sciendo
DOI 10.2478/jazcas-2021-0052

BUILDING AN EDUCATIONAL LANGUAGE PORTAL USING EXISTING DICTIONARY DATA

ANDREJ PERDIH – KOZMA AHAČIČ – JANOŠ JEŽOVNIK – DUŠA RACE
Fran Ramovš Institute of the Slovenian Language, Research Centre of the Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia

PERDIH, Andrej – AHAČIČ, Kozma – JEŽOVNIK, Janoš – RACE, Duša: Building an educational language portal using existing dictionary data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 568 – 578.

Abstract: The article presents the process of building the *Franček* Slovenian language portal aimed at primary- and secondary-school students. We discuss problems and solutions of linking and adapting existing non-pedagogical dictionaries for school use, while overcoming content and structural differences among the dictionaries. We also present some solutions within the process of adaptation to the online medium and visualisation adjustments for three age groups of school users with different content needs and levels of (meta)linguistic knowledge.

Keywords: pedagogical lexicography, language portal, Slovenian language, dictionary linking, children's dictionary

1 INTRODUCTION

Franček is an educational language portal for Slovenian aimed at primary- and secondary-school students. By building the portal we seek to provide a solution to a fundamental obstacle in the early use of dictionaries revealed by studies on the use of electronic resources in the Slovenian educational system ([1], [2]): their lack of adaptation to the users' age.

Since 2014, Slovenian primary- and secondary-school students have used online dictionaries of the Slovenian language only through the *Fran* portal. The *Fran* web portal combines thirty-eight dictionaries (with a total of 689,941 dictionary entries), a dialect atlas, and online language counselling and terminological counselling services, all searchable through a single search engine, displaying results from all the different sources at once ([3], [4]). It was set up in 2014 and quickly became popular in the Slovenian educational system: it is referred to in all recent Slovenian language textbooks, and its use is also promoted by the Slovenian National Education Institute.[1]

[1] E.g., https://www.zrss.si/objava/portal-fran-in-portal-francek-za-solsko-rabo. The *Fran* portal is exceptionally popular, with approximately 200,000 searches recorded daily at the time of writing.

Adaptation of dictionaries to better suit students was the main source of motivation behind designing the new *Franček* portal (https://www.francek.si), where dictionary material is displayed not by individual dictionary, but aggregated to provide wholesome information on individual words with regard to their meaning, synonymy, morphology, pronunciation, phraseology, dialect variation, history, and etymology.



Fig. 1. The main site of the *Franček* portal

Franček combines materials from various lexicographic resources and presents the data in a simplified manner by providing answers to specific questions a student might ask.[2] Information on the data's primary source is clearly marked. This way, via the natural situation of learning about language (i.e., using questions and answers), students are gradually taught how to use and, more so, appropriately understand more complex dictionary content.[3]

Combining different dictionary databases in a single portal is an extremely demanding lexicographic challenge. For example, the label *pogovorno* 'colloquial' alone is used very differently in various Slovenian dictionaries; approaches to labelling

[2] Translations of questions presented in icons in Fig. 1 are as follows: "What does this word mean?"; "Find words with similar meaning."; "How is this word inflected?"; "How do I pronounce this word?"; "Which idioms does this word appear in?"; "How is this word used in dialects?"; "What is the origin of this word?" and "Since when has this word been used?".

[3] Every answer provided on the simplified *Franček* portal is linked to a dictionary entry on the *Fran* portal intended for experienced users.

parts of speech can differ significantly; dictionaries even differ in how they arrange specific headwords in different periods. How should a portal, then, be set up to suitably take account of the numerous differences between dictionaries and combine them reliably? How can conceptually diverse dictionaries be combined into a single format, while at the same time raising young users' awareness of the difference between them?

While designing *Franček*, these questions were addressed at the following three levels:

1) by linking databases to an initially constructed headword list,
2) by displaying dictionary data differently for each age group (i.e., grades 1 to 5, grades 6 to 9, and secondary school), and
3) by making content-related changes to the databases and preparing a suitable supplementary apparatus in the form of tooltips, altered metatext of individual dictionaries, by omitting certain less important information from dictionaries, and by providing links between dictionaries and a pedagogical grammatical description[4] (https://www.francek.si/kje-je-kaj-v-slovnici).

2 SEMI-AUTOMATED LINKING PROCESS

The portal is built around a central headword list (cf. 2.1), with eight modules linked thereto. These provide various linguistic information, namely on the words' meanings, synonyms, morphological paradigms, pronunciation, phraseology, dialect variation, etymology, and information on historical usage. The modules' contents are visualized from underlying databases based on the age group preselected by the user (cf 3.2).

The underlying databases were linked to the headword list using semi-automated linking processes (cf. 2.2). All automated processes were performed on dictionary databases in XML format using XSLT transformations. Manual linking was performed using the *iLex* dictionary writing system (DWS) [6]. Several parts of the data were exported to plain text or Excel files to manually select specific headword IDs to be explicitly included in the XSLT transformation processes.[5] New data, such as the new dictionary for school use *Šolski slovar slovenskega jezika* (ŠSSJ), was manually entered using the *iLex* DWS. Additionally, select dictionary data was modified or enriched to better the end-user experience (cf. 2.3).

2.1 Headword list

The headword list was established on the basis of two general monolingual dictionaries: *eSSKJ – Dictionary of the Slovenian Standard Language, 3rd Edition*

[4] The process of matching lexicographical data to appropriate descriptions within school grammars is presented in [5] in this publication.

[5] Original XML files did not follow a common standard schema, which had to be taken into account and made the preparation more time-consuming.

[7], and *Dictionary of the Slovenian Standard Language, 2nd Edition* (SSKJ2) [8]. While eSSKJ, the newer of the two dictionaries, was prioritized over SSKJ2, it currently contains a lot less than SSKJ2, which represents the vast majority of the headword list. Not all entries from SSKJ2 were accepted into the *Franček* database, as lexemes labelled as *zastarelo* 'obsolete' and *vulgarno* 'vulgar' were excluded. Certain types of SSKJ2 sublemmas were also included in the headword list, such as (non-)reflexive verb pairs and adverbs [9] (5481 out of 12,549 sublemmas, 43.68%). The vast majority of the inclusion/exclusion rules were used during the automatic headword-list building process.

2.2 Linking dictionary data to the headword list

Linking other resources to the headword list was first undertaken as an automated rule-based process, followed by a manual rechecking and linking process to address ambiguities and special cases. Entries were matched according to headwords and, where applicable, part-of-speech data and stress placement. Some caution had to be exercised with regard to POS labels due to different underlying grammatical theories used in different dictionaries; the differences had to be reconciled and the data normalized. While such differences come as no surprise in the case of historical dictionaries, the same issue arose also in the process of linking the *Synonym Dictionary of Slovenian Language* (SSSJ; [10], [11]), even though it is based on SSKJ (1st edition) [12].[6] Stress placement was sometimes also used to automatically differentiate between homographs. Perfect homonyms had to be disambiguated and linked manually.

Links were established at headword level only. We did not attempt to systematically link information at sense level, nor were sense-level gaps filled if the data was available. Consequently, it is possible that a sense not covered in the semantic module may appear in other modules.

This is most evident in the case of the dialect module, since the main source of dialect lexical data used, the *Slovenian Linguistic Atlas* (SLA; [13], [14]), is primarily onomasiological in nature (as opposed to all the other dictionaries, which are classic semasiological dictionaries). The dialect module is divided into two sections: the onomasiological section lists dialect lexemes denoting the meaning of Standard Slovenian lexical forms (i.e. it seeks to provide answers to the question "which words are used to describe this concept and in which dialects?"), and the semasiological section provides alternative dialect senses of Standard Slovenian lexical forms (i.e. "which concepts does this word (also) denote and in which dialects?") [15].

[6] The main reasons for discrepancies are the treatment of the predicative (*povedkovnik*) as a standalone POS (i.e. nouns, adjectives, and adverbs can be interpreted as predicatives depending on their syntactic function and treated as separate lexemes), and the treatment of qualitative and classifying adjectives as separate lexical units (e.g. *zelen* 'green – qualitative adjective': *zeleni* 'green – classifying adjective'); SSKJ does not distinguish between these categories.

Semantic disambiguation proved to be most problematic in the case of the historical module, as the underlying dictionaries describe different lexical systems, dating from mid-16th to late 19th century. Lexemes that semantically greatly differed from their modern Slovenian counterparts, or those whose senses are no longer attested, were manually excluded; e.g. in the case of homonyms *moka* 'flour' and *moka* 'anguish, torment', the latter was excluded, as it had already fallen out of use by the end of the 16th century to eventually be replaced by its Slavic cognate *muka*. Furthermore, differences in orthographic principles among dictionaries had to be taken into account. In cases where the orthographic forms differed from the modern Slovenian ones, the closest form was chosen; e.g. deverbal nomina agentis such as *bravec* 'reader' or 'gatherer', *igravec* 'player' or 'actor', *plezavec* 'climber' etc. were linked to their modern Slovenian counterparts *bralec*, *igralec*, *plezalec* etc. This principle was also adhered to in a limited scope in the case of non-systemic orthographic forms, e.g. *ambašador* was linked to the headword *ambasador* 'ambassador, envoy' (while *bašador*, also attested in the same resource, was excluded).

2.3 Modifications and enhancement of dictionary data

Some data was significantly altered prior to inclusion in the *Franček* database. This was mostly due to the fact that the original resources were less suitable for educational use and thus required simplification. Such was the case with the dialect module, where the main source was the index of the SLA atlas, and a subsection of the historical module, where the main source was a register of all lexemes attested in 16th century Slovenian texts [16]. The latter was used to create a database of the earliest attestations of lexemes in written form.[7]

In SSKJ2, labels pertaining to all or to the majority of senses are presented in the head of the dictionary entry, i.e. at the entry level. As presentation of data from the head is limited, and to give more understandable information to the end-user in the semantic module itself, entry-level labels were transferred to sense level. While the process was automated, complex rule refinement was needed to account for cases where entry-level labels needed to be omitted, usually when the entry-level and sense-level labels belonged to the same type (e.g. stylistic labels or register labels). In some cases, the entry-level labels needed to be placed after the sense-level label to meet the sorting rules in the dictionary (e.g. *ekspresivno* + *pogovorno* 'expressive, colloquial' was changed to *pogovorno* + *ekspresivno*). Some other exceptions also had to be taken into account, as it was not possible to combine the *starinsko* 'archaic'[8] label with the majority of other label types. Special treatment was necessary also

[7] The database excludes a relatively small number of lexemes attested in earlier Slovenian manuscripts due to the unavailability of data in digital form.

[8] The dictionary differentiates between *obsolete* and *archaic* lexis.

with regard to grammatical labels as some combinations are possible, while some labels are mutually exclusive.

Cross-referencing (excluding referential definitions in the semantic module) was reduced as much as possible by inserting the target content in place of the reference's origin, most evidently so in the case of the etymological module.

In the cases of homographs and homonyms, indicators were created to help young users disambiguate among them. While the basic distinction can be done by indicating POS information (1242 entries, automated process), 3213 semantic indicators were also added (out of 4384 total manually added to the underlying database).[9] If neither POS nor semantic indicators could be used for disambiguation, morphological or stress-placement indicators were provided (35 entries).

3 CUSTOMISED DATA VISUALISATION

The *Franček* portal is aimed at students of three age-groups:

1) 1st to 5th grade of primary school,
2) 6th to 9th grade of primary school,
3) secondary school.

ŠSSJ ([17], [18]) seeks to fulfill the needs of the first age group; its extent and concept are adapted to the children's abilities and needs based on the curriculum. This dictionary contains 2000 entries pertaining to basic vocabulary and is displayed in the semantic module.[10] Even though their use in the educational process is expected and planned, other resources used during the creation of *Franček* are not primarily intended for school use.[11] Their excessive complexity, especially in the case of the SSKJ, is well documented ([22], [23]) and strengthens the assumption among students, teachers, and other dictionary users that successful use of dictionaries is something that has to be learned, and that one should practice using dictionaries. As already noted by Tarp [24], a dictionary is not merely a list or a language database; rather, it is primarily a practical tool for language use, which retrieves information from a database as required by the user. The aim of *Franček* is, therefore, to provide language data in a way that will be useful (selection of relevant content) and understandable (adaptive visualization) to students. Since we used existing language resources that had not been created with students in mind, and some of which had not been primarily made for the web, we

[9] The difference stems from the fact that obsolete homonyms were omitted from the *Franček* database.

[10] In other entries, simplified content of SSKJ2 and eSSKJ is displayed.

[11] The 2018 curriculum for Slovene lessons in primary school envisages the use of dictionaries mainly from the 5th grade onwards (in teaching materials, students are most often directed to SSKJ and the *Slovenian Normative Guide* [19]); even before that age, children are expected to be able to at least identify the meanings of words ([20], [21]).

had to find solutions for visualisation of language material that would meet the users' requirements.

Franček features two types of customised visualisation of language data:

1) adjustments due to transfer to the online medium and to the portal design,
2) adjustments due to changed target users.

3.1 Adjustments to the online medium

While eSSKJ and ŠSSJ are primarily online dictionaries, other resources were made as print dictionaries; content is thus structured in a condensed manner due to limited space, which is reflected in the implicit presentation of information with different types of font, abbreviations, symbols, etc. All abbreviations, labels, and symbols were made explicit on *Franček*, e.g. instead of introductory symbols (such as ♦ for the terminological section in SSKJ2) the content is clearly explained (e.g. "*This word is a professional term*"). Labels have not only been fully spelled out (e.g. *pog.* as *pogovorno* 'colloquial') but also explained in tooltips ("A word, multi-word unit, or sense used especially in everyday and less formal communication") and linked to appropriate chapters in school grammars [5]. Visual and audio materials were added. For structuring the content and navigating the portal, standard icons (e.g. the microphone icon indicates the possibility of recording; the map pointer icon prompts users to view a map, etc.) and established web conventions are adhered to (e.g. use of tooltips, links to detailed information, etc.). Styles and colours are consistent throughout the portal (e.g. illustrative material is always green, clickable content is blue, sense numbering is highlighted in blue etc.).

3.2 Adjustments for new target users

Specific age-group requirements and levels of linguistic and metalinguistic knowledge were considered when adapting the display of dictionary data.

A user in the **lowest age group** is, therefore, not overburdened with the dictionary metalanguage and microstructure. Illustrated icons with simple explanations in tooltips are used to provide information on certain stylistic characteristics and grammatical categories (e.g. countable nouns are represented by icons of dice with one (singular), two (dual), and three (plural) dots; outdated synonyms are introduced with an icon of an old man, etc.). In other cases, helping hints were added, e.g. appropriate question words were added next to names of grammatical cases in the morphological module. In this age group, content is limited to semantic, synonymic, morphological, and pronunciation modules.

Visualisation for students of **6th to 9th grade** takes into account that some users in this age group are already familiar with basic metalanguage and use dictionaries for writing. Parts of the dictionary microstructure are explicitly marked (definition, examples, typical constructions, variants). The content is more extensive, as the phraseological, dialect, etymological, and historical modules are included. However,

the number of listed terms, multi-word units, and their variants is limited. Illustrated icons are still used at this stage to symbolize grammatical categories and stylistic characteristics for ease of memorisation, while additional aids (e.g. question words, short explanations of more complex grammatical categories, etc.) have been moved to tooltips.

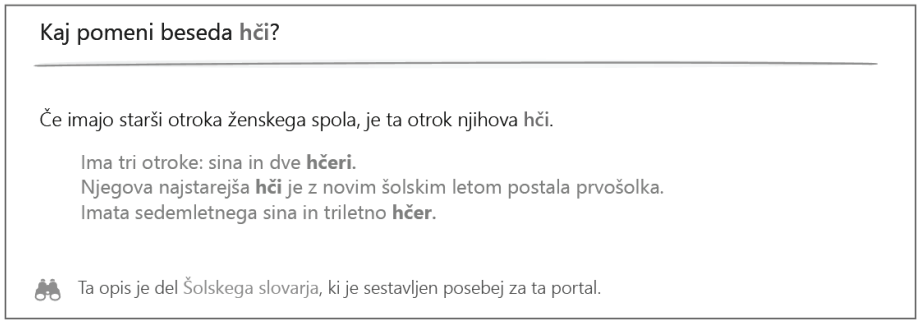


Fig. 2. Visualisation of the semantic module for students of 1st to 5th grade

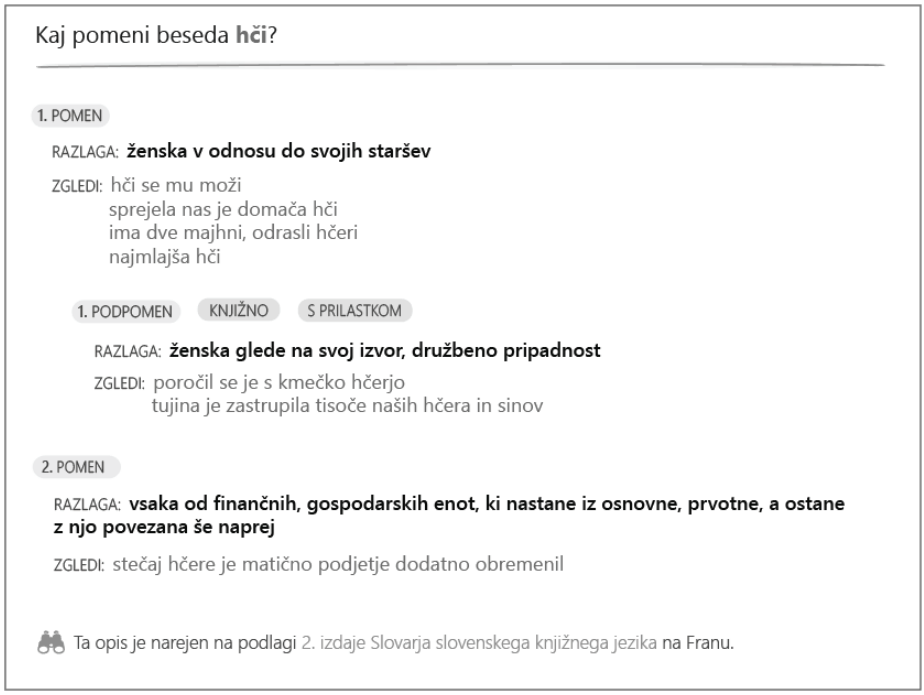


Fig. 3. Visualisation of the semantic module for students of 6th to 9th grade

Visualisation of dictionary content for **secondary-school students** relies on the fact that the users are already familiar with the microstructures of various dictionaries; therefore, dictionary metalanguage is not explicitly presented or graphically illustrated.

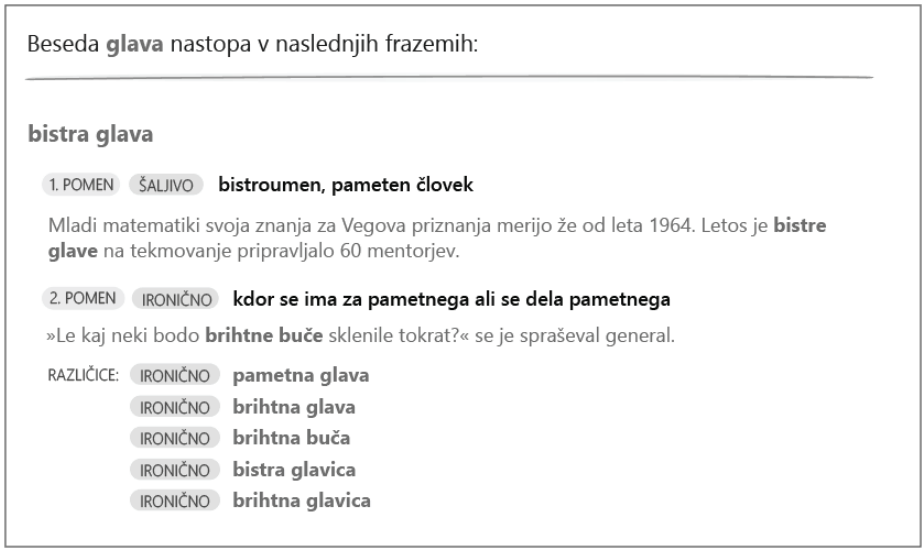


Fig. 4. Visualisation of a phraseological module for secondary-school students

Although visualisation (and content complexity) is only a small step away from that on the *Fran* portal (especially in eSSKJ), explanations in tooltips are still a notable advantage in comparison (e.g. explanation of animacy in the morphological section, links to descriptions within the school grammar, explanations of labels, etc.) and enable proper interpretation of data without detailed knowledge of the concept of the source dictionary.

4 CONCLUSION

Franček, the new educational Slovenian language portal, was built to fill the gap in Slovenian linguistic resources for educational purposes. Lexicographic data on the portal was adapted and linked from existing non-pedagogical dictionaries, while new data was also prepared specifically for this purpose. The lexicographic content is presented from the point of view of individual words, creating a single lexicographic resource. The data is organized in eight modules: semantic, synonymic, morphological, pronunciation, phraseological, dialect, historical, and etymological. Content and visualisation are adjusted to the online medium and adapted to three age groups of

users: primary-school students from 1st to 5th grade and 6th to 9th grade students, and secondary-school students. Additionally, lexicographic data presents a part of a wider ecosystem of linked lexicographic, grammar, and language counselling data.

ACKNOWLEDGEMENTS

This article was produced as part of the project *Portal Franček, Jezikovna svetovalnica za učitelje slovenščine in Šolski slovar slovenskega jezika* (The Franček Portal, the Language Counselling Service for Teachers of Slovenian, and the School Dictionary of Slovenian) co-funded by the Republic of Slovenia and the European Social Fund; part of the research for the project was conducted within the Slovenian Research Agency's P6-0038 program group The Slovenian Language in Synchronic and Diachronic Development.

References

[1] Kosem, I., Stritar, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š., and Rozman, T. (2012). Analiza jezikovnih težav učencev: korpusni pristop. Ljubljana: Trojina, zavod za uporabno humanistiko, 132 p.

[2] Rozman, T., Krapš Vodopivec, I., Stritar, M., and Kosem, I. (2020). Empirični pogled na pouk slovenskega jezika. Ljubljana: Znanstvena založba Filozofske fakultete, 183 p. Accessible at: https://e-knjige.ff.uni-lj.si.

[3] Ahačič, K., Ledinek, N., and Perdih, A. (2015). Fran: the next generation Slovenian dictionary portal. In Natural language processing, corpus linguistics, lexicography: Proceedings, Eighth International Conference, Bratislava, Slovakia, 21–22 October 2015, pages 9–16, RAM-Verlag. Accessible at: https://korpus.sk.

[4] Perdih, A. (2020). Portal Fran: od začetkov do danes. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 46(2), pages 997–1018.

[5] Ahačič, K., Ledinek, N., and Petric Žižić, Š. (manuscript submitted for publication). Presenting and linking grammatical data on the Franček educational language portal.

[6] Erlandsen, J. (2010). iLEX, a general system for traditional dictionaries on paper and adaptive electronic lexical resources. In Proceedings of the XIV EURALEX International Congress. 6–10 July 2010, page 306, Leeuwarden/Ljouwert: Fryske Akademy – Afûk. Accessible at: https://euralex.org.

[7] eSSKJ: Slovar slovenskega knjižnega jezika (2016–). Accessible at: https://www.fran.si.

[8] Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja (2014). Accessible at: https://www.fran.si.

[9] Perdih, A. (manuscript submitted for publication). Učenje o slovarjih v šoli: portal Franček kot most med splošno in pedagoško leksikografijo.

[10] Snoj, J., Ahlin, M., Lazar, B., and Praznik, Z. Sinonimni slovar slovenskega jezika (2018 [2016]). Accessible at: https://www.fran.si.

[11] Snoj, J. (2019). Leksikalna sinonimija v Sinonimnem slovarju slovenskega jezika. Lingua Slovenica 14. Ljubljana: Založba ZRC, 316 p.

[12] Slovar slovenskega knjižnega jezika (2014 [1970–1991]). Accessible at: https://www.fran.si.
[13] Slovenski lingvistični atlas 1 (2014 [2011]). Accessible at: https://www.fran.si.
[14] Slovenski lingvistični atlas 2 (2016). Accessible at: https://www.fran.si.
[15] Ježovnik, J., Kenda-Jež, K., and Škofic, J. (2020). Reduce, Reuse, Recycle: Adaptation of Scientific Dialect Data for Use in a Language Portal for School children. In Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I., pages 31–37, Democritus University of Thrace. Accessible at: https://euralex2020.gr.
[16] Besedje slovenskega knjižnega jezika 16. stoletja (2014 [2011]). Accessible at: www.fran.si.
[17] Godec Soršak, L. (2015). Slovenski otroški šolski slovar. In Slovnica in slovar – aktualni jezikovni opis (1. del), Obdobja 34, pages 243–250. Ljubljana: Znanstvena založba Filozofske fakultete. Accessible at: https://centerslo.si/simpozij-obdobja/zborniki/.
[18] Petric Žižić, Š. (2020). Tipologija razlag v Šolskem slovarju slovenskega jezika. Slavistična revija, 68(3), pages 391–409.
[19] Slovenski pravopis (2014 [2001]). Accessible at: https://www.fran.si.
[20] Godec Soršak, L. (2020). Spodbujanje rabe slovarja v učnem gradivu za slovenski jezik v 1. in 2. vzgojno-izobraževalnem obdobju. In Slovenski jezik in književnost v srednjeevropskem prostoru: Zbornik SDS 30, pages 235–244. Ljubljana: Zveza društev SDS.
[21] Petric Žižić, Š. (2020). Usvajanje besedoslovne jezikovne ravnine in raba slovarjev pri pouku slovenščine v osnovni šoli (pregled učnega gradiva za tretje vzgojno-izobraževalno obdobje). In Slovenski jezik in književnost v srednjeevropskem prostoru: Zbornik SDS 30, pages 245–253. Ljubljana: Zveza društev SDS.
[22] Vrbinc, M. (2004). An empirical study of dictionary use: the case of Slovenia. In ELOPE: English Language Overseas Perspectives and Enquiries, 2(1–2), pages 97–106. Ljubljana: Ljubljana University Press, Faculty of Arts.
[23] Rozman, T. (2010). Vloga enojezičnega slovarja slovenščine pri razvoju jezikovne zmožnosti (PhD thesis). Ljubljana: Filozofska fakulteta, 359 p.
[24] Tarp, S. (2014). Detecting user needs for new online dictionary projects: Business as usual, user research or …? In Research into dictionary use: Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“, pages 16–26. Mannheim: Institut für Deutsche Sprache.