Electronic lexicography in the 21st century:

linking lexical data in the digital age

Proceedings of the eLex 2015 conference

Edited by
Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek

https://elex.link/elex2015/

11-13 August 2015
Herstmonceux Castle, United Kingdom
Electronic lexicography in the 21st century: linking lexical data in the digital age

Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom

Edited by Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek

Published by Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia)
Lexical Computing Ltd. (Brighton, United Kingdom)

Creative Commons Attribution ShareAlike 4.0 International License

Ljubljana/Brighton, August 2015

CIP – cataloguing-in-publication
National and university library, Ljubljana, Slovenia

81’374:004.9(082)(0.034.2)

ELEX Conference (2015; Herstmonceux)

Način dostopa (URL): https://elex.link/elex2015/proceedings/

ISBN 978-961-93594-3-3 (Trojina, pdf)
1. Gl. stv. nasl. 2. Kosem, Iztok
280722944
CONFERENCE COMMITTEES

Organising Committee

Iztok Kosem, chair
Miloš Jakubíček
Jelena Kallas
Simon Krek
Terka Olšanova

Scientific Committee

Andrea Abel
Špela Arhar Holdt
Nicoletta Calzolari
Frantisek Čermak
Patrick Drouin
Darja Fišer
Thierry Fontenelle
Polona Gantar
Alexander Geyken
Patrick Hanks
Ulrich Heid
Kris Heylen
Ilan Kernerman
Adam Kilgarriff
Annette Klosa
Svetla Koeva
Iztok Kosem
Simon Krek
Margit Langemets
Lothar Lemnitzer

Robert Lew
Nikola Ljubešić
Henrik Lorentzen
Amalia Mendes
Rosamund Moon
Christine Möhrs
Carolin Müller-Spitzer
Hilary Nesi
Vincent Ooi
Magali Paquot
Balint Sass
Kristina Štrkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap Jensen
Agnes Tutin
Tamas Varadi
Serge Verlinde
Piotr Zmigrodzki
# TABLE OF CONTENTS

Automatic generation of the Estonian Collocations Dictionary database  
*Jelena KALLAS, Adam KILGARRIFF, Kristina KOPPEL, Elgar KUDRITSKI, Margit LANGEMETS, Jan MICHELFEIT, Maria TUULIK, Ülle VIKS* ........................................... 1

Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German  
*Lothar LEMNITZER, Christian PÖLITZ, Jörg DIDAKOWSKI, Alexander GEYKEN* ........... 21

Making a dictionary app from a lexical database: the case of the Contemporary Dictionary of the Swedish Academy  
*Louise HOLMER, Monica VON MARTENS, Emma SKÖLDBERG* ................................. 32

Towards an Electronic Specialized Dictionary for Learners  
*Marjan ALIPOUR, Benoît ROBICHAUD, Marie-Claude L’HOMME* ............................ 51

The role of crowdsourcing in lexicography  
*Jaka ČIBEJ, Darja FIŠER, Iztok KOSEM* ................................................................. 70

Mobile Lexicography: Let’s Do it Right This Time!  
*Henrik KOHLER SIMONSEN* ................................................................. 84

What can a social network profile be used for in monolingual lexicography? Examples, strategies, desiderata  
*Monika BIESAGA* ......................................................................................... 105

The Construction of Online Health TermFinder and its English–Chinese Bilingualization  
*Jun DING, Pam PETERS, Adam SMITH* ............................................................ 123

Towards the enrichment of terminological resources by scientific corpora analysis  
*Izabella THOMAS, Iana ATANASSOVA* ............................................................. 136

medialatinitas.eu. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin  
*Krzysztof NOWAK, Bruno BON* ........................................................................... 152

Discovering hidden collocations in a bilingual Spanish–English dictionary  
*Margarita ALONSO RAMOS* ................................................................. 170

Management and exploitation of conceptual data and information in technical termbases: the electrotechnical vocabulary  
*Laura GIACOMINI* ......................................................................................... 186
Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography

*Peter MEYER* ................................................................. 198

Using machine learning for semi-automatic expansion of the Historical Thesaurus of the Oxford English Dictionary

*James MCCracken* .......................................................... 211

What is a Target Language in an Electronic Dictionary?

*Anna Helga Hannesdóttir* ................................................. 236

From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography

*Ana Zwitter Vitez, Darja Fiser* ........................................... 250

Editing an automatically-generated index with K Index Editing Tool

*Kseniya Egorova* ............................................................. 268

A study of the users of an online sign language dictionary

*Mireille Vale* ................................................................. 281

Using a Maximum Entropy Classifier to link “good” corpus examples to dictionary senses

*Alexander Geyken, Christian Pöltz, Thomas Bartz* .................... 304

Multilingual lexicography for adult immigrant groups: bringing strange bedfellows together

*Anna Vakalopoulou, Eleni Efthimiou* .................................... 315

Overwriting knowledge: analyzing the dynamics of Wikipedia articles

*Nathalie Mederake* .......................................................... 327

Towards a Pan European Lexicography by Means of Linked (Open) Data

*Thierry Declerck, Eveline Wandl-VoGT, Karlheinz Möorth* ............ 342

Spell-checking on the fly? On the use of a Swedish dictionary app

*Louise Holmer, Ann-Kristin Hult, Emma Sköldberg* .................... 356

A multilingual trilogy: Developing three multi-language lexicographic datasets

*Ilan Kernerman* ................................................................ 372

Multiple Access Paths for Digital Collections of Lexicographic Paper Slips

*Toma Tasovac, Snežana Petrović* .......................................... 384

Longest–commonest Match

*Adam Kilgarriff, Vít Baísa, Pavel Rychlí, Miloš Jakubiček* .............. 397
GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary
   Franck SAJOUS, Nabil HATHOUT ......................................................... 405

Using machine learning for language and structure annotation in an 18th century dictionary
   Petra BAGO, Nikola LJUBEŠIĆ .............................................................. 427

DEBWrite: Free Customizable Web-based Dictionary Writing System
   Adam RAMBOUSEK, Aleš HORÁK .......................................................... 443

Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information
   Pascale RENDERS, Gérard DETHIER, Esther BAIWIR .................................. 452

Improving the use of electronic collocation resources by visual analytics techniques
   Roberto CARLINI, Joan CODINA-FILBA, Leo WANNER .............................. 461

Predicting corpus example quality via supervised machine learning
   Nikola LJUBEŠIĆ, Mario PERONJA ......................................................... 477

Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries
   Ina RÖSIGER, Johannes SCHÄFER, Tanja GEORGE, Simon TANNERT, Ulrich HEID,
   Michael DORNA .............................................................. 486

Linked Terminologies: Applying Linked Data Principles to Terminological Resources
   Philipp CIMIANO, John P. MCCRAE, Víctor RODRÍGUEZ-DONCEL, Tatiana
   GORNOSTAY, Asunción GÓMEZ-PÉREZ, Benjamin SIEMONEIT, Andis LAGZDINS ... 504
Multiple Access Paths for Digital Collections of 
Lexicographic Paper Slips

Toma Tasovac¹, Snežana Petrović²

¹ Belgrade Center for Digital Humanities
² Institute of Serbian Language of the Serbian Academy of Arts and Sciences
E-mail: ttasovac@humanistika.org, snezzanaa@gmail.com

Abstract

The paper describes the process of digitizing and annotating some 23,000 lexicographic paper slips compiled by the amateur lexicographer Dimitrije Čemerikić (1882-1960) to document the Serbian dialect from the historic city of Prizren. This previously unpublished dictionary of the Prizren dialect is an important resource not only for dialectologists and linguists, but also for ethnolinguists and ethnologists who are interested in various aspects of popular culture and urban life in the city of Prizren. The alphabetic arrangement of the macrostructure, however, is not conducive to exploratory searches: if users want to find out which dialect word corresponds to a standard Serbian word, or explore a certain type of vocabulary, they need access paths to the dictionary content that go beyond the indexing of the macrostructure. The paper describes an elaborate annotation strategy based on marking up headwords with standardized orthographic alternatives, providing lexical equivalents and assigning semantic fields to entries in order to achieve robust navigability and searchability of the collection without full-text transcription and/or structural data modeling.

Keywords: digitization; dialect dictionaries; navigation; searchability; access paths

1. Introduction

Despite the dramatic impact which corpus linguistics has had on contemporary lexicographic practice (Sinclair, 1991; Fellbaum, 2009), the history of lexicography cannot be understood without considering the tradition of lexicographic citation slips — the hand-picked excerpts from literary and other sources that are an essential component of the lexicographer’s toolkit (Landau, 1984; Wandl-Vogt, 2005; Bakken, 2006). Collections of lexicographic paper slips are not only an important part of European lexicographic heritage (Considine, 2008), but are research objects in their own right. In this paper, we discuss the process of digitizing and annotating one such collection created by the Serbian amateur lexicographer Dimitrije Čemerikić (1882-1960). Čemerikić’s manuscript, compiled in the middle of the twentieth century using some 23,000 paper slips, contains approximately 16,000 lemmas with definitions and examples that illustrate the variant of Serbian from the historic city of Prizren that is today an endangered dialect (Петровић, 2012; Петровић & Тасовац, 2013).

The main goal we set ourselves for the digital edition of the Čemerikić paper slips was to provide users with improved retrieval possibilities based on multiple access points.
We will show how our decision to implement an elaborate annotation strategy based on marking up headwords, standardizing orthography, providing lexical equivalents and indicating the entry’s semantic fields enabled robust navigability and searchability without full-text transcription and/or structural data modeling.

The paper is structured as follows: Section 2 describes Čemerikić’s manuscript itself in greater detail. Section 3 explains how different methods of digitization (image capture, text capture, data modeling and data enrichment) influence the kinds of access paths that an electronic resource can offer. Section 4 analyzes the need for access paths beyond the dictionary macrostructure, while Section 5 presents in detail how the annotation of the Čemerikić collection has helped us achieve the goal of providing multiple access paths to the collection.

2. The Manuscript

The Čemerikić manuscript is part of the inventory of paper slips collected over a period of almost 100 years for the compilation of the Речник српскохрватског књижевног и народног говора (Dictionary of Serbo-Croatian Literary and Vernacular Language) of the Serbian Academy of Arts and Sciences (Ристић et al., 2011). It is an accident of history that this collection has not been merged with the rest of the Academy’s inventory, but has instead remained physically separate. While a small portion of its valuable content has trickled through to the first 19 volumes of the Academy dictionary that have been published so far, the manuscript contains sufficient interesting material to deserve a publication on its own.

The original of the Čemerikić manuscript is archived at the Institute for the Serbian Language of the Serbian Academy of Arts and Sciences. The digital version has been publicly available since 2013 via Prepis.Org: The Platform for the Transcription and Digital Editions of the Serbian Manuscript Heritage (Тасовац & Петровић, 2013). One small part of the manuscript, dealing with 3,848 entries for words starting with letters а, б и в, has survived in typewritten form on sheets of A4 paper. The bulk of the collection, however, consists of entries written in ink and pencil on paper slips of different sizes and quality, torn-out notebook papers and, in some cases, even cigarette paper. Formally, we can distinguish three types of paper slips: those containing only records of individual word forms (cf. Ѵар, џенем, ептен); those containing only citations (cf. басма шиљте), and those, in the majority, which are already formatted as prototypical dictionary entries with highlighted headwords, grammatical information, definitions, citations etc. Čemerikić used various sources for his work: he excerpted words from various trade records and guild protocols (written in the pre-reform Cyrillic alphabet); ethnographic and historical literature, newspapers, travel literature etc. Most
importantly, however, the manuscript contains an abundance of examples from colloquial, everyday communication as well as numerous descriptions of local cultural traditions. This previously unpublished dictionary of the Prizren dialect is therefore an important resource not only for dialectologists and linguists, but also for ethnolinguists and ethnologists who are interested, for instance, in various aspects of popular culture (customs, superstitions, witchcraft) and urban life (guilds, social and ethnic relations, etc.) in the city of Prizren (Петровић & Тасовац, 2014). We based our approach to digitizing Čemerikić on the premise that electronic access will benefit both scholars (dialectologists, lexicographers and linguists) and the general public interested in the language and culture of the city of Prizren.2

3. Lexicographic Data: From Paper to Screen

Not all digital objects are created equal. We can distinguish four types of methods and activities for creating digital representations of lexical resources: 1) image capture; 2) text capture; 3) (lexicographic) data modeling and 4) (lexicographic) data enrichment. In this section, we will briefly look at these four aspects and their roles in our digitization of the Čemerikić manuscript.

Image capture refers to the process of recording the visual representation of the text by means of digital cameras and scanners and its subsequent delivery to the user as a digital image. Digital images are nowadays quite easy to produce and deliver over the internet but their usability, especially when it comes to lexicographic material, is limited due to a lack of search capabilities. The process of digitizing the Čemerikić manuscript started with the scanning of some 23,000 paper slips. The digital images were made available via the online platform http://prepis.org from the very beginning of the project. Initially, however, the scanned paper slips suffered from some of the same shortcomings as their physical counterparts: identifying and retrieving information about particular words would require browsing hundreds if not thousands of digital images.

Text capture refers to the transposition of textual content into a sequence of alphanumerical characters, which can be accomplished either by human operators who retype the original text; or, automatically, by using an optical character recognition (OCR) software to convert images into searchable strings. Optical Character Recognition (OCR) is widely used in mass digitization efforts, but its application in the realm of recognizing unconstrained hand-written texts is not as successful as it is in cases of printed documents or constrained hand-written domains such as numbers

2We have not conducted specific user surveys with the general public, but our own experience with organizing an exhibition about the Čemerikić manuscript at the Science and Technology Gallery of the Serbian Academy of Arts and Sciences, as well as a previous social media project related to the Serbian Dictionary by Vuk Stefanović Karadžić (1787-1864), which had more than 24,000 followers on Facebook alone, makes us confident that there is a broad interest among the Serbian public for topics related to language history and language diversity.
or postal addresses (Vinciarelli, 2002; Bunke, 2003; Plötz and Fink, 2009). Challenges include low paper quality, ink bleed-thru, line positioning variations (skews), overlapping characters, wide personal variations in glyph formation, and, often, a circular dependency between character segmentation and recognition, sometimes referred to as Sayre’s paradox (Sayre, 1973).

Manually transcribing the full-text of Čemerikić’s paper slips would be a time-consuming and costly process, not just because of the physical qualities of the slips which have not been preserved under ideal archival conditions, but also because of the nature of the material – a dialect with a large number of nonstandard vocabulary items, multilingual content and even nonstandard Cyrillic graphemes. Even if a team of highly-skilled, linguistically-trained transcribers could perform the job, the full-text transcription would not necessarily be sufficient for the creation of robust search and retrieval possibilities.

Lexicographic data modeling refers to the process of explicitly encoding the structural hierarchies and the scope of particular textual components: in the case of lexicographic data, this usually involves marking up both the macrostructure of the dictionary and the microstructure of individual entries (lemmas, grammatical information, senses etc.) A marked-up text increases the information density of the digital surrogate and paves the way for the implementation of more advanced faceted navigation and targeted search capabilities (for instance, retrieving all nouns whose etymology indicates particular linguistic origins; or retrieving all instances of a particular lexeme when it appears in dictionary examples stemming from a particular author). While it would have been ideal to create, for instance, a TEI-encoded ISO-LMF-compatible edition of the Čemerikić manuscript from the outset of the project, this was not a practical choice. With full-text transcription of the entire manuscript remaining beyond our reach due to financial constraints, the structural modeling was also not an option.

Lexicographic data enrichment, on the other hand, does not necessarily depend on the availability of the full text. By data enrichment or annotation, we refer to the process of encoding additional information that specifies, extends or improves upon the information already present in the lexicographic resource. As will be seen in Section 5, entry-level lexical and semantic annotations of the digitized paper slips can increase their use value even without transcription and/or structural modeling of the content.

Before we turn to the analysis of the data enrichment of the Čemerikić collection, one other question remains to be addressed: why do we need multiple access paths in the first place?

4. Access paths

The alphabetical arrangement of entries in a print dictionary functions as a type of index — a retrieval mechanism connecting a known order of symbols to an unknown
order of information (Hass Weinberg, 2010). The user can access dictionary content by consulting the dictionary macrostructure, i.e. the arrangement of lemmas in a given order (see Hausmann & Wiegand, 1989). While alphabetic dictionaries are relatively easy to consult, they are also efficient randomizers of meaning. By grouping lexemes according to their orthography, rather than their sense, standard dictionaries adhere to the abstract convention of alphabetical order, scattering words with similar or related meaning across unpredictable distances. The “psychologically quite unmotivated tyranny of the alphabet” (Makkai, 1980: 127) is both a blessing and a curse. Looking up entries is easy, if one knows precisely what word one is looking for. Discovering unfamiliar words and exploring semantic concepts, however, is considerably more difficult (Tasovac, 2012).

In electronic dictionaries users access lexicographic content not based on a single wordlist but through a search engine: “it may be more appropriate to say that the macrostructure has been replaced by what may be called a data presentation structure.” (Nielsen, 2011: 201; see also Nielsen & Almind, 2011). The lexicographic concept of accessibility needs to be “narrowed down to cover quick and easy access to the specific types of data that can cover a specific type of user’s specific types of need in a specific type of extra-lexicographical situation” (Tarp, 2008: 101). What constitutes quick and easy access, however, depends as much on a particular situation of use as it does on the type of the dictionary being accessed.

Users resort to historical dictionaries, for instance, in roughly three types of situations: (1) when they have difficulties in the reception of historical texts, (2) when they have difficulties in the production of modern translations; and (3) when they have general questions about linguistic and cultural tradition (see Reichmann, 2012: 54). The first two types of situations are text-related: they arise out of the user’s engagement with a particular text. The user can, when reading texts, experience all sorts of semantic difficulties (encounter unknown lexical units; discover gaps in word meaning; raise questions of morphological, syntactic or pragmatic nature). In these cases, the user will use the macrostructure (or the search engine, in the case of an e-dictionary) to locate a specific entry containing the information that he or she needs.

Reichmann’s third situation of use is texttranszendierend [text transcending] (2012: 64). What this means is that lexicographic texts can also be used to study the lexical materialization of cultural and historical relations, processes and transformations. Dictionaries, after all, are not only information-extraction tools: they also serve as texts, models of language and cultural objects deeply embedded in the historical and ideological matrices of their time (Tasovac, 2010). The main difference between the use of dictionaries in specific text reception and text production situations, on the one hand, and more general research situations on the other hand, is the question of initial focus and ultimate scope. In specific, text-related situations of use, the initial focus and ultimate scope are usually the same: extracting the definition of a particular sense of a particular word is usually accomplished by consulting one dictionary entry. In
text-specific situations, the dictionary is used as a look-up tool. In text-transcending situations, it is used as an exploratory tool.

To make the digital edition of the Čemerikić manuscript available in text-specific situations, the images were first digitized and uploaded to Prepis.Org: The Platform for the Transcription and Digital Editions of the Serbian Manuscript Heritage, which uses Omeka, an open-source digital collection management system in its backend (Kucsmá et al., 2010; Tomás, 2011). After merging entries that are written on both sides of individual slips or across several paper slips, we arrived at 16,626 entries. The headwords for all entries were then transcribed and a search plugin implemented with an autocomplete dropdown menu, allowing users to gain a view of the scope of the entire entry list.

![Figure 1: Autocomplete search](image)

The entries are marked in terms of priority for subsequent full-text transcription: priority 1 is given to entries that contain Čemerikić’s citations of spoken sources. These are given the highest priority because of the scarcity of spoken dialectological data for the Prizren dialect, especially from the middle of the century. Editors are also given the freedom to mark with priority 1 entries that are particularly interesting from the point of view of cultural history. Priority 2 is given to entries that contain citations from previously published written sources, more often than not from historical literature; and priority 3 to all other entries. By default, all entries are marked with priority 3 and then manually upgraded to levels 1 or 2 where required. As of this
writing, of the 6820 manually prioritized entries, 3261 were given priority 1; 1826 were
assigned priority 2; and 1724 remained priority 3. Priority 4 is given to transcribed
items, and priority 5 to transcribed entries that have been proofread and approved by
the senior editor. Due to financial constraints, only entries with priority 1 are currently
being transcribed in full.

Direct access to the macrostructure of the Čemerikić collection, while being a sine qua non, would not have been sufficient for a text-transcending, exploratory use. If a user
wants to find out which dialect word corresponds to a standard Serbian word, or
explore a certain type of vocabulary, or certain ethnolinguistic or historical topics, the
alphabetic arrangement of the macrostructure will not be able to provide the answers.
In these types of situation, the user needs access paths to the dictionary content that
go beyond the indexing of the macrostructure.

5. Annotating for multiple access paths

5.1 Standardized Lemmas

The main access structure for the entries in Čemerikić’s manuscript is the headword,
which is usually underlined on the paper slip. In creating our lemma index, we use the
headword, preserving Čemerikić’s original spelling. For each graphemically
non-standard lemma, however, we provide a standardized spelling alternative. For
instance: зъндан > зиндан (semivowel ъ > и); тъмън > таман (semivowel ъ > а); зъмба >
зумба (semivowel ъ > у); чадър > чадор (semivowel ъ > о); дићбек > дибек
(non-standard Cyrillic i-umlaut representing the Turkish vowel û). The standardized
spelling variants are displayed on the page, bellow the lemma (see Picture 1), and
automatically added to the search index so that they appear in the search
autocomplete dropdown menu and point to the original entries.

5.2 Near-Synonyms

The entries are furthermore annotated with standard Serbian lexical equivalents. The
addition of standard synonyms greatly improves the searchability of the collection
because synonyms are also automatically added to the index list. The user can access
the entry зъндан, as aforementioned, by searching for the original spelling, the
standard orthographic representation of the dialect lexeme (зиндан) as well as its
modern standard equivalents затвор or тамница (jail, dungeon).

5.3 Semantic Fields

The collection is furthermore enriched by the application of semantic fields adapted
from Buck (1949) in consultation with the questionnaire of the Serbian Dialect Atlas
These top-level semantic fields were chosen specifically to reflect the semantic categories most prevalent in Serbian dialect dictionaries. They have been tested on a wide range of dialect dictionaries to ensure wide coverage and cross-dictionary applicability.

<table>
<thead>
<tr>
<th>Физички свет (физички и метеорологија)</th>
<th>Physical World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Човек (теленосна и психологичка особина)</td>
<td>Man (body parts, physical and psychological features)</td>
</tr>
<tr>
<td>Родбина (кровно, безкровно и духовно сродство, називи за обраћање)</td>
<td>Kinship (consanguine, affinal and spiritual; terms of address)</td>
</tr>
<tr>
<td>Медицина (болезненост, телесни и душевни недостаци, лекови, ветеринарска медицина)</td>
<td>Medicine (illnesses, physical and mental impairments, medicines, veterinary medicine)</td>
</tr>
<tr>
<td>Животиње (и сточарство)</td>
<td>Animals (and animal husbandry)</td>
</tr>
<tr>
<td>Исхрана (храна и пиће)</td>
<td>Food (and drink)</td>
</tr>
<tr>
<td>Одевање (одећа, обућа, накит, нега, дотеривање)</td>
<td>Clothing &amp; Adornment</td>
</tr>
<tr>
<td>Кућа (покућство, окућница)</td>
<td>Dwellings &amp; Furniture</td>
</tr>
<tr>
<td>Биљке и земљорадња</td>
<td>Vegetation &amp; Agriculture</td>
</tr>
<tr>
<td>Кретање (и превоз)</td>
<td>Motion (&amp; Transportation)</td>
</tr>
<tr>
<td>Глас (говорење, оглашавање, ономатопеје)</td>
<td>Voice (speech, including onomatopoetic sounds)</td>
</tr>
<tr>
<td>Занимања (занати, алати, предмети везани за занимања, материјали, оружје)</td>
<td>Professions (crafts, tools, objects related to professions, materials, weapons)</td>
</tr>
<tr>
<td>Поседовање (имање, трговина)</td>
<td>Possession &amp; Trade</td>
</tr>
<tr>
<td>Простор (односи у простору, положај нечега, место, облици, величина)</td>
<td>Spatial Relations</td>
</tr>
<tr>
<td>Мере (укључујући новац и бројеве)</td>
<td>Quantity &amp; Number (including money)</td>
</tr>
<tr>
<td>Календар (од секунде до века; доба дана, године, месеци, дани у недељи)</td>
<td>Calendar (from second to century; time of the day, seasons, months, days of the week)</td>
</tr>
<tr>
<td>Чулна перцепција</td>
<td>Sense Perception</td>
</tr>
<tr>
<td>Осећања (све везано за субјективни, морални)</td>
<td>Emotion (everything related to the subjective,</td>
</tr>
</tbody>
</table>
Table 1: Semantic fields

The labels for the semantic fields in each entry can be used as a navigational tool to display a list of all entries from the given field, enabling thus a kind of thematic browsing through the collection.

6. Conclusion and Further Work

The agile approach to digitization of the Čemerikić manuscript allows us to deliver rapidly and annotate incrementally, continuously increasing the use value of the collection by providing new access paths for searching and navigation (lemmas, standardized lemmas, synonyms, semantic fields). Since the work on the collection is ongoing, it would be difficult to provide a reliable quantitative overview of the elements added at this point. Once the current process of annotation is complete, however, we will be able not only to assess our own annotations statistically, but also to quantify the distribution of semantic fields across Čemerikić's collection as a whole.

In addition to the semantic fields, which offer a closed set of choices for tagging entries in the Čemerikić collection, we are planning to implement a free-text tagging option as well, to allow for even more flexibility in the tagging process. The multiple access paths will be especially useful in a future iteration of the project, in which we will also open API access to the collection in order to facilitate the integration of the digitized paper slips with other electronic dictionaries and/or multi-dictionary portals.
Figure 2: Entry for нальча
7. Acknowledgements

This article is the result of research on the project Nr. 47016 “Interdisciplinary Research of the Cultural and Linguistic Heritage of the Republic of Serbia and the Development of the Web Lexicon of Serbian Culture” which is fully financed by the Ministry of Education and Science of the Republic of Serbia. Further financing for the advanced annotation of the manuscript has been provided by the Ministry of Culture and Information of the Republic of Serbia.

8. References


Петровић, С. (2012). Турцизми у српском призренском говору: на материјалу из рукописне збирке речи Димитрија Чемерикића. Београд: Институт за српски језик САНУ.

Петровић, С. & Тасовац, Т. (2013). Призрен - живот у речима. Београд: Институт за српски језик САНУ.

Петровић, С. & Тасовац, Т. (2014). Збирка речи Димитрија Чемерикића као извор за
етнолингвистичка и етнолошка истраживања. Гласник Етнографског института, LXII(2), pp. 171-179.


This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

http://creativecommons.org/licenses/by-sa/4.0/