SLOVENŠČINA 2.0: COLLOCATIONS IN LEXICOGRAPHY: EXISTING SOLUTIONS AND FUTURE CHALLENGES

Iztok KOSEM
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute

Polona GANTAR
Faculty of Arts, University of Ljubljana

Kosem, I., Gantar, P. (2020): Slovenščina 2.0: Collocations in Lexicography: existing solutions and future challenges. Slovenščina 2.0, 8(2): i–vi.

DOI: https://doi.org/10.4312/slo2.0.2020.2.i-vi

Collocations have become an increasingly popular topic of lexicographic research and resources in recent years, something that has been also facilitated by the rapid progress in the field of electronic lexicography. There are ongoing debates about what a collocation actually is, what is its relation to other multiword expressions, how much collocational data should be included in the dictionaries and how it should be presented, and how collocational information should be encoded to make it useful for different purposes. This has prompted us to organize a workshop centred around the topic of collocations. The workshop was collocated with the eLex 2019 conference in Sintra, Portugal. 14 different presentations were given at the workshop, offering an insight into the work on collocation at different institutions around the world. The presentations sparked interesting and thought-provoking discussions, and it was clear that a publication was needed to present the state-of-the-art on collocation in more detail. This led to the preparation of this special issue of the journal Slovenščina 2.0, which contains seven contributions based on the workshop presentations. The contributions cover a wide range of topics related to collocations, in six different languages, giving this special issue a truly international focus and relevance.

The first two papers deal with the definition of collocation, but from two different perspectives. **Iztok Kosem**, **Simon Krek** and **Polona Gantar** provide

a definition of collocation, and the classification of collocation in the typology of word combinations. Motivated by the use of collocational data for lexicographic purposes, they present the main criteria that define collocation on the one hand, and describe the main features that distinguish them from other word combinations on the other. Another, but equally important perspective to defining collocation is offered by **Toma Tasovac**, **Ana Salgado** and **Rute Costa** who focus on the modelling and encoding of polylexical units, including collocations, with TEI Lex-0, using the Dictionary of the Portuguese Academy of Sciences as a case study. Given that the existing TEI Guidelines do not address the encoding of polylexical units in sufficient detail, this paper is a very important and much needed contribution to the fields of lexicography and digital humanities.

The next three papers cover three different aspects of collocations in the lexicographic workflow. **Maria Khokhlova** and **Vladimir Benko** present a study on Russian data in which the role of corpus size in the identification of collocations is examined. In addition to determining the minimum size of a corpus for collocational research, they analyse and compare the suitability of four different association measures for extracting collocations from corpora of different sizes. **Lana Hudeček** and **Milica Mihaljević** present the treatment of collocations in the Croatian Web Dictionary called *Mrežnik*, showing detailed examples of the collocational block, with supporting questions and phrases, for different types of headwords. Their paper also addresses methodological questions such as how to define collocation for such a project, and how to address the issues related to the unrepresentative nature of corpus data. **Sanni Nimb**, **Nicolai Hartvig Sørensen** and **Henrik Lorentzen** look at the dictionary post-publication stage, in particular at the role of collocational changes in the detection of new meanings, which can then be translated into the updates of the Danish monolingual dictionary. They present the results of a corpus study in which automatic extraction methods using bigrams were combined with manual annotations.

The paper by **Ene Vainik, Maria Tuulik** and **Kristina Koppel** brings the psycholinguistic perspective by comparing word associations with collocations in the Estonian language, with special emphasis on the role of different parts of speech. They indicate the potential applications of word associations

in lexicography, e.g. in writing definitions, and in language learning. The final paper of the issue by **Eva Pori, Jaka Čibej, Iztok Kosem** and **Špela Arhar Holdt** offers insights into the user evaluation of an automatically compiled Collocations Dictionary of Modern Slovene. Considering that automatic extraction methods are becoming more and more common in modern lexicography, it is useful to learn how different types of users, in this case, teachers, translators, proofreaders, and lexicographers, have reacted to the use of a dictionary containing rich, but sometimes problematic, collocational data.

SLOVENŠČINA 2.0: KOLOKACIJE V LEKSIKOGRAFIJI: OBSTOJEČE REŠITVE IN IZZIVI ZA PRIHODNOST

Kolokacije so v zadnjih letih postale vse bolj priljubljena tema leksikografskih raziskav in z njimi povezanih virov, k čemur je pripomogel tudi hiter razvoj področja elektronske leksikografije. Številne diskusije potekajo o tem, kaj sploh je kolokacija, kako jo opredeliti do drugih večbesednih izrazov, koliko kolokacijskih podatkov vključiti v slovar, kako naj bodo predstavljeni uporabnikom ter kako kodirati kolokacijske podatke, da bodo uporabni za različne namene. Vse to nas je spodbudilo, da smo v okviru konference eLex 2019, ki je potekala v Sintri na Portugalskem, organizirali delavnico na temo kolokacij. Na delavnici je bilo predstavljenih 14 prispevkov, ki so ponudili vpogled v delo s kolokacijami na različnih ustanovah po svetu in sprožili vrsto zanimivih in stimulativnih razprav. Prav te razprave so spodbudile tudi potrebo po podrobnejšem opisu aktualnega stanja na področju kolokacijskih raziskav v samostojni publikaciji. Rezultat teh prizadevanj je pričujoča tematska številka revije Slovenščina 2.0 s sedmimi prispevki, ki izhajajo iz predstavitev na delavnici. Prispevki naslavljajo širok nabor tem v šestih različnih jezikih, zaradi česar je tematska številka res mednarodna, tako v zastopanosti kot relevantnosti obravnavanih tem.

Prva dva prispevka se lotevata opredelitve kolokacije z dveh različnih perspektiv. **Iztok Kosem, Simon Krek** in **Polona Gantar** opredelijo kolokacijo in njeno umestitev v tipologiji besednih kombinacij. Glavno vodilo pri tem je uporaba kolokacijskih podatkov za leksikografske namene, na podlagi katerega predstavijo tri glavne kriterije pri opredelitvi kolokacije in tudi glavne lastnosti, ki ločijo kolokacije od drugih besednih kombinacij. Drugačno, a enako pomembno perspektivo pri opredelitvi kolokacije predstavijo **Toma Tasovac, Ana Salgado** in **Rute Costa** s prispevkom o modeliranju in kodiranju večbesednih leksikalnih enot, vključno s kolokacijami, v formatu TEI Lex-0, pri čemer kot testni primer vzamejo Slovar Portugalske akademije znanosti. Glede na to da v obstoječih smernicah TEI kodiranje večbesednih leksikalnih enot ni dovolj poglobljeno predstavljeno, gre za zelo pomemben in dragocen prispevek tako za leksikografijo kot tudi digitalno humanistiko.

Sledijo trije prispevki, ki predstavljajo tri različne stopnje v postopku izdelave slovarskih virov. **Maria Khokhlova** in **Vladimir Benko** predstavita študijo na podlagi ruščine, v kateri preučujeta vlogo velikosti korpusa pri luščenju kolokacij. Določiti skušata minimalno velikost korpusa, ki je še ustrezna za kolokacijske raziskave, analizirata in primerjata pa tudi ustreznost štirih različnih statističnih mer pri luščenju kolokacij iz korpusov različnih velikosti. **Lana Hudeček** in **Milica Mihaljević** predstavita obravnavo kolokacij v Hrvaškem spletnem slovarju Mrežnik, ki vključuje prikaz različnih vprašanj in fraz za posamezne tipe kolokacij pri iztočnicah različnih besednih vrst. Avtorici se dotakneta tudi metodoloških vprašanj, kot je na primer opredelitev kolokacije za namene splošnega izhodiščno digitalno zasnovanega slovarja in reševanje problemov, povezanih s slabo reprezentativnostjo korpusnih podatkov. **Sanni Nimb, Nicolai Hartvig Sørensen** in **Henrik Lorentzen** raziskujejo možnosti uporabe kolokacijskih podatkov pri posodabljanju obstoječega danskega enojezičnega slovarja, zlasti vlogo sprememb v rabi kolokacij pri prepoznavi novih pomenov z namenom ugotoviti uporabnost postopka pri pripravi slovarskih posodobitev. V prispevku predstavijo rezultate korpusne raziskave, v kateri so uporabili kombinacijo avtomatskega luščenja bigramov in njihove ročne anotacije s strani leksikografov.

Prispevek **Ene Vainik, Marie Tuulik** in **Kristine Koppel** s primerjavo besednih asociacij in kolokacij v estonščini s poudarkom na vlogi besednih vrst prinaša tematski številki psiholingvistično perspektivo. Avtorice med drugim ponudijo razmisleke o izrabi rezultatov študije na področju leksikografije, npr. pri pisanju pomenskih definicij in pri poučevanju tujih jezikov. Tematsko številko sklene prispevek **Eve Pori, Jake Čibeja, Iztoka Kosma** in **Špele Arhar Holdt** o uporabniški evalvaciji avtomatsko izdelanega Kolokacijskega slovarja sodobne slovenščine. Metode avtomatskega luščenja podatkov so v sodobni leksikografiji vse pogosteje uporabljane, zato je koristno opazovati in analizirati odzive različnih tipov uporabnikov, v tem primeru učiteljev, prevajalcev, lektorjev in leksikografov pri uporabi slovarja, ki vsebuje sicer številne, a včasih problematične kolokacijske podatke.