eReL

E-rječnici i e-leksikografija
E-dictionaries and E-lexicography

Knjižica sažetaka
Book of abstracts

Zagreb
10. – 11. svibnja 2019.

Izdavač / Publisher
Institut za hrvatski jezik i jezokoslovlje
Zagreb, Ulica Republike Austrije 16

Skup / Conference
E-rječnici i e-leksikografija

Uredile / Edited by
Lana Hudeček
Milica Mihaljević
Ivana Brač

Recenzenti
Petra Bago
Matea Birtić
Goranka Blagus Bartolec
Polona Gantar
Lana Hudeček
Nataša Jermen
Željko Jozić
Barbara Kovačević
Simon Krek
Kristian Lewis
Nikola Ljubešić
Ivana Matas Ivanković
Milica Mihaljević
Nives Mikelić Preradović
Christine Möhrs
Bruno Nahod
Dinka Pasini
Sanja Seljan
Kristina Štrkalj Despot
Toma Tasovac
Domagoj Vidović

Oblikovanje i priprema
Davor Milašinčić

ISBN 978-953-7967-71-0

Međunarodni znanstveni skup

E-rječnici i e-leksikografija*

Zagreb, 10. – 11. svibnja 2019.

eReL 2019.

*Skup je organiziran u sklopu istraživačkoga projekta *Hrvatski mrežni rječnik – Mrežnik*, koji financira Hrvatska zaklada za znanost (IP-2016-06-2141).

International conference

E-dictionaries and E-lexicography*

Zagreb, 10 – 11 May 2019

eReL 2019

*This conference was organized within the research project *Croatian Web Dictionary—MREŽNIK* (IP-2016-06-2141), financed by the Croatian Science Foundation.

Organizacijski odbor / Organizing Committee

Lana Hudeček (predsjednica/chair)
Željko Jozić
Branko Štefanović
Milica Mihaljević
Kristina Štrkalj Despot
Kristian Lewis (tajnik/secretary)
Ivana Brač (izvršni tajnik / executive secretary)
Jože Horvat (izvršni tajnik / executive secretary)
Daria Lazić (izvršni tajnik / executive secretary)
Josip Mihaljević (izvršni tajnik / executive secretary)
Siniša Runjaić (izvršni tajnik / executive secretary)

E-rječnici i e-leksikografija

10. i 11. svibnja 2019.
Zagreb

Cilj je skupa okupiti istraživače u području e-leksikografije iz zemlje i inozemstva te hrvatsku e-leksikografiju predstaviti široj javnosti i uvesti je u relevantne tijekove suvremene svjetske e-leksikografije.

Skupom su obuhvaćena ova tematska područja:

- e-rječnici
- e-enciklopedije
- korpus u leksikografiji i korpusna leksikografija
- terminološke i jezične baze podataka
- e-terminologija
- retrodigitalizacija
- računalni leksikografski alati.

Pozvani predavači:

- Stefan Engelberg (Institut für Deutsche Sprache, Mannheim)
- Simon Krek (Center za jezikovne vire in tehnologije Univerze v Ljubljani, Ljubljana)
- Polona Gantar (Filozofska fakulteta Univerze v Ljubljani, Ljubljana)
- Nikola Ljubešić (Institut Jožef Stefan, Ljubljana)

mrežnik

HRZZ
Croatian Science Foundation

VERN'

E-dictionaries and e-lexicography

Zagreb
10-11 May 2019

The international conference E-dictionaries and E-lexicography is held on 10-11 May 2019 in Zagreb. The aim of the conference is to bring together Croatian and foreign experts in e-lexicography, to present Croatian e-lexicography to the wider public, as well as to get acquainted with contemporary research and ideas in the field of e-lexicography.

The conference will focus on the following topics:

- e-dictionaries
- e-encyclopedias
- corpora in lexicography
- databases on language and terminology
- e-terminology
- retrodigitalization
- computer tools for lexicography.

Invited speakers:

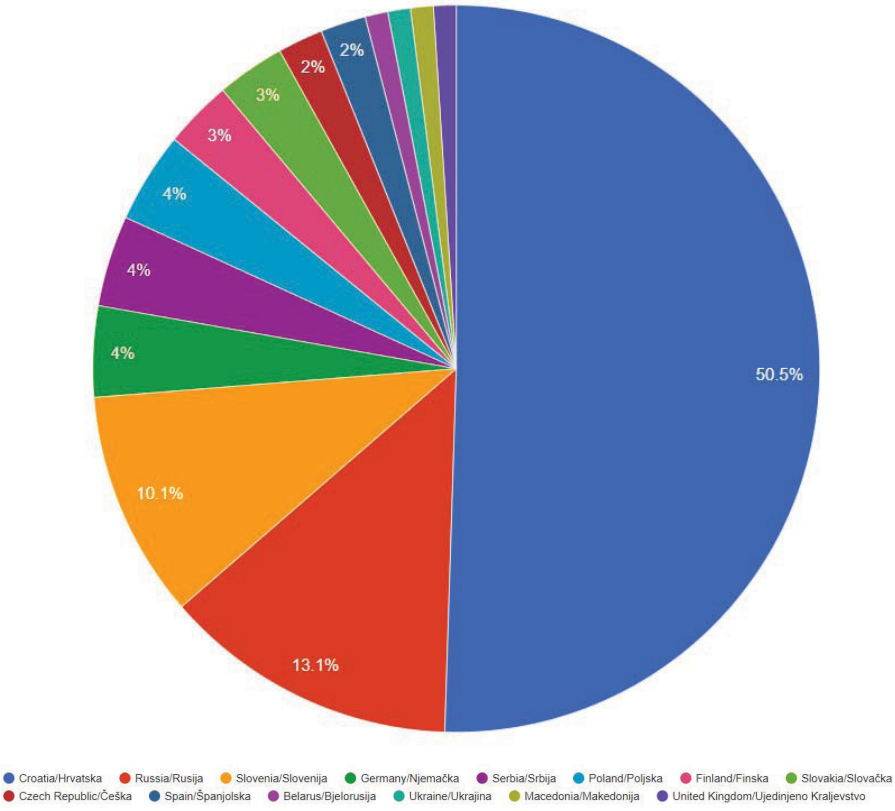
- Stefan Engelberg (Institut für Deutsche Sprache, Mannheim)
- Simon Krek (Center za jezikovne vire in tehnologije Univerze v Ljubljani, Ljubljana)
- Polona Gantar (Filozofska fakulteta Univerze v Ljubljani, Ljubljana)
- Nikola Ljubešić (Institut Jožef Stefan, Ljubljana)

mrežnik

HrZZ
Croatian Science Foundation

VERN'

Broj sudionika po zemljama
Number of participants by countries



mreznik

HrZZ
Croatian Science Foundation

VERN'

Stefan Engelberg
Institute for the German Language Mannheim
engelberg@ids-mannheim.de

What will we need dictionaries for (at all)? The Internet, lexicography and the primacy of the user

Lexicography has seen two important developments during the last twenty years, a medial one and a conceptual one. Firstly, lexicography has switched from printed to electronic dictionaries. Secondly, metalexicography has brought the needs of the dictionary user to the fore of dictionary theory. It seems that internet lexicography is the jubilant culmination of these two developments: multimedia complement the text-laden content of traditional dictionaries; new access structures allow faster and far more diversified access to information; dictionaries can be combined into large digital information bases, etc.

However, most internet dictionaries still look very conventional: here and there an audio file, a frequency chart or an extra corpus example, but that is usually the end of the adventure. And more importantly, not only did lexicography change during the last twenty years, the rest of the world and in particular the internet did too. While some decades ago everybody who wanted to find out something about a word would have to consult a dictionary, nowadays, the Internet can itself be used by everybody as a large lexical resource. It is common now to google words in order to find out about their commonality, to rely on electronic spell checkers, or to turn to random texts on the Internet in order to see which preposition has to go with which particular verb. And this is only the beginning of the development of numerous methods that everybody can apply in order to explore the lexical richness found on the Internet – without using dictionaries!

The talk argues that we, as lexicographers, have to watch these developments carefully, because they will dispense with a lot of services that have been traditionally provided by dictionaries. On the one hand, we will have to integrate lexicography and non-lexicographic internet-based lexical exploring strategies, and on other hand, we will have to focus on strengths that cannot easily be replaced by (semi-)automatic big data exploration.

Polona Gantar
Faculty of Arts, University of Ljubljana
apolonija.gantar@ff.uni-lj.si

Dictionary of Modern Slovene: From the Lexical Database to the Digital Dictionary Database for Slovene

The ability to process language data has become fundamental for the development of technologies in various areas of human life in the digital world. The development of computer-readable linguistic resources, methods and tools is therefore also one of the key challenges for contemporary Slovene language. This challenge has been recognized in the Slovene language community both at the professional as well as the state level, and has been the subject of quite a number of activities over the past ten years, which I would like to present in this paper.

The idea of a comprehensive dictionary database covering all levels of linguistic description of modern Slovene, from the morphological and lexical to the syntactic, had already been formulated in the framework of the European Social Fund project Communication in Slovene (2008-2013), within which the Slovene Lexical Database was created. When designing the Slovene Lexical Database (SLD), we pursued two goals: to create linguistic description of Slovene intended for human users which would also be useful for machine processing of Slovene. Ever since the construction of the first corpus of Slovene, it has become evident that there is a need for a description of modern Slovene based on real language data, and that it is necessary to understand the needs of language users in order to create useful language reference works. In addition, it became obvious that only the digital medium enables the comprehensiveness of language description and that the design of the database must be adapted to it from the start. Also, in terms of formats and international standards the description needs to follow good practices as closely as possible, as this enables the inclusion of Slovene into a wider network of resources, such as Open Linked data, BabelNet, ELEXIS, etc. Given/Due to time pressure and trends in lexicography, we had to consider the procedures for automating the extraction of linguistic data from corpora and the inclusion of crowdsourcing into the lexicographic process.

In accordance with the essential idea of creating an all-inclusive Digital Dictionary Database for Slovene, several independent databases have been created over the past two years. Specifically, the Collocations Dictionary of Modern Slovene, and the automatically generated Thesaurus of Modern Slovene – both also exist as independent online dictionary portals. One of the novelties that we put forward together with both dictionaries is the concept of the so-called 'responsive dictionary', which includes crowdsourcing methods. Ultimately, the Digital Dictionary Database provides all (other) levels of linguistic description: morphological, with the Sloleks database upgrade, phraseological, with the construction of a multi-word expressions lexicon, and syntactic, with the formalization of valency patterns of Slovene verbs. Each of these databases contains its own specific language data that will ultimately be included in the comprehensive Slovene Digital Dictionary Database, which will represent basic linguistic description of Slovene both for the human and machine user.

Simon Krek
Centre for language resources and technologies,
University of Ljubljana
simon.krek@guest.arnes.si

Dictionary Matrix vs. Matrix Dictionary (in ELEXIS)

In ELEXIS Horizon 2020 infrastructure project, the creation of a new lexicographic resource titled "matrix dictionary" was included in the proposal. The concept of "matrix dictionary" incorporates two importantly different aspects that were later re-defined and split into two entities that could be described as "ELEXIS dictionary matrix" and "ELEXIS matrix dictionary". The key aspect of the first – "dictionary matrix" – is providing extensive links between key structural elements found in different types of dictionaries. Therefore, it is focused on (direct or indirect) linking of existing lexicographic resources, either on the lemma level and other "simpler" levels, but also on the sense level by pivoting through one of the existing semantic resources – BabelNet. Ultimately, linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in different types of existing lexicographic resources, monolingual, multilingual, modern, historical etc. from dictionary matrix will be available as part of the infrastructure in ELEXIS.

Building upon the first, the goal of the second – "matrix dictionary" – is more ambitious and involves possible long-term development beyond the ELEXIS project. To quote from the proposal:

"Ultimate goal is the creation of a universal (integrated and enriched) registry/network of semantic relations used as a semantic intermediary language for global knowledge exchange, focused on difficult polysemous vocabulary (single-word and multi-word), modern and historical; the realisation of a universal lexicographic metastructure; a matrix dictionary spanning across languages and time."

To borrow from the goal followed by Universal Dependencies project ("cross-linguistically consistent grammatical annotation"), and extending it from syntax to semantics, one could describe this entity as "Universal

Concepts", empirically verifiable through massive linking available in the ELEXIS dictionary matrix. As part of its research agenda, ELEXIS project will primarily provide a testbed for "Universal Concepts" and – if proven viable and useful – a community for its further development. The talk will briefly present both challenges: how to link existing lexicographic resources and how to understand the (degree of) universality of a "universal concept".

Nikola Ljubešić
Jožef Stefan Institute
Faculty of Computer and Information Science,
University of Ljubljana
nljubesi@gmail.com

"Deep lexicography" - fad or opportunity?

In recent years we are witnessing staggering improvements in various intelligent data processing tasks due to the developments in the area of deep learning, ranging from image and video processing to speech processing and natural language understanding. In this talk I want to discuss the opportunities and challenges that these developments pose for the area of electronic lexicography.

The most of my talk I will discuss the concept of representation learning of various elements of language, namely words, lexemes and utterances, and their applicability to lexicography. I will start with the well known approaches to learning static representations of words, the so called word embeddings, and their usage in tasks such as semantic shift detection and prediction of specific lexical features. I will continue with touching upon the multilingual dimension of representation learning and wrap up with the most recent developments in natural language understanding in form of learning dynamic, context-aware representations of words.

Špela Arhar Holdt
Centre for language resources and technologies,
University of Ljubljana
spela.arhar@cjvt.si

Jaka Čibej
Faculty of Computer and Information Science,
University of Ljubljana
jaka.cibej@cjvt.si

How users responded to a responsive dictionary: The case of the Thesaurus of Modern Slovene

In March 2018, the Centre for Language Resources and Technologies of the University of Ljubljana published the *Thesaurus of Modern Slovene* (Krek et al. 2017a), the largest automatically generated collection of Slovene synonyms. The thesaurus introduced a new type of dictionary called the *responsive dictionary* (Arhar Holdt et al. 2018): it is designed from the onset to be entirely digital, is initially compiled through automatic extraction methods (Krek et al. 2017b), and is freely available as both an online language resource and as a database for the further development of tools and resources. Its most defining characteristic is its ability to quickly and flexibly respond to both language changes and the feedback provided by its users; in the case of the thesaurus, users can contribute by adding suggestions for missing synonyms and by up- or down-voting existing synonym candidates.

As the first example of a responsive dictionary, the thesaurus differs in many aspects from traditional Slovene lexical resources, which raises the question of how dictionary users view these innovations (e.g. data is extracted automatically and includes some noise; non-experts are involved in dictionary compilation; the resource is never truly finished). Has the community embraced this concept or is it apprehensive? With financial support from the Slovene Ministry of Culture, a survey was conducted in order to gauge user opinions and collect suggestions for future dictionary improvements. The results (N=671) show that most users rate the innovations introduced by the responsive dictionary (especially the ones

dealing with user involvement) as positive, but are more apprehensive when it comes to data reliability (which is in line with the findings of similar studies on user preferences and digital dictionaries; see e.g. Müller-Spitzer, 2014, Arhar Holdt 2018). The paper presents the results of the survey together with a discussion on how the implementation of novelties could be improved to better meet user expectations and needs.

Vladimír Benko
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
vladimir.benko@juls.savba.sk

Omnia Slovaca and *Omnia Bohemica*: Big Enough or Too Big?

Since the advent of *Web as Corpus* (*WaC*) technology, the size of corpora for many languages has crossed the gigaword threshold. Moreover, traditional (albeit unbalanced) corpora often also reach the same order of magnitude. Slovak and Czech are languages with both web and traditional corpora of this size. Furthermore, open-source textual data (such as Wikipedia, Open Subtitles Database, etc.) for both languages are also available. The idea of using all of these resources to make even larger corpora is quite natural.

Omnia Slovaca has been created by merging all Slovak corpus data available, i.e. the *Slovak National Corpus* supplemented with web data crawled by several NLP groups (Masaryk University in Brno, Slovak Academy of Sciences in Bratislava and the Technical University in Košice), the Slovak *Wikipedia*, the Slovak component of the *Open Subtitles Database*, and data provided by European institutions). All data has been retokenised, deduplicated, retagged, and processed by *Sketch Engine*, resulting in a homogeneous resource suitable for lexicographic purposes. The current corpus size is approx. 5 gigatokens, and we expect to increase it by 500 million tokens after the newer edition of the *SNC* and data from the 2018 crawl are incorporated.

Omnia Bohemica is a Czech corpus developed in a similar way, the main difference being its size – the second version is extremely large at more than 12 gigatokens.

With such huge amounts of texts, on the one hand, the lexicographers can find enough lexical evidence for low-frequency language phenomena (such as idioms). However, on the other hand, they face the rather serious problem of being "flooded" by data for even medium-frequency lexical items. Our presentation will cover some of the problems and their solutions.

Matea Birtić
Institut za hrvatski jezik i jezikoslovlje
mbirtic@ihjj.hr

Usporedba mrežnih valencijskih rječnika u slavenskim jezicima

Leksikografski prikaz glagolske valentnosti u valencijskim mrežnim rječnicima slavenskih jezika pokazuje katkada neke specifičnosti u odnosu na prikaz glagolske valentnosti ili spojivosti riječi uopće u poznatim mrežnim izvorima (*FrameNet*, *VerbNet*, *PDE*). Razlog tomu je drukčija jezikoslovna tradicija, pa posljedično tomu i odabir modela, ali i sam ustroj slavenskih jezika. U ovome će se izlaganju pokušati istražiti na temelju pet valencijskih e-rječnika stvorenih za opis slavenskih glagola koje su to sastavnice koje uvjetuju drukčiji prikaz valentnosti te je li drukčiji prikaz doista nužan. Mrežni su valencijski rječnici na kojima će se temeljiti istraživanje ruski *FrameBank*, češki *Vallex*, poljski *Walenty* te dva mrežna valencijska rječnika hrvatskoga jezika, *Crovallex* i *e-Glava*. Osobito će se pozorno preispitati odnos semantike, sintakse i morfologije u opisu glagola u nabrojenim e-rječnicima te će se pokušati izdvojiti prednosti i mane određenih modela za opis glagolske valentnosti u slavenskim jezicima.

A comparison of online Slavic valency dictionaries

The lexicographic representation of verb valency in online valency dictionaries of Slavic languages differs from the representation of verb valency and general combinatorial properties of words in other well-known online resources (*FrameNet*, *VerbNet*, *PDE*). The reason for this is a various linguistic tradition, and consequently the choice of model, as well as the very structure of the Slavic languages. On the basis of five valency e-dictionaries created to describe Slavic verbs, this presentation will attempt to explore which are key features of the different representation of valency in Slavic languages, as well as whether this different representation is actually necessary. The analysis will be based

on the following online valency dictionaries: the Russian *FrameBank*, the Czech *Vallex*, the Polish *Walenty*, and two online valency dictionaries of Croatian, *Crovallex* and *e-Glava*. The relationship between a verb's semantic, syntactic, and morphological description will be particularly closely examined for each online dictionary. Likewise, the advantages and disadvantages of certain valency description models for the Slavic languages will be identified.

Goranka Blagus Bartolec
Institut za hrvatski jezik i jezikoslovlje
gblagus@ihjj.hr

Kolokacijski potencijal glagolsko-prijedložnih sveza u hrvatskome jeziku

Imenice ili prijedložno-padežni izrazi kojima mjesto u sintaktičkom nizu otvaraju određeni glagoli u gramatičkim se opisima, ponajprije s aspekta teorije valentnosti, najčešće smatraju obveznim (najčešće objektnim) dopunama kojima upravlja glagol. Na leksičkoj razini, pak, s obzirom na ustaljenost u upotrebi i nezamjenjivost sastavnica, takve strukture imaju obilježja glagolskih kolokacija koje se u upotrebi uzimaju u gotovu obliku. Osnovni tip glagolskih kolokacija u hrvatskome temelji se na strukturi glagol + imenica u akuzativu (preuzeti dužnost, snositi odgovornost, skuhati ručak). Osim takve strukture, kao posebna izdvaja se struktura glagol + prijedložno-padežni izraz (uzeti u obzir, staviti na raspolaganje, pogledati na sat) unutar koje se kao temeljne izdvajaju glagolska i prijedložna sastavnica, dok imenička sastavnica, ovisno o komunikacijskom kontekstu, može biti promjenjiva: otići po (dijete/kruh/lijek/novine/pomoć/stvari). Takvim se glagolsko-prijedložnim svezama u hrvatskome izražavaju različiti konkretni i apstraktni značenjski odnosi, a kao najčešće izdvajaju se dvije strukture: ponavljajuća prefiksalno-prijedložna struktura (doći do /koga, čega/, izići iz /čega/, naići na /koga što/, sastati se s /kim, čim/) i struktura s različitim glagolsko-prijedložnim obrascima s prefiksalnim (podsjećati na /koga, što/, razgovarati o /kom, čem/, zaljubiti se u /koga, što/) ili neprefiksalnim glagolima (misliti o /kom, čem/, temeljiti se na /čemu/). U hrvatskom jeziku takvih je sveza mnogo, a e-rječnici i e-baze zbog neograničenih mogućnosti unosa podataka plodan su prostor za njihovo bilježenje.

U radu će se na primjerima glagolsko-prijedložnih sveza preuzetih iz Kolokacijske baze hrvatskoga jezika Instituta za hrvatski jezik i jezikoslovlje (http://ihjj.hr/kolokacije/) istražiti sljedeće: 1. koji su glagoli s obzirom na svoj primarni značenjski i sintaktički potencijal nositelji takvih struktura, 2. koji su prijedlozi, s obzirom na svoj značenjski potencijal i sintaktičke uloge, najčešće sastavnice takvih sveza, 3. značenjski potencijal glagolsko-prijedložnih kolokacija u hrvatskom jeziku.

The collocational potential of verbal prepositional phrases in Croatian

Grammatical descriptions usually consider nouns or prepositional-case phrases followed by a verb in the sentence order obligatory (usually object) complements governed by the verb, especially in accordance with the valency theory. At the lexical level, with regard to the stability of usage and constancy of the constituents, these structures have the characteristics of verbal collocations. The main type of verbal collocation in Croatian is based on the structure 'verb + accusative noun' (*preuzeti dužnost* 'to take charge of', *snositi odgovornost* 'bear responsibility', *skuhati ručak*, 'to make/cook lunch'). In addition to this structure, 'verb + prepositional case phrase' can also be considered a specific structure (*uzeti u obzir* 'take into consideration', *staviti na raspolaganje* 'put at disposal, *pogledati na sat* 'look at / watch the clock') in which the verb and prepositional constituents serve as the core, while the noun component may vary depending on the communication context: *otići po (dijete/kruh/lijek/novine/pomoć/stvari)* 'to go get (a child/bread/medicine/newspaper/help/things)'. This kind of verbal-prepositional phrase in Croatian is expressed with different concrete and abstract meanings, and these two patterns can be distinguished: repetitive prefixal-prepositional structures (*doći do* 'come to', *izići iz* 'go out', *naići na* 'come across', *sastati se s* 'meet up with') and structures with different verbal-prepositional patterns with prefixal verbs (*podsjećati na* 'remind about/of', *razgovarati o* 'talk about', *zaljubiti se u* 'fall in love with'). Verbal-prepositional phrases are common in Croatian, and e-dictionaries and e-databases are a fruitful space in which to record them.

This research, based on verbal-prepositional examples from the Institute of Croatian Language and Linguistics' Croatian Collocation Database (http://ihjj.hr/kolokacije/), will explore the following: 1) which verbs, given their primary semantic and syntactic potential, are the bearers of these structures, 2) which prepositions, given their primary semantic potential and syntactic roles, are the most common components of these collocational phrases, and 3) the semantic potential of verbal prepositional collocations in Croatian.

Constantine Borissoff
Infracloud Limited
leo@gir.me.uk

Web based dictionary publishing system

A web-based dictionary compiling and publishing software was developed for the Russian-Sanskrit Comparative Dictionary. The back-end core part of the dictionary is a Structured Query Language (SQL) database. It can be presented as a table consisting of columns and rows of unformatted data. This architecture allows data to be entered into an unlimited number of fields. The data can be edited and retrieved in various ways to create programmable formatted text outputs. The front end consists of several screens for entering, editing, and retrieving textual data. The most important of these is the editing interface. For this specific application it has 35 text boxes linked to relative fields in the database. The interface also has tick boxes and drop down menus for adding various useful tags and labels. Specially made software is used to extract information from the designated cells to compile formatted dictionary entries and visualise them in HTML format. One unique feature of the dictionary system is the ability to generate a LaTeX file, which can be converted to produce a ready-to-publish dictionary in PDF format. Currently, the system is set to create output in the Memoir Class configurable typesetting style, but this can be easily changed. Another advantage of the system is the use of an assembly of open source fonts which allows almost any symbol to be printed using XeLaTeX technology. The software is web-based and provides online access for remote input by an unlimited number of users. The system administrator can assign various user permissions and privileges and access an automated log to control the compiling and editing process. This dictionary compiling and publishing software is a useful, user-friendly tool that can be configured to create both conventional and e-dictionaries in a multitude of languages.

Ivana Crljenko
Leksikografski zavod Miroslav Krleža
ivana.crljenko@lzmk.hr

Hrvatski egzonimi na putu prema mrežnom izdanju

U referatu će se predstaviti projekt *Rječnik stranih geografskih imena*, koji se od 2013. provodi u Leksikografskom zavodu Miroslav Krleža. Cilj projekta je iz dvadesetak relevantnih izvora prikupiti i atribuirati prilagođena zemljopisna imena za zemljopisne objekte smještene izvan hrvatskoga jezičnoga područja (egzonime). U mnogim hrvatskim izvorima egzonimi su zapisani različitim likovima, uglavnom nedosljedno, nesustavno i neujednačeno. Stoga *Rječnik* nudi preporučene jedinstvene prilagođene likove za opću uporabu. Svrha *projekta je* olakšati njihov izbor onda kada se rabi prilagođeni, a ne izvorni lik. Namjera je i stvoriti temelje za standardizaciju hrvatskih egzonima.

Premda se građa ne obrađuje na uobičajen rječnički način, rezultati projekta najsličniji su rječničkoj formi. Objavljeni su u dvama tiskanim priručnicima. *Hrvatski egzonimi I.* (2016.) bave se imenima država, pripadnih glavnih gradova i ovisnih područja, jer su mnogi od njih egzonimi. Obrađena su i imena stanovnika (etnici), odnosni pridjevi, genitivi i lokativi imena država i glavnih gradova jer se ti oblici riječi često pojavljuju u općoj uporabi. *Hrvatski egzonimi II.* (2018.) donose popis više od 3000 suvremenih i povijesnih egzonima za sve tipove zemljopisnih objekata, pripadna izvorna imena, jezik (jezike) izvornoga imena, tip i podtip imenovanoga zemljopisnog objekta te njegovu lokacija.

Projekt ima potencijal dalje se razvijati u *E-rječnik hrvatskih egzonima*. Ideja je pretvoriti postojeću internu bazu podataka u lako pretraživu web aplikaciju koja će sadržavati potvrde pronađene u svim analiziranim izvorima, kao i sve ostale atribute. Kako bismo to ostvarili, namjeravamo se oslanjati na znanja, iskustva i prakse e-leksikografa i računalnih jezikoslovca.

Croatian Exonyms toward its Internet Edition

The paper will present the *Dictionary of Foreign Geographical Names* project, which has been conducted at the Miroslav Krleža Institute of Lexicography since 2013. It is aimed at listing and attributing the adapted place names for the geographical features situated outside the Croatian language area (exonyms) collected from approximately twenty relevant sources. In different Croatian sources, exonyms are written in multiple forms, which makes their writing inconsistent and unsystematic. The *Dictionary* thus offers recommended unique forms of the adapted names for general use. The purpose of the project is to facilitate the selection of forms in cases where an exonym is used instead of the original name. An additional goal is to create a foundation for standardisation of Croatian exonyms.

Although lexical items are not treated in the usual dictionary mode, the results of the project are most similar to the form of a dictionary. They have been published in two printed reference books. *Croatian Exonyms I* (2016) deals with the names of countries, capital cities, and dependent territories, many of which are exonyms. Since they frequently appear in common usage, names of inhabitants, relational adjectives, and genitive and locative forms of country names and capitals are also presented in the book. *Croatian Exonyms II* (2018) lists more than 3,000 current and historical adapted names for all types of geographical features, as well as their original names, the language(s) of the original name(s), the type and subtype of the named geographical feature, and its location.

The project has the potential for further development as an e-dictionary of Croatian exonyms. The idea is to convert the existing database into an easily searchable web application containing attestations from all analysed sources, as well as other attributes. To achieve this goal, we intend to rely on the knowledge, experience, and practices of e-lexicographers and computer linguists.

Маргарита И. Чернышева
Институт русского языка им. В.В. Виноградова РАН
chernysheva@bk.ru

Елена И. Державина
Институт русского языка им. В.В. Виноградова РАН
e_derzhavina@mail.ru

Электронный словарь грецизмов в русском языке XI – XVII вв.

Проект ориентирован на создание первого в отечественной науке «Электронного словаря грецизмов в русском языке XI–XVII вв.» (ЭСгр) с базой данных. Работа рассчитана на три года (2018 – 2020). С помощью созданного в рамках проекта современного инструментария впервые будет представлен в электронном виде полный объем грецизмов, которые встретились в памятниках русской письменности XI–XVII вв., однако до сих пор не получили исчерпывающего описания. Создание электронного словаря ЭСгр будет сопровождаться проведением историко-лексикологических, сопоставительно-типологических и культурно-исторических исследований. Многоаспектное представление грецизмов в электронном виде позволит выполнять разнообразные поисковые операции.

На сегодняшний день база данных включает следующие поля: **заголовочное слово (лемма) по-русски** (отдельную позицию занимают дериваты); **заголовочное слово (лемма) по-гречески**; **часть речи**; **фонетическая или графическая вариантность**; **морфологическая вариантность**; **этимология**; **семантика**; **функционирование** – указание на употребление в переводных и/или оригинальных сочинениях; **датировка**; **цитата первого** (или единственного – в случае hapax legomenon) **употребления** в памятниках письменности; **фиксация в исторических словарях русского и славянских языков**; **связывающие отсылки** (для разного рода вариантов); **тематическая группа**; **дополнительная филологическая информация** любого рода (например, наличие переводного эквивалента, описательного, этимологического перевода или глоссированного толкования); **дополнительная историко-**

культурная информация (особенно важная для объяснения редких слов и реалий).

Таким образом, будучи словарем нового типа, ЭСгр станет основой для осуществления дальнейших историко-лексикологических, сопоставительно-типологических и культурно-исторических исследований.

Результаты работы по созданию «Электронного словаря грецизмов в русском языке XI–XVII вв.» будут размещены в свободном доступе в Интернете.

The Electronic Dictionary of Greek words in 11th- to 17th-century Russian

The purpose of this project is to create the first digital dictionary of Greek words in 11th- to 17th-century Russian (EDGr) with a database. Work is envisioned to last three years (2018 – 2020). The modern tools created within the framework of this project will allow the first digital presentation of the extensive volume of Greek words present in 11th-17th-century Russian literary monuments, which have not yet been fully described. Research in the fields of historical lexicology, comparative typology, and cultural history will accompany the creation of the EDGr. The multi-aspectual representation of the accumulated information in digital form will allow a variety of search operations.

The database includes the following fields: **lemma in Russian** (a separate position is occupied by derivatives); **lemma in Greek**; **part of speech**; **phonetic or orthograpic variation**; **morphological variation; etymology; semantics; indication of use** in the translated and/or original works; **dating**; **first attestation** (or the only attestation in the case of hapax legomenon) in written monuments; **attestation in historical dictionaries of Russian and Slavic languages**; **linking references** (for different types of variants); any kind of **additional philological information** (e.g. translated equivalent, descriptive, etymological translation or glossary interpretation); **additional historical and cultural information** (especially important for explaining rare words and concepts.

As a wholly new type of dictionary, the EDGr will serve as the basis for further research in historical lexicology, comparative typology, and cultural history.

The results of work on the "Electronic Dictionary of Greek Words in 11th- to 17th-century Russian" will be posted in the public domain on the Internet.

Владимир Дубичинский
Варшавский университет
v.dubichynskyi@uw.edu.pl

Некоторые современные вопросы терминографии

Основными вопросами современной теории терминографии являются сегодня: разработка методологических принципов создания терминологических словарей, создание научно обоснованной типологии специальных словарей; разработка инвариантного проекта словаря для описания различных специальных пластов лексики; определение основных параметров терминологических словарей; выработка принципиальных требований к терминографическим произведениям; исследование макро- и микроструктуры словаря; анализ путей отбора терминологического словника; выработка основных приёмов описания терминов; применение компьютеризации в создании терминологических словарей; создание современных электронных терминографических произведений.

Современная терминография сегодня уже невозможна без широкой компьютеризации. Постепенно традиционные методы заменяются компьютерной обработкой лексикографических данных. **Компьютеризация терминографической деятельности** заключается, прежде всего, в создании специализированных машинных терминологических банков данных и в разработке методов формирования этих банков, представления информации в банках и её использовании. На этой основе сегодня сформировались новые направления лингвистики - корпусная лингвистика и электронная терминография.

Автоматизированные терминографические системы, т.е. системы автоматизации подготовки и использования терминологических словарей, включают в себя программы и справочные данные, необходимые для лексикографической обработки текстов. В них используются текстовые редакторы для ввода и коррекции данных, программы контроля данных и запросов к системе, программы контроля орфографии и разметки входного текста, программы сегментации текста на слова, словосочетания, предложения и фрагменты словарных статей, программы лемматизации и подсчёта статистики словоупотреблений, программы загрузки, поиска и коррекции данных и др.

Сегодня актуальным становится вопрос о создании так называемого **терминографического автоматизированного рабочего места (ТАРМ)**, которое представляется автору универсальной терминологической базой данных или совокупностью терминологических баз данных, которыми терминолог может оперировать при создании электронных словарей.

Разработка ТАРМ – новый вид лексикографической деятельности, предполагающий творческое объединение усилий терминологов-специалистов определённой области знаний, терминологов-лингвистов, лексикографов-практиков и программистов. В настоящее время термин «ТАРМ» играет роль принципиального понятия компьютерной терминографии, так как на основе определенной универсальной совокупности терминологических баз данных предлагается стандартизировать и унифицировать весь лексикографический процесс, направить терминографическую деятельность в единое русло интернационализации электронно-словарных комплексов.

По идее автора ТАРМ должно включать: 1) инвариантную электронную структуру словарной статьи, разработанную на основе баз данных толковых, терминологических, идеографических, переводных и др. словарей, которые также являются составными частями соответствующих баз данных; 2) в ТАРМ необходимо наличие четко разработанной системы условных знаков и лексикографических помет, которые должны быть снабжены соответствующим программным обеспечением; 3) компьютерную программу автоматического редактирования сканированных текстов и правки орфографических, грамматических, синтаксических и т.п. ошибок; и т.д.

Компьютеризация словарной деятельности существенно расширяет возможности терминографов. Безостановочное развитие компьютерных технологий подсказывает необходимость полной компьютеризации словарных исследований. И это подтверждают сейчас современные направления терминографии: создание словарных картотек и корпусов специальных текстов на основе компьютерных баз данных, электронное построение словарных статей и автоматическая обработка лексического материала, составление печатных словарей на компьютерной основе и создание собственно электронных словарей (без их бумажных аналогов) и мн. др.

Concepts of modern terminography

The main issues in modern terminographic theory are: the development of methodological principles for creating terminological dictionaries; the creation of a scientifically based typology of special dictionaries; the development of an invariant vocabulary project for describing various special vocabulary layers; definitions of basic parameters of terminological dictionaries; the development of fundamental requirements for terminographic works; research into dictionary macro- and micro-structure; analysis of the selection of terminological vocabulary; the development of basic techniques for describing terms; the use of digitisation in creating terminological dictionaries; the creation of modern electronic terminographic works. Modern terminography is no longer possible without digitisation. Traditional methods are gradually being replaced by digital lexicographic data processing. The digitisation of terminographic activities includes the creation of specialised terminological databases, the development of methods to create these databases, and the presentation and usage of the information in these databases. This is the basis for new approaches to linguistics – corpus linguistics and electronic terminography.

Automated terminographic systems (automated systems to prepare and use terminological dictionaries) include programmes and reference data necessary for lexicographic word processing. They use text editors to enter and correct data and programmes to check data, to query the system, to check spelling, to mark input text, to segment text into words, phrases, sentences, and fragments of dictionary entries, as well as to lemmatise and count word usage statistics, load programmes, search and correct data, etc.

The task of creating a terminographic automated workplace (TARM), which is presented to the terminologist by a universal terminological database or a set of terminological databases that a terminologist can operate when creating electronic dictionaries becomes very important. The development of TARM is a new type of lexicographic activity that presupposes a creative unification of the efforts of terminologists from different fields, terminology linguists, lexicographers, and programmers. Currently, the term TARM is a

computer terminography concept, as it is intended to standardise and unify the entire lexicographic process and to direct terminographic activities towards a single internalised channel of electronic dictionary complexes based on a particular universal set of terminological databases. According to the author, TARM should include: 1) the invariant electronic structure of dictionary entries developed on the basis of databases of explanatory, terminological, ideographic, translational, and other dictionaries, which are also an integral part of the databases; 2) a well-developed system of conventional signs and lexicographic marks, as well as appropriate software; 3) a computer programme for automatically editing scanned texts and editing spelling, grammar, syntax, errors, etc.

The digitisation of vocabulary activities significantly expands the abilities of terminographers. The constant development of computer technology suggests the need for complete digitisation of vocabulary research. This has been affirmed by current trends of terminography: the creation of vocabulary card files and special texts based on computer databases, the digital construction of dictionary entries, the automatic processing of lexical material, the compilation of computer-based printed dictionaries, and the creation of web-born electronic dictionaries, etc.

Ivana Filipović Petrović
Zavod za lingvistička istraživanja HAZU
ifilipovic@hazu.hr

Jelena Parizoska
Učiteljski fakultet Sveučilišta u Zagrebu
jelena.parizoska@ufzg.hr

Mogućnosti leksikografske obrade promjenjivih frazema u mrežnome frazeološkom rječniku hrvatskoga jezika

E-leksikografija otvorila je brojne mogućnosti koje su tradicionalnoj nedostupne, a jedna je od najvažnijih ta da nema prostornog ograničenja kao u tiskanim rječnicima. Zbog toga je e-rječnike moguće drukčije organizirati. Na primjer, kod bilježenja frazema nisu potrebne nadnatuknice, već svi izrazi, pa i oni koji imaju istu sastavnicu, mogu biti zasebne natuknice (npr. *plakati kao malo dijete, ne budi dijete, dijete je na putu* itd.). Također, ako frazem ima leksičke varijante (npr. *ljut (bijesan) kao pas (ris)*), u e-rječniku nisu potrebne uputnice jer se različiti oblici iste konstrukcije mogu dobiti pretraživanjem bilo kojeg njezina dijela. Međutim, frazemi se osim u leksičkim javljaju u brojnim drugim vrstama varijanata. Tako podaci iz računalnog korpusa hrWaC pokazuju da isti izraz može imati oblik glagolskog i imenskog spoja riječi (*prodavati maglu* i *prodavač magle*), na lijevoj strani nekih poredbenih frazema javljaju se različite vrste riječi (*crven/pocrvenjeti kao rak*), a neki se glagolski frazemi javljaju u trima ili više leksičko-sintaktičkih oblika (npr. *biti u gabuli, upasti u gabulu, izvući iz gabule* koga). To otvara pitanje organizacije natukničkoga članka u e-rječniku općenito, kao i neka specifična pitanja, poput broja i vrsta varijantnih oblika frazema koji se bilježe u natuknici te redoslijeda njihova navođenja.

Cilj je ovoga rada pokazati mogućnosti leksikografske obrade promjenjivih frazema u mrežnome frazeološkom rječniku hrvatskoga jezika koji je u izradi, a utemeljen je na podacima iz hrWaC-a. Točnije, pokazat ćemo da se varijantni oblici različitih vrsta mogu bilježiti unutar iste natuknice (što je uobičajena praksa u tradicionalnoj ruskoj i engleskoj frazeografiji), a navodi ih se po frekvenciji pojavljivanja u korpusu. Budući da u e-rječniku

nema prostornih ograničenja, on može uključiti veći broj varijantnih oblika frazema i primjera. Takva leksikografska obrada odražava stvarnu uporabu, a korisnici dobivaju potpunu i preciznu informaciju o različitim oblicima i značenjima frazema, što će im omogućiti bolje razumijevanje i sigurniju uporabu.

The treatment of idiom variation in the Online Dictionary of Croatian Idioms

E-lexicography offers a number of advantages over the traditional printed format, the most important of which is the lack of space constraints. Online dictionary entries may thus be structured differently. For example, idiom entries need not be attached to a headword, so all idioms, including those containing the same words, can be listed as separate items (e.g. *cry like a baby, don't be such a baby*, *a baby is on the way*). Furthermore, cross-references are not required for idioms that have lexical variations (e.g. *mad as hell/a hornet*) because each variant form is searchable by key words. However, idioms vary not only lexically, but in many other ways. Data from the Croatian web corpus hrWaC show that an idiom may occur both as a verbal and a nominal construction (*prodavati maglu* 'to blow smoke', lit. 'to sell fog', and *prodavač magle* 'charlatan', lit. 'fog seller'), that adjectival slots in similes can be filled by verbs (*crven/pocrvenjeti kao rak* 'to go red in the face', lit. 'to turn as red as a crab'), and that some verbal idioms have three or more lexico-syntactic forms (*biti u gabuli* 'to be in a bind', *upasti u gabulu* 'to get oneself into a bind', *izvući iz gabule* koga 'to get someone out of a bind'). This raises the general issue of entry structure in online dictionaries and specific issues regarding the number, type, and order of listing of variant forms in idiom entries.

The aim of this presentation is to show how idiom variations are arranged in entries in the corpus-based Online Dictionary of Croatian Idioms (under development). More specifically, we will show that all variant forms are listed in the same entry (which is standard practice in Russian and English dictionaries) and that their order of listing is based on their frequency of occurrence in hrWaC. Since online dictionaries do

not suffer from the space constraints of print media, entries can include a large number of variant forms and usage examples. Treating variability in this way reflects real usage more closely and provides dictionary users with complete, precise information on the forms and meanings of idioms, which helps them understand idioms more fully and use them confidently.

mrežnik
HrZZ
Croatian Science Foundation
VERN'

Matea Filko
Filozofski fakultet Sveučilišta u Zagrebu
matea.filko@ffzg.hr

Krešimir Šojat
Filozofski fakultet Sveučilišta u Zagrebu
ksojat@ffzg.hr

CroDeriv – nadogradnja hrvatskoga tvorbenog leksikona

U ovom izlaganju predstavljamo CroDeriV, računalni leksikon s podatcima o morfološkoj strukturi i tvorbenoj povezanosti hrvatskoga leksika. Ovaj je jezični resurs izrađen s ciljem da prikaže od kojih se morfema sastoji pojedini leksem i koje morfeme dijeli s drugim leksemima. U mrežno dostupnoj verziji prikazan je morfološki sastav 14.500 hrvatskih glagola. Osnovna natuknica leksikona jest infinitiv rastavljen na leksičke i tvorbene morfeme (npr. do-pis-a-ti). Sučelje omogućuje pretragu prema različitim kriterijima: prema leksičkim morfemima (do-*pis*-a-ti, do-*pliv*-a-ti), pojedinim tvorbenim afiksima (npr. *do*-pisati, *do*-plivati) i mogućim kombinacijama tvorbenih afiksa (na-do-pisati, na-do-X-iv-a-ti). Pretraga prema korijenu prikazuje ukupnu tvorbenu porodicu glagola izvedenih od drugih glagola (npr. *pisati* – *napisati*, *nadopisati*, *prepisivati* itd.).

Jezični resursi koji daju uvid u tvorbenu povezanost leksika razvijaju se posljednjih godina za niz jezika, npr. engleski, njemački, francuski, latinski, ruski i češki, i u pravilu se fokusiraju na označivanje tvorbenih procesa unutar tvorbenih porodica. Za razliku od njih, u CroDerivu je svaka lema segmentirana na morfove, a svi su alomorfi povezani na zajednički morfem, no tvorbena se osnova ne navodi zasebno i ne prikazuje se točan slijed derivacijskih procesa (npr. *pisati* – *prepisati* – *prepisivati* – *isprepisivati*).

U ovome izlaganju prikazujemo daljnji razvoj CroDeriva. Najprije se usredotočujemo na preoblikovanje baze za proširenje drugim punoznačnim vrstama riječi slijedeći dosadašnje principe dvorazinske obrade – segmentaciju riječi na morfove te njihovo grupiranje prema

morfemima. U drugome dijelu usredotočujemo se na model analize i eksplicitnoga prikaza tvorbenih procesa unutar tvorbenih porodica. Na kraju prikazujemo alat kojim se koristimo za vizualni prikaz pojedinih tvorbenih osnova i izvedenica, odnosno za označivanje tvorbenih sljedova unutar tvorbenih porodica.

Ovakva vrsta jezičnoga resursa predstavlja vrijedan izvor podataka za različita leksikološka istraživanja i daljnju leksikografsku primjenu, a korisna je i u poučavaju hrvatskoga kao stranoga jezika te u raznim područjima računalne obrade hrvatskoga.

CroDeriv – development of the Croatian derivational lexicon

In this talk we present CroDeriv – a computational lexicon providing data about morphological structure and derivational relatedness of Croatian words. The purpose of this language resource is to demonstrate which morphemes make up a particular word and which morphemes are shared by other words in Croatian. In its on-line version the morphological structure of 14 500 verbs is presented. The entries in the lexicon consist of infinitives segmented into lexical and derivational morphemes (e.g. do-pis-a-ti 'to add by writing'). The interface enables various sorts of queries: the database can be searched for lexical morphemes (*pis*-a-ti 'to write', *pliv*-a-ti 'to swim'), particular derivational affixes (e.g. *do*-pisati, *do*-plivati) as well as possible and attested combinations of derivational affixes (e.g. *na-do*-pisati, *na-do*-X-*iv-a-ti*). Queries based on roots yield entire derivational families of verbs derived from other verbs (e.g. *pisati* – *napisati*, *nadopisati*, *prepisivati* etc).

Language resources with derivational data have been recently developed for numerous languages, e.g. English, German, French, Latin, Russian and Czech. Generally, they focus on derivational processes among words sharing the same root. Unlike these approaches, each lemma in CroDeriv is segmented into morphs and all allomorphs are linked to their morphemes. Presently, CroDeriv does not mark derivational stems and it does not present the sequence of applied derivations (e.g. *pisati* – *prepisati* – *prepisivati* – *isprepisivati*).

Here we deal with further expansion of CroDeriv. We focus on its reorganization and expansion with other POS following the two-level processing – the segmentation into morphs and their linking to morphemes. Thereby we explicitly mark derivational processes. Finally, the tool for the visualization of derivational processes is presented. CroDeriv is a valuable source of data for various research in lexicology, lexicography, in teaching of Croatian as a foreign language and in NLP applications.

Teodora Fonović Cvijanović
Filozofski fakultet Sveučilišta Jurja Dobrile u Puli
tfonov@unipu.hr

Blaženka Martinović
Filozofski fakultet Sveučilišta Jurja Dobrile u Puli
bmartino@unipu.hr

Vanessa Vitković Marčeta
Filozofski fakultet Sveučilišta Jurja Dobrile u Puli
vvitkov@unipu.hr

Bog u leksikografskoj mreži

Suvremena hrvatska leksikografija plodna je i raznovrsna, no postoje dvojbe koje se preslikavaju iz izdanja u izdanje (što je, dijelom, neizbježno u otvorenome sustavu kao što je leksik). Suvremene leksikografske dvojbe množe se i novima kada se leksik izmjesti u novi medij, u e-rječnik (primjerice otvara se pitanje ozvučivanja leksičkih jedinica te korpusne utemeljenosti). U ovome se radu suvremene dvojbe oprimjeruju leksičkim jedinicama sa sastavnicom „bog" jer donose cijelu paletu otvorenih pitanja, naime dotičemo se pravopisne (*kako bog/Bog zapovijeda*), gramatičke (pr. *bog* kao imenica, kao čestica i kao uzvik), leksičke (*žali Bože* / *žalibože*) i prozodijske (*bogtepítaj* ili *bogtepîtaj*) norme. Dakle, iako je riječ o dvojbama koje se odnose zasigurno na opsežniji korpus (kao što je pitanje gramatikalizacije, pragmatike oblika, (de)onimizacije i frazeologizacije), sužavamo pogled na jednu tvorbenu porodicu koji ipak može analogijom iznjedriti općenitija rješenja. Nudimo primjer mogućega ozvučenja pojedinih natuknica u budućim e-rječnicima te preispitujemo korpusnu utemeljenost sadašnjih leksikografskih rješenja.

Bog (God) in the lexicographical network

Modern Croatian lexicography is rich and diverse but not free of dilemmas, passed on from one edition to another (which is, partly, inevitable in an open system such as the lexicon). Modern lexicographical dilemmas increase in number when lexicons are transferred to a new medium such as e-dictionary (e.g. issues surrounding audio recordings of lexical units and attestation in the corpus). This paper exemplifies modern dilemmas through entries based on the word *bog* (god), because they raise numerous questions; we will address orthographic (*žali Bože* / *žalibože*) and prosodic (*bogtepîtaj* or *bogtepîtaj*) norms. Although these dilemmas certainly refer to a broader corpus (like the issues of grammaticalisation, pragmatics of forms, [de-]onymisation and phraseologisation), we have limited our scrutiny to a single derivational family that can still lead to more general solutions through analogy. We offer the example of possible audio versions of particular entries in future e-dictionaries and reassess the attestedness of current lexicographical solutions in the corpus.

Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
radovan.garabik@kassiopeia.juls.savba.sk

Vladimír Benko
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
vladimir.benko@juls.savba.sk

Word Embeddings Based on Large-Scale Web Corpora as a Powerful Lexicographic Tool

The Aranea Project offers a set of comparable corpora for two dozens of (mostly European) languages, providing a convenient dataset for NLP applications that require training on large amounts of data. *Word embeddings* is an umbrella term for various methods of representing words as vectors within a many-dimensional vector space.

Our work presents an online interface for vector models for the main languages in the Aranea corpora for use in lexicographic work. The implementation is somewhat Slovak-centric in that some features are either available only for Slovak or their implementation for other languages has not been tuned for coverage or accuracy as it is in Slovak lexicography.

There are two models available: one focuses on lemmas, the other on (normalised) forms. The lemma model loses information regarding inflected forms (and therefore perhaps interesting syntactic features), but users generally expect to enter lemmas, which carry semantic information. The models use automatic bigram detection, which aids in the lexicographic description of multiword expressions (there are better tools available for collocation analysis however, so this is useful in a supplemental role only). Out-of-dictionary lemmas can be filtered, which is useful in discovering non-obvious meanings of existing words. Otherwise, the lemmas obtained by statistical and heuristic guesser can be erroneous, but their inclusion often displays unexpected relations between words not covered by existing dictionaries (not only neologisms).

On the other hand, the model based on raw word forms is independent

from existing linguistic annotation (apart from tokenisation and normalisation), meaning the models are not tainted with potential systematic errors or the shortcomings of existing tools (especially systematic errors in lemmatisation, which are known to skew results significantly) and can be used even if other NLP processing components are not available for a given language.

Laura García Fernández
University of La Rioja
laura.garciaf@unirioja.es

Towards the lemmatisation of the verbal lexicon of Old English. Automatisation on a relational database.

This paper presents one of three procedures that the *Nerthus Project* is using to make progress with the lemmatisation of available corpora of Old English, which is applied to preterite-present, contracted, anomalous, and strong VII verbs. Although lemmatisation is generally accepted as a necessary tasks in dictionary making, no lemmatised corpus is available for Old English, which makes lemmatisation a pending task of English Historical Linguistics, as well as of Old English Corpus Linguistics and Lexicography. The *Nerthus Project* is working on creating a set of relational lexical Old English databases that are intended to search, analyse, store, and disseminate Old English linguistic data. The goal of the project is to create a database in which the language itself is the object of analysis. At present, the *Nerthus Project* is lemmatising the verbal lexicon of Old English using the lemmatiser *Norna*. The lemmatiser has been implemented through Filemaker database software and is based on a concordance and an index of the *DOEC*. Automatic searches are launched in the lemmatiser so that inflectional forms are assigned a lemma from a reference list. As stated by some authors, a completely automated procedure would not be reliable, so this methodology combines automatic searches and the manual revision of hits with available lexicographic (*Dictionary of Old English*, Bosworth-Toller, Sweet and Clark Hall) and textual sources (*York-Toronto-Helsinki Parsed Corpus of Old English Prose*). The conclusion is that structured data is necessary to reduce the amount of manual revision. The availability of lexicographical and textual sources in database form contributes to the automatisation of the procedure. Therefore, a great deal of attention must be paid to the selection, revision, and unification of sources.

Larisa Grčić Simeunović
University of Zadar
lgrcic@unizd.hr

Uroš Stepišnik
University of Ljubljana
uros.stepisnik@ff.uni-lj.si

Špela Vintar
University of Ljubljana
spela.vintar@ff.uni-lj.si

Framing definitions: the TermFrame approach to multi-layered definition annotation

Our research proposes a methodological rethinking of the knowledge extraction process and terminology representation. Using the theoretical principles of Frame-Based Terminology (Faber 2009, 2012), we aim to illustrate concept relations and their underlying cognitive frames.

The semantic annotation of definitions aims to offer information about underlying conceptual structures and relations that may indicate language-, genre-, and author-dependent differences in the conceptualisation of karst phenomena.

A trilingual (English, Slovene, Croatian) comparable corpus of texts from the karst domain will be compiled and morphosyntactically annotated. The target size of the corpus is approx. 1 million words per language.

First, computational methods will be used to extract definitions and knowledge-rich contexts from the corpus. Then defining structures and relations will be manually annotated for their semantic components. This approach assigns semantic roles to elements in a definition that can be used as training data to improve existing resources for automatic semantic role labelling.

The annotation consists of three layers:

1. Definition elements (definiendum, definitor, genus proximum)
2. Semantic categories that reflect the hierarchical framework of categories

specifically adapted to karstology (landforms, processes, geomes, entities/elements/properties, instruments and methods)

3. Relations used to mark individual parts of definitions in which a certain property or feature of the definiendum is described (has_form, has_origin, has_size, has_function, has_cause, has_result, has_location, has_position, composed_of, measures, defined_as).

The choice of relations in definitions is not arbitrary – certain logical connections between the semantic category and the relations are used to define it. These connections can help us predict which relations can be found in a definition. These findings will allow us to discover how specialised knowledge is encoded in texts and to better understand the cognitive processes underlying expert knowledge. Finally, the results will enable the more efficient design of termbases and knowledge representation resources.

Marijana Horvat
Institut za hrvatski jezik i jezikoslovlje
mhorvat@ihjj.hr

Istraživanje povijesti hrvatskoga jezika u digitalno doba

Hrvatska ima dugu tradiciju priređivanja djela u području povijesnoga jezikoslovlja i raspolaže mnogim vrijednim izdanjima te je veoma važno da se ta tradicija nastavi i prilagodi današnjemu digitalnom dobu.

Digitalizacija starih tekstova u stalnome je porastu, posebice od početka ovoga stoljeća. Razlozi digitalizacije arhivske, knjižnične, muzejske građe te one u znanstvenim ustanovama ili pohranjene u samostanima, višestruki su. Jedan je od osnovnih razloga digitalizacije starih tekstova zaštita izvornika i očuvanje baštine jer se pri aktivnom korištenju rabe digitalne preslike, a ne sami izvornici, pa ih se taj način čuva od mogućih oštećenja. Digitalne su preslike ujedno i način izrade sigurnosne kopije samih izvornika. Važan je razlog digitalizacije građe njezina veća dostupnost i stručnjacima i široj javnosti. Mladima, koji su otvoreni u prihvaćenju elektroničkih medija, daje se veća mogućnost upoznavanja s kulturnom baštinom. U izlaganju će se ponuditi model digitalizacije starih hrvatskih gramatika jer postojeći retrodigitalizirani resursi ne uključuju gramatike.

Croatian Language History Research in the Digital Era

Croatian philology has a long tradition of publishing reference books in the field of historical linguistics; it is very important to continue and foster this tradition in the future, as well as to adapt it to the modern digital era.

Recent large-scale digitisation initiatives have focused heavily on historical texts. Many reasons have led to the increasing digitisation of archival documents, library and museum material, and collections of old books and manuscripts stored in monasteries. One of the main reasons for digitisation of old texts is to protect the original document

from possible damage resulting from inappropriate handling. The use of digital copies instead of original documents is necessary to protect and preserve old written texts, as they are the most reliable backup of the original document. Another very important reason for digitising old texts is to make them more accessible to scholars, experts, and the public. In this way, young people who prefer using digital media will become more easily acquainted with cultural heritage that would otherwise be inaccessible to them. The digitisation model of old Croatian grammar books will be discussed, as existing resources do not include grammars from the pre-standard period of the Croatian language.

Вікторія Людвігівна Іващенко
Київський університет імені Бориса Грінченка
vicivashchenko@ukr.net

Галина Горбенко
Київський університет імені Бориса Грінченка
h.horbenko@kubg.edu.ua

Олександр Кохан
Видавництво "Розумники"
al.kokhan@gmail.com

Термінографічна база даних TERM_IN у міжслов'янському діалозі: архітектоніка та проектування

Дослідницька палітра сучасної е-лексикографії доволі розмаїта. Залежно від об'єкта е-лексикографування сьогодні на часі говорити про розбудову таких її прикладних напрямів, як е-лінгвографія, е-фразеографія, е-термінографія, е-ідеографія, е-концептографія тощо, які досі ще не мають належного теоретичного обґрунтування. Значних обертів набирає е-термінографія, одним із прикладних аспектів якої є проектування та ведення *термінографічних баз даних*, *віртуальних термінографічних лабораторій* як різновидів і/або складників *словникових баз*, *бібліотек словникових праць*, *словникових порталів*, *словникових веб-сторінок*, орієнтованих не лише на національні пріоритети, а й інтеграцію до світових лексикографічних систем, що передбачає створення й ведення багатомовних електронних ресурсів.

Узагальнення слов'янського термінографічного досвіду кінця XX – початку XXI ст. у форматі укладання бібліографічного покажчика *Слов'янська термінографія (1990 – 2017)*, над яким упродовж 2014 – 2018 рр. працювали члени ТК МКС, задля його утримання в інтерактивному актуалізованому стані вимагає створення багатомовної термінографічної бази даних TERM_IN з огляду на інформаційно-знаннєві запити сучасного суспільства. Пропонований для проектування на сайті ТК КМС новий інтегрований ресурс передбачає програмне забезпечення з можливістю вибору користувацького інтерфейсу 10-ма слов'янськими

мовами (українською, польською, російською, білоруською, хорватською, словенською, чеською, словацькою, сербською, македонською), що сприятиме частковому вирішенню проблеми забезпечення фахівців у галузі термінології інформацією про наявні славістичні термінографічні праці, зібрані на одній технологічній платформі в єдиній системі бібліографічного запису.

На початковому етапі проектування термінографічної бази даних TERM_IN передбачено її типову архітектоніку в кожному користувацькому інтерфейсі тією чи іншою мовою, зокрема розподіл термінографічних даних: 1) за алфавітом (латиничним або кириличним); 2) за роками; 3) за галуззю знань; 4) за типологією: а) одномовні словники (в алфавітному порядку); б) багатомовні словники (в алфавітному порядку); для багатомовних словників – розподіл за конкретними назвами мов (в алфавітному порядку). Зосереджено увагу на лінгвотехнологічних особливостях цього етапу проектування.

The TERM_IN terminographic database in cross-Slavonic dialogue: Architecture and design

The research palette of modern e-lexicography is quite diverse. Depending on the object of contemporary e-lexicography, it is time to discuss the development of its applied areas, such as e-linguagraphy, e-phraseography, e-terminology, e-ideography, e-conceptography, etc., which have not yet been properly theoretically justified. E-terminology is gaining momentum; one of its applied aspects is the design and maintenance of *terminological databases*, *virtual terminographic laboratories* as varieties and/or components of *vocabulary bases*, *libraries of dictionaries*, *vocabulary portals*, and *vocabulary web pages*, which are focused not only on national priorities but on integration into international lexicographic systems, which presuppose the creation and maintenance of multilingual electronic resources.

The Slavonic terminographic experience of the end of the 20th – the beginning of the 21st century has been summed up in the form of a bibliographic index *Slavonic Terminology* (1990 - 2017), which has

been compiled by members of the ISC TC during 2014-2018. In order to keep it in an active as well as interactive state, due to information-knowledge demands of the modern society the creation of a multilingual terminology database TERM_IN is required. The new integrated resource offered for designing on the TC KMS site includes software with the ability to select a user interface in 10 Slavonic languages (Ukrainian, Polish, Russian, Belarusian, Croatian, Slovene, Czech, Slovak, Serbian, Macedonian), which will promote a partial solution of the problem of supplying terminology specialists with information on available Slavic terminographic works, which will be stored in one digital platform in a single bibliographic system .

In the initial stage of designing a terminology database the TERM_IN termbase, the typical architecture is provided for each user interface in all languages, including the distribution of terminological data according to: 1) script (Latin or Cyrillic); 2) year of creation; 3) scientific field; 4) typology: a) monolingual dictionaries (in alphabetical order); b) multilingual dictionary division into particular languages (in alphabetical order); for multilingual dictionaries - the division into specific names of languages (in alphabetical order). Attention is focused on the linguistic and technological features of this design stage.

Mirja Jarak
Filozofski fakultet Sveučilišta u Zagrebu
mjarak@ffzg.hr

Tatjana Tkalčec
Institut za arheologiju
ttkalcec@iarh.hr

Tajana Sekelj Ivančan
Institut za arheologiju
tsivancan@iarh.hr

Srednjovjekovna arheološka terminologija i oblikovanje ranosrednjovjekovnih pojmovnika

U izlaganju se, kroz analizu objavljenih pojmovnika i arheološke literature, problematizira hrvatska srednjovjekovna arheološka terminologija. Za tu granu arheologije još nedostaje terminološki priručnik. Jedan od razloga je svakako kompleksnost sadržaja srednjovjekovne arheologije. U suvremenoj arheološkoj literaturi ta se kompleksnost nastoji sagledati kroz definiranje osnovnih tematskih cjelina, koje se u pravilu susreću u svim razdobljima srednjega vijeka. Polazeći i u predloženom izlaganju od osnovnih tematskih cjelina srednjovjekovne arheologije, nastojat će se sagledati problematika terminologije za rani srednji vijek. Takva usmjerenost proizlazi iz prevladavajućeg interesa autorica za rani srednji vijek, ali i iz važnosti ranosrednjovjekovnog razdoblja unutar stvaranja novih srednjovjekovnih struktura i građe, što zahtijeva nadopune pojmovnika i terminologije koji pokrivaju prethodna arheološka razdoblja.

U priopćenju se tematizira izgrađenost osnovnih ranosrednjovjekovnih arheoloških pojmovnika: pojmovnika arhitekture i skulpture, ranosrednjovjekovnih naselja i grobalja. U hrvatskoj arheološkoj literaturi posebni pojmovnici vezani su prvenstveno uz arhitekturu i skulpturu. Kako su iste teme predmet proučavanja i drugih znanosti, osobito povijesti umjetnosti i arhitekture, pojmovnici se mogu pronaći i u literaturi iz tih područja. Već sa stajališta različite metodologije pojedinih znanosti proizlaze razlike koje bi se trebale reflektirati u različitim definicijama

(opisima) istih pojmova u arheološkoj i primjerice povijesno-umjetničkoj literaturi. Slična zapažanja vrijede za područje ranosrednjovjekovnih naselja, iako ovdje postoji i veliki broj specifično arheoloških sadržaja, koji se ne proučavaju u okvirima drugih znanosti. Još je specifičnija arheološka problematika grobalja, koja zahtijeva zasebni pojmovnik.

Intencija izlaganja je ukazivanje na dosadašnja postignuća u području razvoja ranosrednjovjekovne arheološke terminologije, te zacrtavanje daljnjeg rada koji bi vodio u pravcu cjelovitog definiranja arheoloških sadržaja i usvajanja najprikladnijih jezičnih izričaja.

U radu će se koristiti radna računalna aplikacija Filozofskog fakulteta koja omogućuje upravljanje terminologijom.

Mediaeval archaeological terminology and the making of early Mediaeval glossaries

The subject of this presentation is Croatian Mediaeval archaeological terminology viewed through an analysis of published glossaries and archaeological literature. No terminology handbook for Mediaeval archaeology exists in Croatia. One of the reasons surely lies in the complexity of the content of Mediaeval archaeology. In the contemporary literature, this is understood through a division in basic thematic units common to different Mediaeval periods. The authors of the presentation will provide insight into the terminology for the Early Middle Ages. This direction was determined by the specific interest of the authors, as well as by the importance of the Early Middle Ages to the formation of new Mediaeval structures and artefacts, which require appendices to glossaries and terminology used for preceding archaeological periods.

The authors discuss the main groups of Early Mediaeval archaeological terminology: a glossary for architecture and sculpture, as well as those for Early Mediaeval settlements and cemeteries. In the Croatian archaeological literature, special glossaries have been connected primarily to architecture and sculpture. Because the same subjects may be investigated within the framework of other sciences, other glossaries contain the same concepts. The fact that each science uses different methodologies means

differences exist that must be articulated in definitions (descriptions) of the same concepts in archaeological and e.g. arthistory literature. The case is similar in the field of Early Mediaeval settlements, although this field contains a great deal of content specific to archaeology. The field of cemeteries is even more specific, requiring an individual glossary.

The presentation intends to underline achievements to date in the field of Early Mediaeval archaeological terminology and to direct further work on the subject.

A working version of a computer application created at the Faculty of Humanities and Social Sciences will be used to demonstrate the management of terminology.

Zrinka Jelaska
Filozofski fakultet Sveučilišta u Zagrebu
zjelaska@ffzg.hr

Marija Lütze-Miculinić
Filozofski fakultet Sveučilišta u Zagrebu
mlmiculi@ffzg.hr

Pristup obradi dvojezične građe u e-rječniku i tiskanomu rječniku

E-rječnik i tiskani rječnik sastavljeni na temelju iste početne građe međusobno se u mnogome razlikuju. Iako se češće sastavlja e-rječnik na temelju tiskanoga ili rukopisnoga, obratan odnos možda jasnije pokazuje razlike u ustroju jednoga i drugoga, posebno u natuknicama.

U radu se obilježja sastavljanja obaju vrsta rječnika uspoređuju na primjeru hrvatsko-njemačkoga rječnika školskoga jezika čija je prvotna inačica bila e-rječnik. Zbog osebujnosti građe koja je poglavito pragmatički odabrana, odnosno namijenjena omogućavanju stjecanja pragmatičkoga umijeća hrvatskih inojezičnih govornika njemačkoga i radi korisnikova pronalaženja jedinica koje bi mogao tražiti, tiskani je rječnik nametao različite dodatne dijelove i drugačiji, znatno tradicionalniji ustroj natuknica koji u e-rječniku nisu bili nužni isključivo zbog medija u kojemu je oblikovan. Glavna je razlika u ustroju e-rječnika nevidljiva isprepletenost odnosa među riječima, njihovim različnicama i njihovim odnosima prema drugim riječima iz primjera i natuknica koja se očituje u okupljenim prikazima na ekranu, dok u tiskanomu rječniku svaka vrsta odnosa zahtijeva zaseban dio s mnogim ponavljanjima, i to njemačko-hrvatski i hrvatsko-njemački popisnik.

Data structure for building a bilingual e- and printed dictionary

Built on the same corpora, an e-dictionary and its printed version differ in many ways. Although an e-dictionary is more often designed on the basis of a printed one, the opposite relationship may demonstrate more

clearly the differences between the structure of both dictionaries, esp. concerning lemmas or entries.

As an example, the paper will compare two versions of the Croatian-German Dictionary of School Language produced from the same data, collected in order to enable Croatian dictionary users to gain pragmatic competence in German. The first to be completed was a simple e-dictionary, later followed by a printed one. The printed dictionary made different additions necessary, as well as a more traditional data structure of many entries, which were not needed in the e-version due to the nature of e-media. The biggest difference stems from the hidden network of word-relations, their forms and sentences serving as examples, which become obvious when gathered on the screen. Instead, the printed dictionary requires additional parts with many repetitions that can be found in the German-Croatian, as well as the Croatian-German list of words, as special supplements to the main Croatian-German dictionary.

Mateja Jemec Tomazin
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
mjt@zrc-sazu.si

Mojca Žagar Karer
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
mojca.zagar@zrc-sazu.si

Slovenska terminološka spletišča in njihove zasnove

V prispevku primerjamo in analiziramo prosto dostopna slovenska terminološka spletišča (Terminologišče, Evroterm, Termania) in spletne strani posameznih specializiranih slovarskih projektov (npr. Islovar, EZS Glosar, Slovar finančnih in zavarovalnih izrazov), ki se razlikujejo po zasnovi, naslovniku, številu virov in strokovnih področij, številu vsebovanih jezikov, vključenosti opisa pojma oz. definicij. Prav tako bomo opisali različne uporabniške vmesnike, ki jih ta spletišča uporabljajo, pri čemer bomo izpostavili različne možnosti, npr. enostavno in/ali napredno iskanje, filtriranje zadetkov.

Dobro poznavanje prednosti in slabosti obstoječih spletišč je nujno za zasnovo novega skupnega terminološkega portala, kar bi različnim uporabnikom, zlasti področnim strokovnjakom, prevajalcem, lektorjem in drugim, omogočilo, da bi na istem mestu lahko dobili tisti tip informacije, ki ga potrebujejo.

Takšen portal potrebuje primeren vmesnik, ki bi dopuščal tudi stopenjski prikaz terminoloških informacij in v čim večji možni meri tudi povezave v specializirane korpuse. Za potrebe uporabnikov po terminologiji, ki še ni vključena v obstoječe terminološke vire ali tam ni optimalno predstavljena, bi skrbela integrirana terminološka svetovalnica, kjer bi jim odgovarjali na konkretna terminološka vprašanja.

Slovene terminological websites and their design

The paper compares and analyses freely accessible Slovene terminological websites (Terminologišče, Evroterm, Termania) and websites of specific terminographic projects (e.g. Islovar [Dictionary of informatics], EZS Glosar [Electrotechnical Association of Slovenia Glossary], Slovar finančnih in zavarovalnih izrazov [Dictionary of Finance and Insurance Terms]), which differ in their design, target users, number of included terminological sources and subject fields, number of included languages, and inclusion of concept descriptions or definitions. The article also describes the different user interfaces used by these websites and specifically focuses on different functionalities, e.g. simple and/or advanced searching, filtering of search results.

Good knowledge of the strengths and weaknesses of existing websites is a necessary precondition for designing a new joint terminological portal, which would enable different users access to the type of information they need in one place, especially field experts, translators, copy editors, and others.

This kind of portal requires a suitable user interface that enables a step-by-step display of terminological information and as many links as possible to specialised corpora. An integrated terminological counselling site would answer concrete terminological questions so as to address the needs of users interested in terminology that has not yet been included in existing terminological resources or is not presented there in an optimal manner.

Janoš Ježovnik
Fran Ramovš Institute of the Slovenian Language SRC SASA
janos.jezovnik@zrc-sazu.si

Retrodigitizing Jan Baudouin de Courtenay's *Rez'janskij slovar'*

Jan Baudouin de Courtenay's manuscript corpus, which was to serve as the basis for his planned-but-never-realized dictionary of the Resian dialect of Slovene (an endangered idiom spoken by approximately 1,000 inhabitants of the Resia Valley in Italy and protected under minority-protection legislation as part of the Slovene-language minority), consists of 248 two-sided sheets as well as of 2,055 smaller cards containing dialect data gathered between 1872 and 1893. While the sheets had already been at least partially edited into dictionary form by the author, the cards contain dialect data in rawer form. Initial editorial efforts focused on printing the dictionary – which was rediscovered in the 1960s in the archives of the Russian Academy of Sciences – but were halted for various reasons.

Building on this work, we aim to retro-digitise the dictionary, based on the draft of the dictionary in formatted text form and augmenting it with features that had previously been excluded, simplified, or not considered (restoration of the original manuscript's orthography which had been simplified due to typographical limitations, searchable multilingual indices, headwords in Standard Resian that were not established during much of the original editorial work, etc.).

The work will be done in several phases, including the restoration of the original orthography with manual rechecking of the transcribed data and its conversion into XML format, (manual) lemmatization of dictionary examples, (computer-assisted) rearranging of dictionary examples by lemma, etc., all while preserving the structure of the original, the author's own notes, and existing editorial work, which will be tagged in the original texts wherever possible. Due to the specifics of both types of original manuscript dictionary data, two databases will be used, with the contents of the first being injected into the second after the preparatory phases. The main challenge facing the editors is how to coordinate the already edited data in formatted text, the data on the 248 sheets already partially edited by the author, and the aforementioned nearly raw dialectal data on 2,055 cards.

Владан Јовановић
Институт за српски језик САНУ
vladjovanovic@hotmail.com

Ружица Левушкина
Институт за српски језик САНУ
ruzica.bajic@isj.sanu.ac.rs

Електронска форма Речника САНУ као извор за израду речника српске лингвистичке терминологије

У првом делу рада даћемо опште напомене о великом Академијином речнику – Речнику српскохрватског књижевног и народног језика, који се већ шест деценија израђује у Институту за српски језик, а потом, у општим цртама, изнети податке о најновијој електронској (дигиталној) верзији овога речника, која се такође реализује под окриљем поменутог Института.

У другом делу рада анализираћемо лингвистичке термине које смо добили методом екстраховања лексике из дигиталне форме Речника САНУ, коју засад чини 20 објављених књига Речника. Овај, централни део рада обухватиће следеће фазе: 1) издвајање из електронске верзије Речника речи (одреднице) са ознаком грам., лингв., филол., прав. итд., 2) издвајање речи које нису обележене овим квалификаторима, а које на основу других параметара препознајемо као речи које припадају области лингвистике, 3) анализа структуре одреднице и речничког чланка издвојених речи (творбене форме одреднице, њеног значења и наведених примера), 4) разврставање одредница с обзиром на сферу њихове употребе и књижевнојезичку правилност (савремена или застарела реч, варваризам, калк итд.). На крају предлажемо друге, савремене изворе из којих би се даље могли ексцерпирати термини из новијих лингвистичких теорија и дисциплина.

Циљ рада је да се покаже како и у којој мери електронска форма Речника САНУ може послужити као релевантна база за израду речника српске лингвистичке терминологије.

The electronic SASA Dictionary as a resource for a Serbian dictionary of linguistics

The first part of the paper provides general information about the dictionary of the Serbian Academy of Sciences and Arts' entitled *Rečnik srpskohrvatskog književnog i narodnog jezika* and its electronic version. Work on the SASA dictionary has been underway for more than sixty years and at SASA Institute for the Serbian Language; twenty volumes have been published so far. The electronic version is also a project of the Institute.

In the second part of the paper, linguistic terms extracted from the electronic SASA Dictionary are analysed. This main part of the paper deals with the following: 1) the extraction of dictionary entries with lexicographic notes: *лингв., филол., грам., прав.* etc. 2) the extraction of dictionary entries without these notes that can be nevertheless regarded as belonging to linguistics based on some other criteria; 3) the analysis of the structure of these dictionary entries (their forms, meanings, and illustrations); 4) the classification of the entries according to their sphere of use and standard language norms (contemporary or obsolete word, calque, etc.) Finally, other sources are proposed for the extraction of new linguistic terms for the Serbian dictionary of linguistics.

The aim of the paper is to provide an answer to the following questions: Can the electronic SASA dictionary be used as a relevant resource for the creation of a Serbian linguistic terminology dictionary and, if so, to what extent?

Santeri Junttila
University of Helsinki
santeri.junttila@helsinki.fi

Petri Kallio
University of Helsinki
petri.kallio@helsinki.fi

Sampsa Holopainen
University of Helsinki
sampsa.holopainen@helsinki.fi

Juho Pystynen
University of Helsinki
juho.pystynen@helsinki.fi

Digital etymological dictionary of the oldest vocabulary of Finnish

Our three-year project, entitled *Suomen vanhimman sanaston etymologinen verkkosanakirja* (https://sanat.csc.fi/wiki/Etymologiawiki), began at the University of Helsinki in January 2018, funded by the Kone Foundation. The project aims to etymologise all words inherited from Proto-Finnic (the common proto-language of Finnish, Estonian, Karelian, Livonian and the other Finnic languages). The database includes ca. 2,000 headwords; it is based on the manuscript *Itämerensuomalainen yhteissanasto (Common Finnic vocabulary)* by Docent Petri Kallio.

What makes this new project important is its open editing policy: all academic researchers of Finnic etymology can contribute to discussion on and the preparation of etymological word-articles, which means that the dictionary is constantly being updated. This is important, as etymological dictionaries are usually outdated by the time they are published, and this is certainly true of etymological dictionaries of Finnic languages, such as SSA, the most recent etymological dictionary of Finnish, and EES, the Estonian etymological dictionary. The crowdsourcing of etymological research gives research into Finnic etymology new possibilities for straightforward, free publication, criticism, peer-review, and the

popularization of etymologies. This also makes it methodologically more advanced than many other etymological e-dictionaries based on the digitisation of information published in earlier written sources.

The dictionary has a three-fold structure:

The specific information about each etymology will be provided in a database

The full research history of every headword since SSA (1992–2000) is presented

The etymologies will be edited into easily understandable entries for a broader audience

In the future, it will be possible to use the same platform for the etymological description of the more recent lexical strata of Finnish, as well as of the lexicon of the other Uralic languages or other language-families.

Евгения М. Какзанова
Всероссийский институт научной и технической информации РАН
em@kakzanova.ru

Электронный словарь Минералогических эпонимов

На сегодняшний день справочная литература о минералах необъятна. После появления в 1990 году «Словаря минеральных видов» М. Флейшера на русском языке число минералов значительно выросло, и появилась потребность в новом словаре. Таким словарём явился «Минералогический словарь» В.Г. Кривовичева, вышедший в 2009 году. В 2017 году вышел Геологический словарь. Он содержит около 24500 терминов, относящихся к 30 различным разделам. Одним из разделов является минералогия, включающая 6400 терминов.

К сожалению, минералогические словари устаревают крайне быстро, практически в год выпуска словаря. Поэтому мы считаем, что словарь минералогических эпонимов должен соответствовать двум критериям.

Во-первых, он должен быть электронным с непременной возможностью его ежегодного пополнения.

Во-вторых, во всех лексикографических источниках минералогические эпонимы подаются в общем списке и отдельно не выделяются. Такая форма подачи названий минералов не учитывает тот факт, что минералогические эпонимы уникальны тем, что помимо традиционных производящих основ – антропонимов, топонимов и мифонимов – их производящими основами являются также каронимы, документонимы, зоонимы, фирмонимы, космонимы и этнонимы.

Учитывая разнообразие производящих основ минералогических эпонимов, мы предлагаем в качестве второго критерия словаря минералогических эпонимов именно этот словообразовательный формант. Также непременным элементом словаря должен стать год открытия минерала. Поскольку все минералогические эпонимы являются интернационализмами, мы включаем в словарь их название на русском и английском языках.

Таким образом, наш словарь внесет вклад в российскую терминографию и дополнит существующие лексикографические источники новыми минералогическими эпонимами.

An e-dictionary of mineralogical eponyms

Today's reference literature on mineralogy is vast. After M. Fleischer's Glossary of Mineral Species was published in Russian in 1990, the number of known minerals increased significantly, generating a need for a new dictionary. This dictionary – the Mineralogical Dictionary by V.G. Krivovichev – was published in 2009. The Geological Dictionary was published in 2017, containing roughly 24,500 terms in 30 different sections. One of its sections is mineralogy, which contains 6,400 terms.

Unfortunately, mineralogical dictionaries become obsolete extremely quickly – almost as soon as they are published. Therefore, we believe that a dictionary of mineralogical eponyms must meet two criteria.

Firstly, it must be an electronic dictionary with a built-in regular updating mechanism.

Secondly, all lexicographic sources present mineralogical eponyms in a general list without specific indexing. This form of presentation does not take into account the fact that mineralogical eponyms are somewhat idiosyncratic; apart from the traditional eponymous parts – anthroponyms, toponyms and mythonyms – they may also include caronyms, documentonyms, zoonyms, firmonyms, cosmonyms, and ethnonyms.

Given the diversity of eponymous parts of mineralogical terms, we propose that the second criteria of our dictionary of mineralogical eponyms should be this specificity of their formation. The year in which minerals were discovered is also considered an essential element of the dictionary. Since all mineralogical eponyms are internationalisms, we include their names in the dictionary in both Russian and English.

Our dictionary will thus contribute to Russian terminography and supplement existing lexicographic sources with new mineralogical eponyms.

Virna Karlić
Filozofski fakultet Sveučilišta u Zagrebu
virnakarlic@gmail.com

Petra Bago
Filozofski fakultet Sveučilišta u Zagrebu
pbago@ffzg.hr

Pragmatika i leksikografija: deiktici kao izazov suvremene leksikografije

Rad se bavi problemom leksikografskoga opisa deiktičkih izraza u suvremenim jednojezičnim rječnicima hrvatskoga jezika. Opće je poznato da opisi sinsemantičkih riječi, kao i „riječi na granici punoznačnosti" (Kordić 2002), leksikografima predstavljaju velik izazov, o čemu svjedoče mnoge manjkavosti njihovih definicija te cjelokupnog opisa u rječničkim natuknicama. Zbog izostanka pragmatičkog pristupa u hrvatskoj filologiji, koji je u slučaju opisa ovakvih jezičnih pojava neophodan, slična se situacija zatječe i u gramatikama hrvatskoga jezika.

Pragmatika je lingvistička disciplina koja se bavi proučavanjem jezične upotrebe u kontekstu te odnosom konteksta i raznih aspekata jezične interpretacije (Lycan 1995). Deiksa utoliko pripada domeni pragmatike, jer izravno odražava odnos između jezične strukture i konteksta jezične upotrebe (Levinson 1983). Pojam 'deiksa' odnosi se na skupinu jezičnih izraza koji ukazuju na elemente situacijskog i/ili diskursnog konteksta u iskazu – kao što su primjerice sudionici, vrijeme i mjesto govornog događaja (Lyons 1977).

Deiktici pripadaju skupini leksema koji leksikografima predstavljaju poseban izazov jer se njihovo puno značenje ili referent uspostavlja odnosno identificira tek u okviru konkretnoga govornog događaja. Upravo su zbog toga obilježja deikse stalan podsjetnik na činjenicu da su prirodni jezici primarno nastali radi neposredne komunikacije licem u lice te da se jezik tek do određene mjere može analizirati bez uzimanja u obzir te činjenice (Levinson 1983).

U ovom se radu analiziraju definicije značenja i opisi funkcija deiktičkih izraza u hrvatskoj leksikografiji na odabranim reprezentativnim

primjerima osobnih, vremenskih, mjesnih, diskursnih i socijalnih deiktika. Primjenom analize mrežnih korpusa utvrđuju se manjkavosti obrađenih rječničkih natuknica te se sugeriraju moguća rješenja za pojedine utvrđene probleme.

Pragmatics and Lexicography: Deixis as a challenge of contemporary lexicography

The paper deals with the problem of lexicographic description of deictic expressions in contemporary monolingual dictionaries of the Croatian language. It is widely known that the descriptions of synsemantic words, as well as "words on the border between lexicon and grammar" (Kordić 2002), are a great challenge for lexicographers, as evidenced by the many shortcomings of their definitions and the overall description in the dictionary entries. Due to the lack of a pragmatic approach in Croatian linguistics, which is necessary in the case of analyses of such language phenomena, a similar situation can be found in the grammar descriptions of the Croatian language.

Pragmatics is a subfield of linguistics that studies the use of language in context, and the context-dependence of various aspects of linguistic interpretation (Lycan 1995). The deixis belongs to the domain of pragmatics insofar as it directly reflects the relationship between the language structure and the context of language use (Levinson 1983). The term "deixis" refers to a group of language units that indicate elements of situational and/or discourse context in the utterance -- such as participants, time, and place of the speech event (Lyons 1977).

Deictic expressions belong to a group of lexemes that are especially challenging for lexicographers because their full meaning or referent is established or identified only within the specific speech event. For these reasons the characteristics of deixes are a constant reminder of the fact that natural languages were primarily designed for use in face-to-face interaction, and that a language can only be analyzed to a certain extent without taking into account these facts (Levinson 1983).

In this paper we conduct an analysis of the definitions of meanings and the descriptions of functions of deictic expressions in Croatian lexicography on selected representative examples of person, time, place, discourse and social deixis. By analyzing examples in a web corpus, we identify the shortcomings of the selected dictionaries, and suggest possible solutions to the identified problems.

mrežnik

HRZZ
Croatian Science
Foundation

VERN'

Boris Kern
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
boris.kern@zrc-sazu.si

Društveno odgovorna leksikografija

Suvremena leksikografija temelji se na opsežnom korpusu što od leksikografkinje ili leksikografa zahtijeva prije svega kritičku distancu, i to da se izbjegnu stereotipizacije (prvenstveno sa stajališta društvenih uloga) i antropocentričnost te prilikom opisa leksema koji označavaju različite društvene skupine, posebno manjine. Pritom je važno biti svjestan da u rječniku do izrazitih predrasuda i stereotipnih prikaza, kako zbog prevelikog „povjerenja" u korpus tako i zbog netemeljitog rada leksikografa, može doći i kod bilo kojeg drugog leksema. Upravo tome treba posvetiti posebnu pozornost.

Cilj je izlaganja predstaviti pojedine leksikografske nedoumice (povezane s leksemima koje označavaju stanovnike naseljenih mjesta, države, rase, vjerske zajednice, spolni identitet itd.) koje su se pojavile prilikom rada na dvama aktualnim rječnicima slovenskoga jezika, i to na eSSKJ-u, rastućem jednojezičnom objasnidbenom rječniku slovenskoga književnog (standardnog) jezika te ePravopisu, rastućem pravopisnom rječniku.

Posebna pozornost bit će posvećena nedoumicama u rječničkom opisu leksema u kojem dolazi do razilaženja u terminološkoj i općoj uporabi, što je rezultat neujednačenosti u struci, odnosno promjena do kojih dolazi u posljednje vrijeme i koje u korpusu još nisu zabilježene, a što je posebno karakteristično za lekseme koje se odnose na osobe s posebnim potrebama, odnosno na invalidnost.

Socially Responsible Lexicography

Modern dictionaries are based on extensive corpus material that requires lexicographers to apply a special critical distance in order to avoid stereotypes (especially from the point of view of social roles)

and anthropocentricity; this also holds in the description of lexemes associated with different social groups, especially minorities. In this context, prejudices and stereotypes expressed in a dictionary – both due to excessive "loyalty" to corpus data and to the imprudent work of lexicographers – can also occur in the dictionary presentation of every other lexeme. Thus, caution in this respect is important.

This paper aims to present selected lexicographic dilemmas (related to lexemes that designate inhabitants, countries, members of ethnic groups, religious groups, gender identities etc.) from the practical lexicography work of the two most recent dictionaries of the Slovene language: the eSSKJ, a growing monolingual explanatory dictionary of the Slovene literary language, and ePravopis, a growing orthographic dictionary.

Particular attention is dedicated to dilemmas in the lexicographic description of lexemes, whose terminological and general usage is discrepant and, simultaneously, a consequence of terminological non-standardisation or recent changes that have not yet been registered in the corpus. The latter is especially characteristic of lexemes that designate people with disabilities.

Sanja Kiš Žuvela
Music Academy, University of Zagreb
skiszuvela@muza.hr

Daria Lazić
Institute of Croatian Language and Linguistics
dlazic@ihjj.hr

Musical Terminology in Terminology Databases and General E-dictionaries

When creating terminological entries for reference resources, their purpose and target user group should be carefully taken into consideration. While terminological databases primarily address field experts, users well acquainted with the subject, and translators, laymen in the field would be more likely to search for these terms in a general dictionary. This paper will thus study the approach to musical terms in *Struna*, a database of Croatian special field terminology, and the *Croatian Web Dictionary – Mrežnik.*

Within the project of the development of Croatian special field terminology *Struna,* musical terminology is represented at the basic conceptual level of terms taken from the curricula of public music schools and university music programmes. The approach is normative in accordance with standard contemporary principles of terminology work. The database includes recommended terms, synonyms, abbreviations, symbols, and equivalents in several European languages. Most terms are extracted from a specialised corpus, however, in the case of lexical voids or terms that do not comply with terminological principles, appropriate innovations are proposed.

The *Croatian Web Dictionary– Mrežnik* will also include some basic musical terms. This general monolingual e-dictionary features entries that contain diverse information such as collocations, usage examples, and normative indications in addition to definitions. Although the dictionary is corpus-based, other available sources are also taken into account; in the case of terminology, insight into terminology databases like *Struna* can be useful.

This paper will present the principles by which material was selected, as well as the structure of entries in *Struna* and *Mrežnik* from the perspective of the user. It will also discuss definitions, pragmatic information, approach to synonymy, terminological principles, and some particular aspects of the application of common standards of terminology work to terminology in the humanities.

Kristina Kocijan
Faculty of Humanities and Social Sciences, University of Zagreb
krkocijan@ffzg.hr

Silvia Kurolt
Faculty of Humanities and Social Sciences, University of Zagreb
skurolt@ffzg.hr

Linda Mijić
University of Zadar
mijic.linda@gmail.com

Building the Croatian medical dictionary from medical corpus

The 21st century is characterized by immense collections of unstructured data that represent a real challenge from the NLP perspective. The ability to automatically process and understand this data in the medical domain improves analytical abilities in medical care both at the individual and macro levels. Access to large collections of digital health records, as well as the creation of information retrieval tools, would greatly assist in conducting expensive individual clinical trials through a greater variety of samples and faster and more relevant information available to those who need it (from making decisions on patient's treatment, adjusting the drug to a particular population, up to business decisions of health institutions).

The basic objective of the project is to define linguistic models at the lexical and syntactic level that appear in the health domain, depending on the type of corpus (e.g. a pharmaceutical description of a drug or a patient's medical history).

In the first phase of the project, the texts forming the **medical corpus A** (6 500 pharmaceutical instructions for medicines available in Croatia) were collected. The terminology found in this corpus was analyzed and the semantic subdomains (*anatomy, disease, bacteria*, *drug* etc.) within the medical domain were defined and added to the dictionary entries. These dictionary resources were used as the foundation for the second phase in which morphological grammars recognizing Latinisms as well

as Latin expressions written with Croatian declination extensions were built allowing annotation of these terms as well.

Prepared resources will be made available to a broader scientific community for further research in the field of medicine enabling additional research and development of algorithms for, among others, medical documents classification, medical texts information retrieval or machine translation of medical documentation taking into account quality and reliability as well as terminology variability.

Valeriia Kolosova
Institute for Linguistic Studies, Russian Academy of Sciences_
chakra@eu.spb.ru

Kira I. Kovalenko
Institute for Linguistic Studies, Russian Academy of Sciences_
kira.kovalenko@gmail.com

Ksenia A. Zaytseva
Austrian Center for Digital Humanities, Austrian Academy of Sciences_Ksenia.Zaytseva@oeaw.ac.at

Phytonymic database PhytoLex as a research resource and the basis of the Russian phytonymic Dictionary of the 11th-17th cc.

Compiling dictionaries has always been a highly time-consuming work. Modern computer technologies improve this significantly by increasing the base of dictionary sources, speeding up searches for representative quotations, determining the first chronological fixation of lexemes, and tracking changes in meaning and frequency throughout the period under research. With this in mind, we compiled the Russian 11th-17th-century Phytonymic Dictionary using a previously created database. The PhytoLex database includes word usage with context and detailed descriptions of sources. Particular research work was necessary while filling the database. Project members had to recognise phytonyms in texts, identify the plant if possible (using the Catalogue of Life http://www.catalogueoflife.org), and determine the function of the plant. At this stage, it was possible to identify the variability of plant names, the area of their distribution (if there is information about the place where the text was created), their frequency, their genre, and the presence or absence of polysemy. PhytoLex was created not only as a basis for the dictionary, but also as an independent resource for studying the role of plants in culture. The content and structure of the database allows the taking of samples for plant functions (economic, magical, cooking, medicinal, etc.), parts used, and the place where the plants are grown and/or purchased. Thus, PhytoLex will significantly expand our understanding of the history of

the appearance and usage of phytonyms in the Russian language, as well as help to better present the role of plants in Russian society in the 11th-17th century. The results of the project will be of interest to researchers studying the history of medicine and biology, as well as relevant terminology, since the quotes included in the database contain numerous names of diseases, medicines, medical forms, body parts, etc.

This research is supported by RFBR (the Russian Foundation for Basic Research) project 17-06-00376 "Russian Phytonyms in the Diachronic Aspect (11-17 cc.)"

Barbara Kovačević
Institut za hrvatski jezik i jezikoslovlje
bkova@ihjj.hr

Frazeologija u *Hrvatskom mrežnom rječniku – Mrežniku*

U radu se prikazuje uspostava frazemskoga članka i obrada frazema u elektroničkom općem rječniku hrvatskoga jezika (*Hrvatski mrežni rječnik – Mrežnik*) koji se izrađuje u Institutu za hrvatski jezik i jezikoslovlje. Posebna je pozornost posvećena uspostavljanju frazemskih natuknica s obzirom na njihovu potvrđenost u suvremenim korpusima hrvatskoga jezika (*Croatian web corpus – hrWaC*, *Hrvatska jezična riznica*) iako su u obzir uzeti i tiskani jednojezični i višejezični, opći i specijalizirani frazeološki rječnici, opći rječnici suvremenoga hrvatskoga jezika te mrežno dostupne baze: Kolokacijska baza hrvatskoga jezika (http://ihjj.hr/kolokacije/) *Baza frazema hrvatskoga jezika* (http://frazemi.ihjj.hr/). Kako izrada elektroničkoga rječnika nema prostornih ograničenja, pored navedene frazeološkom praksom uspostavljene obrade s donošenjem značenja i oprimjerenja, rječnik će korisnicima pružiti i uvid u frazemsku etimologiju, tj. motivaciju frazema i njihovo podrijetlo, a to je novina u odnosu na dosad objavljene opće suvremene rječnike.

Phraseology in the *Croatian Web Dictionary – Mrežnik*

The paper presents the process of determining and interpreting phraseological units in the online corpus-based dictionary of the Croatian language (*Croatian Web Dictionary – Mrežnik*), which is being compiled at the Institute of Croatian Language and Linguistics. Special attention is paid to the representation of phraseological units considering their attestation in contemporary corpora of the Croatian language (*Croatian Web Corpus hrWaC*, *Croatian Language Repository*), although monolingual, multilingual, general, and specialised phraseological dictionaries, general dictionaries of the contemporary Croatian language, and the online *Croatian Collocation Database* (http://ihjj.hr/kolokacije/) and *Croatian*

Phraseological Unit Database (http://frazemi.ihjj.hr/) are also considered. Since electronic dictionaries do not have spatial limitations, in addition to the meaning and exemplification established in standard phraseology, the dictionary will provide insight into the etymology of phraseological units, i.e. their motivations and origins, which is a novelty in relation to other published general contemporary dictionaries.

Paweł Kowalski
Instytut Slawistyki PAN
kowalion@wp.pl

Wielojęzyczny słownik słów kluczowych jako narzędzie cyfrowej bazy bibliograficznej z zakresu językoznawstwa slawistycznego

W referacie zaprezentowana zostanie struktura wielojęzycznego słownika słów kluczowych, który stanowi integralną część bibliograficznej bazy językoznawstwa slawistycznego iSybislaw prezentującą cyfrowy (internetowy) system informacyjno-wyszukiwawczy (www.isybislaw.ispan.waw.pl). Jednostki leksykalne (słowa kluczowe) języka słów kluczowych zastosowanego w systemie reprezentowane są przede wszystkim przez terminy językoznawcze. Mimo odmiennej denotacji – słowa kluczowe bezpośrednio denotują zbiory dokumentów, pośrednio zaś rzeczywistość pozadokumentacyjną, podczas gdy terminy denotują elementy rzeczywistości językowe – są równokształtne z terminami językoznawczymi, co pozwala za ich pomocą odwzorować pole semantyczne konkretnej dyscypliny, w tym wypadku językoznawstwa slawistycznego. Słownik jest więc dziedzinowym internetowym słownikiem specjalistycznym, który stanowi narzędzie dla użytkowników bibliograficznej bazy językoznawstwa slawistycznego. Słownik skierowany jest do wszystkich zajmujących się językoznawstwem i terminologią językoznawczą, przede wszystkim do naukowców-językoznawców, doktorantów i studentów kierunków filologicznych, a także tłumaczy artykułów naukowych z zakresu językoznawstwa. Omówione zostaną kryteria doboru podstawowych polskich terminów językoznawczych w funkcji słów kluczowych, które organizują wiedzę w systemie iSybislaw, oraz ich ekwiwalentów w pozostałych językach słowiańskich i języku angielskim. Za podstawowe kryteria doboru jednostek przyjmuje się relewancję, frekwencję, aktywność i powszechność użycia. W pewnym sensie kryteria doboru jednostek są więc kompromisem między kryteriami ilościowymi a jakościowymi. Przy doborze jednostek istotne jest wykorzystywanie już dostępnej infrastruktury komputerowej i cyfrowej CLARIN (tak na poziomie europejskim clarin-eu, jak i krajowym clarin-pl), która może służyć do analizy dokumentów językoznawczych

i eksceprcji z nich terminów lingwistycznych, m.in. pierwsza wersja klasyfikatora tematycznego Wikipedia K-Nearest Neighbours dla tekstów polskich i angielskich („PELCRA NLP Tools - WiKNN classifier"); narzędzie do wyznaczania słów kluczowych w tekście („ReSpa"); narzędzie do wykrywania terminów w tekście („TermoPL"). Obecnie zauważalna jest jeszcze duża dysproporcja pomiędzy jednostkami leksykalnymi w obrębie poszczególnych języków, co do pewnego stopnia ogranicza funkcjonalność systemu. Jednym z proponowanych rozwiązań jest wprowadzenie listy pierwszego poziomu rozczłonkowania języka słów kluczowych i uzupełnienie klas ekwiwalentów o innojęzyczne jednostki najbardziej frekwentywne w zbiorze klas słów kluczowych. Taka propozycja poszczególnych jednostek w innych językach słowiańskich ma charakter wstępny i orientacyjny, a wynika przede wszystkim z pilnej potrzeby uzupełniania i rozbudowywania podstawowych klas ekwiwalentów. W wystąpieniu poruszone zostaną również zagadnienia związane z rozwiązywaniem bieżących problemów przy tworzeniu słownika słów kluczowych, jak: synonimia, homonimia i polisemia w obrębie całego wielojęzycznego zbioru słów kluczowych.

A multilingual dictionary of keywords as a tool for a digital bibliographic database of international Slavic linguistics

This paper presents the structure of a multilingual dictionary of keywords, which represents the digital information retrieval system of the iSybislaw bibliographic database of Slavic linguistics (www.isybislaw.ispan.waw.pl). The lexical units (keywords) of the language of keywords used in the system are represented primarily through linguistic terms. In spite of their different denotation – key words directly denote sets of documents, and thus indirectly denote non-documentary reality, while terms denote elements of linguistic reality – they are formally equal with linguistic terms, which allow them to map the semantic field of a particular discipline (Slavic linguistics in this case). The dictionary is therefore a domain-based online specialist dictionary, which is a tool for users of the bibliographic database of Slavic linguistics. The dictionary is addressed to

all those who deal with linguistics and linguistic terminology, primarily to scholar-linguists, PhD students, and students of philology, as well as translators of academic papers in the field of linguistics. Criteria will be discussed for the selection of basic Polish linguistic terms in the function of keywords that organise knowledge in the iSybislaw system and their equivalents in other Slavic languages and in English. The basic criteria for the selection of units are relevance, frequency, activity, and universality of use. Therefore, the criteria for selecting individuals are a compromise between quantitative and qualitative criteria. When selecting the units, it is important to use the available CLARIN computer and digital infrastructure (both at the European level clarin-eu and the national clarin-pl), which can be used to analyse linguistic documents and excess linguistic terms, including: the first version of the thematic Wikipedia K-Nearest Neighbors theme for Polish and English texts ("PELCRA NLP Tools - WiKNN classifier"); a tool for determining keywords in text ("ReSpa"); a tool for detecting terms in text ("TermoPL"). At present, there is still a large disproportion between lexical units within particular languages, which limits the functionality of the system to some extent. One of the proposed solutions is to introduce a list of the first level of keyword language fragmentation and supplement the equivalence classes with the most important units in the set of keyword classes using the aforementioned digital tools. This proposal of individual units in other Slavic languages is preliminary and indicative; it is primarily due to the urgent need to supplement and develop basic classes of equivalents. The presentation will also address issues related to solving current problems when creating a keyword dictionary, such as synonymy, homonymy, and polysemy within the entire multilingual set of keywords.

Михал Коздра
Варшавский университет
m.kozdra@uw.edu.pl

Возможности использования мультимодальности в *Учебном тематическом словаре русско-польских лексических параллелей*

Целью доклада является представление возможностей применения мультимодальности в *Учебном тематическом словаре русско-польских лексических параллелей* (Коздра, Дубичинский 2019). В словаре отражение находят лексические параллели, т.е. сходные по внешней (графической и/или фонетической) форме лексические единицы русского и польского языков с полным/частичным совпадением или несовпадением значений, которые вызывают аналогичные ассоциации у изучающих иностранный язык [Дубичинский, Ройтер 2015; Дубичинский 2017; Дубичинский, Ройтер 2017; Dubichynskyi, Reuther 2017; Коздра 2017; Коздра 2018]. Словарь адресован студентам, аспирантам, ученикам и всем, изучающим русский и польский языки в разных учебных заведениях, частных компаниях или на языковых курсах, переводчикам русского и польского языков, преподавателям русского языка как иностранного. Первая часть словаря посвящена описанию кулинарной лексики. Объем словаря составляет ок. 350 заголовочных единиц – пар русско-польских лексических параллелей. Словарная статья включает в себя: заголовочную единицу русского словарного запаса с ее польскими коррелятами, краткую грамматическую характеристику заголовочной единицы, упрощенное толкование каждого значения русского слова и его польского соответствия, переводной эквивалент, стилистические и другие лексикографические пометы, а также иллюстративные словосочетания лексико-семантических вариантов заголовочной единицы. Дефиниции и иллюстративные примеры вырабатываются с помощью толковых и двуязычных словарей, а также электронных корпусов (в том числе Национального корпуса русского языка).

Применение принципов мультимодальности в учебном словаре подразумевает привлечение поликодовых объектов, созданных из вербальных и невербальных – визуальных и аудиальных – семиотических

кодов (Kozdra 2018; Coccetta 2009; Kress & van Leeuwen 2010; Royce & Terry 2007; Yen-Liang 2017). Мультимодальная семантизация поддерживает процессы запоминания, что существенно для изучения иностранного языка. В докладе будут освещены также способы включения мультимодальных компонентов (фотографий, видеорядов и аудиозаписей) в словарь, а также создания его электронной версии.

The possibility of using multimodality in the *Learner's Thematic Dictionary of Russian-Polish Lexical Parallels*

The purpose of this paper is to present the possibility of applying the principles of multimodality to the *Learner's Thematic Dictionary of Russian-Polish Lexical Parallels* [Коздра, Дубичинский 2019]. The dictionary contains lexical parallels, i.e. lexical units of the Russian and Polish languages with similar (orthographic and/or phonetic) forms and full or partial identity/non-identity of meanings, which cause similar associations for students learning Russian and Polish [Дубичинский, Ройтер 2015; Дубичинский 2017; Дубичинский, Ройтер 2017; Dubichynskyi, Reuther 2017; Коздра 2017; Коздра 2018]. The dictionary is addressed to students, PhD students, pupils, and all learners of Russian and Polish in various educational institutions, private companies, or language courses, translators of Russian and Polish, and teachers of Russian as a foreign language. The first part of the dictionary is devoted to the description of culinary vocabulary. The volume of the dictionary is approx. 350 entries – pairs of Russian-Polish lexical parallels. The dictionary entry includes: the main lexical unit of the Russian language with its Polish correlates; a brief grammatical description of the main lexical unit; a simplified definition of each meaning of the Russian word and its Polish equivalent; translation equivalent; stylistic and other lexicographical notes; illustrative examples of lexico-semantic variants of the main lexical unit. Definitions and illustrative examples are developed with the use of dictionaries and electronic corpora (including the Russian National Corpus).

Applying the principles of multimodality to a learner's dictionary implies the use of multimodal objects created from verbal and nonverbal – visual

and audial – semiotic channels (Kozdra 2018; Coccetta 2009; Kress & van Leeuwen 2010; Royce & Terry 2007; Yen-Liang 2017). Multimodal semantisation supports the memorisation processes, which is essential for learning a foreign language. Ways in which to include multimodal components (photos, video and audio recordings) in the dictionary and create its electronic version will also be presented.

Cvijeta Kraus
Leksikografski zavod Miroslav Krleža
cvijeta.kraus@lzmk.hr

Irina Starčević Stančić
Leksikografski zavod Miroslav Krleža
irinas@lzmk.hr

Leksikografski zavod u digitalnom okružju

Leksikografski zavod Miroslav Krleža (LZMK), ustanova od kulturnog i nacionalnog značaja za Republiku Hrvatsku, jedina je u Hrvatskoj koja se sustavno bavi leksikografskim i enciklopedičkim radom. U gotovo 70 godina postojanja objavila je vrijedna djela općeg znanja kao i ona s pojačanom nacionalnom tematikom. Od 2009. godine LZMK razvija javno dostupni portal (http://enciklopedija.lzmk.hr) čiji je jedan od strateških ciljeva digitalizacija i mrežno objavljivanje leksikografskih sadržaja. Portal znanja repozitorij je objavljene knjižne građe koji trenutačno obuhvaća sadržaj sedam digitaliziranih arhivskih izdanja s više od 60 000 natuknica (Hrvatski obiteljski leksikon, Hrvatski biografski leksikon, Filmski leksikon, Istarska enciklopedija, Medicinski leksikon, Enciklopedija Miroslava Krleže, Nogometni leksikon). LZMK time ostvaruje svoju misiju i viziju u stvaranju jedinstvenoga javno dostupnoga digitalnog repozitorija te omogućivanju pristupa javnom znanju. Razvojem informacijskih i komunikacijskih tehnologija podaci i informacije pristupačniji su i dostupniji u raznim digitalnim oblicima. Enciklopedički i leksikografski sadržaji razvijaju se u digitalnom okruženju, a LZMK se, kao i ostali svjetski izdavači enciklopedijskih i leksikografskih djela, nastoji tome prilagoditi. Temeljna izdanja LZMK su Hrvatska enciklopedija i Hrvatski biografski leksikon koja imaju svoje mrežne stranice što pokazuje kako je Zavod prepoznao važnost prisutnosti leksikografskih sadržaja u digitalnom okruženju. Hrvatska enciklopedija koja se od 2013. godine izdaje kao mrežno izdanje (http://enciklopedija.hr), nastoji postati prvi enciklopedijski izvor znanja na hrvatskome jeziku za sve korisnike koji žele pronaći točne i pouzdane informacije o Hrvatskoj i svijetu. Pred tradicionalnu leksikografiju i same leksikografe postavljen je

izazov za usvajanje novih znanja i leksikografskih vještina koje su potrebne za razvijanje sadržaja u digitalnom okruženju, a čiji se proces uvelike razlikuje od procesa stvaranja tradicionalnih enciklopedijskih sadržaja. LZMK, kao ustanova s tradicijom izdavanja pouzdanih i vjerodostojnih sadržaja, razvija moderne pristupe informacijama, a istovremeno pridonosi očuvanju nacionalne leksikografske i enciklopedičke baštine. A za znatniji iskorak potrebna su dodatna ulaganja u digitalizaciju i objavu vrijednih leksikografskih sadržaja.

The Miroslav Krleža Institute of Lexicography and Digital Environment

The Miroslav Krleža Institute of Lexicography (LZMK) is a public institution of special status for the Republic of Croatia and the only Croatian institution that has been systematically engaged in lexicography and encyclopaedistics. In its nearly 70-year history LZMK has published valuable works of general knowledge as well as those focusing on national topics. Since 2009 LZMK is developing publicly available repository of encyclopaedic knowledge (http://enciklopedija.lzmk.hr), also known as *The Portal of Knowledge*, offering free access to more than 60,000 articles from seven digitised encyclopaedias and lexicons. LZMK realizes its mission and vision by creating a unique, publicly available digital repository and by facilitating access to public knowledge. The growth of ICT enables the availability of data and information in various digital formats. Consequently, encyclopaedic and lexicographic contents are being developed as digital, and LZMK, as well as other world publishers of encyclopaedias and lexicographic content, is actively addressing challenges and changes in the digital environment. The importance of the presence of lexicographic content in the digital environment has been recognized in online editions of the Croatian Encyclopaedia and the Croatian Biographical Lexicon, as two core publications of LZMK. The Croatian Encyclopaedia (http://enciklopedija.hr), online since 2013, is an encyclopaedic source of knowledge in Croatian language for all users interested in obtaining verified and reliable information about Croatia and

the world. The content development in the digital environment pose new challenges for the lexicographers as well as the encyclopaedic profession and significantly differs from traditional concept of lexicography and lexicographers. LZMK as an institution with tradition of issuing credible and reliable contents develops modern ways of access to information and contributes to preservation of national lexicographic and encyclopaedic heritage. For a more significant progress, additional investments are required to digitise and make valuable encyclopaedic and lexicographic content available online.

mrežnik
HrZZ Croatian Science Foundation
VERN'

Magdalena Kroupová
Czech Language Institute, Czech Academy of Sciences
kroupova@ujc.cas.cz

Barbora Štěpánková
Institute of Formal and Applied Linguistics, Charles University, Prague
stepankova@ufal.mff.cuni.cz

How to Reconcile Semantics and Grammar in a Monolingual Dictionary. The Case of the Verb "být" ("to Be")

The treatment of verbs in a monolingual dictionary requires a specific approach even though they are autosemantic words. In Czech (and some other languages), verbs are considered the centre of the sentence because of their function as predicates and their valency. Information on the syntactic configurations a verb enters, i.e. its valency frame, is an essential part of the verb as a lexical unit. It is therefore necessary to combine grammatical characteristics with semantics.

The verb "být" represents a very specific case. This most frequent verb in Czech functions as a lexical verb (primarily with an existential meaning), a light verb, and an auxiliary verb with various grammatical functions. Two different approaches are possible for sentences such as "židle je ze dřeva" (lit. chair is of wood) and "židle je dřevěná" (lit. chair is wooden), either focusing on their semantic similarity or taking their syntactic differences into account.

The diversity of forms of the verb "být" poses a specific issue. According to corpus data, the lemma "být" subsumes 105 word forms. This is related to this verb being used as an auxiliary in its various forms to express most of the grammatical categories of verbs. Some of its word forms are even considered independent words synchronically.

The aim of this paper is to suggest how to combine information about the semantics and grammatical functions of the verb "být" in a monolingual dictionary such that the result provides a user-friendly picture of the typical usage of this word. It also provides a draft of the dictionary entry for the verb "být" in the Academic Dictionary of Contemporary Czech. The treatment in this dictionary is based on synchronic corpora of Czech, as well as Czech monolingual and specialised dictionaries.

Ivana Lalli Paćelat
Fakultet za interdisciplinarne, talijanske i kulturološke studije
Sveučilišta Jurja Dobrile u Puli
ilalli@unipu.hr

Marija Brkić Bakarić
Odjel za informatiku Sveučilišta u Rijeci
mbrkic@uniri.hr

Isabella Matticchio
Fakultet za interdisciplinarne, talijanske i kulturološke studije
Sveučilišta Jurja Dobrile u Puli
imatticchio@unipu.hr

Razvoj prijevodnih tehnologija kao potpora službenoj dvojezičnosti u Istarskoj županiji – početni koraci

Službena dvojezičnost pretpostavlja svakodnevno stvaranje paralelnih tekstova u dvojezičnim područjima. Slučaj je to i Istarske županije, u kojoj se tekstovi obično sastavljaju na hrvatskome, a zatim se prevode na talijanski jezik. Radi službene prirode tekstova i konteksta uporabe talijanskoga jezika, vrlo je važno imati precizno i ujednačeno nazivlje kao i razvijene jezične tehnologije koje bi omogućile brži i kvalitetniji proces prevođenja pri dvojezičnim institucijama u Hrvatskoj. Cilj je ovoga rada prikazati dosadašnju praksu stvaranja nazivlja i prevođenja u Istarskoj županiji te ponudit moguća rješenja u stvaranju novoga odnosno ujednačavanju i provjeravanju postojećega nazivlja. U analizi nazivlja koristit će se korpusnolingvistički pristup, stoga je prvi korak priprema i izgradnja paralelnoga korpusa. Paralelni korpus bit će sastavljen od hrvatskih službenih javno dostupnih tekstova i njihovih prijevoda na talijanskome jeziku. Korpus će biti sravnjen i pripremljen za planirane lingvističke analize. Očekuje se da će analiza postojećega nazivlja ukazati na potrebu sustavnoga planiranja izgradnje, provjeravanja i normiranja talijanskoga nazivlja koji se tiče hrvatskoga pravnog sustava. Budući da se radi o nazivlju na manjinskome jeziku u Hrvatskoj, koji ima status nacionalnoga jezika u Italiji, neophodno je prilikom izgradnje,

usklađivanja i obradbe nazivlja uzeti u obzir razlike i posebnosti dvaju različitih pravnih sustava kao i trendove u prevođenju i prijevodnim tehnologijama pri sličnim dvojezičnim i višejezičnim institucijama u Europi. Na kraju će se rada stoga predstaviti daljnji koraci u izgradnji prijevodnih tehnologija koje bi olakšale prijevodnu aktivnost i omogućile stvaranje prijevoda visoke kvalitete.

The development of translation technologies as support to official bilingualism in Istria County – the first steps

Official bilingualism assumes the daily creation of parallel texts in bilingual areas. This is also the case in Istria County, where texts are usually written in Croatian and then translated into Italian. Due to the official nature of the texts and the context of Italian language usage, it is very important to have precise, uniform terminology, as well as developed language technologies that enable a faster and more accurate translation process in bilingual institutions in Croatia. The aim of this paper is to present the current practice of creating terminology and translations in the Istrian County and to offer possible solutions in creating new ones or unifying and examining existing terminology. The terminology analysis uses a corpus-based approach; the first step is thus to prepare and build a parallel corpus, which will consist of both official and publicly available Croatian texts and their Italian translations. The corpus will be harmonised and prepared for the planned linguistic analysis. We expect the analysis of existing terminology to prove the need for systematic planning in the creation, examination, and standardization of Italian terminology for the Croatian legal system. Since the terminology is written in a minority language in Croatia, which has the status of a national language in Italy, when creating, harmonising, and analysing terminology, it is necessary to consider differences and particularities between the two different legal systems, as well as trends in translation and translation technologies in similar bilingual and plurilingual institutions in the European Union. Lastly, we will present upcoming steps in the development of translation technologies that will facilitate translation activity and enable the creation of high quality translations.

Biljana Lazić
Rudarsko-geološki fakultet, Univerzitet u Beogradu
biljana.lazic@rgf.bg.ac.rs

Olivera Kitanović
Rudarsko-geološki fakultet, Univerzitet u Beogradu
olivera.kitanovic@rgf.bg.ac.rs

Ivan Obradović
Rudarsko-geološki fakultet, Univerzitet u Beogradu
ivan.obradovic@rgf.bg.ac.rs

Mogućnosti retrodigitalizovanog Nemačko-srpskog rudarskog rečnika

U radu će biti prikazan opis procesa retrodigitalizacije dvojezičnog Nemačko-srpskog rudarskog rečnika iz 1923. godine čiji je autor rudarski inženjer Dragutin Stepanović (Степановић, 1923). Ovaj rečnik je zasnovan na skoro 4 000 leksičkih zapisa koji su prevodilački ekvivalenti ili uputnice. Umesto predgovora autor daje uvid u svoje pismo upućeno "Ministru šuma i rudnika" u kome piše o nameri da zabeleži reči koje se koriste u narodu kako bi izbegao upotrebu nemačkih reči. Iako broj odrednica nije toliko veliki, rečnik može biti vredan izvor za različite istraživačke svrhe (istoriju rudarstva, terminologiju itd.). Rečnik će biti anotiran u skladu sa najnovijim TEI smernicama tokom radionice *Lexical Data Masterclass 2018*. Biće izloženi neki od problema prilikom upotrebe programa GROBID-Dictionaries (Khemakhem et al., 2018) nastali usled neujednačene strukture rečničkih zapisa. U drugom delu rada biće analizirana upotreba pojedinih pronađenih srpskih termina u odnosu na definicije i upotrebu u drugim rečnicima i resursima srpskog jezika.

The possibility of retro-digitising a German-Serbian Mining Dictionary

This paper will describe the retro-digitisation process of a bilingual dictionary concerning the domain of mining – the German-Serbian Mining Dictionary (Nemačko-srpski rudarski rečnik) by mining engineer Dragutin Stepanović, published in 1923 (Степановић, 1923). The dictionary consists of nearly 4,000 lexical entries that are translational equivalents or self-references. Instead of a preface, the author provides insight in a letter addressed to the "Minister of Forestry and Mining". He writes about his intent to note Serbian words used by "ordinary people" in order to avoid the usage of German words. Although the number of headwords is not large, the dictionary can potentially serve as a valuable resource for many different research purposes (mining history, terminology, etc.). The dictionary will be encoded in accordance with TEI guidelines during the Lexical Data Masterclass 2018 workshop. The authors will present some of the issues they faced during GROBID Dictionaries processing (Khemakhem et al., 2018). These issues are mainly a consequence of unstructured lexical entries. In the second part of the paper, specific Serbian terms will be analysed in comparison to definitions and usage in other Serbian dictionaries and resources.

Marija Lütze-Miculinić
Filozofski fakultet Sveučilišta u Zagrebu
mlmiculi@ffzg.hr

Zrinka Jelaska
Filozofski fakultet Sveučilišta u Zagrebu
zjelaska@ffzg.hr

Lovorka Zergollern-Miletić
Učiteljski fakultet Sveučilišta u Zagrebu
l.zergollern-miletic@ufzg.hr

Sandra Mardešić
Filozofski fakultet Sveučilišta u Zagrebu
smardesic@gmail.com

Višejezični e-rječnik razrednoga jezika (hrvatski, francuski, engleski, talijanski i njemački)

U radu se predstavljaju načela sastavljanja i oblikovanja višejezičnoga rječnika razrednoga jezika. Riječ je o jeziku kojim se govori u najrazličitijim razrednim situacijama. Razredni jezik uključuje glavnu komunikaciju, usredotočenu na nastavne ciljeve, ali i različite vidove sporedne komunikacije. U većini predmeta odvija se na materinskom jeziku, no kako se u suvremenoj nastavi stranih jezika teži što većem udjelu ciljnoga jezika, nameće se potreba za leksikografskim opisom jezičnoga varijeteta u kojem se ostvaruje glavna i sporedna komunikacija u razrednom okružju.

Rječnik je namijenjen nastavnicima i studentima stranih jezika, budućim nastavnicima koji žele ovladati sporazumijevanjem u razredu kakvo je prirodno u stranoj kulturi. Pritom se polazi od učeničkih potreba i navika unutar hrvatskoga konteksta. Stoga rječnik ima i svojevrsnu međukulturnu vrijednost.

Sam pojam razrednoga jezika novo je područje istraživanja u hrvatskome. Premda se u proteklih pedesetak godina poznavanje osnovnih obrazaca

jezične uporabe u razredu smatra jednom od temeljnih nastavničkih kompetencija, u Hrvatskoj zasad nema odgovarajućih priručnika. Izrazi iz područja razrednoga sporazumijevanja djelomično su uvršteni u opće jezične priručnike, poput dvojezičnih rječnika. Stoga je zajedničkim radom skupine stručnjaka u području inoga jezika (njemačkoga, engleskoga, talijanskoga, francuskoga i hrvatskoga) sastavljen višejezični e-rječnik razrednoga jezika s otprilike 500 natuknica, a u pripremi je još 1500 natuknica.

Kako je namjena rječnika zahtijevala predradnje u prikupljanju građe koje su uz uobičajene natuknice uključivale i prikupljanje odstupanja ili češćih pogrešaka hrvatskih učenika pojedinih stranih jezika, konačna se građa sastoji dijelom od neuobičajenih natuknica. Za svaki od jezika natuknice imaju raspon od jedne do više riječi, a neke čine i cijelu rečenicu. Uza svaku se navodi bar jedan primjer. Zasad se odjednom može pretraživati samo po jedan strani jezik uz hrvatski, a usporedba jezika moguća je samo u slijedu. Usporedbi pomaže činjenica da se u sastavljanju primjera na stranim jezicima pokušalo odabrati za svaki jezik što sličnije rečenice.

Multilingual dictionary of classroom language (Croatian, French, English, Italian and German)

The paper discusses the principles of compiling a multilingual dictionary of classroom language. The term 'classroom language' implies a type of communication in various classroom situations. Classroom language refers to the main classroom communication, focused on teaching goals, but also to various other aspects of communication in the classroom. In most classes this type of communication is carried out in L1. Since modern teaching of foreign languages encourages the use of the target language, it is necessary to provide a lexicographical approach to the language variety which is used for the primary and secondary communication in the classroom context.

The target audience of the dictionary are teachers of foreign languages, as well as students of foreign languages – future teachers, who wish to master the type of classroom communication that is specific to a foreign

culture. Nevertheless, the starting point are learners' needs and habits that exist within the Croatian educational context. Therefore, the dictionary may be useful in the development of intercultural competence.

The concept of classroom language represents a new field within the studies of Croatian. Over the past fifty years mastery of the basic patterns of classroom communication has been considered to be one of the basic teacher competences. Nevertheless, no books or manuals regarding classroom communication exist in Croatia. Reference to classroom communication may be found in various general reference books, such as bilingual dictionaries. For all the above reasons, a group of experts on foreign language acquisition (German, English, Italian, French and Croatian) decided to compile a multilingual e-dictionary of classroom language, which contains about 500 entries, with another 1500 in preparation.

In addition to compiling usual dictionary material, the envisaged use of the dictionary required the compilation of examples of deviations and common mistakes by Croatian learners of foreign languages. As a result, the dictionary contains a number of rather uncommon entries. Each entry for each language may include one or more words, or even sentences.

For the time being only one foreign language can be searched alongside Croatian, and a parallel comparison is not as yet possible. The comparison has been rendered easier by the fact that the collaborators tried to find similar sentences for each language.

,

Екатерина Петровна Любецкая
Белорусский государственный университет
katerina_lingvo@mail.ru

Переводной компьютерный лексикон: способы достижения эквивалентной соотносительности

Разработка переводного немецко-белорусского компьютерного терминологического словаря nebelex.ru соответствует современным тенденциям лексикографирования: возможности компьютерных систем значительно упрощают сбор, обработку и анализ языковых данных. Данные немецко-белорусского словаря значительно упрощают организацию межкультурной коммуникации, в доступной форме презентуют широкий выбор переводных решений специальной лексики, содержат богатый материал для решения разноаспектных лексикографических задач.

A translational computer lexicon: ways to achieve equivalent correlation

The development of the German-Belarusian translational computer terminology dictionary nebelex.ru corresponds to current trends in lexicography; the capabilities of computer systems greatly simplify the collection, processing, and analysis of language data. The data in the German-Belarusian dictionary greatly simplify the organization of intercultural communication, presenting a wide choice of translational solutions of special vocabulary and rich material for solving multidimensional lexicographic tasks in an accessible form.

Nikola Ljubešić
Jožef Stefan Institute
Faculty of Computer and Information Science,
University of Ljubljana
nljubesi@gmail.com

Tanja Samardžić
University of Zürich
tanja.samardzic@uzh.ch

Tomaž Erjavec
Jožef Stefan Institute
Faculty of Computer and Information Science,
University of Ljubljanatomaz.erjavec@ijs.si

Darja Fišer
University of Ljubljana
darja.fiser@ff.uni-lj.si

Maja Miličević Petrović
University of Belgrade
m.milicevic@fil.bg.ac.rs

Simon Krek
Centre for language resources and technologies,
University of Ljubljana simon.krek@guest.arnes.si

Vuk Batanović
University of Belgrade
vukbatanovic@sbb.rs

The "ReLDI effect": Collaborative development of manually annotated datasets for Slovene, Croatian and Serbian

With the rapid development and increasing accessibility of natural language processing (NLP) techniques, the exploitation of NLP inside electronic lexicography is on a rise. Textual datasets manually annotated with linguistic information are a backbone of the currently dominating

paradigm in NLP based on supervised machine learning. However, developing such manually annotated datasets is a very costly activity, which is one of the reasons for limited availability of NLP technologies for languages with fewer speakers, and especially for less dominant language varieties such as the language of the Internet.

In this talk we present a series of collaborations between researchers developing such datasets for Slovene, Croatian and Serbian, three languages with just a few million speakers each. Close relatedness of these languages brings an opportunity for a synchronized approach to the development of resources and technologies, to the benefit of all parties. Due to the complex political environment, however, such an approach has not been established until the start of the ReLDI (Regional Linguistic Data Initiative) project. The main synergistic effect of the collaborations presented here is achieved by drastically lowering the efforts required to produce datasets in additional languages, primarily in the areas of (1) the development of annotation guidelines, (2) setting up the technical requirements for the annotation campaigns and (3) pre-annotation of data with models trained for another, but very close language.

The linguistic levels covered in the resulting datasets are those of tokenisation, sentence segmentation, normalisation, morphosyntax, lemmatisation, dependency parsing, semantic role labeling, named entity recognition and coreference resolution. Two varieties of each of the three languages are covered: the standard variety and the variety of the language of the Internet.

Ivana Matas Ivanković
Institut za hrvatski jezik i jezikoslovlje
imatas@ihjj.hr

Brojevi u *Hrvatskome mrežnom rječniku – Mrežniku*

Brojevi se obično definiraju kao riječi koje nam govore koliko jedinica ima onoga što znači riječ uz koju stoje ili koliko jedinica treba odbrojiti da se dođe do nekog predmeta. Brojevi su raznolika skupina, mogu biti jednočlani i višečlani, neki su sklonjivi, neki su nesklonjivi, u brojeve se ubrajaju i brojevne imenice i brojevni pridjevi pa se i tipovi sklonidbe razlikuju. Unatoč tome, opisu njihove tvorbe posvećeno je malo prostora. Babić u *Tvorbi riječi u hrvatskome književnome jeziku* to tumači ovako: „Koliko među zamjenicama, brojevima, prijedlozima i veznicima i ima tvorbenih riječi, to ih je samo nekoliko i tipovi više nisu plodni. To su većinom za tvorbu riječi zatvoreni sustavi, jer nema novih osnova, često s malim brojem jedinica…" (Babić 2002: 15). U *Hrvatskome mrežnom rječniku – Mrežniku* predviđeno je da se obradi i tvorba riječi, no pri obradi brojeva i riječi motiviranih brojevima pokazalo se da dosadašnji opisi tvorbe koji se nalaze u gramatikama ne nude sva rješenja, što ih čini manjkavima. Problem je određivanje tvorbe izvedenih brojeva (npr. *jedanaest*, *dvanaest*…), ali i tvorenica koje se mogu uvrstiti u više tvorbenih modela (npr. *četrdesetogodišnj-ica* / *četrdeset-o-godišnj-ica*). U radu će se analizirati tvorbeni modeli, odredit će se osnova i tvorbeni nastavak brojeva i riječi tvorenih od brojeva koji su obrađeni u *Mrežniku*. Neki tvorbeni modeli moći će se prikazati kao nizovi (npr. *trogodišnji* > *trogodišnjak* > *trogodišnjakinja* > *trogodišnjakinjin*), dok neki nisu tako plodni (npr. *tri* > *trojka*).

Numbers in the *Croatian Web Dictionary – Mrežnik*

Numbers are usually defined as words that describe how many units there are of what the word denotes, or how many units should be counted to come to some object. They are a heterogeneous group; they can be simple

or complex, some are declinable while some are indeclinable. As a word group, numbers include nouns and adjectives, so the types of declination are different. Despite this diversity, the description of their word formation is not as detailed. In *Tvorba riječi u hrvatskome književnome jeziku*, Babić explains: "Among the pronouns, numbers, prepositions, and conjunctions, there are words formed from others, but these are just a few, and the types of formation are no longer productive. From the aspect of word formation, these are mostly closed systems, because there are no new root words, and there is a small number of units…" (Babić 2002: 15). Word formation data will be provided in the *Croatian Web Dictionary – Mrežnik*, but when describing numbers and words motivated by numbers, it has been shown that current descriptions in grammars are not sufficient. Problems lie in determining the formation of some derived numbers (e.g. *jedanaest* 'eleven', *dvanaest* 'twelve'...), as well as word forms that can be interpreted in more than one way (e.g. *četrdesetogodišnj-ica* / *četrdeset-o-godišnj-ica* 'fortieth anniversary'). In this paper, the word formatting models of numbers and words formed from the numbers included in *Mrežnik* will be analysed and roots and suffixes will be determined. Some formations can be shown as strings (e.g. *trogodišnji* > *trogodišnjak* > *trogodišnjakinja* > *trogodišnjakinjin*), whereas some are not so fruitful (npr. *tri* > *trojka*).

Maja Matijević
Filozofski fakultet Sveučilišta u Zagrebu
majamatijevic5@gmail.com

Krešimir Pavlina
Filozofski fakultet Sveučilišta u Zagrebu
kpavlina@ffzg.hr

Bernardina Petrović
Filozofski fakultet Sveučilišta u Zagrebu
bernardina.petrovic@ffzg.hr

Emotnik kao poticaj interaktivnoj nastavi

E-rječnik *Emotnik* jednojezični je, korpusno utemeljen, dinamičan, jednostavno pretraživ mrežni rječnik hrvatskoga jezika koji će nakon prve faze rada biti dostupan vanjskim korisnicima. Izrađuje se od ak. god. 2017./2018. u okviru studentskoga projektnog zadatka *Rječnički opis izraza za emocije u hrvatskome jeziku* na Odsjeku za kroatistiku Filozofskoga fakulteta Sveučilišta u Zagrebu i u suradnji s Odsjekom za informacijske i komunikacijske znanosti istoga fakulteta. U prvoj se fazi izrade rječnika obrađuju jezični izrazi za dvanaest emocija – šest primarnih (radost, žalost, strah, ljutnja, iznenađenje i gađenje) i šest odabranih sekundarnih (nada, ljubav, mržnja, ljubomora, zavist i stid). Rječnički se članak sastoji od jednorječne ili višerječne natuknice, gramatičke obavijesti, semantičkoga opisa i potvrda prikupljenih iz pisanih i govorenih izvora. Svi se podaci unose izravno u mrežnu bazu koja je nastala u okviru ovoga projektnog zadatka, i to radi bolje preglednosti i sigurnosti prikupljenih podataka, poticanja studentske suradnje i kasnijega jednostavnijeg kreiranja sučelja za vanjske korisnike. Metapodaci korišteni u sustavu omogućuju studentima jednostavno unošenje pronađene građe, a vanjskim korisnicima jednostavno pregledavanje i pretraživanje prema više kriterija. *Emotnik* je zamišljen ne samo kao poticaj interaktivnoj nastavi u kojemu su studenti aktivno uključeni u sve faze izrade rječnika nego i kao poticaj suradnji odsječkih sastavnica istoga fakulteta. Slobodna dostupnost korisnicima i dinamičnost kojom se otvara mogućnost daljnje

dorade dvije su važne pretpostavke za pronalaženje mjesta ovomu rječniku u hrvatskoj e-leksikografiji.

Emotnik as stimulus for interactive teaching

The *Emotnik* e-dictionary is a monolingual, corpus-based, dynamic, easily searchable online dictionary of the Croatian language that will be available to users after the first phase of the work. It has been in developed since the 2017/2018 academic year as a part of a student project assignment called *Dictionary Description of Emotional Expressions in the Croatian Language* at the Department of Croatian Language and Literature at the Faculty of Humanities and Social Sciences, in cooperation with the Department of Information and Communication Sciences at the same faculty. In the first phase of the project, language expressions for twelve emotions were processed – six primary ones (joy, sorrow, fear, anger, surprise, and disgust) and six selected secondary ones (hope, love, hatred, jealousy, envy, and shame). Dictionary entries consist of one-word or multi-word entry words, grammatical information, semantic descriptions, and examples found in written and spoken sources. All data were entered directly into an online database created within the framework of this project: for better visibility and security of the collected data, encouraging student collaboration and later simplifying the creation of user interface. The metadata used in the system allow students to easily enter the corpus, and also allow users to easily browse and search by multiple criteria. *Emotnik* is conceived not only as a stimulus for interactive teaching in which students have been actively involved in all phases of development, but also as a stimulus for collaboration between departments of the faculty. Free accessibility for users and openness with the possibility of further processing, changing, and improving are two important prerequisites for this dictionary to find its place in Croatian e-lexicography.

Peter Meyer
Institute for the German Language Mannheim
meyer@ids-mannheim.de

Thinking in networks: Towards a graph-augmented lexicography

Graphs can be used as a complementary data and access structure superimposed on lexicographical XML documents, leading to what Měchura (2016) calls graph-augmented trees – in contradistinction to a strictly "graph-only" approach more suitable for NLP applications (Gracia, Kernerman & Bosque-Gil 2017). A property graph can be used to represent intra- and cross-document relations as well as supplementary or meta-information in what would otherwise just be an unstructured set of entries or XML documents.

However, the complexity of even moderately sized graphs is a barrier to efficiently handling such graph-augmented resources for several reasons: The editing process involves highly non-linear graph operations; graph traversal concepts are challenging for end users and add computational complexity; the two loosely coupled data layers (XML and graph) make change management and presentation more difficult.

The paper proposes a set of practices to overcome these difficulties and illustrates them through the example of a collection of heterogeneous loanword dictionaries (where the graph expresses, inter alia, arbitrarily long chains of lexical borrowings): (1) Hierarchically structured human-readable markup remains at the center of lexicographical activity and of the default presentation. The graph serves mainly as an access structure. (2) A real-time graph query system powers systematic editing and proofreading cycles in the lexicographic process and also gives end users advanced data access options. The paper uses the example of an open-source project (Meyer & Eppinger 2018) to show how a visual query builder can be used to express almost arbitrarily complex graph constellations in an intuitive way. (3) Bookkeeping of data changes should itself be graph-driven, e.g. by (a) dynamically flagging search results with 'todo' attributes and (b) separating concerns (such as raw vs. edited data, versioning, user comments etc.) through systematically interconnected subgraphs.

Josip Mihaljević
Institute of Croatian Language and Linguistics
jmihalj@ihjj.hr, jomihx@gmail.com

Gamification of *Croatian Web Dictionary – Mrežnik*

Web dictionaries sometimes contain multimedia such as games. Games usually accompany lexicographic works for children and can appear on special optical media. Now, most multimedia content accompanying the content of lexicographic works is published online. Examples of dictionaries that have games for learning vocabulary are Merriam-Webster dictionaries; in Croatia, games come on CD with the printed *First School Dictionary*, while online games accompany the *First Croatian Orthography Manual*. There is also a link to a typing game on the site of the *Croatian Archive Dictionary*. Unfortunately, most dictionary websites do not contain entertaining educational content such as games to help non-native speakers and children learn certain words, different forms of words, semantic and syntactic relations between words, spelling, pronunciation, etc. Gamification is a process in which gaming mechanics and elements are applied to a non-gaming situation to make it more entertaining. This paper will explain the process of gamifying the dictionary content of the *Croatian Web Dictionary – Mrežnik*, especially dictionary modules for children and non-native speakers. Gamification will be conducted by using games for learning vocabulary. There will be a demonstration of many different games that can be implemented into the dictionary structure. These games include quizzes, memory games, crossword puzzles, drag-and-drop games in which pictures and words are connected, etc. The technological solutions used to develop these games will be explained. Other gamification elements that can be implemented into online dictionaries, such as badges for successfully finding words or leaderboards in competitive games, will also be presented.

Nives Mikelić Preradović
Faculty of Humanities and Social Sciences, University of Zagreb
nmikelic@ffzg.hr

Error-tagging of CroLTeC (computer learner corpus of Croatian as a foreign language)

The paper describes the error-tagging scheme developed for the CroLTeC learner corpus (http://teitok.iltec.pt/croltec/index.php?action=home) - the first computer learner corpus of Croatian as a foreign language. CroLTeC contains essays collected from 755 students with 36 different mother tongues, among which the most prominent were Spanish, English, German, Polish, Chinese, French and Arabic. It consists of 6,213 essays, out of which 1,217 were digitally born, while 4,996 essays were scanned, transcribed in RTF format and converted into XML format. CroLTeC has a total of 1,054,287 tokens, and essays have been collected on all 6 CEFR levels of language learning at Croaticum – Center for Croatian as Second and Foreign Language at the Faculty of Humanities and Social Sciences in Zagreb. All CroLTeC essays contain metadata about the title, number and type of essay (homework, part of exam or field class, etc.). Data were lemmatized and annotated with morphosyntactic tags with the RELDI tagger (Ljubesic et al., 2016). Also, the corpus ise searchable by age, sex, language proficiency level and the mother tongue of the learner.

The error-tagging scheme is partially based on Solar (the scheme of Slovene's developmental corpus) and the error-coding of the Cambridge Learner Corpus and further tailored to Croatian language. The goal of the development of the error-annotation scheme is to build a sub-corpus that will serve as a repository of authentic data about the learner's interlanguage. It should enable researchers and teachers of Croatian as a foreign language to explore the interlanguage, to discover the aspects of the grammar that are the most difficult to master and to tailor teaching materials to different groups of learners (not only according to their Croatian language proficiency level, but also to their first language). Finally, the error-tagged sub-corpus should also serve as a starting point for designing computer-aided tools to correct lexical errors, misuse of verbal tenses, phrasal verbs and collocations.

Валерия Морозова
Национальный исследовательский университет, Высшая школа экономики tito_alba@mail.ru

Разработка и сопровождение Электронного словаря грецизмов и полонизмов в русском языке XI–XVII вв.

Проект ориентирован на создание первого в российской науке «Электронного словаря грецизмов и полонизмов в русском языке XI–XVII вв.» на основе материалов, созданных М.И. Чернышевой и Е.И. Державиной (ИРЯ РАН).

Информация для генерации словарных статей скомпонована в виде реляционной базы данных на основе системы управления SQLite. Данная система выбрана исходя из отсутствия необходимости активного редактирования содержимого базы данных; после конечной реализации сайта и наполнения базы данных в случае необходимости правок планируется замена файла базы данных на сервере. База данных содержит таблицы *word, phonetics, morphology, sources, philol_info*, в которых находится 18 полей. Греческие и польские заимствования дифференцируются с помощью тега *lang*, также внесенного в базу данных.

Разработка сайта производится на базе фреймворка Django на языке программирования Python (версия 3.5+). Интерфейс сайта скомпонован на основе шаблонов из библиотеки Bootstrap и располагает разделами “информация о проекте”, “контакты” и “словарный поиск”. Планируется разработка поиска по заголовочным леммам, который для удобства будет восприимчив к фонетическим и морфологическим вариантам слов. Например, при отправлении запроса, для которого не существует отдельной словарной статьи, будут выдаваться те статьи, в которых данный запрос указан как зафиксированный фонетический вариант. Наибольшую ценность электронного сайта составит продвинутый поиск по широкому диапазону полей: поиск по дате создания источников, в которых зафиксированы леммы; поиск по цитатам и собственно источникам, поиск по тематическим группам. Планируется поддержка гибкого поиска: поиск по началу и/или концу слова, выдача похожих на запрос лемм. Данные функции реализованы на базе регулярных выражений, которые применяются на диапазоне заголовочных лемм, из фонетических и морфологических вариантов.

Созданный в результате данного проекта сайт словаря будет доступен для общего пользования в сети Интернет под оригинальным доменным именем.

The development and processing of the Electronic Dictionary of Greek Words in 11th- to 17th-century Russian

The purpose of this project is to create the first digital dictionary of Greek words in 11th- to 17th-century Russian on the basis of materials created by E. Derzhavina and M. Chernysheva (Institute of Russian Language).

The data for generating dictionary entries is structured as a relational database using the SQLite management system. This system was chosen as there is no need to frequently edit database entries. If corrections are needed after the website is launched, the database file will be replaced with an updated one. The database contains the tables *word, phonetics, morphology, sources, philol_info*, including 18 columns. Polish and Greek loanwords are differentiated with a special *lang* tag, which is also present in the database.

Website development is based on the Django framework and programmed in Python (3.5+ version). The website interface was created on the basis of Bootstrap library templates; it will contain "contacts", "project information", and "search" sections. As regards the search system, search by lemmas is planned, which will include matches with morphological and phonetic variants. For instance, if a search query does not result in a match with a dictionary entry, the user will receive links to articles in which the lemma is mentioned as a phonetic or morphological variant. The most valuable function of the website is an advanced search function that will allow queries for bounded periods of source text creation, in which lemmas are mentioned in source content and thematic groups. Flexible searches including search by beginning and/or end of word is also planned. This feature will be applied to the use of regular expressions on a range of lemmas and morphological and phonetic variants.

The website will be available for public use on the Internetvia an original root domain name.

Christine Möhrs
Institute for the German Language Mannheim
moehrs@ids-mannheim.de

Sarah Torres Cajo
Institute for the German Language Mannheim
torres@ids-mannheim.de

The microstructure of a lexicographical resource of spoken German: meanings and functions of the lemma *eben*

Within the project "Lexik des gesprochenen Deutsch" ('lexis of spoken German') we aim to develop a corpus-based dictionary of standard spoken German. Currently, the consideration of authentic spoken language and its specific interactional characteristics is a desideratum in lexicographic works. For instance, the label "spoken" is usually assigned to a lemma or a construction as part of pragmatic information (style/register). It is unclear which data and methods are employed to derive this kind of information. Thus, our project additionally intends to develop adequate methods to analyze, structure, and describe items of spoken German.

All lexicographic information is based on publicly available authentic data of everyday communication which is accessible through the DGD ('Database of Spoken German'). Specifically we are employing FOLK ('Research and Teaching Corpus of Spoken German'), a corpus of 281 transcribed private, institutional and public interactions.

In order to adequately present the interactional functions of a lemma or a construction we had to develop a microstructure containing new and innovative types of lexicographic information. Within a function-based type of article we employ specific examples to illustrate the interactional role of the described item as well as sequential, syntactic and prosodic attributes. Basic meanings of lemmas are described in a more traditional type of article. In a summary article, function- and meaning-based articles are interlinked so that users have an overview on the use of a specific lemma in spoken German. In this talk we will

present our innovative microstructure on the basis of the lemma *eben* which can appear as an adverb (engl. *just* [*now*]) as well as a modal or discourse particle with various interactional functions. For instance, speakers can use *eben* to build coherence by referring to a pragmatic pretext in the prior interaction and thus update a conversational topic across a considerable length of time.

Anja Nikolić-Hoyt
Zavod za lingvistička istraživanja HAZU
anhoyt@hotmail.com

Karlo Schubert
karlo.schubert@gmail.com

Konstrukcija i struktura *Somatskoga tezaurusa hrvatskoga jezika*

Tema izlaganja je *Somatski tezaurus hrvatskoga jezika*, konceptualno ustrojen leksikografski resurs motiviran čovjekom i dijelovima čovječjega tijela, koji se izrađuje u Zavodu za lingvistička istraživanja HAZU. U prvom se dijelu sažeto govori o teorijskim postavkama i arhitekturi *Somatskoga tezaurusa*, a u drugom se također u kratkim crtama iznose pitanja i problemi njegove računalne aplikacije.

Naime, zamišljen kao *riznica* u kojoj se grupira zajedno i čuva ili *tezaurira* sve što je vezano uz tijelo, *Tezaurus* objedinjuje različite tipove odnosa koji uz semantičke odnose u užem smislu kao što su sinonimija/antonimija te hijerarhijsko-hiponimijski i meronimijski odnosi vertikalne podređenosti, uključuje i druge tipove odnosa poput atribucije svojstava, funkcionalnih ili pak asocijativnih odnosa, čime odstupa od ustaljenih tezaurusnih struktura. Navode se i karakteristične frazeološke sveze. Primjerice, uz krovni somatizam KOSA tek po jedan primjer za svaki tip odnosa: *čupa* (stilistička sinonimija), *pokrovni sustav* (hiperonimija), *vlas* (meronimija), dakle tradicionalni odnosi, ali i novi *plava kosa* (atribucijski), *češljati kosu* (funkcionalni), te *ukosnica* (asocijativni); također i karakteristične frazeološke sveze (*diže se kosa <na glavi> komu* [zaprepašten je, prestravljen je tko] i drugi izrazi prenesenog značenja (*sijeda glava* [zrela, mudra osoba]), koji elaboriraju svojstva i funkcije somatskog koncepta KOSA. U primjeni to znači da se svaki od ukupno 55 krovnih somatskih koncepata oprimjeruje odnosno orječuje što većim brojem riječi odnosno leksičkih dokaza različitih aspekata svoga konceptualnog sadržaja, pri čemu svakom aspektu toga sadržaja odgovara jedna sastavnica tezaurusne arhitekture: (1) ulazni članak (2) hiperonimija (3) meronimija (4) hiponimija (5) atribucija (6) funkcionalni odnosi (7) frazeološke sveze i drugi izrazi prenesenog značenja (8) poslovice. Također treba naglasiti

da su uz semantičku motivaciju neki odnosi u tezaurusu utemeljeni na morfološkoj derivaciji, na primjer: *kosica, kosurina, kosat, češljati/počešljati, raščešljati, začešljati* i drugi.

U drugom će se dijelu govoriti o računalnoj podršci odnosno o modulu unosa polistrukturirane tezaurusne baze te o modulu njezine objave.

The Construction and Structure of the *Somatic Thesaurus of the Croatian Language*

The topic of this paper is the *Somatic Thesaurus of the Croatian Language,* a conceptually organized lexicographic resource motivated by the human body and its parts that is currently being compiled at the Linguistic Research Institute of the Croatian Academy of Sciences and Arts. In the first half of the presentation, we will discuss the theoretical postulates and architecture of the *Somatic Thesaurus*, and in the second half, we will briefly present questions and problems concerning its computer application.

Conceived as a *treasury* in which everything concerning the body is kept, this thesaurus unites different types of relations that, in addition to semantic relations in the narrow sense, such as synonymy/antonymy and hierarchical-hyponymic and meronymic relations of vertical subordination, also includes other types of relations such as the attribution of properties, functional or associative relations, by which it deviates from traditional thesaural structures. Characteristic idiomatic expressions are also discussed. For example, in addition to the macro somatism KOSA 'hair', one example of each type of relation are as follows: *čupa* (stylistic synonymy), *pokrovni sustav* (hyperonym), *vlas* 'strand of hair' (meronymy), that is, traditional relations, but also newer ones such as *plava kosa* 'blond hair' (attributional), *češljati kosu* 'to comb one's hair' (functional), and *ukosnica* (associative); as well as characteristic idiomatic expressions (*diže se kosa* <*na glavi*> *komu* 'someone's hair stood on end' [someone was shocked] and other expressions (*sijeda glava* 'gray head' [wise person]), which elaborate properties and functions of the somatic concept KOSA. In practice, this means that each of the 55 macro

somatic concepts generate a great number of words or lexical evidence of different aspects of their conceptual meaning, where each aspect of that meaning corresponds to one component of the thesaural architecture: (1) entry (2) hyperonymy (3) meronymy (4) hyponymy (5) attribution (6) functional relations (7) idiomatic and other expressions (8) proverbs. It is also important to point out that, in addition to semantic motivation, some relations in the thesaurus are based on morphological derivation (e.g., diminutive *kosica,* augmentative *kosurina,* etc.*)*.

In the second part, we will discuss the computer support, that is, the module of entering the polystructured thesaural database and the module of its publication.

Mariia Novak
Kazan Federal University
Kazan Scientific Centre of Russian Academy of Sciences
mariaonovak@gmail.com

Yana Penkova
V.V. Vinogradov Russian Language Institute of Russian Academy of Sciences
amoena@inbox.ru

Nataliya Kuleva
V.V. Vinogradov Russian Language Institute of Russian Academy of Sciences
kulevana@mail.ru

Polonisms in an Electronic Historical Dictionary of Borrowings in the 11-17-Century Russian[1]

This paper represents a project for a new electronic historical dictionary of Greek and Polish loanwords in 11th-17th-century Russian, specifically the part that describes loanwords from or via the Polish language. Polish influence on the Russian vocabulary was most significant in the 15-17th centuries. During this period, the Russian language actively borrowed native Polish vocabulary and continued to learn Greek and Latin lexis, but not directly as before, but through the intermediary of western European languages, including Polish.

The electronic historical dictionary of loanwords is based on 11th-17th-century Russian-language dictionary materials, supplemented by other sources (of which a dictionary of Polonisms by W. Witkowski [Kraków, 2006] is essential). Its concept implies information on the chronology, thematic diversity, and complex adaptation of foreign language vocabulary. Data on morphological characteristics, phonetic and morphological variability, etymology, date, quotation and sources, semantics, attestation

[1] This study is supported by the Russian Foundation for Basic Research (research project no. 7-29-09113 офи-м)

in other historical dictionaries of East Slavic languages, and thematic groups will be placed in various zones of dictionary entries, as will additional philological or historical and cultural commentaries. The dictionary should become a fundamentally new tool for research on historical lexicology.

We focus here on the following lexicographical and theoretical problems related to the preliminary stage of the database compilation:

Etymological problems (establishing vocabulary content, taking into account a critical review of existing lexicographic sources, separating loanwords from cognates; identifying ways of borrowing, interpreting phonetic and grammatical adaptation of borrowings, including cases in which different phonetic and morphological variants are borrowed by different routes and from different languages);

Chronological problems (clarification of dates, identification of second borrowing [re-borrowing], which initially came to the Old Russian from the Greek language, semantic borrowings);

Systemic problems (establishing the cultural and historical context of borrowing, taking into account the thematic classification of vocabulary), and others.

Jacek Nowakowski
Lingwistyczna Szkoła Wyższa w Warszawie

Jan Franciszek Nosowicz
Lingwistyczna Szkoła Wyższa w Warszawie
jnosowicz@lingwistyka.edu.pl

W kwestii fiksacji onimów w e-słownikach, e-encyklopediach i w leksykografii korpusowej

Jednym z czynników przynależności kulturowej tekstu decyduje istniejąca w każdym języku bardzo duża grupa nazw własnych, nazywanych w onomastyce onimami. Najnowszą dziedziną onomastyki najbardziej dyskusyjną i najszybciej się rozwijającą jest – ergonimia. Najczęściej uważa się, że ergonimy to nazwy instytucji, przedsiębiorstw, partii, organizacji, stowarzyszeń itp. Spośród nich specyficzną i całkowicie zwartą grupę tworzą nazwy instytucji, których zadaniem jest kształcenie. Natomiast niniejszy tekst dotyczy tylko nazewnictwa instytucji szkolnictwa wyższego w Polsce.

Podstawowym źródłem omawianych w niniejszym artykule onimów jest Rejestr instytucji szkolnictwa wyższego opublikowany w Zintegrowany system informacji o nauce i szkolnictwie wyższym POLon (http://polon.nauka.gov.pl/).

Zawiera on 527 nazw wszystkich jednostek szkolnictwa wyższego w Polsce (bez jednostek filialnych), w tym 144 nazw uczelni państwowych i 383 prywatnych. Zebrane nazwy wydawałoby się, że są dość schematyczne, niemniej z różnych względów zasługują na uwagę.

Analiza nazw własnych, a szczególnie nazewnictwa uczelni, wymaga specyficznego podejścia. Ogólnie rzecz biorąc można stwierdzić, że nie tworzą one kategorii gramatycznej, nie są też typową kategorią leksykalną. Świadczy o tym ich wyjątkowo wysoka skłonność do tworzenia wyrażeń syntaktycznych, odporność na przekształcenia w procesie translacji i stosunkowo niewielka podatność na zmiany, wywołane uwarunkowaniami pozajęzykowymi.

W niniejszym referacie chcemy również przedstawić tendencje zmian w

nazewnictwie szkół wyższych w Polsce szczególnie z punktu widzenia dostosowania ich do wymogów obowiązujących od 1.X.2018 r. w Prawie o szkolnictwie wyższym.

On the fixation of onyms in e-dictionaries, e-encyclopedias and corpus lexicography

The cultural affiliation and cultural membership is predominantly determined by a significant group of proper names, knows as onyms. One of the most debatable and fastest growing domains within the realm of onomastics is ergonymy. The family of ergonyms includes the names of institutions, political parties, organisations, associations; this family defines the names of educational establishments. The presentation discusses the nomenclature of institutions of higher education in Poland.

The names of institutions of higher education covered by the presentation have been extracted from the Register of Institutions of Higher Education released in the Integrated System of Information on Science and Higher Education, otherwise known as POLon (http://polon.nauka.gov.pl/).

This register contains 527 names of institutions of higher education, including 144 state universities and 383 private establishments. Although the collected names appear to adhere to highly schematic linguistic principles, their linguistic properties deserve particular attention.

The analysis of proper names, including the nomenclature of universities, requires a specific approach. Viewed generally, they form neither a separate grammatical category nor a lexical class. Syntactic valency, resistance to translational transformation, and low responsiveness to linguistic modification are the major features of ergonyms.

This paper aims to present current trends and tendencies in the nomenclature of universities and institutions of higher education, based on the Polish Law on Higher Education which entered into force on October 1, 2018.

Marko Orešković
MathOS Osijek
moreskovic@nsk.hr

Domagoj Ševerdija
Odjel za matematiku Sveučilišta Josipa J. Strossmayera u Osijeku
dseverdi@mathos.hr

Mario Essert
Fakultet strojarstva I brodogradnje Sveučilišta u Zagrebu
messert@fsb.hr

Tri hrvatska e-rječnika u svjetskome LLOD oblaku

Jezikoslovni povezani podatci (LLOD - http://linguistic-lod.org/llod-cloud) moderan su način mrežnog objavljivanja lingvističkih podataka i mrežne obradbe prirodnoga jezika, a koji je utemeljen na nekoliko načela, od kojih su za jezikoslovce najvažniji sljedeći: a) podatci se mogu slobodno dohvaćati uz uobičajena autorska prava (npr. uz Creative Commons licenses); b) dio su općeg, tj. javnog repozitorija jer koriste iste mrežne standarde (RDF, JSON-LD i sl.) pa se mogu s lakoćom povezivati s drugim podatcima; c) sve se aktivnosti mogu pratiti preko običnog mrežnog preglednika. Povezani grafovi jezikoslovnih podataka najfleksibilniji su format za interoperabilnost, jer se RDF modeli iz različitih jezikoslovnih resursa, čak i različitih jezika, mogu s lakoćom integrirati. Povezanost jednog LLOD izvora s drugim izvorima otvara dodatnu semantiku, tj. pomaže korisnicima otkrivanje dodatnih izvora u globalnoj mreži. Razlog tomu je što su izvori, na koje je povezan neki model, na identičan način povezani i s drugima u mreži, za koje korisnik u početku niti ne zna da postoje. Radi se o jezikoslovnom podgrafu globalnoga mrežnoga oblaka (LOD) s više od 90 milijardi trojaca, tj. grana povezanih podataka u grafu. Od travnja 2018. godine sintaksno-semantički je okvir SSF kao čvor CroLLOD u DataHub-u postao dio globalnog LLODa s preko 20.000 svojih SSF-poveznica na BabelNet ontologiju i preko 67.000 poveznica na LexInfo. Mrežni podatci hrvatskih leksikona svakodnevno se upotpunjuju i

odgovarajuće opisuju preko OWL, Lemon ili NIF modela. Međutim, nisu samo povezanosti riječi iz triju vrsta SSF e-rječnika s drugim riječima globalnoga oblaka jedina vrijednost ovoga rada. Snaga leži u dohvatljivim semantičkim uzorcima ('sentic patterns') koji su zadnji iskorak u svijetu semantičke analize velikih ('big data') podataka, a koje SSF ima dobro razvijene u svojoj O-strukturi – sintaktičkim uzorcima gramatičko-semantičkih obilježja. Radi se, dakle, o kvalitetnom ulasku i uključenju hrvatskoga jezika među druge svjetske jezike, na mjesto koje mu već stoljećima pripada.

Three Croatian e-dictionaries in the worldwide LLOD Cloud

The goal of the Linguistic Linked Open Data movement (LLOD; http://linguistic-lod.oeg/llod-cloud) is to publish data for linguistics and natural language processing based on the following principles: a) data should be openly licensed, using licenses such as the Creative Commons licenses; b) resolving an LLOD resource should return results using web standards, such as HTML, RDF or JSON-LD, and the elements in a dataset should be uniquely identified by means of a Uniform Resource Identifier (URI); and c) users can access more information using web browsers. Including links to other resources can help researcher discover additional resources and provide semantics, by linking to other models which they may not have been aware existed. LLOD is a linguistic subgraph of the global network cloud (LOD), with more than 90 billion triples (branches of related data in the graph). Finally, to be visible to others, the data should be registered in a public repository. The Resource description framework (RDF) data for the Syntactic and Semantic Framework (SSF) is maintained in a DataHub repository. Since becoming part of the global linguistic linked data network in April 2018, the SSF has added over 20,000 links to the BabelNet ontology and over 67,000 links to LexInfo. The network data of Croatian lexicons is appropriately described over OWL, Lemon or NIF models. The interconnection between three types of SSF lexicons (MSY, LEX and MWE) and other words in the global cloud, however, is not the only value of this work. Including the

SSF in the global LLOD cloud enables researchers to work with the so-called 'sentic patterns' in a 'big data' environment, which the SSF has implemented using 'O-structures' – patterns of grammatical (WOS) and semantic (SOW) features. The result has given the Croatian language a long-awaited place among other world languages.

Benedikt Perak
Faculty of Humanities and Social Sciences, University of Rijeka
bperak@uniri.hr

Application of the graph algorithms for the semantic-syntactic construction analysis

From the systemic perspective of construction grammar (Langacker 2008), syntactic constructions express semantic functions of the conventionalized conceptualizations. The paper shows the use of graph algorithms for analyzing and enriching the lexical data obtained from the network patterns of semantic-syntactic constructions, also known as Word Sketches on the SketchEngine service (Kilgarriff et al. 2010), in a corpus. The study of enTenTen and hrWac corpora explains computational tools for API data harvesting procedures from SketchEngine using the Py2neo and Neo4j graph property database (Webber and Robinson 2018), the creation of the data model for storing lexical items and constructions that can be queried, analyzed using the graph algorithms for intersection, shortest path, centrality, community detection and visualization.

The paper particularly demonstrates the network analysis of the syntactically connected lexemes within a textual corpus for identifying semantically related lexical items and semantic applications of the coordination (and/or) and identification (is_a) grammatical relations. The network analysis of these syntactic relations produces a valuable functional insight into the conventionalization of the lexical use, their semantic domains, polysemous structure of related domains, metonymy and metaphoricity.

Andrej Perdih
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
andrej.perdih@zrc-sazu.si

Portal Fran: od začetkov do danes

Portal Fran: Slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU je ob svojem nastanku oktobra 2014 obsegal 21 slovarjev in jezikovnih atlasov, dve jezikovni svetovalnici in povezave do korpusov in drugih jezikovnih zbirk. Vsebine portala prikazujejo slovenski jezik v njegovi sodobni knjižni, narečni in zgodovinski podobi. Prispevek opisuje razvoj portala po njegovem nastanku, pri čemer so bili v ospredju zlasti vključevanje novih vsebin ter izboljšanje funkcionalnosti, prikaza in uporabniške izkušnje.

Po vsebinski strani razvoj portala poteka v smeri dodajanja prej neobjavljenih slovarjev in jezikovnih atlasov, pa tudi drugih zbirk in programske opreme. Slovarje, ki so bili dodani na portal Fran po njegovem nastanku, lahko uvrstimo v dve skupini. Prvo predstavljajo tisti slovarji, ki so v preteklosti že bili javno objavljeni, navadno v tiskani obliki, za objavo na portalu pa je bila potrebna pretvorba iz različnih digitalnih formatov. Drugo skupino predstavljajo novi slovarji, ki šele nastajajo. Pri teh, t. i. rastočih slovarjih, se novi sestavki načeloma enkrat letno dodajajo že objavljenim. Do konca leta 2018 se je število objavljenih slovarjev povečalo na 35. Poleg slovarjev so bile na portal dodane različne zbirke, med drugim orodje za vnos posebnih znakov *ZRCola* ter podportal slovenskih slovnic in pravopisov od l. 1584 do danes.

Pomemben vidik sodobnih spletnih slovarjev je sodelovanje uporabnikov. Za uporabnike je bila dodana možnost komentiranja slovarskih sestavkov in predlaganja novih besed. Uporabniški predlogi se uredniško presojajo tako za vključitev v katerega od rastočih slovarjev kot za popravke obstoječih slovarjev.

Zaradi velika zanimanja tujih uporabnikov in raziskovalcev je bila izdelana angleška različica vmesnika. Pri nekaterih slovarjih se kaže potreba po inovativnih grafičnih načinih prikazovanja podatkov, te so prišle najbolj do izraza pri prikazu *Sinonimnega slovarja slovenskega jezika* in *Vezljivostnega slovarja slovenskih glagolov*.

Za učinkovitejšo uporabo portala kot celote so bila dodana e-gradiva za šolsko rabo in informacije o Franu.

Slovenian Language Portal Fran: From the Beginning to the Present

The *Fran: Dictionaries of the Fran Ramovš Institute of the Slovenian Language ZRC SAZU* portal was published in October 2014. At the time it consisted of 21 dictionaries and linguistic atlases, two language counselling services and hyperlinks to corpora and other linguistic resources. The content of the portal describes the Slovenian language in its contemporary literary, dialectal and historical forms. The article focuses on later development of the portal where the main goals were inclusion of new content and improvement of its functionality, visualization, and user experience.

Content-wise, several previously unpublished dictionaries, other language resources and software have been added. Some of the dictionaries had already been published before, usually in printed form, and had to be transformed from various digital formats before inclusion on the portal. Other dictionaries are currently being built at the institute and new lexical entries are added to these so called growing dictionaries once a year. In December 2018, the number of published dictionaries on the portal rose to 35. In addition to dictionaries, various resources have been added to the portal, including the special characters input system ZRCola, and the Slovenian grammars and orthographic dictionaries sub-portal consisting of publications from 1584 until the present.

An important aspect of online dictionaries is user involvement. The options for users comment on dictionary entries and to suggest new words were added. Due to great interest from foreign users and researchers, an English interface was also added. In some dictionaries, innovative graphic representation of data was developed, most notably in the Dictionary of Synonymy of the Slovene Language and the Valency Dictionary of Slovenian Verbs. For more efficient use of the portal as a whole, digital materials for school use and materials including information about Fran are available on the portal.

Anita Peti-Stantić
Faculty of Humanities and Social Sciences, University of Zagreb
anita.peti-stantic@gmail.com

Mirjana Tonković
Faculty of Humanities and Social Sciences, University of Zagreb
mirjana.tonkovic@ffzg.hr

Croatian Psycholingiuistic Database: Challenges and Achievements

The categories of concreteness and imageability, as well as the categories of relative frequency and the age of acquisition, have just relatively recently become a topic in linguistic research. This type of research gained prominence since linguists and psycholinguists started to deal with language processing, especially through experimental work in which they set out to determine the relevance of specific variables during language processing. This line of research demonstrated that the categories of concreteness and imageability, along with frequency (a well-established measure in this line of research) play a significant role in comprehension and memorizing. This was the reason for building Croatian Psycholinguistic Database within the CSF scientific project HRZZ-IP-2016-06-1210 *The Building Blocks of Mental Grammar in Croatian: Constraints of Information Structure*. The database to date contains empirically established values for the categories of concreteness, imageability, relative frequency and age of acquisition for 3000 lexemes (nouns, verbs and adjectives) of Croatian extracted from the hrLex lexicon. The lexeme frequency was extracted from the hrWaC corpus. During the second phase of the project, the database was supplemented by additional 3000 nouns, verbs, adjectives, and adverbs. The additional lexemes were extracted from the *Frequency Dictionary of Croatian* (1500 most frequent lexemes) and from elementary-school textbooks with the intention to identify lexemes belonging to general academic vocabulary. Here we present the methodology of constructing the database openly available at http://megahr.ffzg.unizg.hr/ and discuss the decisions we made during the selection and assessment of lexemes. The functionality of the database will be presented and compared to similar databases

created for English (Bird, Franklin and Howard 2001; Coltheart 1981; Cortese and Fugett 2004; Schock, Cortese and Khanna 2012), French (Desrochers and Thompson 2009), Italian (Della Rosa et al. 2010; Rofes, Aguiar and Miceli 2015) and Norwegian (Linde et al. 2015; Simonsen et al. 2013).

Along with the differences in concreteness and imageability concerning parts of speech which have already been analyzed (Peti-Stantić et al. 2017), here we present the statistical and qualitative data comparing the assessment of concreteness and imageability, as well as the correlations of those values in comparison to the AoA and relative frequency.

Снежана Петровић
Институт за српски језик САНУ
snezzanaa@gmail.com

Тома Тасовац
Центар за дигиталне хуманистичке науке
ttasovac@humanistika.org

Расковник: речничка платформа Института за српски језик САНУ

Током последње деценије у Институту за српски језик САНУ је у оквиру различитих пројеката започета систематска и институционална дигитализација речничке грађе и речничких извора, као и припрема за писање речника у дигиталном окружењу. Најзапаженији резултати постигнути су на пољу дигитализације рукописне грађе и ретродигитализације речника. Након кратког историјата ових активности, највећа пажња у раду биће посвећена речничкој платформи „Расковник“, која се развија у сарадњи са Центром за дигиталне хуманистичке науке (ЦДХН) и на којој се тренутно налази пет ретродигитализованих речника - Српски рјечник / Lexicon SerbicoGermanico-Latinum Вука Стефановића Караџића и четири дијалекатска речника (Речник косовско-метохиског дијалекта Глише Елезовића, Речник говора јужне Србије Момчила Златановића, Рјечник говора Прошћења (код Мојковца) Милоша Вујичића и Рјечник дубровачког говора Михаила Бојанића и Растиславе Тривунац). У раду ће бити описани како радни процес дигитализације (захват текста, коректура, структурно обележавање у XML-у, у складу са Смерницама Иницијативе за кодирање текста (TEI)), тако и системска архитектура саме платформе коју чине изворна XML база података eXist-db у позадини, апликациони програмерски интерфејс (API) за машински приступ лексичким подацима и динамични интерфејс за крајње кориснике који је израђен у PHP-у и JavaScript-у. Функционалности саме платформе (претрага по лемама, претрага по пуном тексту, филтери за претрагу, могућности употребе језика за претраживање корпуса (Corpus Query Language, CQL), георефернцирање и сл.) произилазе из доследног и детаљног обележавања речничке макро-, микро- и медиоструктуре и системске архитектуре која у потпуности користи предности структурираног текста. Посебна пажња биће посвећена начину на који су иновативни

концепти речничких исечака (“dictionary slices”) и прелиставању по обележјима (“feature-based browsing”) преточени у конкретна решења којима се јача истраживачки потенцијал ремедијатизованих историјских речника.

Raskovnik: The dictionary platform of the SASA Institute of the Serbian Language

Through a number of projects over the last decade, the SASA Institute of the Serbian Language has initiated systematic digitisation of lexical resources, as well as preparations for the adoption of dictionary-writing systems. The most prominent results have been realised in the fields of digitising manuscript sources and retro-digitising dictionaries.

After providing a brief overview of these activities, the paper will focus on the dictionary platform Raskovnik, which is being developed in cooperation with the Belgrade Center for Digital Humanities (BCDH) and which currently features five retro-digitised dictionaries: Srpski rječnik / Lexicon Serbico-Germanico-Latinum by Vuk Stefanović Karadžić, and four dialectal dictionaries (Rečnik kosovsko-metohiskog dijalekta by Gliša Elezović, Rečnik govora južne Srbije by Momčilo Zlatanović, Rječnik govora Prošćenja (kod Mojkovca) by Miloš Vujičić, and Rječnik dubrovačkog govora by Mihailo Bojanić and Rastislava Trivunac). The paper will describe both the digitisation workflow (text capture, corrections, structural markup in XML according to the Guidelines of the Text Encoding Initiative (TEI), and the platform’s system architecture which consists of eXist-db, a native XML database, in the back-end, an xQuery-based application programming interface (API) for machine access to lexical data, and a dynamic PHP and JavaScript-based interface for end-users. The functionalities of the platform (lemma search, full-text search, search filters, the use of Corpus Query Language (CQL), georeferencing, etc.) are directly related to the consistent and detailed tagging of lexical macro-, micro-, and mediostructure as well as the system architecture which takes full advantage of highly structured texts. Special attention will be paid to the implementation of innovative concepts such as dictionary slices and feature-based browsing as solutions for expanding the exploratory potential of re-mediatised historical dictionaries.

Людмила Рычкова
Гродненский государственный университет имени Янки Купалы
l.v.rychkova@mail.ru

Инфологическая модель многоязычной объяснительной терминологической базы данных по лингвистике

Электронная терминография традиционно фиксирует термины различных языков для специальных целей (LSP) и ориентирована на специалистов соответствующих предметных областей знания. В то же время уже сложившееся в практике преподавания английского, но все еще не достаточно представленное для других языков выделение в рамках каждого LSP такого идиома, как LSP для академических целей, не нашло пока должного отражения при создании специализированных терминологических баз данных (ТБД). Многоязычная объяснительная ТБД по лингвистике создается как инициативный проект студенческого научного кружка «Современная лингвистика», действующего при кафедре перевода и межкультурной коммуникации Гродненского государственного университета имени Янки Купалы, объединяющего студентов – будущих переводчиков и специалистов по межкультурной коммуникации, изучающих такие иностранные языки, как английский, испанский, итальянский, китайский, немецкий, польский. В докладе будет представлена инфологическая модель данной ТБД, а также обоснованы принципы, положенные в основу ее разработки, обусловленные необходимостью учета специфики многопрофильности современной лингвистики, наличия развитой внутриотраслевой межпарадигмальной полисемии и омонимии, а также особенностей прагматики потенциальных пользователей и датологического наполнения ТБД, осуществление которого предполагается силами самих студентов, что будет способствовать как упрочению их лингвистических знаний, так и формированию профессиональных компетенций в области терминологического менеджмента.

The infological model of a multilingual explanatory linguistics terminology database

Electronic terminography traditionally records the terms of various languages for special purposes (LSP) and is aimed at specialists in the relevant subject areas of knowledge. However, it is already established in teaching English, but still not sufficiently present in teaching other languages. The possibility of selection within each LSP of such an idiom as LSP for academic purposes, has not yet been adequately reflected in the creation of specialised terminological databases (TDB). This multilingual explanatory linguistics TDB was created as an initiative project of the "Modern Linguistics" student scientific circle, which operates at the Department of Translation and Intercultural Communication at Yanka Kupala State University in Grodno, uniting students, future translators and specialists in intercultural communication, who study foreign languages such as English, Spanish, Italian, Chinese, German, and Polish. The presentation will present an infological model of the TDB, as well as the principles the model is based upon – the versatility of modern linguistics, the presence of highly developed interparadigmatic terminological polysemy and homonymy, as well as the pragmatics of potential users and the process of introducing data into the TDB that should be implemented by the students themselves, which will contribute both to the strengthening of their linguistic knowledge and the formation of professional competence in the field of terminology management.

Kirill Semenov
Higher School of Economics, Moscow
kir.semenow@yandex.ru

Statistical Approach for Recognition and Detection of Phonetic Borrowings in Chinese (Case Study of the Russian Loanwords)

The detection of phonetic borrowings in written Chinese faces several significant problems. They are: the absence of spaces between words and the lack of an independent set of symbols for foreign vocabulary; significant homonymy within the language; dramatic differences in standard average European and Chinese phonetic systems. All these facts complicate the processing of Chinese texts for different NLP tasks – from named entity recognition to word segmentation and machine translation. In our work, we assume that there is a strategy of transliteration of hieroglyphs for a given European language (Russian, in our case). Theoretically, we try to interpret the strategy of Chinese transliteration in terms of comparison of Chinese and Russian phonetical inventories. Using a dataset based on multilingual Wikipedia articles, we analyse the most frequent n-grams in Russian loanwords and compare them with frequent n-grams in authentic Chinese texts. In doing so, we study sequences of characters that are statistically likely to show that a loanword is present. We also implement an automatic transliteration programme that generates the transcriptions of loanwords based on Xinhua News Agency's officially accepted transliteration rules (1993) and compare the attested use of hieroglyphs in translations to the prescribed transliterations. Using machine learning methods, we investigate which features (both linguistic and extralinguistic) are significant to different transliteration strategies. The results of our work will be used to adjust existing algorithms of named entity recognition for Chinese so they perform better on phonetic borrowings from European languages.

Alexander Shaposhnikov
Vinogradov Institute of Russian Language,
Russian Academy of Sciences
possidima@gmail.com

The experience of creating a standard architecture of a dictionary entry on the example of an electronic Russian-Sanskrit Comparative Dictionary

In the presentation, the author shares his thoughts on the general principles of creating diachronic comparative dictionaries on the example of a Russian-Sanskrit comparative dictionary (1). A unique feature of the RusSktComparDct system is the built-in automatic layout and formatting of dictionary entries using the XeTeX layout system, which allows not only the display of all sorts of fonts, complex diacritics, and unified transliteration, but also generates a PDF format print layout.

Since the dictionary under discussion is comparative, the Russian and Sanskrit headwords are not only clearly separated, but also converge with each other in the centre of the dictionary entry. The original orthographic form (Cyrillic & Devanagari) of the words are placed along the margins. However, the Proto-Slavic etymon or "internal" reconstruction of the Russian word (Latin) and transliteration (Latin) of the Sanskrit word are located at the centre. Interpretations and definitions for both languages are placed under the headwords. This unconventional entry format makes it significantly easier to compare and parse the text.

This is followed by a clear-cut construction of lexical commentary subdivided into two parts – Russian historical & etymological comments and Sanskrit word-formation and etymological comments. These articles convey the historical and dialectal variability of Russian and Sanskrit vocabulary and present data on historical word-formation morphology and etymology, making apparent not only external similarities between Russian and Sanskrit words, but also their homogeneity and common origins.

Lexical commentary is completed with related words from other languages in the Indo-European family.

This is followed by an original rating system; morphological features and word-formation models are rated. As a result of this methodical division of the related and common words in Russian and Sanskrit into groups according to ratings, the characteristics of the lexical layers of different chronological and spatial origin are manifested clearly.

mrežnik
HrZZ
Croatian Science Foundation
VERN'

Сергей Д. Шелов
Институт русского языка им. В. В. Виноградова РАН
Volehs@mail.ru

Концептуальный уровень термина: Опыт формализации и компьютерного вычисления

В докладе утверждается, что определимость термина может лежать в основе идеи концептуального уровня термина, которая может учитывать совершенно разные концеп-туальные структуры терминологии. Для разработки формальной модели понятийного уровня термина я применю некоторые понятия теории графов.

Я буду говорить о графе непосредственной определимости < (x, y), означающем, что x < y тогда и только тогда, когда термин x непосредственно определяется через (термин) y.

Используя граф непосредственной определимости, можно применить к нему некоторые стандартные понятия теории графов и, в частности, понятия “путь” и “длина пути”. Так, я определяю концептуальный уровень L(x) термина x, как максимальную длину всех путей от любого термина до термина X в графе < (X, Y). Для вычисления самого длинного (самого короткого) пути в графе до термина X можно воспользоваться стандартными алгоритмами вычисления путей в графе до вершины X.

Если мы ограничиваемся только родовыми отношениями между терминами, мы сразу же приходим к традиционной родо-видовой структуре, которая сохраняет традиционное понятие иерархического уровня родо-видовой иерархии неизменным.

Согласно общепринятым логическим требованиям, система явных определений не должна иметь кругов (circulus vitiosus) и, кроме того, ни один из терминов не может определяться сам через себя. Действительно, если в системе терминологических определений нет порочных кругов, то идея концептуального уровня каждого термина дает наиболее убедительные результаты. Однако даже если в этой системе и существуют логические круги, формализация одной и той же идеи может быть обобщена и привести к вполне удовлетворительным результатам. В своем выступлении я предполагаю дать несколько различных примеров, иллюстрирующих предложенную идею.

The conceptual level of the term: The experience of formalising and computing

In my presentation, I claim that term definability could underlie the idea of the **conceptual level of a term**, which may take entirely different conceptual structures of terminology into account. To develop a formal model of the conceptual level of a term, I will apply some concepts of **graph theory.**

Thus, I will discuss the direct definability graph < *(x, y)* meaning that *x* < *y* if and only if term *x* is directly defined through term *y*.

Having introduced the concept of the direct definability graph, some standard concepts in graph theory may be used – the concepts of "path" and "path length" in particular. **Allow me to define the conceptual level *L(x)* of term *x* as the maximum length of all paths from any term to term *x* in < *(x, y)*.**

To calculate the longest (or shortest) path in a graph to terms, one can use standard algorithms to calculate paths in a graph to the vertex X.

As soon as we confine ourselves only to generic relations among terms, we immediately come to the traditional generic structure that preserves the traditional concept of hierarchical level with no changes.

According to the common logical requirements, a system of explicit definitions must not have circles (*circulus vitiosus*) and no terms may be defined through themselves. In fact, if there are no vicious circles in the system of terminological definitions, the idea of the conceptual level of each term provides the most convincing results. However, even if there are logical circles within this system, the formalisation of the same idea can be generalised and may lead to quite satisfactory results. In my presentation, I propose to illustrate the idea through several different examples.

Anita Skelin Horvat
Filozofski fakultet Sveučilišta u Zagrebu
askelin@ffzg.hr

Diana Hriberski
Filozofski fakultet Sveučilišta u Zagrebu
dhribers@ffzg.hr

O nekim aspektima izrade (e-)rječnika neologizama u hrvatskome jeziku

Neologizam se može definirati kao nova jezična tvorevina u smislu nove riječi, značenja ili izraza u nekom jeziku u danom vremenskom okviru što predstavlja vrlo široko shvaćenu definiciju neologizama (c.f. Barnhart i Barnhart 1990, Muhvić-Dimanovski 2007). Pojedini autori neologizmima poimaju posuđenice i oživljenice (Mounin 1974) pa se uz brojnost i neodređenost definicija javlja i problem kategorizacije neologizama. Pri Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu radi se na prikupljanju neologizama u hrvatskome jeziku primarno ekscerpirajući primjere iz medijskih tekstova, tiskanih novina, magazina, tjednika i mjesečnika te mrežnih izdanja dnevnih i inih novina kao i raznolikih portala na hrvatskome jeziku s ciljem izrade e-rječnika neologizama. S obzirom na specifičnost ovakvoga rječnika, osobitim se problemom pokazalo utvrđivanje kriterija za definiranje neologizma, a nadalje i odabir natuknica koje bi postale dijelom e-rječnika. Za odabir natuknica koristi se usporedba primjera sa suvremenim rječnicima hrvatskoga jezika, a potvrde se traže i u stranim (e-)rječnicima, kako neologizama, tako i standardnoga jezika. Pri utvrđivanju neologizama u obzir treba uzeti i mnoge druge pojedinosti poput (ne)ograničenosti uporabe, učestalost, izvor u kojem se riječ pojavila, vremenski okvir u kojemu se koristi i sl. Osim toga, javlja se problem utvrđivanja podrijetla i problem kategorizacije primjera prema tome radi li se o posuđenici, tuđici, prevedenici i sl. Proučavanje neologizama u nekom jeziku indikativno je za produktivnost i inovacije, a isto tako i za međujezične dodire i utjecaj jezika, a utjecaj engleskoga jezika na hrvatski, na raznim jezičnim razinama, odražavaju mnogobrojni primjeri iz prikupljenoga korpusa.

Korpus sadrži gotovo 2000 natuknica koje su zabilježene u kontekstu upotrebe, a obuhvaćaju i višerječne izraze i fraze, uključujući i one na stranom jeziku.

On some aspects of compiling an (e-)dictionary of Croatian neologisms

Neologisms are often defined as any new linguistic coinage, including new words, meanings, or expressions in a language in a given time frame (c.f. Barnhart i Barnhart 1990, Muhvić-Dimanovski 2007). Additionally, some authors consider loanwords and revived words to be neologisms (Mounin 1974); in addition to the number and uncertainty of definitions, there is the issue of categorisation. The Institute of Linguistics at the Faculty of Humanities and Social Sciences in Zagreb is collecting neologisms with the aim of creating an e-dictionary. The primary collection method is the excerption of examples from media texts, magazines, periodicals, online editions of newspapers, as well as from various web-portals in Croatian. Due to the specificity of this dictionary, the issue appeared of how to determine the criteria for defining neologisms and, furthermore, how to select which lexemes would become a part of the e-dictionary. In order to choose the entries for the e-dictionary, the examples are compared with contemporary dictionaries of Croatian and foreign languages. In deciding whether a word is a neologism or not, a number of factors should be considered, such as potential usage restrictions, frequency, the source in which the word occurs, the time frame in which it is used, etc. Additionally, there are issues in determining the origin of the word and categorising whether it is a loanword, a foreign word, a calque, etc. The study of neologisms is indicative of productivity and innovation, as well as language contact. The influence of English on Croatian on different linguistic levels is reflected in numerous examples in the corpus. The corpus contains almost 2,000 examples recorded in context; it also includes multi-word expressions and phrases, including those in foreign languages.

Ivan Smolčić
Miroslav Krleža Institute of Lexicography
ivan.smolcic@lzmk.hr

Petra Bago
Faculty of Humanities and Social Sciences, University of Zagreb
pbago@ffzg.hr

Zdenko Jecić
Miroslav Krleža Institute of Lexicography
zdenko.jecic@lzmk.hr

Tablični prikazi kao prilozi strukturiranosti enciklopedičkih djela

Enciklopedička djela čine sintezu znanja određenoga područja od njihova interesa, nerijetko i čitavih ljudskih dosega u vremenu u kojemu nastaju. Kao visokoinformativna izdanja tercijarnoga tipa, važan dio njihova koncepta čini strukturiranost sadržaja u cilju što lakšega razumijevanja i korištenja. Ovaj rad istražuje zastupljenost, način organizacije i funkcionalnost tabličnih prikaza u enciklopedičkim djelima. Ispitivana su općepoznata tradicionalna (tiskana) enciklopedička izdanja poput Encyclopaedie Britannice, Encyclopedie Americane, Chambers's Encyclopaedie, World Book Encyclopedie, Brockhaus Enzyklopädie, la Grande Encyclopédie, kao i projekata hrvatske enciklopedike: Hrvatske enciklopedije, Proleksis enciklopedije, Hrvatske tehničke enciklopedije. Sadržajnom analizom samih enciklopedičih projekata i njihovih tabličnih prikaza kao važnih segmenata strukturiranosti relevantnih podataka donosi se zastupljenost i sadržaj tabličnih prikaza u pojedinim enciklopedičkim projektima. Usporedbom rezultata sadržajne analize više enciklopedičkih izdanja prikazane su različitosti i sličnosti vrsta tabličnih prikaza i njihova sadržaja kako bi se ukazalo koje su mogućnosti i potrebe sastavljanja ovakvoga visokostrukturiranoga sadržaja. Također, donosi se poseban osvrt na pregled funkcionalnosti takve vrste strukturiranoga sadržaja u mrežnim izdanjima nekih od navedenih enciklopedičkih projekata, kao i Wikipedije kao najraširenije i najviše korištene mrežne enciklopedije. Konačno, cilj rada je prikazati evoluciju tabličnih prikaza u

enciklopedičkim izdanjima, od pukih prikaza sistematiziranih podataka u tradicionalnim izdanjima, do sastavnica elemenata strukture mrežnih izdanja koje se koriste i za razvoj ontologija, čime se otvaraju nove mogućnosti u razvoju enciklopedike i promjeni pristupa prilikom izvedbe mrežnih enciklopedičkih projekata.

Tables as means to enhance the structure of encyclopedic works

Encyclopedic works are a synthesis of knowledge of a certain field of interest, and often of the whole human reach of the time in which they arise. As highly informative works of tertiary type, an important part of their concept is content structure with the purpose of making them easier to understand and use. This paper explores the extent of tables, their means of organization and their functionality in encyclopedic works. The research was conducted on well known traditional (printed) encyclopedic works, such as Encyclopaedia Britannica, Encyclopedia Americana, Chambers's Encyclopaedia, World Book Encyclopedia, Brockhaus Enzyklopädie, La Grande Encyclopédie, as well as on Croatian encyclopedic works: Croatian Encyclopedia, Proleksis Encyclopedia, Croatian Encyclopedia of Technology. A content analysis was conducted on tables in encyclopedic projects as important components of the structure of relevant data, giving an overview of the extent of tables and their content in selected encyclopedic works. By comparing results of the content analysis of multiple encyclopedic works, differences and similarities of table types and their contents are shown in order to point out possibilities and requisites when compiling such high-structured content. Furthermore, an insight is provided into functionalities of such structured content in online editions of some of these encyclopedic works, as well as of Wikipedia as it is the most widely spread and used online encyclopedia. The aim of this paper is to present the evolution of tables in encyclopedic works, from mere rendering of systematized data in traditional works, to tables as elements of structure in online works that can be used for developing ontologies, which opens new possibilities in the development of encyclopedic studies, and a new approach when creating online encyclopedic works.

Vera Smole
Filozofska fakulteta, Univerza v Ljubljani
vera.smole@ff.uni-lj.si

Helena Gabrijelčič Tomc
Naravoslovnotehniška fakulteta, Univerza v Ljubljani
helena.gabrijelcic@ntf.uni-lj.si

Alenka Kavčič
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
alenka.kavcic@fri.uni-lj.si

Interaktivni tematski narečni slovar – poskus uporabe novih medijev

Elektronski medij nudi povsem drugačne možnosti izdelave slovarjev, kot smo jih bili vajeni v knjižni obliki. Poleg lažje dostopnosti slovarja širom sveta, preprostejšega iskanja besed v slovarju ter možnosti sprotnega dodajanja in posodabljanja vnosov, medij vpliva tudi na možnost interakcije in prikaza podrobnosti slovarskega sestavka na zahtevo, preprosto medsebojno povezovanje leksemov ter z vključitvijo multimedije tudi bolj slikovit opis slovarskega sestavka. Tako smo v okviru študentskega interdisciplinarnega projekta jezikoslovci, računalniški programerji in grafični oblikovalci zasnovali tematski narečni slovar s poudarkom na interaktivnosti in napredni uporabniški izkušnji ter estetski celostni grafični podobi. Slovar (http://slovar-orodja.si/), na primeru teme »staro orodje« v enem od slovenskih dolenjskih govorov, v zgradbi slovarskega sestavka izkorišča možnosti interaktivne spletne aplikacije in poleg klasične sestave vsebuje še vrsto dodatnih sestavin. V okviru ponazarjalnega gradiva so to slike, avdio in video posnetki s transkibiranimi prepisi in prevodi v knjižno slovenščino, v slovarskem sestavku pa razdelki z neobveznimi medleksemskimi povezavami do novih slovarskih sestavkov: poleg na morebitne sinonime in etimološke osvetlitve manj znanih besed še na besede z istim korenom ter na besede, ki tvorijo ozke teme (npr. sestavni deli nekega orodja, orodja za njegovo vzdrževanje ali vrste določenega orodja). Pri oblikovanju smo izhajali

iz uporabnika in z izdelavo grafične identitete dosegli odprto rešitev, ki omogoča nadaljnje nadgradnje tako vsebine kot funkcionalnosti. Učinkovito vizualno in interaktivno komunikacijo smo dosegli z uporabo grafičnega jezika, navdihnjenega v staranju in večnosti. Barvna paleta obsega naravne odtenke, različne nasičenosti sive in rjave barve, tekstura vstopne strani pa je oblikovana kot porumenel papir, kar simbolno predstavlja učinek sprememb, ki jih ima čas na fizični svet. Po drugi strani pa ohranitveno vrednost slovarja predstavlja kolo v logotipu in na sliki vstopne strani. Kolo kot simbol ponavljanja, večnosti in življenja. V prispevku bo prikazana zgradba slovarskega sestavka v interakciji z računalniškimi in grafičnimi rešitvami v spletni aplikaciji.

An interactive thematic dialectal dictionary – an attempt to use new media

Electronic media offers completely different possibilities in the creation of dictionaries as compared to printed media. In addition to the greater global accessibility of the dictionary, easier searching for dictionary words, and the ability to add and update entries, the choice of media influences the ability to interact and display details of the dictionary composition on demand, enables interconnections between lexemes, and offers clearer descriptions of the dictionary composition through multimedia. Thus, as a part of a student interdisciplinary project, a group of linguists, computer programmers, and graphic designers have designed a thematic dialectal dictionary with an emphasis on interactivity, advanced user experience, and aesthetic graphical identity. The dictionary (http://slovar-orodja.si/) containing expressions from the semantic field 'old tools' in a Slovene Dolenjska dialect takes advantage of the possibilities offered by an interactive web application in the structure of dictionary composition; in addition to the classical structure, it contains several additional components. Illustrative material includes images, audio and video recordings with transcriptions and translations into literary Slovene. The dictionary also contains sections with optional inter-lexeme links to new dictionary compositions; in addition to possible synonyms and

etiological explanations of lesser-known words, links to words with the same root and to words that form more specific themes are also provided (i.e. components of a tool, tools for its maintenance or types of a particular tool). The design is user-centric, the design of a visual identity provides an open solution that may offer the possibility to further expand its content and functionality. Efficient visual and interactive communication was obtained through a graphical language, inspired by the phenomena of aging and eternity. The colour palette includes natural colours and different saturation values of grey and brown, while the texture of the main page was designed to look like aged paper, expressing the effect of temporal changes on the physical world. On the other hand, the preservation value of the dictionary is expressed through the wheel in the logo and on the main page as a symbol of cyclicity, eternity, and life. This presentation shows the structure of the dictionary composition in interaction with the computer and graphic solutions in the web application.

Irena Srdanović
Faculty of Philosophy, Juraj Dobrila University of Pula
isrdanovic@unipu.hr

From specialized web corpora of tourism to a learner's dictionary

Japanese language education at Juraj Dobrila University of Pula aims to cover both general and tourism-oriented specialized Japanese language learning. As there is a need for Japanese speaking tourist guides and other personnel in the tourism industry, the program offers the Japanese language for tourism and other courses that attempt to assist students in acquiring the needed skills and specialized vocabulary. The students are involved in the process of creating a specialized web corpora of Croatian tourism in Japanese language, which is then used as a source for building a bilingual Japanese-Croatian electronic learner's dictionary of terminology related to Croatian tourism.

This paper presents two approaches in creating the specialized web corpora of Croatian tourism in Japanese language and their usages in building the specialized learners' dictionary. Both approaches use WebBootCat technology (Baroni et al. 2006, Kilgarriff et al. 2014) to automatically create specialized web corpora. The first approach creates the corpora from selected seed words most relevant to the topic. The second approach specifies a number of web pages that cover tourism oriented information for specified regions, cities, and sites in Croatia available in Japanese language, which are then used for web corpora creation inside the Sketch Engine platform. Both approaches provide specialized web corpora small in size but rather useful for lexical profiling in the specific field of tourism. In the process of dictionary creation, the second approach proved to be especially useful for selection of lexical items, while both approaches proved to be very useful for exploration and selection of authentic examples from corpora. The dictionary is currently in a pilot phase and is being used and built by learners and teachers through the open-source dictionary platform Lexonomy (Mechura 2017).

Rada Stijović
Institut za srpski jezik SANU
rada.stijovic@isj.sanu.ac.rs

Ranka Stanković
Rudarsko-geološki fakultet, Univerzitet u Beogradu
ranka.stankovic@rgf.bg.ac.rs

Mihailo Škorić
Rudarsko-geološki fakultet, Univerzitet u Beogradu
mihailo.skoric@rgf.bg.ac.rs

Veb alat za upravljanje građom Rečnika SANU i anotacija listića

Građa na osnovu koje se izrađuje Rečnik srpskohrvatskog književnog i narodnog jezika SANU, ekscerpirana iz preko 4.500 pisanih izvora i sakupljana više od 160 godina u svim govorima štokavskog narečja, zabeležena je na oko 5.000.000 listića. Njen značaj je višestruk. Na osnovu nje treba da se napiše još najmanje 15 tomova Rečnika. Ona je dobra osnova za razna fonetska, morfološka i sintaksička istraživanja. Na osnovu nje se može pratiti razvoj jezika tokom protekla dva veka, mogu se saznati brojne dijalekatske crte iz vremena kada su zbirke nastale (većina primera jedini su podaci o govoru nekog kraja toga vremena), mogu se praviti etimološke studije itd. Građa ima i kulturološku vrednost – na njoj su radila mnoga znamenita imena naše kulturne istorije – Jovan Jovanović Zmaj, Jovan Skerlić, Radoje Domanović, Isidora Sekulić, Milan Rešetar i dr.

Ova dragocena, a veoma trošna građa skenirana je u periodu 2016–2018. godine. Za upravljanje elektronskim verzijama listića, 2017. godine kreirana je veb aplikacija za efikasnu anotaciju listića odrednicama. Dalji razvoj aplikacije pratio je informacione potrebe korisnika – osim skraćene anotacije sa obeležavanjem samo odrednice zabeležene na listiću, omogućeno je i puno anotiranje, koje podrazumeva dodavanje odrednice u obliku koji će stajati u Rečniku, zatim oznake za homonim, tekst primera, reference izvora iz koga je primer uzet sa skraćenicom pod kojom se donosi u Rečniku, kao i tip listića (rukopis, kucani tekst).

U radu će biti predstavljen veb portal i rezultati anotacije, koja je krenula od slova P prema slovu Š. Na anotaciji je do sada povremeno radilo 12 različitih anotatora, ali u svakom trenutku 3-5 anotira istovremeno sa različitim intenzitetom rada. Od 813 sekcija sa 2,010,508 listića je skraćenom anotacijom obrađeno 795 sekcija sa 1,934,583 listića, među kojima je 201,487 različitih odrednica. Rezultati anotacije omogućiće procenu broja reči do kraja Rečnika, kao i okvirni popis odrednica.

A web tool for managing material for the SASA dictionary and the annotation of lexicographic card files

The material for the development of the Dictionary of the Serbo-Croatian Standard and Vernacular Language was collected across 160 years and is recorded on roughly 5,000,000 lexicographic citation cards. It was manually excerpted from over 4,500 written sources and collected in the field in all pronunciations of the Štokavian dialect. At least 15 new volumes of the dictionary are planned based on these card files. They can also serve as the basis for various phonetic, morphological, and syntactic research, as well as for analysing language development over the past two centuries, dialectal features from the time when the collections were created (often the only data on the speech of a region at the time of collection), and etymological studies. Its cultural value is also exceptional, as it includes contributions from many outstanding names in Serbian cultural history – Jovan Jovanović Zmaj, Jovan Skerlić, Radoje Domanović, Isidora Sekulić, Milan Rešetar, etc.

This precious, delicate material was scanned from 2016-2018, and in 2017, a web application was developed to efficiently annotate the electronic cards. It was further enhanced based on user needs, enabling (in addition to constricted annotation, where only card headwords were marked) a more detailed annotation including dictionary entry form, homonym tag, attestation and bibliographic reference, abbreviation in the dictionary, and card type (handwritten, typed).

This paper will present the web tool and annotation results from the letter P to Š. So far, 12 different annotators have been working on

the annotations, 3-5 annotating simultaneously at any moment with varying intensity. Of the 813 sections with 2,010,508 cards, 795 sections with 1,934,583 cards were processed with constricted annotation, including 201,487 different headwords. Annotation results will offer an estimation of the remaining number of words for the dictionary with the headword list.

Alicja Sztuk
University of Warsaw
alicja.sztuk@uw.edu.pl

New Perspectives in the Field of Terminology Management

Terminology constitutes a fundamental part of languages and communication for special purposes. A great deal of research has been conducted worldwide in terminology from many different perspectives using various scientific approaches and methods. Terminology constitutes in my opinion, a key issue for entire societies, and it thus requires thorough research, a great deal of scientific discussion, and – most importantly – regular exchange of scientific experience in the field of terminology.

As mentioned above, more and more approaches and methods are appearing with regards to terminology. Some of these use advanced computational technologies (tools, programmes, applications) and therefore enable qualitative and temporal improvements in the field of terminology management. As the term 'terminology management' is quite general, I will concentrate only on the most crucial aspects thereof, e.g. term extraction, collection, and representation, including the representation of relations (vertical or horizontal) between particular terms (taxonomies). Another important issue in the terminology management process is the successful exchange of data (terms) without data loss or distortion.

I will elaborate on this subject by presenting some practical examples of tools and programmes that are designed for successful terminology management and data exchange. Furthermore, I intend to expound upon the expression 'intelligent linguistics tools and applications' through practical examples.

Mirjana Šnjarić
Filozofski fakultet Sveučilišta u Zagrebu
msnjaric@ffzg.hr

Mirjana Borucinsky
Pomorski fakultet Sveučilišta u Rijeci
mborucin@pfri.hr

Glagolsko-imeničke kolokacije hrvatskoga, njemačkoga i engleskoga općeznanstvenog jezika u općoj dvojezičnoj e-leksikografiji

Ovaj rad bavi se leksičkim sredstvima standardnoga jezika preuzetim u jezik znanstvene komunikacije. Zbog razlike u kolokacijskim značenjima koja se ostvaruju u hrvatskome, njemačkome i engleskome jeziku, čestih značenjskih dvojbi prevoditelja znanstvene literature i autora znanstvenih tekstova kao i nedovoljne pokrivenosti ovog područja leksika i leksičkih relacija u hrvatsko-njemačkim i hrvatsko-engleskim općim dvojezičnim rječnicima rad upozorava na činjenicu da za korisnika s hrvatskim kao izvornim jezikom ostaje nedostupan jedan važan sloj leksika znanstvenoga jezika.

Autorice upućuju na nedostatak pokrivenosti općeznanstvenih glagolsko-imeničkih kolokacija u postojećim tiskanim općim dvojezičnim rječnicima s hrvatskim jezikom kao izvornim jezikom rječnika te na potrebu poboljšanja leksikografskog prikaza kolokacija u strukturi rječničkoga članka. Prijedlozi autorica mogu pridonijeti poboljšanju leksikografskog prikaza kolokacija u postojećim dvojezičnim rječnicima, a pohranjivanje popisanih i opisanih općeznanstvenih glagolsko-imeničkih kolokacijskih obrazaca u elektroničkom repozitoriju podataka odabranih za prikaz u trojezičnoj usporedbi hrvatskoga, njemačkoga i engleskoga jezika može poslužiti kao temelj dorade postojećih dvojezičnih rječnika, ali i kao temelj izrade budućih dvojezičnih i trojezičnih mrežnih rječnika, specijaliziranih za općeznanstveni jezik u kojima bi kolokacijama pripadalo izdvojeno mjesto.

Na odabranim primjerima glagolsko-imeničkih kolokacija iz područja humanističkih znanosti (npr. *Belege heranziehen, razmotriti znanstvene*

dokaze, to consider facts; Theorie herausarbeiten, razviti teoriju, to develop / formulate a theory) i njihovoj međujezičnoj usporedbi ilustrirat će se razlike u značenjima i nejasnoće s kojima se suočava hrvatski korisnik. Takve bi se značenjske nedoumice i nedorečenosti leksikografskog prikaza mogle izbjeći izradom budućih novih mrežnih dvojezičnih i trojezičnih rječnika općeznanstvenoga hrvatskog, njemačkog i engleskog jezika u kojima bi glagolsko-imeničke općeznanstvene kolokacije bile bolje obuhvaćene i leksikografski prikazane. Takva otvorena elektronička baza podataka dozvoljavala bi stalnu aktualizaciju i doradu rječnika, moderniziranje čestom provjerom u suvremenom korpusu hrvatskoga znanstvenog jezika i nadopunom novim pojavama u jeziku znanstvene komunikacije.

Verb-noun collocations of the Croatian, German and English common language of science in general bilingual e-lexicography

This paper deals with lexical means of general language which are also a part of the common language of science. Due to the difference in collocational meanings between Croatian, German and English, authors of scientific texts as well as translators are facing many difficulties when trying to find proper functional translation equivalents in the target language. This part of the lexicon and lexical relations are not adequately covered in Croatian-German and Croatian-English general bilingual dictionaries and as a result, an important layer of the common language of science remains unavailable for dictionary users whose L1 is Croatian.

The authors argue that there is a need to improve the lexicographic representation of collocations in general bilingual dictionaries and provide arguments and suggestions as how to improve the presentation of collocations in printed bilingual dictionaries by relying on computer-based methods. Furthermore, a list of verb-noun collocations for the three aforementioned languages created for the purposes of this study can serve as a basis for improving existing bilingual dictionaries but also for bilingual and multilingual e-dictionaries specializing in the general language of science in which collocations should take a more prominent place.

Selected examples of verb-noun collocations typically found in texts on arts and humanities (e.g. *Belege heranziehen, razmotriti znanstvene dokaze, to consider facts; Theorie herausarbeiten, razviti teoriju, to develop / formulate a theory*) will show that these three languages complement each other. Furthermore, a contrastive analysis will illustrate how different meanings can be avoided by improving the representation of collocations typically found in the common language of science in bilingual and multilingual dictionaries of Croatian, German and English, thus making this specific layer of the lexicon more available to the dictionary user.

Vanja Štefanec
Faculty of Humanities and Social Sciences, University of Zagreb
vstefane@ffzg.hr

Matea Filko
Faculty of Humanities and Social Sciences, University of Zagreb
matea.filko@ffzg.hr

Marko Tadić
Faculty of Humanities and Social Sciences, University of Zagreb
marko.tadic@ffzg.hr

Croatian-English parallel corpus of legislative documents

In this paper, we will present a Croatian-English parallel corpus of legislative documents. The parallel corpus consists of ca. 1,800 documents of Croatian national legislation and their official English translations. All texts to be included in the Croatian–English parallel corpus have been previously annotated with EUROVOC descriptors. The corpus will be processed for sentence alignment and provided in aligned document collection format and as a translation memory usable for the training of MT systems.

First, we will present the corpus development methodology. Using a limited number of texts, we will evaluate the accuracy of automatic sentence aligners freely available for Croatian: LF Aligner and CORAL (CORpus ALigner). Based on these results, we will choose the better tool for the alignment. The structure of the obtained corpus will be presented, i.e. the relation between sentence pairs, as well as the final accuracy of 1) sentence segmentation, and 2) automatic alignment.

We will then present the results of two experiments. The obtained parallel corpus will be divided into a training set to train the statistical MT tool and a testing set to evaluate it. An additional statistical MT tool will be trained and evaluated using an equally divided sub-corpus of texts limited to a narrower domain defined by EUROVOC descriptors. Finally, we will present the results of the evaluation based on the texts in 1) the general legislative domain, and 2) a narrower domain.

We believe that the presented results can improve MT systems in the legislative domain. As Croatian is an official EU language and a large number of texts are translated from Croatian to English and vice versa on a daily basis, it is reasonable to believe that there is a need for translation memories limited to narrower domains.

Kristina Štrkalj Despot
Institute of Croatian Language and Linguistics
kdespot@ihjj.hr

Ana Ostroški Anić
Institute of Croatian Language and Linguistics
aostrosk@ihjj.hr

Towards a multilingual figurative thought and language repository

This paper presents a method for language-specific metaphorical conceptualisation analysis developed within the MetaNet.HR project (a repository of conceptual metaphors, semantic frames, image schemas, and cognitive primitives of the Croatian language). Following a theory-driven introspective top-down approach in metaphor analysis (the MetaNet method), this database schematically represents how human reasoning is based on human experience through primary metaphors, how primary metaphors contribute to complex conceptual metaphors, how metaphors can be decomposed into relations among semantic frames, image schemas, and cognitive primitives, and how all this contributes to the meaning of expressions and grammatical constructions.

To build the database, a bottom-up corpus-based analysis is used (the MetaNet.HR method for metaphorical conceptualization analysis) that enables a language-specific and English independent analysis of conceptual systems. For each concept, a web corpus is analysed using Sketch Engine. First, a list of target words for which the corpus is queried is compiled using the coordination column in Word Sketches and the Thesaurus in Sketch Engine. A Word Sketch and a random concordance sample (300 lines per target word) is then analysed for each of the target words, as well as annotated on the linguistic (using the MIPVU procedure) and conceptual level using more annotators and measuring IAA.

This procedure is ready to be applied to other languages to achieve the ultimate goal of creating a large-scale multilingual figurative thought and language repository. This kind of repository has the potential to

revolutionise the field by enabling further non-speculative comparative analyses and answering many unanswered questions concerning metaphor and linguistic diversity, most importantly that of what is universal and what is culturally specific in the ways humans conceptualise.

Marta Tío-Sáenz
University of La Rioja
marta.tio@unirioja.es

The limits of a semiautomatic lemmatisation procedure for Old English verbs in a lexical database

The aim of this presentiation is to present a semiautomatic lemmatisation procedure implemented through database software. The language analysed is Old English, for which parsed but unlemmatised corpora are available. The lexicographical sources are also unlemmatised, including standard dictionaries of Old English by Hall, Bosworth and Toller, and Sweet, but constitute valuable sources of philological data. However, these are not based on an extensive corpus of the language, but rather on the partial list of texts listed in their prefaces. *The Dictionary of Old English* (DOE) is still in progress. The scope of this presentation is restricted to weak verbs in Old English. The lemmatisation procedure has been implemented using Filemaker database software in the lemmatiser *Norna*. *Norna* is one of the building blocks of the *Nerthus* lexical database of Old English; it includes an index with 190,000 inflectional forms. This paper illustrates the results of automatic searches, the comparison and validation of forms A-H with the DOE, and the subsequent addition of non-canonical forms found there to the inventory. This is followed by the generation of non-canonical > canonical patterns of spelling specific to inflectional endings, stem vowels, and prefixes of weak verbs. These patterns are then applied to I-Y forms and cross-validated with other sources. The results of the lemmatisation of Old English weak verbs led to the identification of more than 30,000 inflectional forms of weak verbs. The conclusions insist on the limits of automatic lemmatisation. There are many unexpected spellings within the paradigm of weak verbs, unforeseeable abbreviations, and many ambiguities that can only be ultimately lemmatised with the help of dictionaries. However, the sets of correspondences of non-canonical spellings have proven to be successful not only in finding more forms, but also in reducing the amount of manual revision necessary.

Катерина Велјановска
Филолошки факултет "Блаже Конески"
k.veljanovska@gmail.com

За развојот на македонската е-лексикографија

Развојот на технологијата ја потенцира важноста од компјутеризација и на лексикографските истражувања, создавање на речнички картотеки врз основа на дигиталните ресурси и создавање електронски речници според принципите на корпусната лексикографија. Во овој прилог ќе се фокусираме на електронските речници како моќна алатка која го забрзува и го олеснува процесот на анализа, превод или учење на конкретен јазик.

Имајќи предвид дека не постои македонски национален корпус, лингвистите, преведувачите и другите заинтересирани за оваа проблематика може да користат, покрај печатени изданија, и други извори. Во оваа прилика нашето внимание ќе го задржиме на две групи.

Во првата група се е-речници во кои е застапена лексиката на македонскиот јазик:

- корпус што се базира на дела од македонската литераратура и кој е достапен на адресата http://www.makedonski.info

- македонскиот корпус развиен во рамки на проектот GRALIS раководство на Бранко Тошовиќ кој содржи дел од творештвото на Блаже Конески , и преводи на македонски јазик на делата на Иво Андриќ и се наоѓа на следната адреса: http://www-gewi.uni-graz.at/gralis

- македонско-чешкиот паралелен корпус во кои се застапени книги кои се преведени и достапни на двата јазика: http://kontext.cz

Во втората група се специјалните е-речници кои се однесуваат на одреден лексички слој:

- за пословиците во македонскиот јазик: http://macedonia.auburn.edu/proverbs/php

- кратенките во македонскиот јазик може да се најдат на адресата: http://macedonia.auburn.edu/abbreviations/kratenki/htm1

- како резултат на проектот Паралелни фразеолошки корпуси на Филолошкиот факултет во кои се застапени фраземи во повеќе јазици и

нивните значења е направен овој корпус и се наоѓа на адресата: http://www.frazemi.ukim.finki.mk

- платформа за автоматско менување на македонски глаголи наспрема француски, шпански и англиски се наоѓа на адресата: www.fleximac.free.fr/mkd/

On the development of Macedonian e-lexicography

The development of technology emphasises the importance of computerisation and lexicographic research, the creation of dictionary files based on digital resources, and the creation of electronic dictionaries according to the principles of corpus lexicography. In this article, we will focus on electronic dictionaries as a powerful tool to speed up and facilitate the process of analysing, translating, or learning a particular language.

Bearing in mind that there is no Macedonian national corpus, linguists, translators, and others interested in this issue can use e-dictionaries in addition to printed editions, and other sources. On this occasion, we will restrict our attention to two groups.

The first group contains e-dictionaries in which the lexicon of the Macedonian language is represented by examples from these corpora:

- a corpus based on works of Macedonian literature: http://www.makedonski.info

- a Macedonian Corpus developed within the GRALIS project of Branko Tošović, which contains a part of the work of Blaže Koneski and Macedonian translations of the works of Ivo Andrić: http: //www-gewi.uni-graz.at/gralis

- a Macedonian-Czech parallel corpus, which contains books translated into and available in both languages: http://kontext.cz

The second group consists of special e-dictionaries that refer to a certain lexical layer:

- a dictionary of proverbs in Macedonian: http://macedonia.auburn.edu/proverbs/php

- a dictionary of abbreviations in Macedonian: http://macedonia.auburn.edu/abbreviations/kratenki/htm1

- the Parallel Phraseological Corpus project conducted at the Faculty of Philology resulted in a corpus thatcontains phrases in several languages and their meanings: http://www.frazemi.ukim.finki.mk

- a platform for automatically translating Macedonian verbs into French, Spanish, and English: www.fleximac.free.fr/mkd/

Domagoj Vidović
Institut za hrvatski jezik i jezikoslovlje
dvidovic@ihjj.hr

Morfološko-naglasna obrada glagola u *Mrežniku*

U radu se opisuje morfološko-naglasna obrada glagola u *Hrvatskome mrežnom rječniku* – *Mrežniku* te se uspoređuje s obradom u drugim suvremenim hrvatskim rječnicima ponajprije s obzirom na izbor glagolskih oblika i naglasne tipove. Na temelju neposrednoga uvida zaključuje se kako je obradba glagola u e-rječnicima znatno razrađenija te obuhvaća veći broj kategorija ponajprije zbog toga što nije prostorno ograničena. Primjerice, u *Mrežniku* su navedeni svi prezentski oblici, a u Školskome rječniku hrvatskoga jezika samo morfološko-naglasno relevantni. U radu se ujedno objašnjavaju postupci koji su utjecali na izbor obrađenih glagolskih oblika, od normativnih (naglasni oblici u skladu s pravilima standardnojezičnoga naglašivanja) do uporabnih (manje složena obradba za aorist i imperfekt).

Verbal accentuated inflectional forms in *Mrežnik*

This paper analyses the approach to accentuated verbal inflectional forms in the *Croatian Web Dictionary* – *Mrežnik* and compares it with the approach to verbal morphology and accentuation in other contemporary Croatian dictionaries, focusing on the choice of accentuated verbal inflectional forms. On the basis of his analysis, the author concludes that the compilation of verbal inflectional forms in e-dictionaries is more complex and comprises more categories, mainly due to the fact that e-dictionaries have no space limitations. For example, *Mrežnik* provides all present forms, while the *Croatian School Dictionary* provides only distinctive and representative forms. The paper also explains the reasons that influence the choice of certain verbal forms, which range from normative (i.e. accentuated forms are in accordance with standard Croatian rules of accentuation) to pragmatic (fewer forms are given for aorist and imperfect).

Zvjezdana Vrzić
University of Rijeka
zvrzic@ffri.hr, zv2@nyu.edu

Small languages with big dictionaries: How to get to it with FLEx and Webonary

The paper will present and discuss the features of two open-source applications available for the development and online presentation and search of dictionaries for small, minority and/or endangered languages without large financial resources and teams of lexicographers. Both applications, the Fieldworks Language Explorer (FLEx) and the Webonary, were developed by SIL Language Technology. They can greatly aid the work of a lone dictionary developer, typically, a linguist involved with language documentation building a corpus-based dictionary or a group of community members who are not professional lexicographers.

FLEx allows building of a bilingual or multilingual lexicon as well as the printing of professionally laid-out dictionaries. In addition, it enables cooperation among two or more displaced collaborators. Through the process of dictionary-building, FLEx also makes it possible for linguists to be more efficient when interlinearizing texts. Finally, FLEx also allows users to export dictionary files in the format that can then be imported into Webonary, an application devised for simple and free publishing of searchable dictionaries online.

After introducing the features of the two dictionary-building tools, the author of the presentation will focus on FLEx and her own work experience with it in order to discuss the advantages and difficulties of using the tool while working on a long-term project of language documentation of an endangered language spoken in Croatia.