SEITSMES „GLOBAL WORDNET" KONVERENTS

Rahvusvahelist keeletehnoloogia konverentsi, mille keskmes on mõisteline arvutisõnastik ehk *wordnet*, on korraldatud alates 2002. aastast. Järjekorras seitsmes „Global WordNet Conference" toimus seekord Tartu Ülikoolis 25.–29. jaanuarini 2014. Varasemad on toimunud Indias (2002, 2010), Tšehhis (2004), Koreas (2006), Ungaris (2008) ja Jaapanis (2012). Tänavune konverents tõi rõõmustavalt palju osalejaid kogu maailmast, kokku 40 riigist. Viiel päeval kuulati üle 55 ettekande, korraldati uue juhatuse valimiskoosolek ja lahutati ühiselt meelt eesti rahvamuusikat tutvustaval tantsupeol.

Wordnet'ide koostamisega on tegeldud alates 1980. aastatest, kui rühm teadlasi prof George Milleri juhtimisel alustas ingliskeelse mentaalse leksikoni mudeli loomist, mis kannab nime Princeton WordNet (PWN). Nüüdseks on maailmas eri keelte *wordnet*'e palju. Nende arendamist ja sel teemal toimuvaid konverentse koordineerib organisatsioon Global WordNet Association (GWA, koduleht http://globalwordnet.org), koondades rahvusvahelist keeletehnoloogide, informaatikute ja lingvistide võrgustikku, kokku *ca* 60 riigist. *Wordnet*'i konverentside eesmärk on iga kahe aasta tagant kokku tuua spetsialiste ja huvilisi, nii neid, kes ise teevad *wordnet*-tüüpi sõnastikke, kui ka neid, kes mingi muu keeletehnoloogilise rakenduse osana *wordnet*'i kasutavad. Eestlased on kuulunud GWA-sse algusest saadik ja osalenud ettekannetega Eesti Wordnetist kõigil konverentsidel.

Konverentsi kavas olid jätkuvalt kesksel kohal *wordnet*'i kui mõistelise arvutisõnastiku koostamise küsimused ja lahendused eri keelte jaoks: ettekandeid peeti ligi 30 erineva keele *wordnet*'idest (sh kurdi, bulgaaria, ungari, serbia, amhara, sotho, assami, vene, horvaadi, hiina, jaapani jpt). Räägiti nende keelte erisustest, aga ka vahenditest, kuidas keeleandmete õigsust kontrollida. Lisaks oli hulganisti ettekandeid mitmesugustest keeletehnoloogilistest tööriistadest, mis ühe osana kasutavad *wordnet*'ide andmebaase.

Hetkel on üheks populaarsemaks rakenduseks sõnatähenduste automaatne kindlakstegemine tekstis. Sõnatähenduste automaatse ühestamisega haakuval teemal pidas konverentsi kutsutud esineja, Pisa ülikooli professor Alessandro Lenci ettekande „Will Distributional Semantics Ever Become Semantic?" Ta ühendas teoreetilise keeleteaduse ja rakendusliku iseloomuga keeletehnoloogia valdkonna, keskendudes eeskätt distributiivse semantikaga lahendatavatele polüseemia ja järelduste (ingl *inference*) probleemidele. Distributiivne semantika on arvutilingvistikas leksikaalse semantika tööriist, mis võimaldab konstrueerida tähenduskirjeldusi sõnade esinemuse põhjal ulatuslikes tekstikorpustes. Huvilistele on prof Lenci loeng järelvaadatav http://www.uttv.ee/.

Teine kutsutud esineja oli prof Bolette Pedersen Kopenhaageni ülikoolist, kes tutvustas mitmekeelse (taani, rootsi, norra, islandi, soome ja eesti keel) andmebaasi loomist aastatel 2011–2012 Põhjamaade projekti „META-NORD – Euroopa avatud lingvistilise infrastruktuuri Balti- ja Põhjamaade haru"

raames. Projekti üheks alamülesandeks oli eri keelte (taani, rootsi, soome, eesti) *wordnet*'ide hindamine ja nende keelte omavaheline sidumine läbi ingliskeelse *wordnet*'i. Projekti ühe tulemusena valmis WordTies (wordties.cst.dk), mis visualiseerib mõiste nii ükskeelsest *wordnet*'ist koos tema semantiliste seostega kui ka näitab mõiste vastet teistes keeltes.

Kolmas kutsutud esineja oli Kadri Vider Tartu Ülikoolist, kes rääkis Eesti Keeleressursside Keskusest, selle põhimõtetest, digiajastu vajadusest ressursside koondamise ning ühtlustamise järele, aga ka tegevustest CLARIN-i, META-NET-i ja DASHISH-i projektides.

Enam kui 55 ettekannet oli jaotatud temaatilistesse plokkidesse, mis liigendasid tiheda programmi keeleteaduslike küsimuste, keeletehnoloogiliste meetodite ja rakenduste ning uute töövahendite tutvustamise osadeks.

Wordnet'i kogukond on alati suhtunud toetavalt iga uue *wordnet*'i loomisse. Konverentsidel esitletakse tavapäraselt nii uusi kui ka juba vanemaid ja rohkem arendatud *wordnet*'e üldsektsioonides. See võimaldab uutel tulijatel kogenumatelt väärt tagasisidet saada ja ühtlasi edasiarendustest eeskuju võtta. Sel aastal esitleti kolme uut *wordnet*'i. Purya Aliabadi andis ülevaate kurdi keele *wordnet*'i (KurdNet) koostamisest, Mampaka L. Mojapelo sotho keele *wordnet*'i loomisest. Semiidi keelkonda kuuluva ja põhiliselt Etioopias kasutatava amhara keele *wordnet*'i loomist on alustatud üsna hiljuti. Neist kõnelemisel keskenduti üldiselt koostamispõhimõtetele, kvaliteedi hindamisele ja mahu suurendamise võimalustele.

Laiema pildi saksa keele *wordnet*'ist lõid Erhard Hinrichs ja Verena Henrich, kes võrdlesid kaht keeleressurssi, GermaNet'i ja saksa keele digitaalset sõnaraamatut. Selgus, et sõnatähenduste esitus neis sõnaraamatuis suurel osal juhtudest küll kattub, kuid esineb ka erinevusi. Viimased tulenevad ilmselt sõnaraamatute koostamise erinevatest eesmärkidest ja sõnatähenduste detailse esitamise vajalikkusest. *Wordnet*-tüüpi sõnastiku suurendamisest olemasolevate sobivate sõnaraamatute abil ja nendes sisalduvate sõnatähenduste joondamisest itaalia keele *wordnet*'is kõneles ka Tommaso Caselli. Neis ettekannetes esitatud arenemisvõimalused saavad kindlasti teerajajaiks teistele. Saksa keele kohta oli veel teinegi ettekanne: Christina Hoppermann keskendus saksa keele ühend- ja väljendverbidele, püüdes lahendada nende leksikaalsete suhete – eeskätt hüponüümia – esitamise probleemi. Lahendusena tutvustas ta vastavaid kriteeriume, mille järgi andmebaasis olevaid ja veel juurde lisatavaid morfoloogiliselt kompleksseid verbe võiks käsitleda.

Keele morfoloogiale keskenduti ka bulgaaria, rumeenia ja horvaadi keelest kõneldes. Borislav Rizov andis ülevaate bulgaaria *wordnet*'ist BulNet ning analüüsis selles andmebaasis esinevaid nimisõnade ja tegusõnade vahelisi tuletussuhteid. Verginica Mititelu andis ülevaate rumeenia RoWN-i edenemisest ja selle suurendamisest sõnadevahelise tuletussuhte kaudu. Krešimir Šojat keskendus ettekandes horvaatia CroWN-ist samuti sõnadevahelistele tuletussuhetele ja nende olemusele horvaadi keeles. CroWN on koostatud n-ö laiendatud (adapteeritud) mudelina, st aluseks on PWN, mida on tõlgitud ning kohandatud oma keele jaoks sobivaks. Selline loomisvõte kipub esile tooma aga mitmeid probleeme, eriti keele morfoloogiast tulenevate semantiliste suhete esitamisel. Selle ettekande teema ja probleemidega haakus Veronika Vincze ettekanne ungari HuWN-ist, mis on koostatud samuti n-ö laiendatud mudelina. Vincze võrdles HuWN-i mõistelist hierarhiat PWN-iga ja tõi välja mitmeid lahendusi edaspidiseks tööks.

Markus Forsberg andis ülevaate rootsi *wordnet*'i Swesaurus koostamis-

põhimõtetest ja keeleressurssidest, millele nad toetunud on. Swesaurus on loodud eelkõige keele spetsiifikat arvestades käsitsi, kuid projekti META-NORD raames osutus vajalikuks suurendada andmebaasi PWN-i tõlkimise ja kohandamise abil.

Poola keeles on koostamisel korraga kaks *wordnet*'i: plWN ja PolWN. Marta Dobrowolska rääkis plWN-ist ja võrdles selles kasutatud termineid PWN-iga. Maciej Piasecki ja Elżbieta Hajnicz keskendusid plWN-ist kõneldes selle rakendusvõimalustele leksikaal-semantilise ressursina. Bartłomiej Kochanowski andis ülevaate väiksema mahuga, kuid samuti jõudsalt arenevast teisest poola keele *wordnet*'ist polWN.

Marissa Griesel ja Sonja Bosch rääkisid, kuidas on viie aasta jooksul edenenud Aafrika Wordnet, millesse on koondatud neli keelt. Projekt sai alguse 2008. aastal ja andmebaas on praeguseks kasvanud 42 000 mõisteni. Griesel ja Bosch tõdesid, et andmebaasi mahtu peaks oluliselt suurendama, tutvustades selleks kasutatavaid keelekorpusi ja tuues välja keelte spetsiifikast johtuvad väljakutsed.

Indias kõneldavate keelte *wordnet*'ide kohta on konverentsidel olnud mitmeid ettekandeid. Tänavu sai enim tähelepanu assami *wordnet*, mida tutvustas Shikhar Kr. Sarma. Kuulajad said ülevaate assami *wordnet*'i sünonüümiasuhetest ja masintõlkest.

Üks *wordnet*'iga töötamise väljakutseid on tähenduste esitamise moodused. Keeleteaduslik teemadeplokk puudutaski ennekõike arutlusi leksikaalsetest suhetest, kuna sisuseosed teevad *wordnet*-tüüpi sõnastiku ahvatlevaks loomuliku keele semantikaga kokkupuutuvatele keeletehnoloogidele ja informaatikutele. Marek Maziarz keskendus sellele, kuidas tavasõnastike pragmaatiline info (nagu stiili-, eriala- jms märgendid) saaks rikastada ka *wordnet*'i põhimõtetega sõnastikku. Kontrollimaks suhte kehtivust, olid poolakad välja töötanud omakeelsed lingvistilised kontrolltestid. Alice Zhangi ettekande aluseks oli mõte, et omavahel sünonüümsed omadussõnad võivad üksteisest erineda tähenduse intensiivsuse poolest ja et *wordnet*'is pole skalaari siiani kajastatud. Ettekandes näidati, kuidas teatud algoritmiliste vahenditega seda puudujääki vähendada.

Ettekannetes käsitletud tehnilisemad teemad tutvustasid nii erinevaid tehnoloogilisi rakendusi *wordnet*'i jaoks kui ka rakendusi, mis põhinevad *wordnet*'il. Antoni Oliver tutvustas automaatseks *wordnet*'i laiendamiseks loodud tööriista WN-Toolkit, mis sorteerib keeleandmeid nii semantiliselt märgendatud korpusest kui ka olemasolevatest sõnastikest. Tööriista kasutati mitme keele peal ja teostati kvaliteedihindamine. Mark Finlayson USA-st demonstreeris Java tarkvarapaketi võimalusi tööks *wordnet*'iga. Hugo G. Oliveira ja Paulo Gomesi ettekanne Onto.PT-st põhines esimese autori doktoritööl ja kirjeldas erinevate portugalikeelsete ressursside automaatset ühendamist, mille tulemuseks on rikkaliku suhetesüsteemiga ja vabalt kättesaadav *wordnet*'i-sarnane leksikon (portugali keele viies *wordnet*).

Viimasel ajal tähelepanu keskmesse tõusnud ontoloogiate teema oli esindatud paari ettekandega ka sel konverentsil. Natalia Loukachevitch ja Boris Dobrov tutvustasid neljandat *wordnet*-tüüpi sõnastikku Venemaal (olemas on: RussNet – u 20 000 mõistega Peterburis käsitsi koostatav sõnastik ning kaks Moskvas tehtavat Russian WN-i, mis mõlemad on PWN-i tõlked). Oma uut andmebaasi nimetavad nad pigem lingvistiliseks ontoloogiaks, kus tehakse vahet üldontoloogial, mis sisaldab üldkeele andmeid, ning valdkonnaontoloogial – neil on selleks sotsiaal-poliitiline valdkond. Mustafa Jarrari ettekantu puudutas lingvistilise ontoloogia tegemise meetodit erinevate keelte andmebaaside ühitamise abil. Kergete killast

see ülesanne ei ole, sest on ju teada, et mõisted pole eri keeltes ühtviisi leksikaliseerunud. Pakuti välja idee standardi loomiseks, et kontrollida kvaliteeti ja võrrelda erinevaid meetodeid.

Esimesed katsetused ehitada mitmekeelset *wordnet*'i kakskeelsete sõnastike põhjal pärinevad 1990. aastate keskelt. Sel konverentsil jagas Quentin Pradet kogemusi tööst prantsuskeelse *wordnet*'iga, kus lisaks kakskeelsest sõnastikust automaatselt saadud andmetele arvestatakse ka keele süntaktiliste reeglitega. Sellega haakus Martin Benjamini intrigeeriva pealkirjaga ettekanne „Elephant Beer and Shinto Gates: Managing Concepts in a Multilingual Database" – ta andis idee, kuidas kasutada tõlkimise käigus semantilisi suhteid siis, kui eri keeltes pole täpseid tõlkevasteid võimalik leida.

Marten Postma sõnavõtt koos juhendaja Piek Vosseniga keskendus semantilisele sarnasusele ja sünonüümiale kui *wordnet*'i võtmemõistetele. Tutvustati töövahendit, millega saab mõõta semantiliselt sarnaste mõistete kaalu. Selline keelest sõltumatu mõõtmisviis toetab paljuski inimese keelelist intuitsiooni. Ka Jaapani uurija räägitu haakus intuitsiooni teemaga: nimelt kõneles Hitoshi Isahara katsest ühitada intuitsioonil põhinev jaapani keele *wordnet* ja tekstist automaatselt (toetudes keele morfoloogilisele infole) leitud mõistete vaheliste suhetega sõnavõrgustik. Kontrollimaks suhete kehtivust, viidi läbi assotsiatsioonikatsed õpilastega ja nii saadi täiesti uue kvaliteediga ressurss.

Eestlastelt oli konverentsil kaks ettekannet. Esimese pidas Indrek Jentson, kes kõneles VerbNet'i Workbench'ist kui vahendist, mis aitab eesti keele verbe klassifitseerida nii nende tähenduse kui ka predikaatargument-struktuuri järgi. Teise ettekande pidas Ahti Lohk *wordnet*-tüüpi sõnastiku semantiliste suhete kontrollimise graafiteooriast lähtuvast meetodist.

(Paralleel)korpuste kasulikkusest *wordnet*'ide rikastamisel pidas mitu ettekannet Francis Bond. Nii näiteks on tegemisel inglise-hiina paralleelkorpus, kus tähendused on omavahel joondatud ja mida saab kasutada hiina või inglise *wordnet*'i täiendamiseks.

Siintoodu on vaid ülevaatlik teemadering. Huvilised saavad konverentsi artiklikogumikku sirvida konverentsi kodulehel http://gwc2014.ut.ee/. Konverentsi korraldas Tartu Ülikooli arvutiteaduse instituut koostöös eesti ja üldkeeleteaduse instituudiga, toetajad olid Arvutiteaduse tippkeskus, Eesti Keeleressursside Keskus, CLARIN ERIC ning Tartu linnavalitsus.

HEILI ORAV, SIRLI PARM