Dejan Gabrovšek

Delavnica o leksikografiji in korpusnem jezikoslovju Lexicom 2022

Cobiss: 1.19
https://doi.org/10.3986/JZ.29.1.12

Lexicom – an intensive workshop in lexicography and lexical computing, 13.–17. 6. 2022

Enotedenska delavnica Lexicom 2022 je potekala med 13. in 17. 6. 2022 v češkem mestu Telč. Osredotočala se je na računalniško podprto leksikografijo, uporabo korpusov ter na zbiranje in interpretacijo večjih količin jezikovnih podatkov. Lexicom organizira podjetje Sketch Engine, ki se ukvarja s korpusno in širše računalniško obdelavo jezikov.

Predavali so: izkušeni angleški leksikograf Michael Rundell, Miloš Jakubíček, ki se ukvarja predvsem z gradnjo korpusov in slovarjev, ter Ondřej Matuška, ki je vodil praktični del, npr. uporabo CQL. Predavatelji odlično obvladajo svoja področja, tako da smo lahko dobili poglobljen in sodoben vpogled v vsako obravnavano tematiko.

V primerjavi z drugimi leti je bilo udeležencev dokaj malo, zgolj devet, kar je najverjetneje posledica pandemije koronavirusa. Vsi razen udeleženke z Maldivov smo bili iz Evrope. Delavnica je bila že enaindvajseta po vrsti.

Prvi dan je bil namenjen predvsem uvodu v to, kaj je slovar danes, kaj je slovar bil v dobi pred digitalizacijo in kako se je razvijalo slovaropisje. Prikazani so bili trendi v slovaropisju, ki se večinoma nanašajo na splet, npr. večji poudarek na vizualni predstavitvi (kar je bilo za slovenščino do neke mere narejeno pri projektu Franček). Drugi del dneva je bil namenjen seznanitvi s korpusi, zlasti z njihovo sestavo: zaradi visokih frekvenc manjšina besed zaseda večino korpusa (besede kot *biti*, *imeti*, *in*, *da*, *v*, *za*), večina besed pa je redkih, zato je pomembno imeti čim večji korpus. Z večjim korpusom namreč lažje pridobimo več redkih besed, prepoznamo enkratne oziroma priložnostne besede in jih ločimo od jedrnega besedja.

Dejan Gabrovšek ▪ ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana ▪ dejan.gabrovsek@zrc-sazu.si ▪ https://orcid.org/0000-0002-4995-1814

Prispevek je nastal v okviru programa P6-0038, ki ga financira ARRS.

CC BY NC ND

Začetek drugega dne je bil namenjen pregledu teorij, ki bodisi obravnavajo slovaropisje (metaleksikografija) bodisi prispevajo k bolj sistematičnemu opisu v slovarjih, to so zlasti teorije o pomenu besed. Od obravnavanih je v slovenščini verjetno najbolj znana Apresjanova šola, na kratko pa so bile obravnavani še teoretiki Charles J. Fillmore, Eleanor Rosch, George Lakoff in Igor Melčuk. V nadaljevanju je bil poudarek na računalniški obdelavi jezikov in predstavitvi statističnih metod za obdelavo jezika, npr. formule logDice za računanje trdnosti kolokacij.

Prvi del tretjega dne je bil posvečen regularnim izrazom in CQL (*Corpus Query Language*), računalniškemu jeziku, ki omogoča iskanje besednih oblik in skladenjskih enot v korpusu. CQL se je v dosedanjih skladenjskih raziskavah izkazal za zelo uporabnega in škoda je, da se ga ne uporablja pogosteje. Sledila je predstavitev dvojezične leksikografije, z vidika, da se od enojezične razlikuje po tem, da gesla nimajo razlag, ampak prevode in da se pomeni med besedami ne prekrivajo vedno, prav tako pa se ne prekrivajo kulture, zato določenih besed ni mogoče (natančno) prevesti, ampak jih je treba parafrazirati. Sledilo je predavanje o označevanju korpusov: prezapleten sistem oznak zelo zniža natančnost označevanja, zato mora biti relativno preprost, seveda pa se moramo raziskovalci teh poenostavitev zavedati in jih pri raziskovanju upoštevati. Zavedati se je tudi treba, da bodo v označevanju vedno napake, kljub temu pa jih je tako malo, da to večine analiz ne moti.

Četrti dan je bil v celoti posvečen slovaropisju. V prvem delu smo obravnavali razlage, pri čemer je bilo opozorjeno, da je treba najti razlago, ki bo dobro pokrila vse prototipne zglede, ne pa tudi vseh obrobij oziroma redkih rab ali celo potencialnih rab oziroma pomenskih odtenkov neke besede, saj je nemogoče pokriti celotno obrobje, ne da bi razlaga postala predolga in s tem nerazumljiva. Obrobni zgledi se namreč praviloma dobro asociativno nanašajo na osrednji pomen. Razlaga mora upoštevati tudi okoliščine rabe neke besede. Pomen je sicer zelo izmuzljiva entiteta, zato lahko isto besedo opišemo na različne načine. Raba sinonimov v razlagi je le redko uporabna. Vse to kaže, da je pisanje razlag zelo zahtevno. Sodobni trendi so, da mora biti razlaga kratka in razumljiva, nanašalne razlage pa se opuščajo (eSSKJ jih sicer do neke mere še vedno ohranja, nima pa razlag kot *glagolnik od X* ali *ženska oblika od X*).

Naslednje predavanje je bilo posvečeno kvalificiranju in kvalifikatorjem. Večina besedja je nezaznamovanega, zato posebnih oznak nima, kvalifikatorji pa označujejo odklone od nezaznamovanega.

Poleg kolokacij, ki so že dobro raziskan koncept, je bil omenjen še dokaj nov termin *koligacija*. Koligacija je sistem omejitev rabe določene besede in skladenjskih nizov. Večja količina podatkov je omogočila, da lahko raziskujemo, s katerimi drugimi besedami se določena beseda pogosteje druži, v katerih oblikah nastopa najpogosteje in kakšne daljše nize (razširjene kolokacije) tvori. Korpusni

pristop je prinesel spoznanje, da je v jeziku veliko ponavljanja daljših struktur. Koligacija se vsekakor kaže kot svež koncept, ki ga je smiselno prenesti tudi na slovenščino.

Posebno predavanje je bilo namenjeno zgledom v slovarjih. Dobre zglede iščemo predvsem s pomočjo GDEX-a (*good dictionary examples*), ki na vrh razvršča krajše zglede, take, ki ne vsebujejo sovražnega govora, nimajo redkih besed ali netipičnih znakov, seveda pa je presoja, kaj je dober zgled, še vedno zelo subjektivna in je računalnik ne more opraviti povsem zanesljivo.

Kot primer enostavnega spletnega programa za pisanje kratkih slovarjev smo spoznali program *Lexonomy* (https://www.lexonomy.eu/), ki je uporaben za manjše projekte, verjetno pa bi težko podpiral večje slovarje, ki hočejo prikazati več informacij.

Peti dan je bil namenjen spletnim virom, kot je *Wikipedija*, in slovarjem, ki jih urejajo uporabniki sami, npr. *Urban Dictionary*. Predstavljene so bile njihove prednosti (hitra odzivnost, dobra pokritost specializiranih področij) in slabosti (nekatera področja niso pokrita, možnost zlorab). Mnogo slovarjev (tudi portal Fran) ima možnost, da uporabniki predlagajo nove besede. Tako je, na primer, v enega od angleških slovarjev prišla beseda *selfie*.

Naslednje predavanje je bilo namenjeno avtorskim pravicam: te so v spletnem okolju povsem neurejene, zakonodaja pa je zastarela. Zadnje poglavje je bilo posvečeno računalniškemu tvorjenju slovarjev, ki se kaže kot perspektivno, čeprav še ne dovolj natančno. Skupna ugotovitev je bila, da dela za leksikografe ne bo zmanjkalo, računalniki pa nam pomagajo predvsem pri enostavnejših in monotonih opravilih.

Enotedenska delavnica je bila zame zelo koristna, saj sem lahko spoznal najnovejše trende v slovaropisju in korpusnem jezikoslovju ter dobil širši pogled na celotno slovaropisno pokrajino, ki ga doslej nisem imel. Predavanja so se enakomerno dopolnjevala z vajami, tako da znanje ni ostalo le teoretično, ampak smo ga lahko utrdili tudi praktično.

Naslednji Lexicom bo potekal od 11. do 15. septembra 2023 v Cambridgeu. Priporočam udeležbo vsem, ki se zanimajo za slovaropisje in korpusno jezikoslovje.

Povezave

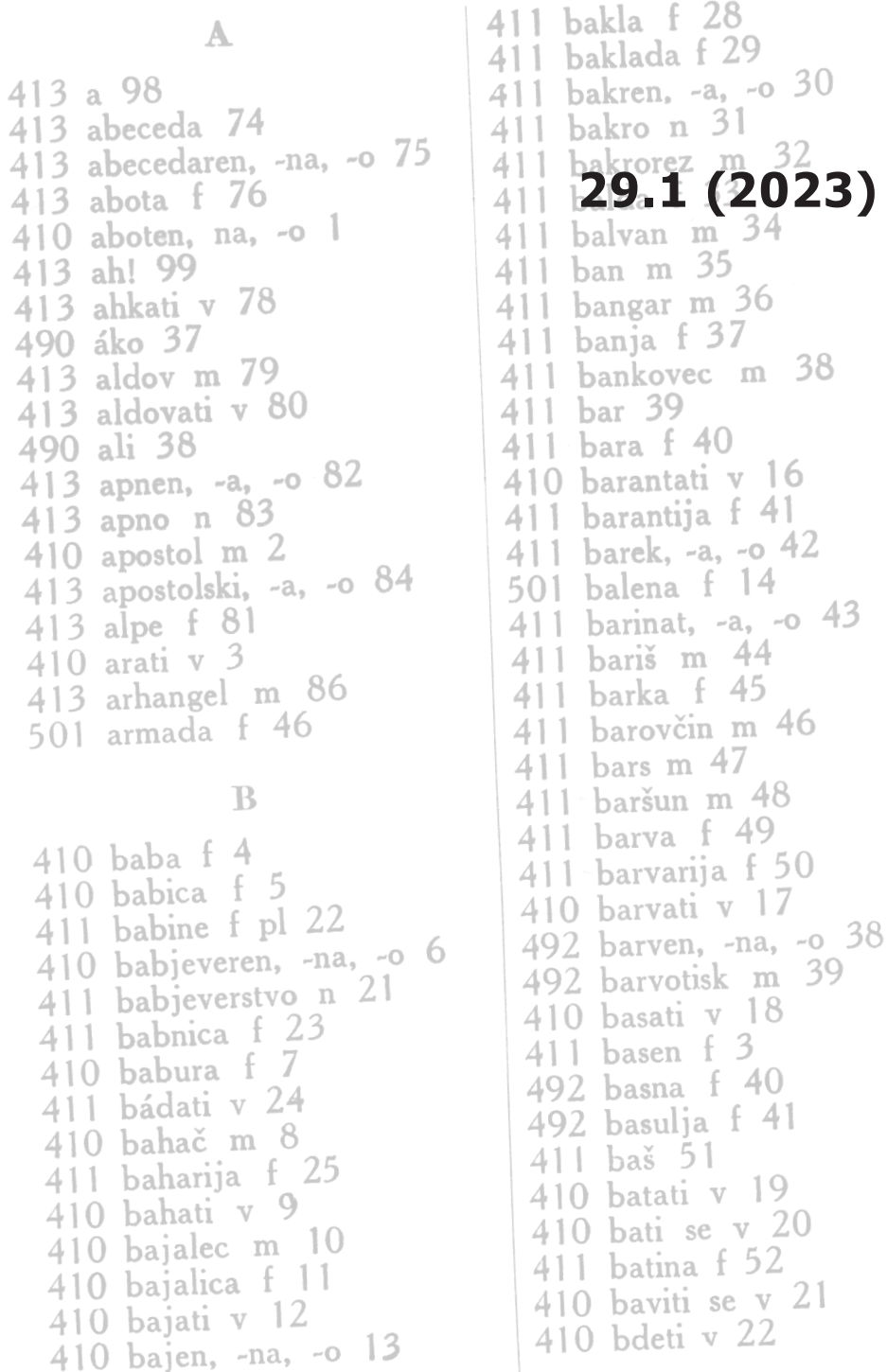
Lexonomy: https://www.lexonomy.eu/.
Sketch Engine: https://www.sketchengine.eu/.
Stran z informacijami za Lexicom 2022: https://lexicom.courses/upcoming-lexicom/.
Stran z informacijami za Lexicom 2023: https://lexicom.courses/upcoming-lexicom/.

Založba ZRC

Revija Inštituta za slovenski jezik Frana Ramovša ZRC SAZU

Jezikoslovni zapiski

29.1 (2023)



Jezikoslovni zapiski 29.1 (2023)

ISSN 0354-0448

Uredniški odbor **Hubert Bergmann, Metka Furlan, Alenka Jelovšek, Mateja Jemec Tomazin, Karmen Kenda-Jež, Valerij M. Mokijenko, Alenka Šivic-Dular, Andreja Žele**
Urednik **Peter Weiss**
Tehnična urednica **Alenka Jelovšek**
Prevod izvlečkov in povzetkov v angleščino **Donald Reindl, DEKS, d. o. o.**

Naslov uredništva **Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti (ZRC SAZU)**
Inštitut za slovenski jezik Frana Ramovša
Novi trg 4, SI-1000 Ljubljana, Slovenija
Telefon **+386 1 4706 160**
peter.weiss@zrc-sazu.si, isj@zrc-sazu.si
http://ojs.zrc-sazu.si/jz
http://bos.zrc-sazu.si/knjige/index.html

Izdal **ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša**
Založila **Založba ZRC**
Zanju **Oto Luthar, Kozma Ahačič**
Glavni urednik **Aleš Pogačnik**

Prelom **Simon Atelšek**
Oblikovanje **Evita Lukež**
Tisk **Cicero, Begunje, d. o. o.**
Naklada **200 izvodov**

Letna naročnina **10 €**
Letna naročnina za študente **8 €**
Cena posamezne številke **7 €**
Cena dvojne številke **12 €**
Naročila sprejema **Založba ZRC, p. p. 306, 1001 Ljubljana, Slovenija**
Telefon **+386 1 4706 464**
zalozba@zrc-sazu.si

JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE

Revija izhaja s podporo
Javne agencije za raziskovalno dejavnost Republike Slovenije.

Jezikoslovni zapiski so uvrščeni v mednarodne zbirke podatkov MLA International Bibliography of Books and Articles on the Modern Languages and Literatures, New York, ZDA; Bibliographie linguistique / Linguistic bibliography, The Hague, Nizozemska; IBZ, K. G. Saur Verlag, Osnabrück, Nemčija; New Contents Slavistics, Staatsbibliothek zu Berlin, Nemčija.

To delo je na voljo pod pogoji slovenske licence Creative Commons 4.0, ki ob priznavanju avtorstva dopušča nekomercialno uporabo, ne dovoljuje pa nobene predelave.

CC BY NC ND