SAATEKS

Käesolev artiklikogumik põhineb ettekannetel, mis peeti 2006. aasta sügisseminaril „Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine“. Seminari idee sündis kolme projektiga tegelevatel uurijatel: „Koodivahetuse, eesti vahekeele ning lastekeele andmekorpuste koostamine ja üldkirjeldus“ (2005–2008, ETFi grant nr 6151), „Koodivahetuse, vahe- ja lastekeele korpuste töötlemine ja haldamine” (2005–2008, riiklik programm „Eesti keel ja rahvuslik mälu”, grant R05/01) ning „Eesti keelekeskkonna arengu analüüs, modelleerimine ja juhtimine“ (2003–2007, sihtfinantseeritav teema nr 0132493s03). Olles tegelnud vahekeele, lapsekeele ja koodivahetuse korpuste koostamisprintsiipide väljatöötamise, keeleainese sisestamise ja korpustepõhiste uurimustega, tundus töörühma liikmetele, et on paras aeg vestelda kolleegidega saavutustest, probleemidest ja plaanidest ning ühtlasi teha väike vahekokkuvõte.

Kogenud uurijate kõrval osalesid seminaril ka algajad – magistrandid ja doktorandid. Et mitmesuguste keelekorpuste koostamine on suuremal või vähemal määral seotud infotehnoloogia arenguga, siis võtsid seminaril sõna mõlema valdkonna asjatundjad. Kogumikus on esindatud vaid keeleinimeste artiklid.

Kuna kõik kolm valmivat korpust on olemuselt erinevad, on erinevad ka keeletehnoloogilised probleemid.

Koodivahetuse korpuses kasutatakse CHILDESi võrgustikku. Kui mõned üldpõhimõtted välja arvata, siis pole kakskeelse kõne kodeerimine ja märgendamine eriti formaliseeritud ning erinevates koodivahetuse korpustes esinevad väga erinevad süsteemid. Vähene standardiseeritus on tingitud sellest,

et uurijate teoreetilised lähtekohad ja huvid ei kattu. Näiteks need, kes tegelevad eelkõige koodivahetuse pragmaatikaga, ei pruugi pöörata erilist tähelepanu morfoloogilise info kodeerimisele. Kui uurija on veendunud, et kakskeelne kõne pole kahe ükskeelse kõne summa ja et seetõttu mõningaid keelendeid ei olegi võimalik liigitada, tuleb mõelda n-ö ambivalentsete elementide tähistamisest, kusjuures ambivalentsus võib olla erinevat liiki.

Lapsekeele korpuse töötlemise aluseks on samuti CHILDES, mille põhjalik ja võimalusterohke märgendussüsteem jääb aga eesti keele morfoloogia kirjeldamisel hätta. Lapsekeele korpuse koostamist ja sellega töötamist kergendaks eriline, just selleks otstarbeks mugandatud morfoanalüsaator.

Eesti vahekeele korpuse vealeidja loomine sõltub ilmselt kõige rohkem originaalsetest keeletehnoloogilistest lahendustest; korpuse kasutamise alus on täpne, aga samas paindlik vealiklassifikatsioon.

Suurem osa kogumiku artiklitest on kirjutatud Eesti vahekeele korpuse (EVKK) teemadel. EVKK töörühm tegutseb Pille Esloni juhendamisel; Mare Kitsnik on teinud korpuse põhjal uurimuse eituse vigadest ning ta kaitses sel teemal magistritöö 2007. aasta kevadel. Peatselt valmib Anastassia Šmõreitšiku magistritöö. EVKK koostamisel ja uurimisel osaleb ka Tartu Ülikooli doktorant Helena Metslang. Lugeja leiab kogumikust kõigi nimetatute artiklid.

Pille Eslon käsitleb õppijakeelekorpusi üldisemalt. Autor näitab, missugune koht kuulub nende seas Eesti vahekeele korpusele. Ta peatub EVKK võimalikel rakendustel e-keeleõppes ja teaduslikus uurimistöös. Lähemalt tutvustatakse korpuse märgendamist ning sõna- ja vormisagedusmooduli kasutamist.

Anastassia Šmõreitšik keskendub meetoditel, mille abil võib leida korrelatsiooni erinevate vealiikide vahel. See on oluline, kuna mõnda tüüpi vead põhjustavad teisi vigu. Samas leidub korpuses 57% ± 5% vigu, mida igal juhul interpreteeritakse mitmeti. Nendevaheliste seoste väljaselgitamisel on ilmselt õigem hakata otsima mustreid (*patterns*).

Mare Kitsniku artikkel annab EVKK-põhise veaanalüüsi näite – süstemaatilise ülevaate eituse korrektsest ja vigasest kasutusest eesti õppijakeeles. Analüüsi tulemusel osutus võimalikuks välja selgitada, millised eituse aspektid on õpetamise seisukohalt olulised ja vajavad õpikutesse sisseviimist, senisest rohkemat tähelepanu ja erinevaid harjutusi.

Helena Metslang kirjeldab õppijakeele korpuspõhise vea- ja kontrastiivanalüüsi võimalusi ning näitab, kuidas neid analüüsimeetodeid rakendada EVKK alusel. Korpus lubab läbi viia nii kvalitatiivseid kui kvantitatiivseid uuringuid, aidates leida nähtusi, mis muidu märkamatuks jäävad. Samuti saab korpuse põhjal kontrollida traditsiooniliste tõekspidamiste paikapidavust eesti õppijakeele kohta.

Reili Argus keskendub vigade transkribeerimis- ja kodeerimisprobleemidele ning näitab, miks keeleandmed peaksid tingimata olema hõlpsasti automaatselt töödeldavad. Käesolev uurimus on osa lähiajal kaitsmisele minevast doktoritööst.

Vene-eesti koodivahetuse korpusega tegeleb Anastassia Zabrodskaja. Oma artiklis kirjutab ta kontaktidest tulenevast keelemuutusest, mida on võimalik jälgida vene-eesti korpuse põhjal. Korpus lubab empiiriliselt tõestada, et koodivahetus ja konvergents käivad käsikäes. Ka see artikkel on osa valmivast doktoriväitekirjast.

2007. aastal on plaanis jätkata sügisseminari ilusat tava. Loodetavasti saame lugejale lähitulevikus pakkuda ka järgmise seminari kogumiku.

Anna Verschik

Sügisseminari „Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine“ ajakava

Aeg: 20. oktoober 2006
Koht: Senatiruum U-648
Ettekandeks ja aruteluks kokku 20 minutit (15 + 5)

10.00–10.20
Avasõna
Martin Ehala, Anna Verschik, Mart Laanpere

Vene – eesti koodivahetuse korpus (juh. Mart Laanpere)

10.20–10.40
Anastassia Zabrodskaja. Vene-eesti koodivahetuse korpuse piiridest