Väitekiri sõnastiku õppijasõbralikust näitelausest

> Kristina Koppel. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. (Dissertationes linguisticae Universitatis Tartuensis 38.) Tartu: Tartu Ülikooli Kirjastus, 2020. 248 lk.

Kristina Koppeli väitekiri käsitleb (õppe)sõnastike näitelausete automaatse tuvastamise võimalusi, täpsemalt on uurimisobjektiks sõnastiku hea näitelause tunnused. Nüüdisajal, kui järjest enam on tähelepanu keskmes keeleõppijate autonoomsus ning individuaalne eripära ja keeleoskuse arenemise kulg, muutub aina olulisemaks õpet toetavate vahendite, muu hulgas elektrooniliste sõnastike ja keeleõppekeskkondade mitmekesisus, kvaliteet ja õppijasõbralikkus. Väitekirja teema on seega eesti e-leksikograafias mitmes mõttes vajalik ja uudne. Näitelausete automaatset tuvastamist on siiani suhteliselt vähe kasutatud ning selle ala uurimistöö on alles algusjärgus. Samal ajal on väitekirja teema oluline ka rahvusvaheliselt: hea näitelause parameetrite hulgas on nii universaalseid kui ka keelespetsiifilisi tunnuseid. Seos rahvusvahelise uurimistööga on kahtlemata üks väitekirja tugevusi.

Väitekirja moodustavad kokkuvõttev osa ja aastatel 2016–2019 avaldatud viis artiklit. Eestikeelsetest artiklitest on kaks ilmunud Eesti Rakenduslingvistika Ühingu aastaraamatus ja kaks ajakirjas Lähivõrdlusi. Lähivertailuja; ingliskeelne viie autori ühisartikkel on publitseeritud ajakirjas International Journal of Lexicography.

Kokkuvõte koosneb kuuest peatükist ning lisadest, mis sisaldavad konfiguratsioonifaile, musta ja halli nimekirja. Mustas nimekirjas on loetletud sõnu, mis on näitelauses keelatud, nt vulgarismid, hallis nimekirjas aga sõnu, mille eest vähendatakse näitelause kandidaadi üldskoori, nt

kõnekeelsed sõnad. Sissejuhatusele eelneb väitekirja kesksete mõistete ja lühendite loend, mis toetab lugejat teksti süvenemisel. Väitekiri kuulub autori määratluse järgi kahe uue leksikograafia haru, korpusleksikograafia ja automaatse leksikograafia valdkonda. Koppel on asetanud väitekirjale nii teoreetilised kui ka rakenduslikud eesmärgid, mis uurimuse tugevat rakenduslikku suundumust arvestades on kindlasti põhjendatud.

Teooriapeatükis tutvustab Koppel sõnastiku näitelausete tüpoloogiat, funktsioone, valiku põhimõtteid ja hea näitelause tunnuseid, toetudes peamiselt Bo Svenséni ning Sue Atkinsi ja Michael Rundelli tüpoloogiatele. Väitekirja keskmes on korpuspäringusüsteemi Sketch Engine'i integreeritud tööriist GDEX ehk Good Dictionary Examples, mille eesti mooduli arendamiseks on Koppel analüüsinud eesti keele (õppe)sõnastike ja eesti keele kui teise keele õpikute lausete parameetreid ning loonud analüüsi tulemusi arvestades eesti moodulile kuus erinevat versiooni. Mooduli arendamisel on Koppel õnnestunult kombineerinud reeglipõhist ja masinõppe meetodit, lisaks on moodulit pilootuurimusena hinnatud vastava ülesande abil, mis annab väärtuslikku teavet edasiseks tööks.

GDEX-i eri keelte (sh eesti keele) moodulite arendamise oluline tulemus on hea näitelause universaalsete ja keelespetsiifiliste parameetrite väljaselgitamine. Parameetreid on Koppel käsitlenud kahes artiklis ja kokkuvõtva osa tabelis 1 (lk 31–32). Esimeses, koos Jelena Kallasega kirjutatud artiklis „Õppijasõbralik korpuslause: automaatse valiku võimalusi" [P1] on öeldud, et teatud kriteeriumid (lause peab olema täislause, lauses ei esine pronoomeneid ega anafoore) on keeleülesed, teatud keelespetsiifilised (lause pikkus, märksõna asukoht lauses). Teises artiklis „Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks" [P3] on lühidalt välja toodud mooduli arendamise käigus lisatud eesti keele hea näitelause parameetrid. Põhjalikum kokkuvõttev analüüs sellest, millised parameetrid on universaalsed, millised keelespetsiifilised ning millised on eesti keele spetsiifilised, väitekirjas siiski puudub.

Väitekirja artiklid edenevad loogiliselt ja peegeldavad kujukalt autori arengut. Teemat sissejuhatav ühisartikkel [P1], mille põhiautor on Koppel, annab ülevaate autentsete korpuslausete kasutusvõimalustest õppeleksikograafias ja keeleõppes ning kirjeldab meetodeid, mis võimaldavad õppijasõbralike korpuslausete automaatset valikut. Artikkel on referatiivne, kuid teemapüstitusega hästi põhjendatav ja täidab sellisena oma eesmärgi, selgitades valdkonna hetkeseisu ja tutvustades lugejale uusi keeletehnoloogilisi vahendeid. Väitekirja artiklitest on see ainus, milles käsitletakse riivamisi korpuste kasutamist keeleõppes. Selle sissejuhatuses on öeldud: „Õppijad teevad korpusmaterjaliga töötades ise keele kohta järeldusi. Laused, mis sisaldavad vihjeid konteksti kohta, aitavad mõista uute sõnade tähendust, ning laused, mis sisaldavad kollokatsioone ja esindavad süntaktilisi mustreid, aitavad ennetada vigu, mida teist keelt õppides tüüpiliselt tehakse." Lugejat huvitanuks, millisena näeb Koppel keeleõppeprotsessi ja milline arusaam keele omandamisest on tema arvates korpuspõhise keeleõppe taustal. Artiklis keskendutakse siiski vaid õppeleksikograafiale, mille ülesanne on kahtlemata kaudselt keeleõppe toetamine ja selleks eelduste loomine, kuid mis ei kuulu otseselt keeleõppe valdkonda. Seega ei saa artikli puhul rääkida erinevalt autorite väidetust autentsete näitelausete kasutusvõimaluste analüüsist keeleõppes, küll aga on õnnestunud autentsete näitelausete kasutusvõimaluste analüüs õppeleksikograafias. Oluline on järeldus, et näite-

lausete automaatse tuvastamise tulemused on ainult nii head, kui hea on korpus, st kui hästi on korpus tasakaalustatud.

Ingliskeelne ühisartikkel „Identification and automatic extraction of good dictionary examples: The case(s) of GDEX" [P2] asetab reeglipõhisel valemil töötava Sketch Engine'i tööriista GDEX eesti mooduli arendamise rahvusvahelisse konteksti. Artiklis arutletakse hea näitelause tunnuste üle ning antakse ülevaade automaatselt tuvastatud näitelauseid kasutavatest leksikograafia- ja keeleõppeprojektidest. Tulenevalt eesmärgist on suur osa ka sellest artiklist tutvustav-kirjeldav. Väitekirja seisukohalt olulist uurimuslikku lisa pakuvad nelja keele, sh eesti keele hea näitelause keelespetsiifilised parameetrid (tabel 1).

Väitekirja kolmandas artiklis [P3] keskendub Koppel GDEX-i eesti mooduli versioon 1.4 arendamisele „Eesti keele naabersõnade sõnastiku" andmebaasi näitelausete põhjal. Lugejale tutvustatakse GDEX-i tööpõhimõtteid ning kirjeldatakse lühidalt eesti keele GDEX-i vanemaid versioone 1.2 ja 1.3. Teema avamine sellega mitte kursis olevale lugejale on üldiselt hästi õnnestunud. Artikli uurimuslik osa on väitekirja seisukohalt keskne: uue konfiguratsiooni arendamisel testiti põhjalikult klassifikaatoreid, mida GDEX näitelausete tuvastamiseks kasutab, ning selgitati välja näitelausete tugevad ja nõrgad parameetrid. Arendusprotsessi ja selle etappe on artiklis üksikasjalikult ja arusaadavalt käsitletud.

Neljanda artikli „Leksikograafide ja keeleõppijate hinnangud automaatselt tuvastatud korpuslausete sobivusele õppesõnastiku näitelauseks" [P4] eesmärk on hinnata eelnevat uurimistööd, selgitades välja, kas automaatselt valitud autentsed korpuslaused sobivad leksikograafide ja keeleõppijate hinnangul eesti keele B2–C1-keeleoskustaseme õppesõnastiku näitelauseteks. Hüpoteeside tõestamiseks valiti „Eesti keele naabersõnade sõnastiku" andmebaasist 40 juhuslikku märksõna. Artiklis tutvustatud uurimus on esimesi, milles kasutatakse rahvahanke (ka rahvateaduse) põhimõtet. Seda uurimust tasub siiski võtta prooviuuringuna, sest eriti leksikograafide valim on üldistuste tegemiseks liiga väike. Kuna küsitluses osales 14 hindajat (viis leksikograafi ja üheksa keeleõppijat) ning igale lausele andis hinnangu kümme inimest, siis järeldub, et kõik keeleõppijad ei hinnanud kõiki lauseid. Artiklist ei selgu, kuivõrd võisid individuaalsed erinevused mõjutada hindamistulemusi. Küsimusi tekitab, kas osalenud leksikograafide seas oli ka sõnastiku koostajaid ning kas nad olid seetõttu teadlikud sellest, millised hindamisülesande laused on pärit sõnastikust ja millised korpusest. Need tegurid võisid samuti uurimistulemusi mõjutada. Selgemalt oleks võinud välja tuua seosed näitelausete tunnuste ja nende sobilikuks/mittesobilikuks hindamise vahel.

Väitekirja viimases artiklis „Eesti keele kui teise keele õpikute lausete analüüs ja selle rakendamine eri keeleoskustasemete sõnastike näitelausete automaatsel valikul" [P5] analüüsitakse eesti keele kui teise keele õpikute lausete parameetreid. Näitelausete parameetreid on küll analüüsitud keeleoskuse alltasemete kaupa (A1, A2, B1, B2, C1), kuid GDEX-i eesti mooduli versioonid on loodud üldistele keeleoskustasemetele: versioon etBasic-v1 A-tasemele, versioon etIndependent-v1 B-tasemele ja versioon etProficient-v1 C-tasemele. Selline lahendus on ehk esimese etapina põhjendatud A- ja C-taseme puhul, kuid B-alltasemete suure erinevuse tõttu väga küsitav. Tehtud otsus vajaks seetõttu põhjendamist, eriti arvestades, et väitekirja kokkuvõtva osa sissejuhatuses nimetab autor õppesõnastike sihtgrupina peamiselt B2–C1-keeleoskustasemel keeleõppijat. Õpikulaused on

õppesõnastiku lausete alusena kahtlemata üks võimalus, kuid kindlasti mitte ainus. Lugejal olnuks huvitav teada, kas kaaluti ka muid võimalusi ning kui jah, siis millised võimalused olid veel kõne all ning miks otsustati just õpikute korpuse kasuks.

Üks keskne arutlemist vääriv teema selles väitekirjas on terminoloogia. Kuna tegemist on üsna uudse valdkonnaga, põhineb terminikasutus paljuski (toor)tõlgetel ega ole veel kuigivõrd kinnistunud ning võib alaga vähem kursis olevas lugejas tekitada hulgaliselt küsimusi. Seetõttu oleks olnud põhjust veelgi selgemalt välja tuua ühelt poolt õppekorpuse ja õppesõnastiku omavaheline seos, aga ka erinevused, teisalt seletada õppekorpuse erinevust õppijakorpusest. Samuti tekib küsimus, kas korpuspäringusüsteemi lihtsustatud versiooni etSkELL-i ehk Sketch Engine for Estonian Language Learning on ikka põhjendatud nimetada automaatseks keeleõppekeskkonnaks või keeleõppeportaaliks või hoopis keeleõpperakenduseks või kasutajaliideseks. Kui selle abil saab lugeda näiteid, vaadata sõnavisandeid (sh naabersõnu ehk kollokatsioone) ja sarnaseid sõnu ehk tesaurust, siis on tegemist pigem e-õppesõnastikuga. Automaatne keeleõppekeskkond peaks võimaldama oma keeleoskuse hindamist, harjutamist, tagasisidet jne. Vähe läbinähtavad on ilma täiendava seletuseta ka terminid *sõnavisand* ja *kroolimine*. Tähendab ju esimene üldkeeles üldjoonelist kavandit, teine aga seostub keskmisel eestlasel esmajoones ujumisega. Sobivate ja läbipaistvate uute terminite kasutuselevõtt ja kinnistumine võtavad mõistagi aega.

Kristina Koppeli väitekiri on oluline uurimus e-leksikograafia alal, millel on esmajoones rakenduslik, kuid ka teoreetiline väärtus. Väitekirja autori uurimistööle seatud eesmärgid saavad enamasti täidetud. Uurimistulemused pakuvad näitelausete automaattuvastuse edasiarendamiseks mitmeid võimalusi, mida väitekirja autor kokkuvõttes ka põhjalikult analüüsib. Väitekiri annab seega arvestatava panuse e-leksikograafia arendamisse Eestis.

ANNEKATRIN KAIVAPALU