

TÄHELEPANEKUID KONVERENTSILT „ELEX 2009”

Ülevaateartikkel eelmisest suuremast sõnastikurahva kogunemisest (Euralex 2008) lõppes sõnadega: „Järgmisel korral kogunevad Euroopa leksikograafid 2010. aastal Friisimaal Leeuwardenis.”¹ Tegelikult kogunesid paljud leksikograafid juba varem: 22.–24. oktoobri 2009 toimus Belgia ühes tähtsamas ülikoolilinnas Louvain-la-Neuve’is konverents pealkirjaga „Elektroonilise leksikograafia uued ülesanded ja uued rakendused XXI sajandil”. Esinejaid oli üle kuuekümne, osalejaid kokku 30 riigist. Korraldajaks Louvain’i Katoliikliku Ülikooli juures tegutsev inglise korpuslingvistika töörühm. Lisaks sellele, et konverents oli suurepäraselt korraldatud, olid ka ettekanded väga põnevad. Kirjeldame mõnd huvitavamat.

E-sõnastikest üldiselt. Esimese plenaarettekande pidas Michael Rundell (Suurbritannia), kes andis ülevaate leksikograafia ja arvutite seoste oluliseimatest arengutest alates 1960. aastatest kuni tänapäevani. Arvutite kanda jäetakse üha rohkem ülesandeid. Inimtöö panus on leksikograafias viimasel ajal oluliselt vähenenud: näiteks, kui varem tuli näitelauseid käsitsi sedeldada, siis nüüd on võimalik neid peaaegu automaatselt tohtu suurtest korpus-test SketchEngine’i ja GDEX-i tarkvara

abil välja võtta ja sõnastikukirjetesse lisada. Leksikograafia on muutumas sedavõrd arvutipõhiseks, et ei ole kindel, kas kümne aasta pärast enam ongi leksikograafe ja üldse sõnastikke. Kui ei, siis kas miski muu võiks neid asendada, ja mis see „muu” võiks olla? Kui jah, siis igal juhul erinevad sõnastikutegijate ülesanded oluliselt praegustest ja sõnastikud on täiesti uut tüüpi sisuga.

Ulrich Heid (Saksamaa, Stuttgart) kõneles sõnastikumudelitest. Praegused e-sõnastikud kipuvad üldjuhul kopeerima pabersõnastikke, s.t on lineaarsed ja neis esitatakse vaid üht tüüpi infot, mis on kasutajale korraga nähtav. E-keskkonna võimalused on siin suurel määral ära kasutamata: näiteks võiks kasutajal olla võimalik kuvada eri infot vastavalt sellele, milliseid andmeid ta parasjagu vajab (verbi-konstruktsioone, morfoloogiat vmt). Sellist sõnastiku ülesehitust toetab näiteks sealsamas Stuttgardis koostatav andmemudel OWL-DL, mis keskendub just eri tüüpi infokihtide eristamisele. Oluline on kindlaks teha, mis andmeid täpselt (ja kuidas) kasutajale anda.

Sama liini jätkas Dennis Spohr (Stuttgart), kes visandas multifunktsionaalse sõnastikutüübi, mida iseloomustab modulaarne ülesehitus. See teeb võimalikuks ühest elektroonilisest andmebaasist moodustada mitmesuguseid erinevaid sõnastikke vastavalt kasutaja vajadustele.

¹ M. Jürviste, M. Langemets, M. Tiits, T. Valdre, P. Voll, XIII leksikograafiakongress Barcelonas. – Keel ja Kirjandus 2009, nr 2, lk 157.

Piek Vossen (Holland) esitas põgusa ülevaate Wordneti ajaloost kuni tänapäevani. Varasemate üksikprojektide asemel on keskendutud EuroWordNeti loomisele (eri keelte Wordnetid on inglise keele vahendusel omavahel seotud), mille järgmiseks sammuks on GlobalWordnet Grid: selles oleksid omavahel seotud keeled üle kogu maailma, aga ilma inglise keele vahenduse ta, tuginedes formaalsele andmeskeemile. Mõistete tähendusi on selles skeemis võimalik kirjeldada loogikaaksioomidena. GlobalWordneti esimeseks projektiks on KYOTO, mis töötab sisuliselt nn Wiki-põhimõttel, eri valdkondades pädevad kasutajad saavad andmeid töödelda, lähtudes automaatselt etteantud terminivalikust (see saadakse valdkonna dokumentide arvutianalüüsil).

Sõnastike kasutamisest. Anna Dziemianko (Poola) vaatles, mis mõju on sellel, kas õpilased kasutavad traditsioonilist pabersõnastikku või elektroonilist arvutisõnastikku. Selgus, et e-sõnastike kasutajad läbisid testi märkimisväärselt paremate tulemustega (seda nii teksti mõistmise kui ka tekstiloomes ülesandes). A. Dziemianko arvates on e-sõnastike põhiline eelis siiski just vormis, kuidas infot esitatakse (arvuti juhib vähem tähelepanu ebaolulistele asjadele), mitte niivõrd sõnastiku kasutamise kiiruses.

Robert Lew (Poola) kõneles sõnastikukirje struktuurimenüüdest, mida kasutatakse pikemate artiklite juures mõnikord ka traditsioonilistes pabersõnastikes, rohkem aga e-sõnastikes. Tema uuringust selgub, et selliste struktuurimenüüde kasutamine (kakskeelsetes) e-sõnastikes on väga kasulik info leidmise kiiruse ja täpsuse seisukohalt, aga eeskätt siis, kui lisaks struktuurimenüüle markeeritakse põhikirjes selgelt eristuva värviga menüüst valitud tähendusloki info.

Tadeusz Piotrowski (Poola, Wrocław) rääkis oma kogemusest mobiilisõnasti-

kega, mis on kasutusel taskuarvuti funktsioonidega mobiiltelefonis. Nende puhul on eriti tähtis, et info oleks lihtsasti haaratav, kiiresti ligipääsetav ja selge esitusega. Seetõttu on mobiilisõnastikud eriti sobivad kakskeelsete sõnastike lühendatud versioonide jaoks, kus esitatakse ainult kõige olulisem info (selliste seadmete ekraan on tavaliselt väike). Tulevikus võiks olla võimalik teha suvalisest tekstist pilt mobiiltelefoni kaameraga, kasutada tekstituvastusprogrammi ja seejärel tõlkida automaatselt kas kogu tekst või üksiksõnu sõnastikutarkvara abil. Firma ABBYY ongi asunud seda lahendust välja töötama.

Isabel Duran-Muñoz (Hispaania, Málaga) vaatles, missugused on sõnastike kasutamisel kutseliste tõlkijate vajadused, tuginedes nende seas korraldatud küsitlusele. Selgus, et tõlkimisel peetakse kõige vajalikumaks *online*-sõnastikke (56%), sagedamini kakskeelseid sõnastikke (19%), aga mõneti ka Wikipediat ja Google'it. Kõige tähtsamaks kriteeriumiks peetakse sõnastiku autorit ja eriala, kõige tähtsusetumaks kasutusjuhendeid (see, et sõnastike kasutajad ei loe kasutusjuhendeid, on leksikograafias juba peaaegu käibetõeks saanud). I. Duran-Muñozi hinnangul võiks e-sõnastikes olla võimalus valida, millist infot soovitakse kuvada (näiteks semantilist infot ei ole tõlkijatel sageli vaja ja see muudab sõnastikukirje tarbetult mahukaks).

Hillary Nesi (Suurbritannia) käsitles oma kõnes e-sõnastike keeleõppes kasutamise negatiivseid külgi. Ida-Aasias on väikesed elektroonilised taskusõnastikud saanud (üli)õpilaste seas väga populaarseks ja neid kasutatakse väga aktiivselt. Õpetajad (õppejõud) suhtuvad neisse siiski väga skeptiliselt, sest sageli on nende sõnastike kvaliteet küsitava väärtusega. Kui trükis ja isegi CD-l avaldatud sõnastikel on teatav kvaliteedigarantii (neid on toimetatud),

siis muudel e-sõnastikel on tihti palju kummalisi vigu. Teine skepsise allikas on see, et õppejõudude arvates ei peaks õppimine olema liiga lihtne: uute sõnade õppimiseks peaks vaeva nägema, et see ka päriselt tulemuslik oleks. Kõige selle juures tuleks siiski arvestada, et e-sõnastikud pakuvad hulganisti uusi õppimisvõimalusi. Õpetajad võiksid innustada õpilasi rohkem rääkima sellest, milliseid sõnastikke nad kasutavad (seda peetakse tihti eraasjaks, millest teistega ei räägita). Tuleks välja juurida mentaliteet, et elektrooniliste taskusõnastike kasutamine on poolelnisti illegaalne. Ka kirjastajad võiksid sellele kaasa aidata, tutvustades selliseid sõnastikke senisest rohkem.

Multifunktsionaalsetest e-keskkondadest. Toma Tasovac (Serbia) tutvustas veebiportaali Transpoetika, semantilise sõnaklassi kesket serbiainglise õppesõnastikku, mis vastandub tavapärasele leksikotsentristlikule (sõna-kesksele) sõnastikule: keel tugineb tema meelet (ja muuhulgas ka N. Švedova arvates, kellele Tasovac oma ettekandes viitab) just semantilistele sõnaklassidele, mitte sõnadele. Seetõttu on Transpoetika üles ehitatud Wordneti põhimõttel: kasutaja saab peale põhikirje uurida märksõnaga seotud teisi sõnu (sünonüümia, antonüümia, hüpero- ja hüponüümia ja muudes suhetes) ning pääseda neile ligi erinevatest digiteeritud tekstidest. Nagu Ulrich Heid, Dennis Spohr ja Isabel Duran-Muñoz, rõhutas ka T. Tasovac, et sõnastik ei tohiks olla ühekihiline, vaid peaks arvestama erinevate semantiliste seostega, mille põhjal on võimalik sõnastikku mitmeti struktureerida. E-sõnastik ei tohiks olla pelgalt pabersõnastiku digitaliseeritud versioon, see peaks olema pigem *online*-teenus.

Ka Serge Verlinde (Belgia) rõhutas, kui tähtis on, et praegusel e-ajastul ei jäädaks enam traditsioonilise sõnastikumudeli juurde. Tema osalusel koosta-

tud prantsuse keele õppekeskkond BLF (Base lexicale du français) püüab ajaga kaasas käia ja pakkuda keeleõppijale senisest mugavamalt õppekeskkonda. BLF-i veebilehel on kuus alajaotust, mis kõik vastavad õppija eri vajadustele: laiem info saamine sõna kohta, sõna tõlkevastete leidmine, sõnade kasutamise ja tõlkevastete kontrollimine, abi otsimine teksti mõistmisel ja tekstiloomes, harjutused eri teemadel. Niisiis on see keskkond, mis pakub kasutajatele eri vajadustest lähtuvalt eri sisuga leksikograafilist infot.

Lars Trap-Jensen, Henrik Lorentzen ja Jørg Asmussen (Taani) tutvustasid taani sõnastikuprojekti, mis kasutab erinevaid leksikaalseid ressursse: taani keele seletussõnastikku (1950–, 100 000 sõna), suurt taani keele sõnastikku (1700–1950, 250 000 sõna), tekstikorpust (56 milj), DanNet'i (taani WordNet). Üks päring annab vastuseks andmed korraga kõigist allikatest.

Uuemast ja juba tuntud tarkvarast. Konverentsil tutvustati ka mitmesugust leksikograafide tarbeks loodud sõnastikutarkvara. David Baines (Belgia) rääkis vabavarast Sil FieldWorks Language Explorer (FLEEx), mis on väga sobilik etnoloogilise sisuga leksikograafiliste projektide jaoks, sest võimaldab mugavalt analüüsida ka niisuguseid keeli, millel ei ole veel välja kujunenud ortograafiat. Ilan Kernerman (Iisrael) tõi näiteid sõnastikukeskkonnast KXD Shell, mis suudab suvalisi XML-formaadis kirjutatud sõnastikuandmeid mugandada sobivateks arvutiliidestest kasutatavateks sõnastikeks. David Joffe (Lõuna-Aafrika) reklaamis juba tuntud Tshwanelexi sõnastikusüsteemi (TLEEx): selle uute raketustena tõi ta välja näiteks sõnade järjestamise korpuseinfo alusel ning võimaluse eksportida sõnastikuandmed trükkimiseks sobivasse vormi (nt doc-formaati). Christos Tsalidis, Elena Mantzari ja Mavina Pantazara (Kreeka)

tutvustasid kreeka leksikograafilist toimetamissüsteemi, mis võimaldab lisada sõnaartiklisse grammatilist infot, muudab automaatselt tähendusnumbreid, kontrollib ristviiteid jne. Vera Kuzmina ja Anna Rõlova (Venemaa) esitlesid ABBYY Lingvo sõnastikke ja sõnastike koostamise tarkvara, mis on lubaduste kohaselt äärmiselt paindlik [kirjete koostamisel saab kasutada nn teksti lohistamist (*drag and drop*); leksikograafid saavad määrata väga palju endale sobilikke koostamiskriteeriume, märgendeid jms, kogu tööd on võimalik teha kaugmeetodil, serveripõhiselt]. FLEx, TLEx ja ABBYY lubavad kõik väga suurt paindlikkust. Iseasi, kas see ka tegelikus elus nii lihtne ja mugav on: eri sõnastikel on väga erinevad koostamispõhimõtted. Stendiettekandega (Margit Langemets, Andres Loopmann ja Ülle Viks) oli esindatud Eesti Keele Instituudis valminud sõnastikutarkvara EELex, mille avalik versioon on mõeldud tavakasutajale, kes tahab veebis koostada oma kakskeelset sõnastikku. Soovi korral (kokkuleppel EKI-ga) saab ta kasutada ka eesti-X-keele andmebaasi: kas märksõnade loendit või kogu sõnastiku eesti poolt.

Adam Kilgarriff (Suurbritannia) tutvustas veel üht töövahendit, mis leksikograafide elu lihtsamaks peaks tegema: selleks on SketchEngine, mis tuvastab etteantud korpusest automaatselt sõna tähtsamad kollokatsioonid erinevate kriteeriumide alusel. Sellelega on seotud omaette tarkvara GDEX, mis leiab SketchEngine'i andmete põhjal korpusest sõnastikele kõige sobivamad näited. Päris automaatselt see valik siiski toimuda ei saa, sest korpuse näitelauseid algkujul enamasti sõnastikunäiteks ei sobi, isegi kui korpus on väga suur. Aga häid ideid annab see leksikograafile kindlasti.

Üldpilt. Ettekannete temaatika oli avar, põhiliste valdkondadena võiks välja tuua järgmised:

- e-leksikograafia üldiselt: ajalugu ja perspektiivid, sõnastike mudelid;
- uued rakendused: mobiili- ja veebi-sõnastikud, õppe- ja õpikeskkonnad;
- sõnastike kasutamine ja kasutajagruppide vajadused;
- sõnastike mikrostruktuur: erinevad infokihid, sõnastikuülesed seosed;
- sõnastikes navigeerimine ja päringuvõtted: olulised on otsingu kiirus ja täpsus, vaatamise mugavus;
- sõnastike koostamise ja toimetamise vahendid;
- korpuste kasutamine: näidete saamiseks, sõna kasutamisinfo leidmiseks, neologismide, fraseologismide, terminite jne tuvastamiseks;
- keeletarkvara kasutamine: korpusest morfoloogilise, süntaktilise, semantilise info leidmine ja selle sõnastikus esitamine;
- olemasolevate (nn vanade) sõnastike arvutivormingusse teisendamine;
- konkreetsed probleemid: esiplaanil olid mitmesõnalised väljendid, kollokatsioonid, grammatilised konstruktsioonid;
- konkreetsed sõnastikuprojektid: viipekeele, murde-, etümoloogiasõnastik, õppesõnastikud, WordNet, eri keelte sõnastikuprojektid.

Peamised tunnussõnad, mis kordusid paljudes ettekannetes, olid *info mitmekihilisus*, *paindlikkus* ja *integreeritud süsteemid*. Elektrooniline sõnastik ei ole lineaarse pabersõnastiku analoog, tal on mitmekihiline hierarhiline struktuur, mis võimaldab hoopis paremat ligipääsu sõnastikule, näiteks päringuid erinevate infoliikide järgi või erineva mahu ja sisuga väljavõtteid samast sõnastikust. Sõnastiku koostamise vahendid on paindlikud, nad ei sõltu keelest ja võimaldavad luua erineva struktuuriga sõnastikke ja samast sõnastikust erinevaid vaateid. Sõnastik ei ela eraldatud maailmas, vaid integreeritud süsteemis: sõnastiku koostamisel saab kasutada tekstikorpusi ja

teisi sõnastikke, samuti keeletarkvara; klikates tekstisõnal, on võimalik vadata vastavat kirjet sõnastikus; sõnastikud töötavad eri keskkondades (veebis, mobiiltelefonis jne).

Kokkuvõtteks võib öelda, et konverentsi sisu vastas päris hästi pealkirja-

le ja pakkus uusi vaatenurki ja võimalikke perspektiive elektroonilises leksikograafias. Kõik tutvustatu ei ole veel üldiselt kättesaadav, kuid areng on elektroonilises leksikograafias kiire.

MADIS JÜRVISTE, ÜLLE VIKS