Janoš Ježovnik

Orodja in metode za ustvarjanje inovativnih e-slovarjev, Ljubljana, 17.–20. maj 2016

Cobiss: 1.25

Od 17. do 20. maja 2016 je v Ljubljani v okviru dejavnosti Evropske mreže za e-leksikografijo (ENeL) potekalo izobraževanje Orodja in metode za ustvarjanje inovativnih e-slovarjev, ki sta ga soorganizirala Center za jezikovne vire in tehnologije Univerze v Ljubljani in Trojina, zavod za uporabno slovenistiko. Udeležilo se ga je 28 udeležencev z raziskovalnih ustanov iz 14 različnih držav, med njimi tudi štirje sodelavci Inštituta za slovenski jezik Frana Ramovša ZRC SAZU v Ljubljani. Namen izobraževanja je bil udeležence seznaniti z nekaterimi načini gradnje in analize korpusov, avtomatskega izvoza in urejanja podatkov, potrebnih za oblikovanje slovarskega sestavka, ter objave urejenih slovarskih sestavkov s pomočjo spletnih in drugih slovaropisnih sistemov, in sicer tako v teoriji kot v praksi.

Dogodek se je začel z uvodnim nagovorom in predstavitvijo izvajalcev posameznih delavnic ter s kratko predstavitvijo vsakega od udeležencev. Zatem sta Carole Tiberius (Inštitut za nizozemsko leksikologijo, Leiden) in Simon Krek (Institut »Jožef Stefan«, Ljubljana) predstavila slovarska projekta, izdelana na podlagi korpusne metode, in sicer Nizozemski splošni slovar (*Algemeen Nederlands Woordenboek*, ANW) oziroma Slovensko leksikalno bazo (SLB). V popoldanskem delu je Egon Stemle (Inštitut za specializirano komunikacijo in večjezičnost, EURAC, Bolzano/Bozen) predstavil koncept svetovnega spleta kot korpusa (*Web as Corpus*) ter metode avtomatskega pridobivanja korpusnega gradiva s spleta na podlagi ključnih besed in njegove obdelave (izločanje korpusnega šuma, čiščenje odvečnih vsebin ...), ki smo jih udeleženci v praktičnem delu tudi preizkusili.

Drugi dan je bil namenjen spoznavanju oblikovanja in analize pridobljenih podatkov. Carole Tiberius je uvodoma predstavila načela načrtovanja strukture slovarskih sestavkov in njenega prikaza s pomočjo diagramskega jezika UML. Sledila je predstavitev razširljivega označevalnega jezika XML, v katerem je napisana glavnina shem sodobnih e-slovarjev, in standarda TEI, ki opisuje nabor in načela uporabe oznak v jeziku XML za potrebe oblikovanja različnih strojno berljivih besedil, tudi e-slovarjev. Michal Měchura (Univerza Dublin City, Dublin) je predstavil Lexonomy, spletno okolje za pisanje in objavljanje e-slovarjev. Ta uporabniku omogoča enostavno oblikovanje slovarske sheme, vnos in oblikovanje slovarskih sestavkov ter njihovo objavo, podpira pa tudi avtomatski uvoz podatkov v ustreznem formatu. Brezplačni sistem, ki zahteva le registracijo pri

avtorju in se bo v prihodnosti še dograjeval, je intuitiven in uporaben za različne vrste leksikografskih projektov, ne omogoča pa (še) podvajanja gnezdenih oznak XML, kar nekoliko zmanjšuje preglednost ustvarjene sheme.

V praktičnem delu smo udeleženci oblikovali svojo slovarsko shemo in jo vnesli v lasten slovarski projekt v sistemu Lexonomy. Sledila je predstavitev sistemov za korpusne analize in poizvedbe. Miloš Jakubíček (Lexical Computing, Brighton – Brno) je predstavil arhitekturo in delovanje korpusnega orodja SketchEngine, skupaj s Carole Tiberius pa v nadaljevanju jezik za korpusne poizvedbe CQL (Corpus Query Language) in načela za pisanje slovnice besednih skic v tem jeziku. Besedne skice so razširitev orodja SketchEngine in omogočajo prikaz kolokacij iskane besede, ki se pojavljajo v korpusu, glede na slovnične relacije, vnaprej definirane s slovnico besednih skic; kolokacije je mogoče razvrstiti tako po pogostnosti kot po relativni statistični relevantnosti. V zaključnem delu drugega dneva je Iztok Kosem (Trojina, Ljubljana) opisal delovanje in uporabo orodja GDEX (Good Dictionary EXample), prav tako implementiranega v okolje SketchEngine, ki služi razvrščanju konkordanc glede na njihovo primernost za vključitev v slovarski sestavek. Stavčni zgledi v slovarju bi morali težiti k čim večji avtentičnosti, informativnosti in razumljivosti ter k prikazu čim bolj tipične rabe slovarske iztočnice. GDEX zglede v konkordancah ovrednoti po vnaprej določenih merilih (npr. dolžina povedi, pojavljanje pogosto ali redko rabljenih besed, število velikih začetnic, položaj leme v stavku itd.) in bolje ocenjene konkordance uvrsti na vrh seznama, s čimer leksikografu olajša iskanje ustreznih stavčnih zgledov.

V naslednjem delu smo udeleženci spoznavali avtomatsko luščenje podatkov iz besedil. Izvajalci so predstavili nekaj idej in pobud na tem področju ter njihov potencial za uporabo v e-leksikografiji:

- prizadevanja delovne skupine, ki se znotraj ENeL ukvarja s t. i. inovativnimi e-slovarji;
- platformo za obdelavo strukturiranih in nestrukturiranih podatkov velikega obsega v realnem času QMiner in njeno implementacijo na spletni strani EventRegistry (http://eventregistry.org), ki iz novičarskih spletnih virov v več jezikih pridobiva podatke v realnem času in ponuja strnjene opise različnih svetovnih dogodkov;
- projekt v nastajanju Elexis, v okviru katerega bi povezali obstoječe e-slovarske opise in njihove dele ter jih integrirali v obliki portala z multimedijskim prikazom rabe besed, temelječim na prepletu podatkov iz več jezikov;
- opis dejavnosti akcije PARSEME, ki deluje v okviru iniciative COST in se ukvarja z razčlenjevanjem naravnih jezikov in večbesednimi izrazi, in skupnega srečanja predstavnikov PARSEME in ENeL, namenjenega vzpostavitvi interdisciplinarnega sodelovanja med obema akcijama.

Po predstavitvah smo spoznali še eno od orodij okolja SketchEngine, ki omogoča enostavno izbiro slovarskih zgledov s klikanjem (*TickBox lexicography*, v

slovenskem prostoru t. i. kliksikografija) in njihov avtomatski izvoz v poljubno ciljno delovno okolje. Dotaknili smo se tudi naprednejše metode izvoza korpusnih podatkov v surovi obliki s pomočjo formata JSON.

Sklepni dan izobraževanja je bil namenjen seznanjanju z načini in načeli objavljanja e-slovarjev. Michal Měchura je uvodoma izpostavil posebnosti, na katere je treba biti pozoren pri oblikovanju spletnega slovarja ali slovarskega portala. Prikazal je uvoz slovarskih podatkov v spletno okolje Lexonomy in njihovo nadaljnje oblikovanje. Udeleženci smo podatke, ki smo jih avtomatsko izvozili prejšnji dan, uvozili v slovarske sheme, oblikovane drugi dan izobraževanja. V nadaljevanju smo spoznali še postopek objave tako ustvarjenega spletnega slovarja na portalu Lexonomy. Kot zaključek izobraževanja sta sledila predstavitev rezultatov udeležencev in podajanje povratne informacije izvajalcem.

Izobraževanje s strukturirano predstavitvijo tako osnovnih kot naprednejših metod in načel e-leksikografije je kljub različnim izhodiščnim ravnem znanja in področjem zanimanja udeležencev poskrbelo za kvalitetno seznanitev s sodobnimi težnjami na omenjenem področju ali vsaj za osvežitev že obstoječega znanja.