, nt

   %glo: today who is on duty

3. Sõltridade jätk algab järgmisel tekstireal tabulaatoriga.
4. Tähtis on, et sõltrea lõppu ei pandaks kirjavahemärke.

Pakutakse ka üldkasutatavaid võimalusi (vt tabel 4).

Tabel 4. Transkribeeritud faili sõltridade näiteid (LIPPS Group 2000: 152)

%add:	näitab lause adressaati
%glo:	lause glossimine sõna-sõnalt või morfeem-morfeemilt
%gpx:	mitteverbaalne informatsioon: noogutab pead
%mor:	sõnade morfoloogiline kirjeldus
%spa:	kõneakt, kõnetegu
%tra:	lause vaba tõlge (vene-eesti koodivahetuse korpuses tõlgitakse inglise keelde)
%com:	uurija kommentaarid

Kuna vene-eesti koodivahetuse korpuse loojaid huvitab koodivahetatud sõnade morfoloogiline integratsioon ja/või selle puudumine, mitteverbaalne info ei ole esmatähtis, siis otsustati %gpx rida kasutatada vaid vajaduse korral. Seevastu kommentaaride rida (%com) on väga tähtis. Iga näite juures esinevad ka %glo ja %tra read.

3. Sisestava koodivahetuse morfoloogiline analüüs

Järgmise alaosa põhiteema on sisestava (valdavalt ühesõnalise) koodivahetuse morfoloogia analüüsimine. Korpuses seletatakse seda kommentaaride real (%com). Korpuses on ka vahelduva koodivahetuse näiteid (vt lähemalt Zabrodskaja 2005: 58–61), kuid siin on keskendutud just sisestavale koodivahetusele.

Senini on jäänud lahtiseks küsimus, kas iga näidet peab glossima või ei pea. Kui tegu on aglutinatiivsete keeltega (näiteks türgi keel), siis soovitatakse morfeem-morfeemilist glossimist. Glossi eesmärk on kergendada korpusest arusaamist, eriti juhul, kui kasutaja ei oska kumbagi keelt (LIPPS Group 2000: 159). Nagu teada, on eesti keel aglutineeriv-flekteeriv ja vene keel flekteeriv. Et mõlema keele morfoloogia on arenenud ja keeruline, ning analüüsi seisukohalt on just morfoloogiline integratsioon väga tähtis, püütakse glossida kõiki lauseid.

Analüüsi põhimõisted on järgmised. Et kõige sagedamad koodivahetatud sõnaliigid on eesti nimisõnad, analüüsitakse nende kahel tasandil (või kahes järgus) loomulikus vestluses toimuvat integreerimist vene maatriksisse: soo määramine ja vastavate käändelõppude lisamine. Tihtilugu määravad koodivahetajad vaid sugu, kuid teatud positsioonides (näiteks 2. käändkonna elututel nimisõnadel ainsuse nominatiivis ja akusatiivis) pole see võimalik. Näiteks siis, kui puuduvad nimi-

sõnaga soos ja käändes ühilduvad omadus- või asesõnad ning soos ühilduvad verbi mineviku ainsuse vormid. Näidete kodeerimisel tuli välja, et %com rida sisaldab liiga pikki seletusi. Nende asemel on välja töötatud tabelis 5 esitatud süsteem, mida täiendatakse iga uue üksikjuhu kodeerimisvajadustest lähtuvalt.

Tabel 5. Integreerimata elementide tähistused

	Koodivahetatud sõnal:	**Tähistus %glo real**	**%com rida**
1	puudub morfoloogiline integratsioon vene maatriksisse, sest süntaktiline positsioon ei nõua seda (NOM=ACC=Ø)	Ø	Ø = *mingit selgitust ei ole*
2	puudub morfoloogiline integratsioon vene maatriksisse, kuigi potentsiaalselt on see võimalik	Ø+?	Ø+? = *antakse võimalik variant või variandid*
3	on morfoloogiline integratsioon nii soos kui ka käändes	Ø+	Ø+ = *selgitatakse vene morfoloogia seisukohalt*
4	on osaline morfoloogiline integratsioon vene maatriksisse (sugu +, kääne –)	Ø ½ G (***G****ender*)	Ø ½ S = *mingit selgitust ei ole*
5	on osaline morfoloogiline integratsioon vene maatriksisse (sugu –, kääne +)	Ø ½ C (***C****ase*)	Ø ½ C = *mingit selgitust ei ole*
6	on kakskeelse homofooni kuju: realisatsioon on eesti	HE (***H****omophone*, ***E****stonian*)	HE = *mingit selgitust ei ole*
7	on kakskeelse homofooni kuju: realisatsioon on vene	HR (***H****omophone*, ***R****ussian*)	HR = *mingit selgitust ei ole*
8	on kakskeelse homofooni kuju: realisatsioon on vahepealne	HB (***H****omophone, on* ***B****order line*)	HB = *näidatakse, mis keele foneetilisele kujule on konkreetne hääldatud sõna(osa) lähedane*

Kõik tabelis 5 toodud %glo reale lisatud märgendused tuli lisada täiendfaili ning teabefailis lahti mõtestada. Eestikeelsete nimisõnade (täieliku/osalise/potentsiaalse) morfoloogilise integratsiooni selgitamisel on vaja kasutada vene käändkondade süsteemi, näidata käändelõppe ja määrata ka sugu. Ühes täiendfailis selgitatakse lühendeid, millega tähistatakse korpuses vene kolme käändkonda. Lühidalt võib seda kirjeldada järgmiselt:

a) esimene käändkond – 1st D(eclension) C(lass),
b) teine käändkond – 2nd DC,
c) kolmas käändkond – 3rd DC.

Üksikasjalikult kirjeldatakse vene käänete süsteemi, tuues vene nimisõnu sisaldavaid näiteid. Märgitakse, et vene keeles on 6 käänet ja eesti keeles 14. Erilist tähelepanu pööratakse soo aspektile. Rõhutatakse, et erinevalt eesti keelest on vene keeles kolm sugu: nais-, mees- ja kesksugu, olemas on ka üldsugu (*umnica* 'tarkpea', *jabeda* 'kaebupunn' jne). Sedalaadi keelekirjeldused on vajalikud eriti neile uurijatele, kes tunnevad huvi vene-eesti koodivahetatud sõnade morfoloogia vastu, kuid ei oska kumbagi keelt.
Järgmisena kirjeldatakse ja kommenteeritakse korpuse valmisnäiteid. Näide (5) on kõnekas juhtum.

(5)
@Begin
@Participants: RES_2 Researcher_two, T1 teacher_one
@Sex of Participants: female, female
@Languages: Russian (1), Estonian (2)
@Location: A school in Narva with Russian as a language of instruction
@Date of boarding: 14-MAR-02
@Date of coding: 11-APR-06
@Transcriber: Verschik, Anna
@Coder: Zabrodskaja, Anastassia
@Warning: this is only a small sample

*RES_2: kus@2 on@2 kolmas@2 klass@2?
%tra: where is the third form
*T1: Nadja@1, otvedi@1 učitelnicu@1 k@1 õpetaja@2 Monika@2 .
%glo: Nad-ja take teacher to teacher-Ø+? Monika-Ø+?
%tra: Nadja, take the teacher to the teacher Monika
%com: _k õpetaja Monika_ 'to the teacher Monika' Ø+? = _k õpetaje Monik-e_ or _k õpetaj-e Monik-a_ (1st DC, DAT).
@End

Nagu näeme, on *õpetaja Monika* jagamatu sõnaühend, millel puudub morfoloogiline integratsioon. Näide (5) paneb mõtlema, kuidas tähistada morfoloogilise integratsiooni puudumist neil juhtudel, kui potentsiaalselt on see võimalik. *Monika* võib olla nii eesti kui ka vene nimi. Ent juhtum on keerulisem kui konsonandilõpuliste sõnade puhul: täishäälik *-a* kuulub eesti keeles tüvve (*Monik<u>a</u>*), vene keeles on see aga 1. käändkonna ainsuse nimetava lõpp (*Monik-<u>a</u>*). Selles konkreetses näites on õpetaja Monika eestlanna (lisaks on see sõnaühendi osa), mistõttu interpreteeritakse nime eesti sõnana. %com-real pakutakse võimalikud variandid, kus mõlemad või üks nimisõna oleksid integreeritud vene maatriksisse.

Näide (5) erineb näitest (6), kus integratsioon puudub, sest eesti nimisõna süntaktiline positsioon vene maatriksis seda ei nõua. Eesti *juust* sobib vene 2. käändkonda, kuhu kuuluvad meessoost null-lõpuga elutud nimisõnad, mis ei nõua akusatiivis mingit lõppu: *pokupali stol* 'ostsime laua' jne.

(6)
@Begin
@Participants: B1 buyer_one, B2 buyer_two
@Sex of Participants: female, male
@Languages: Russian (1), Estonian (2)
@Location: Jaama market
@Date of boarding: 12-SEP-03
@Date of coding: 10-AUG-06
@Transcriber: Verschik, Anna

@Coder: Zabrodskaja, Anastassia
@Warning: this is only a small sample

*B1: v@1 tot@1 raz@1 my@1 etot@1 juust@2 tebe@1 pokupali@1 ?
%glo: that time we this cheeseØACC youDAT buy 1 SG PAST
%tra: did we buy this cheese for you that time
%com: _juust_ 'cheese' Ø.
@End

Vahel võib eesti sõna olla integreeritud nii soos kui ka käändes. Võrdluseks vaatame näidet (7).

(7)
@Begin
@Participants: O official
@Sex of Participants: female
@Mother tongue of Participant: Russian
@Languages: Russian (1), Estonian (2)
@Location: Tourist service centre
@Date of boarding: JUL-02
@Date of coding: 11-APR-06
@Transcriber: Verschik, Anna
@Coder: Zabrodskaja, Anastassia
@Warning: this is only a small sample

*O: no@1 eto@1 bez@1 käibemaks@2a@1 .
%glo: but this without VAT-GEN-Ø+
%tra: but this is without VAT
%com: _bez käibemaksa_ 'without value added tax'
Ø+ = 2nd DC, GEN.
@End

Selle kõneleja idiolektis võib *käibemaks* olla nii täiesti aktsepteeritud ja integreeritud laen kui ka lihtsalt sisestav koodivahetus. Kas see on laen või koodivahetus, on võimatu otsustavalt öelda, sest põhimõttelist vormilist vahet ei ole, küsimus on ainult konventsionaliseerumise astmes. Kuigi mitmed formaallingvistiliselt orienteeritud koodivahetuse uurijad on püüdnud luua teoreetilist alust tüüpilise laenu ja tüüpilise koodivahetuse eristamiseks, pole see siiamaani õnnestunud. Seda enam, et on teisi uurijaid, kes arvavad, et see on põhimõtteliselt võimatu, pigem on tegemist kontiinumiga. Mõlema koolkonna esindajate püüdlustega võib tutvuda Anneli Sarhimaa (1999: 127–130) väitekirjas. Kuna sõna *käibemaks* fikseeriti mitmel kõnelejal, võib ettevaatlikult pakkuda, et koodivahetuse ~ laenu kontiinumil on see lähedane just viimasele. See sõna käändub vastavalt vene ükskeelse grammatika reeglitele: 2. käändkonna genitiivi lõpp on *-a*. Samas on võimalik varieerumine integreeritud ja mitteintegreeritud kuju vahel, sest ei saa ennustada, kas see sõna esineb integreerituna või integreerimata sama kõneleja kõnepruugis järgmisel päeval.

Näites (8), mis fikseeriti lindistatud telesaate "Sputnik" (06.04.2001) transkribeerimisel, esineb eesti sõna *linnavalitsus* vene lauses morfoloogiliselt integreerimata, kuigi see sobib vene 2. käändkonda (meessoost null-lõpuga nimisõna), mille instrumentaal (*s chem?* ehk *millega?*, eesti vaste on kaasaütlev) oleks *om*-lõpuga: *s linnavalitsus-om*.

(8)
@Begin
@Participants: H TV_show_host
@Languages: Russian (1), Estonian (2)
*H: konkurs@1 organizovan@1 Estonskim@1 tanceval´nym@1 agenstvom@1 sovmestno@1 s@1 linnavalitsus@2 .
%glo: competition hold-PS-PART Estonian Dance agency together with municipal administration-Ø+?
%tra: the competition was held by Estonian Dance agency together with the municipal administration
%com: _linnavalitsus_ 'municipal administration' Ø+? = _s linnavalitsus-om_ (2nd DC, INSTR) .
@End

Sisestava koodivahetuse piiripealsete nähtuste analüüsi eesmärk on tuvastada ja uurida morfoloogia ning prosoodia/fonoloogia piiril asuvaid nähtusi, mis võivad osutuda määravaks otsustamisel, mis keelde antud element kuulub. Tihtilugu on aga tegemist vahepealsete variantidega, mis ei luba elementi selgelt klassifitseerida. A. Verschik (2005) näitab, et mitmekeelses kõnes on selliseid asju küll ja küll. Siia kuulub eestipärane kahe rõhu säilitamine liitsõnades (kas *´infotehnol´oogia* või *infotehnol´o(o)gia*), aga ka vokaali pikkuse säilitamine (kas *jaanipäev* või *janipjaev*). Sellistel puhkudel sünnivad nn kompromissivormid, mis ei kuulu kummassegi ükskeelsesse varianti, seepärast tõstetakse neid eraldi esile ja kommenteeritakse %com real. Vaatame näidet (9).

(9)
@Begin
@Participants: S_1 Salesperson_one, S_2 Salesperson_two
@Sex of Participants: female, female
@Mother tongue of Participants: Russian, Estonian
@Languages: Russian (1), Estonian (2), Undecidable (0)
@Location: Railway Station market
@Date of boarding: 04-JUL-03
@Date of coding: 11-APR-06
@Transcriber: Verschik, Anna
@Coder: Zabrodskaja, Anastassia
@Warning: this is only a small sample

*S_1: Ja@1 na@1 kilogramm@1 popravilas´@1 (.) posle@1 janipäev@0 .
%glo: I one kilogramØ-ACC put on 3SG-PAST after Midsummer Day-HB-Ø+?

%tra: I put on one kilogram after Midsummer Day
%com: _janipäev_ 'Midsummer Day' HB = jani@1 + päev @2 _janipäev_ 'Midsummer Day' Ø+? = _posle janipäeva_ (GEN).
@End

Ülaltoodud näites (9) oli *jaanipäev* ühelt poolt morfoloogiliselt integreerimata laen, teiselt poolt võib seda juhtumit analüüsida ka foneetilisest küljest: pikka vokaali ei säilitata, hääldatakse venepäraselt. Sõnas *janipäev* on rõhk viimasel silbil, %com-real näitab seda viimase silbi allakriipsutamine. Kui oleks *janipjaiva*, siis oleks täielik foneetiline integratsioon. Kuid %com-rida ei tasu liiga pikaks teha, sest detailidega ülekoormatus häirib lugemist ja mõistmist. Kui interpreteerida koodivahetatud nimisõna *janipäev* vene grammatika terminites, siis käitub see 2. käändkonna meessoost elutut objekti tähistava nimisõnana, mille ainsuse genitiivi lõpp on *-a*: *posle janipäev-**a*** 'pärast jaanipäeva'.

Lõppsõna

Vene-eesti koodivahetuse korpuse huvid keskenduvad praegu peamiselt eesti üksuste (valdavalt nimisõnade) morfoloogilisele ja fonoloogilisele integratsioonile vene maatriksraamist vaadatult või selle puudumisele. Erilist tähelepanu pälvivad kompromissivormid ja uued konstruktsioonid, mida pole kummaski keeles.

Korpuses leidub hulk näiteid, mis nõuavad pikemat ja keerukamat analüüsi või mille puhul tekib mitu analüüsivõimalust. Näidete analüüs nõuab pidevalt uute märgenduste loomist. Seda illustreerib eriti ilmekalt artikli 3. peatükk. Kuna üheski valmiskorpuses pole neile adekvaatseid vasteid, püütakse märgendusprotsessi teha võimalikult lihtsaks, et neil, kes tahavad oma keeleainest vene-eesti omaga võrrelda, tekiks selline võimalus.

Tulevikus kavatsetakse pühenduda ka koodivahetatud sõnade foneetilisele analüüsile. Siis saab välja pakkuda eesti keelest vene keelde ümberasunud nimisõnade koodivahetuse ~ laenu kontiinumi.

Kirjandus

Backus, Ad 1996. Two in One. Bilingual Speech of Turkish Immigrants in The Netherlands. Studies in Multilingualism 1. Tilburg: Tilburg University Press.

Baškirova, Tatjana 2006. Koodivahetus eesti keele kui teise keele tunnis. Magistritöö. Tallinn: Tallinna Ülikool.

Bucholtz, Mary 2000. The politics of transcription. – Journal of Pragmatics 32 (8), 1439–1465.

Gardner-Chloros, Penelope; Moyer, Melissa; Sebba, Mark; van Hout, Roeland 1999. Towards standardizing and sharing bilingual data. – International Journal of Bilingualism 3, 395–424.

Hennoste, Tiit 2000a. Sissejuhatus suulisesse eesti keelde VI: lausung suulises kõnes I. – Akadeemia 10, 2221–2254.

Hennoste, Tiit 2000b. Sissejuhatus suulisesse eesti keelde VII: lausung suulises kõnes II. – Akadeemia 11, 2463–2486.

Hennoste, Tiit 2000c. Sissejuhatus suulisesse eesti keelde VIII: lausung suulises kõnes III: eneseparandused. – Akadeemia 12, 2687–2710.
Karlsson, Fred 2002. Üldkeeleteadus. Renate Pajusalu, Jüri Valge, Ilona Tragel (tõlkinud ja kohandanud). Tallinn: Eesti Keele Sihtasutus.
Konnapere, Indrek 2006. Vene-eesti koodivahetuse konversatsioonilised funktsioonid vene emakeelega ajateenijatel. Bakalaureusetöö. Tallinn: Tallinna Ülikool.
LIPPS Group, The 2000. The LIDES Coding Manual: A Document for Preparing and Analyzing Language Interaction Data. – International Journal of Bilingualism 4/2, 131–278.
MacWhinney, Brian 1995. The CHILDES Project: Tools for Analyzing Talk. 2nd edition. Hillsdale, NJ: Erlbaum.
Milroy, Lesley; Muysken, Pieter 1995. Introduction: code-switching and bilingualism research. – L. Milroy, P. Muysken (Eds.). One Speaker, Two Languages: Cross-Disciplinary Perspective on Code-Switching. Cambridge: Cambridge University Press, 1–14.
Myers-Scotton, Carol 1993. Social Motivations of Code-Switching. Oxford: Clarendon Press.
Ochs, Elinor 1979. Transcription as a theory. – Elinor Ochs, Bambi Schieffelin (Eds.). Developmental Pragmatics. New York: Academic Press, 43–72.
Ozernova, Jekaterina 2005. Eesti laenud venelaste kõnes. Bakalaureusetöö. Tallinn: Tallinna Ülikool.
Rannut, Mart 1994. Beyond linguistic policy: The Soviet Union versus Estonia. – Tove Skutnabb-Kangas, Robert Phillipson (Eds.). Linguistic Human Rights. Overcoming Linguistic Discrimination. Berlin, New York: Mouton de Gruyter, 179–208.
Sarhimaa, Anneli 1999. Syntactic transfer, contact-induced change, and the evolution of bilingual mixed codes. Focus on Karelian-Russian language alternation. Studia Fennica Linguistica 9. Helsinki: Finnish Literature Society.
Zabrodskaja, Anastassia 2005. Vene-eesti koodivahetus Kohtla-Järve vene emakeelega algkoolilastel. Tallinna Ülikooli eesti filoloogia osakonna toimetised 6. Tallinn: Tallinna Ülikooli Kirjastus.
Zabrodskaja, Anastassia 2006. Eestivene keel(evariant): kas samm segakoodi poole? – Keel ja Kirjandus 9, 736–750.
Verschik, Anna 2001. Interferentsi mehhanismidest ja vene-eesti kontaktidest. – Keel ja Kirjandus 8, 529–542.
Verschik, Anna 2005. Russian-Estonian language contacts, linguistic creativity, and convergence: New rules in the making. – Multilingua 24, 413–429.
Verschik, Anna; Zabrodskaja, Anastassia 2006. The Russian-Estonian LIDES corpus: Elaboration of morphological encoding principles. – Abstract book. Sociolinguistics Symposium 16 (6–8 July 2006). Limerick: University of Limerick, 39–40.
Viikberg, Jüri 1999. Eesti rahvaste raamat. Rahvusvähemused, -rühmad ja -killud. Tallinn: Eesti Entsüklopeediakirjastus.
Vseviov, David 2002. Kirde-Eesti urbaanse anomaalia kujunemine ning struktuur pärast Teist maailmasõda. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid 8. Tallinn: TPÜ Kirjastus.

Anastassia Zabrodskaja (Tallinna Ülikool) on uurinud keelekümblust, laste kakskeelsust ja vene-eesti koodivahetust. Viimasel ajal on peamiseks huviobjektiks vene-eesti koodivahetuse korpus.
anastaza@tlu.ee

RUSSIAN-ESTONIAN CODE-SWITCHING CORPUS: ELABORATION OF ENCODING PRINCIPLES

Anastassia Zabrodskaja
Tallinn University

The paper has several aims: 1) to introduce the goals of the LIPPS group and the Russian-Estonian code-switching corpus (LIDES) in the Estonian context, 2) to give an overview of the Russian-Estonian code-switching corpus with its sub-corpora (in preparation at Tallinn University), 3) to make an overview of the standards used to transcribe and encode multilingual data in the LIDES database, and 4) to formulate some principles of morphological encoding.

Several sub-corpora are planned within the corpus: (a) bilingual TV talk shows; (b) data from bilingual Tallinn; (c) data from the predominantly Russian-speaking North East (Narva and Kohtla-Järve).

The encoding of Russian-Estonian code-switching probably requires a special approach: Russian is written with Cyrillic letterswhereas Estonian uses the Roman script. The different alphabets may lead to a different treatment of Estonian elements in writing and in oral communication.

Both Estonian and Russian have a developed inflectional morphology. Full integration of nouns means gender assignment and adding of inflectional morphology (case, number, or case and number). Empirical observations show that full morphological integration of an Estonian single noun into the Russian matrix is not always the case. The authors of the corpus are interested in instances where Russian inflectional morphology is absent, although the noun fits structurally into Russian declension classes.

The focus is also on items whose belonging to either language is not clear. If a speaker speaks Estonian with a Russian accent, common internationalisms as well as Estonian proper names are ambiguous. Retention of two stresses in Estonian compound nouns in the Russian matrix is one of the relevant features to be encoded.

As far as morphology in the Russian-Estonian LIDES Corpus is discussed in the second part of the article, here primary attention is given to: 1) morphological and phonic integration or lack thereof, and 2) compromise forms, new creations.

It is necessary to introduce a special encoding system for in-between items and think of a way of encoding lack of integration in order to distinguish it from zero-endings. Numerous relevant examples are presented in the paper.

Keywords: corpus linguistics, code-switching, Estonian, Russian