

## ***Lexicom 2010: jezikoslovni korpusi po meri***

Med 7. in 11. junijem 2010 je na Filozofski fakulteti Univerze v Ljubljani potekala delavnica *Lexicom* na temo leksikografije in računalniškega jezikoslovja. Delavnico vsako leto prireja podjetje Lexicography MasterClass<sup>1</sup> iz Velike Britanije. Tokratna gostiteljica *Lexicom* 2010 sta bila zavod za uporabno slovenistiko Trojina,<sup>2</sup> ki je poskrbel za organizacijo, in Filozofska fakulteta v Ljubljani, ki je udeležencem odstopila svoje prostore in uporabo računalnikov.

Na *Lexicomu* se teorija venomer prepleta s prakso, udeleženci se usposablajo za leksikografsko delo na podlagi korpusov in se seznanjajo z računalniškimi programi kot podlago za leksikografsko dejavnost.

Prvi *Lexicom* se je odvijal leta 2001 na Univerzi v Brightonu, tako da je delavnica, ki je letos potekala v Ljubljani, že deseta po vrsti. Udeležujejo se je strokovnjaki različnih profilov, od leksikografov, jezikoslovcev, računalniških strokovnjakov, prevajalcev, terminologov do podiplomskih študentov. Letošnje delavnice se je udeležilo 27 udeležencev iz Francije, Španije, Norveške, Indije, Burundija, Estonije, Romunije, Poljske, Srbije, Hrvaške in Slovenije.

Petdnevni program je bil intenziven in poln novosti na področju leksikografije in računalniške obdelave naravnega jezika. Pripravili so ga predavatelji podjetja Lexicography MasterClass **Adam Kilgarriff**, **Michael Rundell** in **Sue Atkins**, ki pa na tokratnem *Lexicomu* ni predavala, ter **Simon Krek** z Instituta Jožef Stefan in podjetja za razvoj jezikovnih tehnologij Amebis.<sup>3</sup>

Adam Kilgarriff je direktor podjetja Lexical Computing Ltd. in eden izmed treh direktorjev podjetja Lexicography MasterClass. Razvil je spletno orodje Corpus Architect<sup>4</sup> (Sketch Engine), o katerem bo govora v nadaljevanju. Njegovo znanstveno in strokovno delo obsega računalniško jezikoslovje in korpusno analizo. Je gostujoči profesor na Univerzah v Leedsu in Sussexu. Na njegovo pobudo je nastala projektna skupina SENSEVAL,<sup>5</sup> ki razvija orodja za avtomatično razdvoumljanje besed, v zadnjem času pa raziskuje, kako iz svetovnega spleta pridobiti jezikoslovni korpus, zaradi česar je ustanovil projektno skupino ACL SIGWAC.<sup>6</sup>

Michael Rundell ima za seboj že trideset let ustvarjanja na področju leksikografije, sprva kot urednik pri Longmanu, nato pa kot glavni urednik serije Macmillanovih enojezičnih pedagoških slovarjev.<sup>7</sup> Je eden izmed vodilnih evropskih leksikografov, zlasti za enojezično leksikografijo, ki že od vsega začetka v svoje delo uvaja računalniške tehnike in delo s korpusi. Je soavtor mnogih korpusov, med njimi tudi Britanskega nacionalnega korpusa. V zadnjem času se posveča vprašanju, kakšna je prihodnost slovarjev v obdobju prevladujočih spletnih tehnologij. Skupaj s Sue Atkins je napisal knjigo *The Oxford Guide to Practical Lexicography* (2008).

Angleška strokovnjaka sta k sodelovanju pri izvedbi programa zaradi izkušenj z dvojezično leksikografijo povabila Simona Kreka, ki je med drugim glavni urednik *Velikega angleško-slovenskega slovarja Oxlikega*.

<sup>1</sup> <<http://www.lexmasterclass.com/>>.

<sup>2</sup> <<http://www.trojina.si/Vsebine/Sl/Domov/Domov.aspx>>.

<sup>3</sup> <<http://www.amebis.si/>>.

<sup>4</sup> <<http://www.sketchengine.co.uk/>>.

<sup>5</sup> <<http://www.senseval.org/>>.

<sup>6</sup> <<http://www.sigwac.org.uk/>>.

<sup>7</sup> <<http://www.macmillandictionary.com/>>.

ford (2005–2006), s katerim je v slovenski prostor uvedel moderno korpusno leksikografijo, saj je vzporedno s slovarjem začrtal tudi razvoj slovenskega referenčnega korpusa. Je koordinator projekta *Sporazumevanje v slovenskem jeziku 2008–2013*,<sup>8</sup> s katerim bo slovenski jezik na podlagi obsežnega korpusa dobil leksikalno bazo, slovnični analizator, izdelan bo samostojni korpus usvajanja slovenščine, sestavljena bosta tudi pedagoška korpusna slovnica ter slogovni priročnik.

Predavatelji so poleg teoretično podprtih dejstev predstavili svoje izsledke na veliko primerih, ki se niso nanašali samo na enojezično angleško slovaropisje, temveč so se posvečali splošni leksikografiji in praksi, uporabni za vse jezike, kar je tudi namen delavnice. Ob aktivni udeležbi slušateljev so bila njihova predavanja vselej interaktivna. Dovolj časa so posvetili tudi praktičnim vajam.

Leksikografija, če izvzamemo metaleksikografijo, je praktično naravnana veda, ki se v zadnjih dveh desetletjih silovito spreminja, saj so se z uporabo korpusnega jezikoslovja povsem zamajali temelji tradicionalnega zbiranja in obdelovanja podatkov za uporabo v slovarjih. Osnovno izhodišče za kakršnokoli učinkovito leksikografsko delo je računalniško podprti jezikovni korpus oz. analiza korpusnih podatkov.

### Korpusna leksikografija

Lexicomova delavnica se je tako tudi začela s predavanjem Michaela Rundella na temo korpusne leksikografije in zasnove jezikovnega korpusa. Udeleženci smo prejeli odgovor na vprašanje, zakaj so potrebni veliki oz. vedno večji korpusi. Skoraj 90 odstotkov najpogostejših besed v nekem

jeziku, tj. približno 7500, je zelo običajnih, kar pomeni, da je velika večina besed redka in ravno zato potrebujemo čimveč podatkov za njihovo analizo. Britanci se bodo, na primer, po napovedih do leta 2011 s svojim korpusom povzpeli na 20 milijard besed. Slovenski korpus po številu nič ne zaostaja za vodilnimi svetovnimi korpusi, saj bo, kot že omenjeno, kmalu vseboval milijardo besed. Spoznali smo parametre za gradnjo čimbolj reprezentativnih korpusov, kamor med drugim sodijo tudi korpusi govornega jezika.

Skupna predavanja so obsegala še teme, kot so: evalacija besednih skic v smislu kakovosti in uporabnosti dobljenih rezultatov za leksikografsko delo, uporaba teoretičnega znanja pri leksikografovem delu, prednosti, ki jih uvaja procesiranje naravnih jezikov v korpusno jezikoslovje (označevanje besed in stavkov v besedilih, segmentacija, skladišna analiza in oblikoslovno označevanje), razdvoumljanje besed. V okviru iskanja meje med pomeni se je Kilgarrieff dotaknil teme, kako razviti samodejno združevanje pomenov s pomočjo t. i. podatkovne baze CoCoDo, ki bi bila podlaga za slovar. S takšno bazo je mogoče samodejno ekstrahirati stalne besedne zveze, slovnične konstrukcije in domene.

### Corpus Architect/Sketch Engine (SkE)

Predstavitvi splošnih dejstev o korpusu je sledil uvod v uporabo programskega modula Corpus Architect/Sketch Engine<sup>9</sup> za analizo različnih jezikovnih funkcij. Gre za eno najpomembnejših orodij za korpusno raziskovanje s funkcijami, kot so: konkordanč-

<sup>8</sup> <<http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx>>.

<sup>9</sup> Februarja 2010 je v Ljubljani v okviru projekta *Sporazumevanje v slovenskem jeziku* potekala delavnica za uporabo orodja Corpus Architect (SkE) *Slovnica besednih skic*. (<http://www.slovenscina.eu/Vsebine/Sl/Dogodki/BesedneSkice/Program.aspx>)

nik, možnost izdelave besednega seznama, tezavra ter lastnega korpusa. V modul lahko vnesemo korpus kateregakoli jezika z njegovimi slovničnimi vzorci, ta pa iz njih ustvari besedne skice (Krek in Kilgarriff 2006). To so samodejni, na korpusu temelječi sežetki slovničnega in kolokacijskega okolja neke besede. Besedne skice delo leksikografov znatno olajšajo, mogoče pa jih je uporabiti tudi v druge jezikoslovne namene, kot npr. za jezikovno analizo, odkrivanje večbesednih enot, razdvoumljanje besed. Spoznali smo novo orodje WebBotCat za gradnjo korpusov po meri. Govora pa je bilo tudi o spletu kot podatkovni bazi za jezikoslovni korpus.

Praktične vaje iz tega vsebinskega sklopa so vključevale izdelavo geselskega članka, primerne za enojezični pedagoški slovar. Pri delu se nismo zgledovali po kakšnem obstoječem slovarju, pač pa smo uporabili edinole korpusno orodje SkE.

Sicer smo bili pri praktičnih vajah razdeljeni v dve skupini, od katerih se je ena ukvarjala s praktično leksikografijo in svoje delo predstavila zadnji dan delavnice. Druga skupina se je medtem posvečala računalniškemu jezikoslovju in načinu procesiranja naravnih jezikov.

### Programi za izdelavo slovarjev

Simon Krek je predstavil programe oz. sisteme za izdelavo slovarjev. Kompleksna struktura geselskih člankov v takšnem sistemu dobi novo, pregledno podobo, čeprav je podatkovna slovarska baza prvotno shranjena v formatu XML. Krek je ilustriral uporabo sistema na podlagi nekaterih obdelanih slovenskih gesel iz nastajajoče leksikalne baze za slovenščino. Sistem omogoča zelo pregledno in za oko privlačno strukturo geselskega članka. Gre za celovit opis leksikalnih enot s podrobno pomensko čle-

nitvijo, ki je odraz sodobnega stanja v korpusu, z vključitvijo slovničnih informacij in kolokacij, s frazeologijo, stalnimi besednimi zvezami in korpusnimi zgledi.

### Ontološki pristopi

V okviru prispevka o področnih ontologijah in terminologijah je Kilgarriff med drugim za primer navedel splošni tezaver *WordNet*,<sup>10</sup> ki nastaja tudi za slovenščino in trenutno obsega opis 5000 najpomembnejših konceptov.

Predstavljena je bila teorija pomenskih shem Charlesa Fillmora *FrameNet*<sup>11</sup> in njena uporaba v leksikografiji. *FrameNet* temelji za izločanju informacij o povezanih pomenskih in skladijskih lastnostih leksikalnih enot in je bil prvotno namenjen obdelavi angleških besed. V tem času se uporablja tudi že za nekatere druge jezike, v dvojezičnem kontekstu pa tudi za slovenski jezik.

### Dvojezična leksikografija: angleško-slovenska izkušnja

Izpostavila bi rada Krekov prispevek na temo dvojezične leksikografije, v katerem je obravnaval vrste dvojezičnih slovarjev in njihovih uporabikov, način zbiranja podatkov za izhodiščno podatkovno bazo, analizo baze in njen prevod v ciljni jezik, iskanje primernih ustreznice, pri čemer je treba upoštevati pogosto nesimetrično uje-manje semantičnih vsebin med jezikoma in kolokacijsko sobesedilo. Odgovoril je na vprašanje, kaj storiti, kadar v ciljnem jeziku ustreznice ni na voljo. V dvojezični leksikografiji je neizogibna uporaba vzporednih oz. dvojezičnih korpusov, njihova uporaba pa ima tako dobre plati kot nekatere pasti. Spregovoril je o končni fazi oblikovanja

<sup>10</sup> <<http://wordnet.princeton.edu/>>.

<sup>11</sup> <<http://framenet.icsi.berkeley.edu/>>.

geselskega članka. V idealnem primeru ga uredita dva urednika, ki vsak posebej poskrbita za interese uporabnikovega izhodiščenega oz. ciljnega jezika. Kateri so koraki pri sestavljanju geselskega članka, kakšne prevajalske tehnike pri tem uporabljamo, zakaj je pomembno razlikovanje med prevodom, ki je odvisen od sobesedila in tistim, ki to ni, kako je sestavljen geselski članek, kaj so indikatorji in kolokatorji, kako iz prvega zapisa članka, ki je zmeraj predolg, izluščimo bistvene sestavine, na vsa ta vprašanja in še več je Krek podal izčrpne odgovore.

### **Avtomatizacija leksikografskih postopkov**

Leksikografsko delo je po tradiciji zelo zamudno, zato je treba v času računalniške tehnologije tudi na tem področju poskrbeti za avtomatizacijo leksikografskih opravil v največji možni meri. Najprej smo se sprehodili skozi zgodovino takšnih rešitev, za konec pa se seznanili s projektom Dante,<sup>12</sup> ki je podatkovna baza obdelanih angleških besedil, v kateri je zelo podrobno opisano osnovno besedišče angleškega jezika.

Iskanje po korpusu je večinoma možno le s pomočjo formaliziranega jezika, ki vsebuje zakodirane informacije v taki obliki, da je mogoče iskano informacijo dobiti iz korpusa. Takšen jezik je za leksikografa začetnika praviloma zelo zapleten, vendar ga je zelo koristno poznati.

### **Leksikografske definicije**

Kako se lotevamo oblikovanja definicij in kaj moramo pri tem upoštevati, je teoretično ter z mnogimi primeri podprl Michael Rundell. Obravnaval je navajanje zgledov v slovarju in izpostavil kriterije za dober zgled. Predstavljen je bil GDEX, samodej-

ni načina iskanja dobrih slovarskih primerov iz korpusa, ki ga prav tako omogoča že omenjeni programski modul SkE.

Obširno smo obdelali kvalifikatorje, njihove vrste in kriterije za uvrščanje v slovarje. Do zdaj je njihovo vnašanje v slovar ali korpus potekalo ročno, s prihodom modula Sketch Engine pa dodelitev kvalifikatorjev poteka samodejno.

### **Izdelava lastnega korpusa**

Kilgariff je ponazoril še eno praktično vrednost programa SkE, in sicer pripravo korpusa po lastnih kriterijih ter njegov vnos v Sketch Engine za nadaljnjo analizo. Med praktično vajo je vsak udeleženec sestavil svoj korpus.

### **Izbor iztočnic**

Pri izboru iztočnic, ki bodo zastopane v slovarju, je treba upoštevati vrsto meril, ki se od slovarja do slovarja razlikujejo, saj so odvisne od njegove velikosti, uporabnika, namena uporabe ter nenazadnje od založniške politike. Rundell je povzel najvažnejša dejstva in posvaril pred verodostojnostjo besedil na slovarskih platnicah, saj so vse prej kot točna.

### **Najbolj pogoste besede**

Avtomatični izbor najbolj pogostih besed v korpusu se na prvi pogled zdi enostavno opravilo, vendar pa pri tem naletimo na vrsto težav zaradi nejasnega odgovora na vprašanje, kaj je beseda in kje so njene meje, kaj je razlika med besedo in lemo, kako označiti slovnične kategorije, kaj storiti s številkami, imeni, večbesednimi izrazi, homonimijo. Z raziskovanjem učinkovitega izbora besed po pogostosti se ukvarja projekt Evropske unije KELLY, v katerega je vključenih devet držav in še ni zaključen.

<sup>12</sup> <<http://www.webdante.net/>>.

## Zanimivosti

Lexicoma so se udeležili tudi štirje računalniški strokovnjaki z Univerze Hyderabad v Indiji, ki so Kilgarriffu asistirali pri praktičnih vajah procesiranja naravnega jezika. Pri delu s Sketch Engine so pomagali udeležencem pri njihovih specifičnih nalogah. **Siva Reddy** je predstavil t. i. korpusno tovarno, ki sočasno vsebuje korpusne različnih jezikov s sto in več milijoni besed.

Zanimivo je bilo slišati tudi predstavitev nacionalno pomembnih in izjemno obsežnih leksikografskih projektov udeležencev iz Srbije in Norveške.

Vsi predstavljeni prispevki, gradivo ter praktične naloge so zbrani v zajetnem zvezku, ki ga je prejel vsak udeleženec, in na koncu katerega je navedena obsežna bibliografija ter koristni naslovi spletnih strani, ki napotijo na številne korpusne, spletne slovarje in druge uporabne leksikografske vire.

Delavnico je z izводи najnovejšega slovarja *Macmillan English Dictionary for Advanced Learners*, ki ga je prejel vsak udeleženec, podprla založba Macmillan Education, vsebinsko pa so jo podprle akademske ustanove ACL-SIGLEX, DSNA, ELSNET in EACL.

## Sklep

Po vsem povedanem je bil vsebinski del Lexicoma vsekakor uspešen, za kar gre v prvi vrsti zahvala direktorjema Lexicography MasterClassa Adamu Kilgarriffu in Michaelu Rundellu ter Simonu Kreku. Mojca Šorli je v imenu zavoda Trojina poskrbela za brežibno organizacijo delavnice, v organizacijskem odboru pa so bili še Iztok Kosem, Karmen Kosem, Tadeja Rozman, Simon Rigač, Simon Šuster in Ana Zwitter Vitez.

Vendar se ljubljanski *Lexicom* 2010 še nadaljuje. Udeleženci smo ugotovili, da imamo veliko skupnih točk pri raziskovalnem in praktičnem delu, zato si še zmeraj prek elektronske pošte izmenjujemo izkušnje, napotke in nasvete za izboljšavo svojih leksikografskih orodij. Adam Kilgarriff nas je presenetil z darilom, predvsem vse nas, ki smo se zagreli za orodje SkE. Kmalu po delavnici je posodobil in prenovil njegovo podobo in dodal nove funkcije. Hvala, Adam, svojega prevajalskega in leksikografskega dela si brez tega programa ne znam več predstavljati!

Zahvaljujoč organizaciji slovenske ekipe je uspel tudi družabni del delavnice, ki je bil prav tako pester, saj so udeleženci pod strokovnim vodstvom spoznali Ljubljano, se odpeljali na Bled, na nočnem pohodu uživali ob razgledu na Ljubljano in se razvajali ob dobrotah slovenske kulinarike.

Podobnih delavnic, ki bi prispevale k izmenjavi tako bogatih izkušenj in tako obsežnega in poglobljenega znanja s področja leksikografije si še želimo.

Anita Srebnik  
Univerza v Ljubljani  
Filozofska fakulteta  
anita.srebnik@guest.arnes.si

## Literatura

Krek, Simon, in Kilgarriff, Adam, 2006: *Slovene Word Sketches*. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije. Mrežni zbornik*. Ljubljana: Institut »Jožef Stefan«.

Atkins, B. T. Sue, in Rundell, Michael, 2008: *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

<<http://www.lexmasterclass.com/>>. (Dostop 13. 07. 2010.)

<<http://www.trojina.si/Vsebine/Sl/Domov/Domov.aspx>>. (Dostop 13. 07. 2010)

<<http://www.amebis.si/>>. (Dostop 13. 07. 2010.)

<<http://www.sketchengine.co.uk/>>. (Dostop 15. 07. 2010.)

<<http://www.senseval.org/>>. (Dostop 14. 07. 2010.)

<<http://www.sigwac.org.uk/>>. (Dostop 14. 07. 2010.)

<<http://www.macmillandictionary.com/>>. (Dostop 14. 07. 2010.)

<<http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx>>. (Dostop 13. 07. 2010.)

<<http://wordnet.princeton.edu/>>. (Dostop 14. 07. 2010.)

<<http://framenet.icsi.berkeley.edu/>>. (Dostop 14. 07. 2010.)

<<http://www.webdante.net/>>. (Dostop 15. 07. 2010.)