https://doi.org/10.14618/ids-pub-11180

Iztok Kosem

TRENDI – A MONITOR CORPUS OF SLOVENE

Abstract In this paper we present Trendi, a monitor corpus of written Slovene, which has been compiled recently as part of the SLED (Monitor corpus and related resources) project. The methodology and the contents of the corpus are presented, as well as the findings of the survey that aimed to identify the needs of potential users related to topical language use. The Trendi corpus currently contains news articles and other web content from 110 different sources, with the texts being collected and linguistically annotated on a daily basis. The corpus complements Gigafida 2.0, a 1.13-billion-word reference corpus of standard written Slovene. Also discussed are the ways in which the corpus will be integrated into various lexicographic projects, helping not only in the identification of neologisms but also in monitoring changes in already identified language phenomena.

Abstract Monitor corpus; language use; trends; Slovene; neologisms; lexicography; newsfeed

1. Introduction

One of the challenges of lexicographers has always been staying on top of changes in a language. This has become even more crucial in recent years as the technological progress and the shift of dictionaries to the online media has raised the expectations of the users; as research shows (e. g. Kosem et al. 2019) uptodatedness is among the highest valued dictionary features. As a result, new words and senses have started to enter dictionaries at a much faster rate than ever before. The field of neology has attracted more and more attention, with the recent COVID-19 pandemic with all its new vocabulary being a good case in point.

In order to be able to obtain the information on new words and uses of words, one needs a monitor corpus. The main characteristic of monitor corpora is that new (recently published) texts are added on a regular basis, thus enabling monitoring language change. This is also one of the main challenges of monitor corpus compilation, namely being able to regularly obtain, annotate and upload the texts in the shortest time span possible. A monitor corpus for Slovene has long been called for by Slovenian lexicographers, linguists, and other users in need of information about current language use. This has become possible with the introduction of the JSI (Jozef Stefan Institute) Newsfeed service (Trampuš/Novak 2012), which collects news articles from websites across the world, covering over 35 languages. The service has already been used in various projects, including in the compilation of the Gigafida reference corpus of Slovene, version 2 (Krek et al. 2019).[1]

In September 2021, we started a new project called SLED (Monitor Corpus for Slovene and related language resources), which is funded by the Ministry of Culture of the Republic of Slovenia. The project has three aims, the main one being the development of the monitor corpus for Slovene, including the methodology for its regular updating. The second project aim is to regularly provide statistical datasets that will include information of interest to the wider public, for example trending words, new words, words with decreased usage, etc. The third project aim is to develop a tool for topic modelling that can be used on Slovene texts; in this way, we want to provide some topic categorization for each text included in the Monitor Corpus for Slovene, thus enabling more detailed analyses.

[1] https://viri.cjvt.si/gigafida/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

In this paper, we first make an overview of related projects for languages other than Slovene, and then focus on various aspects of the first monitor corpus of (written) Slovene, called Trendi, including the methodology behind its compilation. We point out some of the important decisions that had to be made during corpus conceptualization. We also present the results of a user survey, which was used to get a better understanding of user needs and expectations. We demonstrate some of the ways in which the corpus will be integrated into the dictionary-making workflows in Slovenia. We conclude by presenting future plans related to the monitor corpus of Slovene, and related resources.

2. Related work

The concept of a monitor corpus is far from new. One of the first monitor corpora was the Bank of English,[2] which was first published in 1991. It contains over 650 million words and was used in the compilation of the COBUILD dictionary. Today, it is still a representative subset of the 4.5-billion-word COBUILD corpus. There is no information on when the corpus was last updated. Access to the corpus is very limited, with only the staff and students at the University of Birmingham having access.

Another influential corpus for English, in this case American English, is the Corpus of Contemporary American English (COCA; Davies 2008-),[3] which covers the period from 1990 onwards and contains over 1 billion words. It is a genre-balanced corpus, containing texts from eight different genres (spoken, fiction, popular magazines, newspapers, academic texts, TV and movies subtitles, blogs, and other web pages). The corpus was last updated in March 2020, which somewhat limits its "monitor" status.

Also part of the corpora at English-corpus.org is a regularly updated monitor corpus NOW (News on the web; Davies 2016–),[4] which contains nearly 15 billion words from web-based newspapers and magazines from 2010 to "yesterday" (if we borrow the wording from Mark Davies). As it is mentioned on the website, the corpus grows by about 180-200 million words per month.

Part of the same family as NOW and COCA is a more specialized Coronavirus corpus (Davies 2019-),[5] which spans the period from January 2020 to yesterday and contains over 1.4 billion words. Limited to the genre of web news in English, it grows about 3-4 million words per day.

There are also corpus resources for monitoring languages other than English, for example Timestamped JSI web corpora, which are available in 18 different languages and contain news articles collected by the JSI Newsfeed service. The corpora are available in the Sketch Engine corpus tool (Kilgarriff et al. 2004) and in addition to the usual Sketch Engine functions, the users can also use Trends (Herman 2013), a feature focused on identifying trends in word usage. The corpora contain texts from 2014 to April 2021 (time of the last update) and are of different sizes, with the English corpus containing approx. 60 billion words.

Similarly multilingual is the Google Books Ngram Viewer, which offers searching and various visualizations of word/ngram use over time (1500 to 2019). The resource could loosely

[2] https://cqpweb.bham.ac.uk/.

[3] https://www.english-corpora.org/coca/.

[4] https://www.english-corpora.org/now/.

[5] https://www.english-corpora.org/corona/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

be called a corpus, as the data are based on texts and the user is able to get to the parts of the documents; however, the usual functions for linguistic investigation of corpora such as concordancers, collocations, etc. are not available.

There are many other monitor corpora in existence, which are used by lexicographic institutions and available only internally. An example of such a resource is ONLINE, a dynamic monitor of Czech, compiled by the Czech National Corpus. It contains approx. 6.3 billion words, coming from web news, discussions (under news articles), forums, and social networks (Facebook, Twitter, Instagram). The ONLINE corpus is in fact divided into two complementary corpora – ONLINE_NOW and ONLINE_ARCHIVE. ONLINE-NOW, which is updated daily, covers the period of the current month + the last six months, whereas the ONLINE_ARCHIVE covers the preceding period back to February 2017. At the beginning of each month, the contents of the oldest month of the ONLINE_NOW corpus are moved to ONLINE_ARCHIVE.

Until now, there were no monitor corpora of Slovene in existence. Nonetheless, recently, a resource called Language Monitor (Kosem et al. 2021) has been developed, which indicates trending words and N-grams in recent periods and is updated on a regular basis. Like Timestamped JSI web corpora, the Language Monitor also uses the IJS Newsfeed service to export news articles. After linguistic annotation (tokenization, lemmatization, morphosyntactic annotation, parsing) of texts, word lists are generated, and statistical calculations are conducted. This basically means that whenever Language Monitor is being updated, a sort of temporary monitor corpus is being created and a considerable manual effort is needed. Moreover, word lists provided are not linked to examples of use (e. g. corpus concordance), limiting their usefulness.

3. Trendi – a monitor corpus of Slovene

Services such as Language Monitor, which offer already prepared statistics for users, are more suitable for the general public; while lexicographers, linguists, and other language experts may find some of these options useful, they also need direct access to corpus data for their analyses. This is the motivation behind the SLED project, in which the first monitor corpus of Slovene, called Trendi, will be compiled.

3.1 Methodology

One of the main decisions in preparing Trendi was determining the time period covered by the corpus, and the regularity of its updates. As we learned from analysing a selection of monitor corpora of other languages, there was no uniform approach used. Our main principle was for Trendi to fill the gap not covered by the most recent version of the Slovene reference corpus, i. e. the Gigafida corpus, version 2.0 (Krek et al. 2019). Thus, with Gigafida 2.0 (1.13 billion words) covering the period from 1991 to 2018, the first version of Trendi covers the period from 2019 onwards. There are plans to publish Gigafida 3.0 towards the end of 2022, and to then make much more regular updates to the corpus, which will result in the monitor corpus covering a shorter period, and also being smaller in size.

Maintaining a close compatibility with the Gigafida corpus also means that the Trendi corpus covers (or monitors) the standard written Slovene language. The decision was based mainly on the needs of potential users of the corpus (translators, linguists, researchers,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

computational linguists, etc.) but also on the fact that non-standard Slovene is being covered by other projects such as JANES (Jezikoslovna analiza nestandardne slovenščine, 'Linguistic Analysis of Nonstandard Slovene'; Fišer/Ljubesic/Erjavec 2018).

As far as updates of the Trendi corpus are concerned, a new version will be released every month, uploaded both to the CLARIN repository and the relevant concordancers.

3.2 Contents

All the contents of the Trendi corpus are at the moment obtained by using the IJS Newsfeed service (Trampuš/Novak 2012). The Newsfeed has already been used for linguistic projects like Language Monitor (Kosem et al. 2021), which indicates trending words and N-grams in recent periods and is updated on a monthly basis. The selection of newsfeed sources for Language Monitor was very inclusive, taking all Slovenian sources with at least 10 articles per year.

In the selection of the sources for the Trendi corpus, we wanted to be more rigorous. Also, we had to consider the fact that we wanted Trendi to represent standard written Slovene. For this reason, we joined forces with the compilers of the Gigafida corpus. We made a list of all Slovenian sources that were part of the newsfeed since 2019 and made an analysis of their contents. The initial list included 243 sources. 90 sources were immediately excluded because they were mostly foreign websites o websites with non-Slovenian content. A further 34 sources were excluded for various reasons: not being a news source (e. g. blogs, government and company websites), not covering standard Slovene (e. g. repositories of academic publications such as diplomas and theses), and being an aggregator of news from news sources which were already on the list. The final list included 110 sources, with the top 15 and the number of news items from 2019 to 2021 shown in Table 1.

Source	Number of articles
sta.si	260,080
rtvslo.si	97,924
siol.net	69,471
delo.si	65,415
24ur.com	61,623
dnevnik.si	47,749
vecer.com	45,548
novice.svet24.si	42,049
vestnik.si	41,525
zurnal24.si	39,220
ekipa.svet24.si	35,326
demokracija.si	26,604
gorenjskiglas.si	22,883
nova24tv.si	20,153
slovenskenovice.si	18,622

Table 1: Top 15 news sources by a number of articles (2019–2021) in the Trendi corpus

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

One thing to note is that some of the sources, e. g. sta.si, delo.si and dnevnik.si, have some of their content available only through subscription. As a result, such news items collected from their websites contain only a title, sometimes a subheading, and the first paragraph. This issue has been resolved by forming a close collaboration with the Gigafida corpus team, as they are in the process of signing contracts with source providers to send them the full contents on a regular basis. Once this procedure is established, the contents of Trendi (as well as Gigafida of course) will become even richer.

At the time of writing this paper, the first version of the corpus was being prepared, with the intention of including the data up to May 2022, so we did not yet have the exact details on its size. We have already made some preliminary calculations for 2019–2021 data, and the 2019 subcorpus contains nearly 12,5 million words per month, the 2020 subcorpus nearly 15 million words per month, and the 2021 subcorpus nearly 21 million words per month. One of the reasons for the continuous increase in size per year is the regular appearance of new websites, for example necenzurirano.si was launched in 2020 and is already 28th on the list of sources (per number of news items) with 8,494 news items. This finding also underlines the importance of continuously monitoring the Slovenian web space for new websites, and adding the relevant websites to the Trendi corpus.

3.3 Article collection and annotation pipeline

The texts for the Trendi corpus are being downloaded on a daily basis, in the JSON format. All the articles from each individual source are merged into a single daily file before annotation. The deduplication check, i. e. ensuring (via URL) that the same article is not downloaded more than once, is already performed by the JSI Newsfeed service. No further deduplication is conducted at the moment, although we are aware that very similar articles can be found in various sources, especially media ones. This is because we want to make it possible for the users of the corpus to analyse the contents of individual sources, compare two or more sources, etc. A different approach will probably be taken for the Gigafida corpus where the deduplication is done on a paragraph level (Krek et al. 2019). Such a step for a reference corpus is very much needed, also considering the fact that STA (sta.si) is a service for the distribution of original press releases, which means that many media websites prepare articles based on these pieces of information and often use a considerable portion of the contents.

During the annotation of the files, the processes of tokenization, lemmatization, morphosyntactic tagging, dependency parsing, and named entity recognition are performed. The annotation output, provided in the CONNL-U format, is converted into the TEI format, the format needed for the calculations of various statistics, and for the conversion into the VERT format, used by the KonText and NoSketchEngine concordancing tools.

At the moment we are still getting the data from the JSI Newsfeed, therefore the TEI files need to be put through an additional step of source filtering, using the list of 110 sources as described in section 3.2. In the near future, we intend to limit the newsfeed extraction to only the sources selected for the Trendi corpus (and relatedly Gigafida).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

3.4 Accessibility

The Trendi corpus will be accessible via two concordancers: KonText CLARIN.SI (https://www.clarin.si/kontext), originally developed for the Czech National Corpus (Machálek 2020), and NoSketchEngine CLARIN.SI (https://www.clarin.si/noske/). The concordancers are somewhat complementary: they share many features, but KonText offers the option of registration and with that saving of searches and favourite corpora, whereas NoSketchEngine offers certain additional features such as Keyword extraction.

The Trendi corpus will also be uploaded to the CLARIN.SI repository, in both CONNL-U and TEI formats. Normally, corpora are provided only in the TEI format, however, our computational team has advised us to include CONNL-U as well, as this format is often preferred for processing tasks. The corpus will be made available under the Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. Due to copyright restrictions, we intend to publish paragraph samples from each text, the procedure already used for other corpora of Slovene such as the ccGigafida 1.0 corpus (Logar et al. 2013). We are also considering providing the full version of the corpus to individual researchers, under the condition of a signed agreement.

3.5 User survey

In addition to providing access to the Trendi corpus, the SLED project also aims to provide users interested in most current language use with various statistics and similar information on word usage. In order to get a good understanding of user needs and preferences regarding current language trends, we conducted a survey. The survey was prepared on the 1KA platform (http://www.1ka.si/) and included eight questions; four questions were content-related, and four questions collected information on the respondents (gender, age, occupation, and field(s) of activity).

The survey was completed by 100 respondents, 82 females and 18 males. The majority of the respondents were between 26–55 years old, with groups of 26–35 (33%) and 46–55 (32%) having the highest shares. 61% of the respondents were employed in the public sector (e. g. education, health organizations, public administration), and 20% were self-employed. Other groups such as company employees (6), retired persons (4), unemployed (5), and students (3) were much smaller. In terms of the field of activity, where the respondents could choose more than one option, the following groups dominated: proofreading (60%), translating (46%), language lover (38%), academic research writing (34%), language research (32%), and creative writing and blogs (22%). 40% in total is represented by the respondents from various categories of language education (Slovene as L1 in elementary or secondary school, Slovene as L2, language subjects at a university level).

In the first question, the respondents were provided with six different scenarios[6] and they had to express their level of interest (not interested at all, not interested, neither interested nor interested, interested, very interested, don't know). As Table 2 shows, they were interested in all the scenarios, with "interested" and "very interested" covering between 74–88% of responses. The highest interest was expressed for the last scenario where the trends in usage for two or more words or word combinations could be compared. Also high was the

[6] An example of a particular scenario was provided for clarity.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

level of interest in the information whether the usage of a word or word combination is increasing or decreasing recently.

On the question of whether the information on current language use would be helpful for their work, over three quarters (76%) of the respondents replied with Yes, with only 9% answering No, and 15% opting for "I don't know".

The third question asked the respondents about the importance of various visualizations of language data: diagrams, tables, and word lists. The answers indicate that diagrams, tables, and word lists are all considered important for the respondents, with up to 87% (for word lists) of the respondents considering them important or very important. A closer inspection of the results reveals that the respondents tended to slightly prefer a more simplistic presentation of language data, with tables with figures attributed the lowest combined importance (64%).

Scenario	Not interested at all	Not interested	Neither	Interested	Very interested	Don't know
Words or word combinations typical of a certain period compared to another period (e.g. which words are much more frequent in February 2020 compared to February 2021)	2%	8%	12%	**48%**	30%	0%
The period in which a certain word or word combination is the most frequent (e.g. was the word "tycoon" really the most frequent word in the period 2008–2009?)	5%	6%	14%	**44%**	30%	1%
Is the use of a word or word combination increasing or decreasing? (e.g. is the use of the word "epidemic" on the rise or is it decreasing)	2%	5%	5%	42%	**46%**	0%
In which texts (by topic) is a word or word combination more frequent? (e.g. is the noun "forward" really most frequent in sports texts)	1%	5%	9%	37%	**46%**	2%
What is the category distribution of the use of a word or a word combination? (e.g. is the word combination "collective immunity" found only in medical texts or not?)	2%	4%	12%	**51%**	30%	1%
Which of two (or more) words or word combinations is used more frequently in recent years/months? (e.g. which of the words "anti-vaxxer" or "countervaxxer" is used more frequently'?)	1%	5%	6%	34%	**53%**	1%

Table 2: Answers to six selected scenarios

The last question was an open-ended one and offered the respondents an opportunity to provide their own suggestions or scenarios for providing information on current language trends. The suggestions can be grouped into the following categories:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

- linkability or integration with other language resources and data access via API
- comparison of synonyms or related words (foreign words or loanwords vs Slovene equivalents)
- inclusion of examples of word usage, e. g. via links to corpus concordances
- monitoring different senses of words over time
- monitoring syntactic behaviour of words over time
- monitor multiword units (e. g. phrases) over time

Although the survey confirmed several objectives of the SLED project, the findings also made it clear that the community needs both online access to the Trendi corpus, as well as an online tool that facilitates the analyses of language trends beyond the scope of the Trendi corpus, and offers simple visualizations of complex statistical data.

Given that the project only promised statistics regularly uploaded to the CLARIN repository, we had to rethink the approach and have started preparing an online service that will be linked to various corpora, including Trendi, and will offer the users different options of analysing language data and exporting the results. The service, planned to be completed by the second half of 2022, will get the data via API from a data warehouse where we will store statistical information on word forms, lemmas, collocations, and other linguistic phenomena. The statistical information is calculated using the pipeline extension based on the corpus extraction tool LIST (Krsnik et al. 2019).

4. Integration of Trendi into the lexicographic workflow

Over the past year, while working on the Language Monitor and later on the conceptualization of the monitor corpus, we have already started working on the infrastructure that would support the needs of lexicographers. Of course, this goes beyond identifying neologisms, which is indeed the most common use of monitor corpora; lexicographers also need to identify and monitor potential future neologisms (words with a frequency below the threshold for inclusion into a dictionary), identify new uses of existing meanings (e. g. via collocations), and identify meanings, collocations and other phenomena which are already included in dictionaries and are used less and less frequently or not at all. At the moment, Slovenian lexicographers are very much hindered by the fact that they do not have direct access to language data beyond 2018 - this makes any language description immediately, at least to a certain extent, outdated.

One of the important pieces of the planned infrastructure is the data warehouse mentioned in section 3.5, which will serve as a repository of all possible information from corpora. The data warehouse will be linked with corpora and dictionary tools, and indirectly (after the lexicographers analyse the data) with the Digital Database for Slovene, which is being developed at the Centre for Language Resources and Technologies at the University of Ljubljana. Most importantly, the data warehouse will contain information that is at the moment not relevant or not yet relevant – for example, potential neologisms, which are at the moment not yet frequent enough or limited to too few sources, can be saved there (but not in the Digital Database). Furthermore, all identified bad collocation candidates from automatic extractions can be recorded there to avoid duplication of work in the future – based on our experience, getting rid of repeated inspection of bad data could save a considerable amount of time, especially in this day and age when corpora are very large.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

5. Conclusions and future plans

The Trendi monitor corpus for Slovene described in this paper is a very much needed resource for tracking trends in Slovene language use. Because of the various purposes, the corpus will be used for, it was paramount to prepare a sound and sustainable methodology for text collection and annotation, as well as for linguistic data extraction. This will facilitate lexicographic, linguistic, and other analyses, thus benefitting end-users of dictionaries and similar resources. In addition, the fact that Trendi will complement the Gigafida reference corpus will mean that there will now be corpus data on the Slovenian language from 1991 to yesterday.

More challenging tasks lie ahead. Among them is ensuring regular updates to the corpus, by which we mean both uploading new versions to the concordancers, but also identifying and adding new web sources. A detailed evaluation and possible improvement of the article collection procedure will be made and will include selecting a sample of articles from each ***source and identifying potential issues such as unwanted content*** **(*e.g. menus*)** ***being included,*** only part of the article being collected, etc.

Finally, the activity that is currently underway and which will improve Trendi, but also other corpora, even more, is the development of an automatic text categorization program. At the time of writing, we have been finalizing the list of text categories (e. g. politics, sport) and preparing the training corpora. In the coming month, the algorithm using supervised training will be developed and then tested on newly acquired articles, and more importantly on the Gigafida corpus. This development means that in the future lexicographers could also be provided with the category dispersion of different language phenomena.

References

Davies, M. (2008-): The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/ (last access: 23-03-2022).

Davies, M. (2016-): Corpus of News on the Web (NOW). https://www.english-corpora.org/now/ (last access: 23-03-2022).

Davies, M. (2019-): The Coronavirus Corpus. https://www.english-corpora.org/corona/ (last access: 23-03-2022).

Fišer, D./Ljubešić, N./Erjavec, T. (2018): The Janes project: language resources and tools for Slovene user generated content. Language Resources & Evaluation. https://doi.org/10.1007/s10579-018-9425-z.

Herman, O. (2013): Automatic methods for detection of word usage in time. Bachelor thesis. Masaryk University.

Kilgarriff, A./Rychlý, P./Smrz, P./Tugwell, D. (2004): The Sketch Engine. In: Williams, G./Vessier, S. (eds.): Proceedings of the Eleventh EURALEX International Congress, Lorient, France. Lorient, pp. 105–116.

Kosem, I./Krek, S./Gantar, P./Arhar Holdt, Š./Čibej, J. (2021): Language monitor: tracking the use of words in contemporary Slovene. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual. Brno, pp. 514-527. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_33_pp514-528.pdf.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag. *https://doi.org/10.14618/phpy-6r66*

Kosem, I. et al. (2019): The image of the monolingual dictionary across Europe: results of the European survey of dictionary use and culture. In: International Journal of Lexicography 32 (1), pp. 92–114. https://doi.org/10.1093/ijl/ecy022.

Krek, S. et al. (2019): Corpus of Written Standard Slovene Gigafida 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1320.

Krek, S./Arhar Holdt, Š./Erjavec, T./Čibej, J./Repar, A./Gantar, P./Ljubešić, N./Kosem, I./Dobrovoljc, K. (2020): Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ed.): LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11–16, 2020, Marseille, France. Paris: ELRA – European Language Resources Association. 2020, pp. 3340–3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krsnik, L. et al. (2019): Corpus extraction tool LIST 1.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276.

Logar, N./Erjavec, T./Krek, S./Grčar, M./Holozan, P. (2013): Written corpus ccGigafida 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. http://hdl.handle.net/11356/1035.

Machálek, T. (2020): KonText: Advanced and Flexible Corpus Query Interface. In: Proceedings of LREC 2020, pp. 7005–7010.

Trampuš, M./Novak, B. (2012): The internals of an aggregated web news feed. In: Proceedings of 15th Multiconference on Information Society 2012 (IS-2012). http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf.

Contact information

Iztok Kosem
Jožef Stefan Institute & Faculty of Arts, University of Ljubljana
iztok.kosem@ijs.si

Acknowledgements

The project *Spremljevalni korpus in spremljajoči podatkovni viri* (SLED) is funded by the Ministry of Culture of the Republic of Slovenia.

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, *Language Resources and Technologies for Slovene*)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66