Language Monitor: tracking the use of words in contemporary Slovene

Iztok Kosem[1,2], Simon Krek[1,2], Polona Gantar[1], Špela Arhar Holdt[1], Jaka Čibej[1]

[1] Centre for Language Resources and Technologies (CJVT), University of Ljubljana
[2] Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: iztok.kosem@cjvt.si, simon.krek@ijs.si, apolonija.gantar@guest.arnes.si, spela.arhar@cjvt.si, jaka.cibej@ff.uni-lj.si

Abstract

In this paper, we present Language Monitor 1.0, a new online resource for monitoring language changes in Slovene, developed at the Centre for Language Resources and Technologies at the University of Ljubljana. The resource is another part of the newly developed infrastructure for researching and describing contemporary Slovene. Language Monitor 1.0 offers four sections to observe word usage: (1) a single-word list, (2) word groups, (2) a neologism section, and (4) word comparisons. The words for a single-word list are manually validated from a list of salient word candidates, which are identified using the Simple Maths method. The paper also describes future plans, including the setup of a relational database linked with a data warehouse solution for analysis purposes, which will include various statistical information on different language phenomena relevant for researchers, lexicographers, and other users, and will provide possibilities for adding several new features to the Language Monitor.

Keywords: Language Monitor; trends; neologisms; language change; corpus

1. Introduction

One of the most challenging tasks of dictionary makers has always been ensuring that the dictionary content remains up-to-date. Modern lexicography now has all the means to address this – large corpora that can be updated on a daily basis, advanced tools for analysing the use of words over time, etc. As a result, the duration of periods between dictionary updates has decreased dramatically, from several years to months. This change has also been driven by user expectations, and the perception of dictionaries, or rather lexical resources, in modern society. The COVID-19 pandemic has exposed such a need even more – new words and word meanings have been entered into dictionaries more rapidly than ever before. It should be noted that updating the dictionary with neologisms solves only part of the problem. What about updating collocations, examples, spelling, and even definitions? It could be argued that having outdated content in a dictionary is just as problematic as lacking information on contemporary language use.

There is another element of language change that dictionary entries do not cover, namely trends in the use of existing vocabulary. Some words or their meanings, which are already established in the language, can suddenly be used much more frequently, or can be replaced by another word for a certain period. Such information can also be relevant for users, both language experts and the general public.

Another challenge brought on by monitoring language change is data modelling, as one wants to ensure that information on different language phenomena can be constantly updated, and at the same time remain compatible with databases of dictionaries and other relevant resources. Furthermore, all this information needs to be made (immediately) available to different interested parties and propagated across different resources in order to reach as many user groups as possible.

The challenges above are those faced by the Slovene lexicographic community, and probably many others, with an additional problematic factor, which is that the entire Slovene language description is in need of a significant update. This means that the language changes that need to be described may reach as far as 30 years in the past (the last general dictionary of Slovene was published in 1991[1]), and such efforts are underway. However, other solutions and methodologies have been developed to partially address this issue. These solutions include responsive dictionaries (Arhar Holdt et al., 2018), using a combination of automatic lexical data extraction and ongoing validation (e.g. Collocations Dictionary of Modern Slovene; Kosem et al., 2018[2]), and resources that focus on temporal information such as the resource presented in this paper.

In this paper, we present a new free online resource for Slovene, Language Monitor, which has been developed at the Centre for Languages and Resources at the University of Ljubljana. First, we make an overview of existing research and dictionary practices of monitoring language use. Then, we present Language Monitor 1.0, both the backend, i.e. data collection and processing procedures, and frontend, i.e. the interface. Next, we outline future plans, which include the development of a data warehouse that will be used by not only the Language Monitor, but also other resources and tools. We conclude by summarising the main points and considering potential future challenges.

2. Monitoring language use

There is a great deal of research on detecting changes in language (see e. g. Geeraerts, 2014 for an overview), with much more focus being on new words and meanings, i.e. lexical and semantic neologisms, than on changes in usage of existing meanings. Relatedly, a number of corpus-based statistical approaches and tools have been developed for neologism detection in longitudinal corpora, for example NeoCrawler

[1] There was an updated version published in 2014, but as the reviews (Ahlin et al., 2014; Krek, 2014) have pointed out, the changes introduced were not that significant.

[2] https://viri.cjvt.si/kolokacije/eng/

(Kerremans et al., 2012), NeoTrack (Janssen, 2008), ZeitGeist (Veale, 2006), Neoveille (Cartier, 2019). Similar functionality is offered by the Trends feature (Herman & Kovar, 2013) in the Sketch Engine corpus tool (Kilgarriff et al., 2004). However, the main aim of Trends is to flesh out words with significant increase or decrease in use over time.

Specifically in the area of semantic neology, a number of corpus-based techniques have been developed in the distributional semantic framework to detect semantic changes in large corpora (Sagi et al., 2011; Cook et al., 2014; Gulordava & Baroni, 2011). Such studies approach semantic neologisms from a computational perspective, while Heylen et al. (2015) present a more lexicologically oriented approach based on word space models. A similar study for Slovene was done by Fišer and Ljubešić (2016), who explored semantic shifts in Slovene tweets.

N-grams and collocations can play a pivotal role in the detection of semantic neologisms, as shown for example by projects such as AVIATOR (Renouf, 1993) and WebCorpLSE (Kehoe & Gee, 2009; Renouf, 2009). Nimb et al. (2020) used bigrams to detect new meanings of existing words in Danish for the purposes of updating the Danish dictionary. Pollak et al. (2019) conducted a similar study for Slovene, using collocations to detect new meanings in computer-mediated communication. But as Renouf (2013) points out, collocations can also help us track the life-cycle of a word, i.e. phenomena such as birth, increased use (through productivity, creativity, etc.), death, and possible revival. These aspects of collocations in Slovene have been explored in the Collocations in Slovene project (KOLOS; Kosem et al. 2020).

Translating linguistic methods into lexicographic practice, several authors have discussed the criteria of including neologisms into dictionaries (e.g. Barnhart, 1985; Metcalf, 2002; Ishikawa, 2006; O'Donovan & O'Neill, 2008; Cook, 2010; Freixa & Torner 2020). In this respect, the study by Nimb et al. (2020) is particularly valuable as it describes the decisions made and criteria used on a concrete dictionary project. What is particularly noteworthy is that Nimb et al. report (ibid. 2020: 122) that the results of their analyses lead not only to the addition of new meanings, compounds, and collocations to the dictionary, but also to the revisions of definitions and the inclusion of new usage examples.

Dictionaries use different methods and different types of data in conveying the information on language change to their users. First and foremost, announcements on newly added words and word meanings are made by dictionary publishers. The periods between these announcements have become increasingly shorter. They can now be made every few months, depending on the amount of new vocabulary that needs to be explained. During the COVID-19 pandemic, we have seen dictionaries all around the world react in an unprecedentedly rapid manner, introducing pandemic-related vocabulary within months if not weeks of the start of the pandemic.

The second approach used by dictionaries is to include the information on word usage over time directly in dictionary entries. An example of such an approach can be found

in the *Digitales Wörterbuch der deutschen Sprache* (DWDS)[3] where each headword is accompanied with a line graph showing its use from 1946 (or 1600) onwards, with the frequency data coming from German corpora. This approach in principle shows the change in usage for every word (but not its individual meanings), but the information needs to compete with other more often consulted information in an entry such as definitions, collocations, etc. A different method is used by Dictionary.com where the information on trends is displayed only for words whose usage has recently increased significantly; however, this information is displayed much more prominently, on the dictionary homepage, in a manner similar to that used by stock-exchanges (Figure 1).

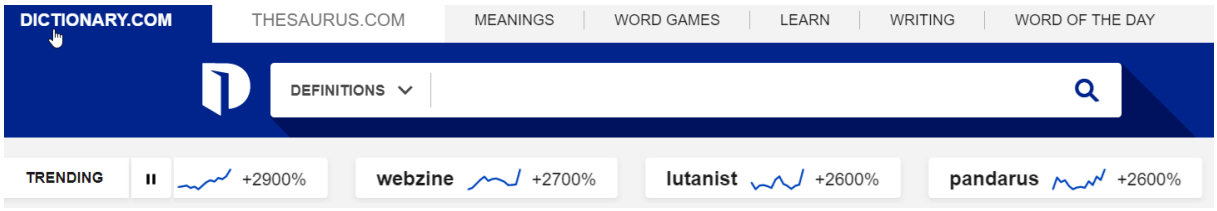

Figure 1: Trending words offered by Dictionary.com.

Some dictionaries rely on user provided information to detect language change, either indirectly or directly. An example of such practice is exhibited by *Merriam-Webster Dictionary*,[4] showing a list of the 10 most frequent words looked up by users, which is refreshed every 30 seconds.[5] While the users may not necessarily look up only words trending in frequency of use or new words (Table 1), it can be argued that many of the words from the list are probably a reaction to a current event or trending topic. As such, they not only reflect the individual's personal activities (e.g. reading), but a general topic that is relevant in a given language community at that moment.

love, infrastructure, racism, erotic, watering hole, fore, fascism, consort, hi, integrity, ambivalent, nonce, perseverance, drub, anti-sex, nexus, joke, berate, nickname, cisgender, sexi-, countenance, inclination, democracy, humility, answer, pandemic, diversity, esoteric, cognitive, autonomous, obtuse, innovation, fraud, insight, et al., pron, communism

Table 1: Words featured in the top 10 looked up by users in *Merriam-Webster Dictionary* (over a 10-minute span).

One shortcoming of the approaches mentioned so far is that they mainly promote the content already included in the dictionary. In other words, lexical or semantic

[3] https://www.dwds.de/
[4] https://www.merriam-webster.com/

[5] A similar approach is used by Oxford Dictionary at https://lexico.com, although it is not completely clear whether "Trending words (most popular in the world)" is a list of searches or corpus frequency.

neologisms that may have been detected by lexicographers but still need to be described are not included. Some dictionaries address this gap by using the crowdsourcing approach, asking users for suggestions for new words to be added to the dictionary. This approach is used by *Collins English Dictionary* in its *Word submissions* section. What is particularly commendable in the case of this particular dictionary is that the users are given publicly visible feedback on their suggestions in the form of a status note (Pending Investigation, Rejected, or Published).

As for dictionaries of Slovene, the coverage of language change has been focussed mainly on neologisms through the *Growing Dictionary of the Slovenian Language* (*Sprotni slovar slovenskega jezika*; Krvina, 2014-). Changes in the usage of existing Slovene vocabulary are much less documented, and the data has so far not been available to the general public. We decided to address this gap by developing a new resource – the Language Monitor.

3. Language Monitor 1.0

Version 1.0 of the Language Monitor (*Jezikovni sledilnik* in Slovene, or *Sledilnik* for short; https://viri.cjvt.si/sledilnik/slv/) was published in January 2021 and offers an overview of a number of salient words that have significantly impacted the language of Slovene online media in 2020 by visualising the information on temporal trends of words, i.e. the changes in their relative frequencies over a period of time. The main aim of Language Monitor in the current version is to inform users about the most prominent words in a certain period, and about new words coming into the language.

In the following subsections, we describe the data used (Section 3.1) and the process of obtaining the most salient words (Section 3.2), as well as the features of the Language Monitor 1.0 (Section 3.3).

3.1 Data

Language Monitor uses the data from the Gigafida 2.0 Reference Corpus of Written Standard Slovene (Krek et al., 2020), which covers the period between 1991 and 2018, and from the IJS NewsFeed service (Trampuš & Novak, 2012), which has been used since 2019 for daily extraction of texts from over 100 Slovene online sources, including the website of the main national television station MMC RTV Slovenija and the Slovene newspaper with the largest readership, *Delo*. The top 10 sources (by number of articles in 2020) are listed in Table 2. The output of the IJS NewsFeed service is processed through a custom pipeline that tokenises, lemmatises, morphosyntactically annotates, and segments the texts into sentences, resulting in XML files in TEI P5 format.[6]

[6] TEI P5 Guidelines - https://tei-c.org/guidelines/p5/

Our list of NewsFeed sources currently contains 102 sources. Only the sources providing at least 10 news items per year are included, but new sources or sources exceeding the minimum limit are regularly added to the list. The size of the yearly corpus from these sources was approx. 130 million tokens for 2019 and approx. 146 million tokens for 2020. Monthly subcorpora thus contain between 10 and 12 million tokens, with daily sizes ranging from 200,000 to 400,000 tokens. For reference, the yearly subcorpora from Gigafida 2.0 (1991-2018) contain an average of almost 46 million tokens, which is three times less than the yearly corpora from NewsFeed.

Source	Description	URL-domain	IJS Newsfeed articles from 2020
Slovenska tiskovna agencija (STA)	Slovenian Press Agency news portal	sta.si	101,060
MMC RTV Slovenija	National radio and television news portal	rtvslo.si	35,723
Siol.net Novice	Online news portal	siol.net	23,968
Delo	Newspaper website	delo.si	22,765
24ur.com	Commercial radio and television news portal	24ur.com	21,293
Žurnal24	Newspaper website	zurnal24.si	18,082
preberi.si	News aggregator	preberi.si	17,079
Večer	Newspaper website	vecer.com	17,054
Dnevnik	Newspaper website	dnevnik.si	15,400
Svet24	Newspaper website	novice.svet24.si	15,243

Table 2: List of sources providing most news texts in 2020.

3.2 Extraction of Salient Words

The salient words included in the Language Monitor 1.0 are obtained by comparing two corpora representing the reference period and the current period, respectively. For the most salient words of 2020, the reference corpus used was the amalgamation of Gigafida 2.0 (covering the period between 1991 and 2018) and the IJS Newsfeed output from 2019. The contemporary corpus contained the IJS Newsfeed output from 2020 (January-December).

Frequency lists of word forms[7] were extracted from both corpora using LIST (Krsnik

[7] Word forms were extracted instead of lemmas in order to prevent the merging of potential homonyms in the vein of *pot* (masculine noun, 'sweat') and *pot* (feminine noun, 'path'). Lists of word forms extracted with LIST contain lemmas and full morphosyntactic descriptions using the MTE-6 annotation schema (http://nl.ijs.si/ME/V6/msd/html/msd-sl.html), while lists of

et al. 2019), a custom-made open-source software tool for the extraction of corpus data that can be used to generate frequency lists of characters, word parts, word forms/lemmas, or word sets (n-grams). LIST supports the TEI P5 XML format and the VERT format, and outputs .TSV files.

The extracted frequency lists of word forms were then converted to frequency lists of lemmas (keeping the relevant discriminatory information such as gender for nouns and aspect for verbs). Next, the entries from both frequency lists were compared in terms of their relative frequencies using the Simple Maths formula (Kilgarriff, 2009), where fr1 is the relative frequency of a word in the reference corpus, fr2 is the relative frequency of the word in the contemporary corpus, and N is the smoothing parameter (in case the word is not found in the contemporary corpus and fr2 equals zero; the smoothing parameter was set to 1 in our case):

$$sm = (f_{r2} + N) / (f_{r1} + N)$$

Table 3 shows the top 10 most salient words of 2020, along with their MTE-6 lexical features, absolute and relative frequencies, and Simple Maths scores.

Lemma	MTE-6 Lexical Features	f_a (1991-2019)	f_a (2020)	f_r (1991-2019)	f_r (2020)	Simple Maths Score
koronavirus	Som	175	214,947	0.120	1463.444	1307.997
covid	Som	0	90,054	0	613.123	614.123
pandemija	Soz	1,668	76,873	1.140	523.382	245.034
covid	Kag	0	22,870	0	155.708	156.708
karantena	Soz	2,852	48,976	1.949	333.448	113.400
epidemija	Soz	11,028	118,082	7.537	803.949	94.285
protikoronski	Pp	0	11,880	0	80.884	81.884
koronavirusen	Pp	1	10,148	0.000683	69.092	70.044
epidemiološki	Pp	1,771	21,253	1.210	144.700	65.914
Covid	Slm	0	9,419	0	64.128	65.128

Table 3: The top 10 most salient words of 2020 compared to 1991-2019.

The list of most salient words of 2020 contains neologisms (*covid-19*, *protikoronski* 'anti-corona (adjective)') as well as existing words with a significant increase in usage during 2020 (*epidemiološki* 'epidemiological', *karantena* 'quarantine', *pandemija* ('pandemic', noun), *koronavirusen* 'adjective; related to coronavirus'), *epidemija* 'epidemic'). However, the list also contains a number of problems caused by errors in automatic lemmatisation and morphosyntactic tagging. For instance, 'covid' is lemmatized as both

lemmas contain only parts-of-speech, which would merge the frequencies for *pot* (masculine) and *pot* (feminine).

'covid' and 'Covid' and tagged as a common noun (Som), a proper noun (Slm) or even as a numeral (Kag). There is also the problem of the overlap with n-grams: 'covid' mostly often occurs as 'covid-19', which is treated as a 3-gram by our tokeniser ('covid', '-', '19'). We have amended this during manual analysis (changing *covid* to *covid-19*), as version 1.0 of the Language Monitor only focuses on single words. N-grams will be treated in future versions (as described in Section 4).

The obtained lists of salient words were manually analysed to remove noise. The relevant words were then included in the Language Monitor 1.0 database along with their frequencies.

3.3 Features

The Language Monitor 1.0 offers four sections to observe word usage: (1) a single-word list, (2) word groups, (2) a neologism section, and (4) word comparisons.

The first option (shown in Figure 2) features a list of 100 words that have been identified as the most salient in 2020 compared to the period between 1991 and 2019. The user can click on a word in the list and is provided with a line graph showing the trend of the word's relative frequency between January 2020 and December 2020. Figure 2 shows the temporal trend of the word *koronavirus* ('coronavirus'), the most salient word of 2020. The line graph shows a steep increase of usage between February and March 2020, when an epidemic was officially declared in Slovenia. After the initial surge, the usage of *koronavirus* stabilises and remains relatively unchanged in the period between June and December 2020.

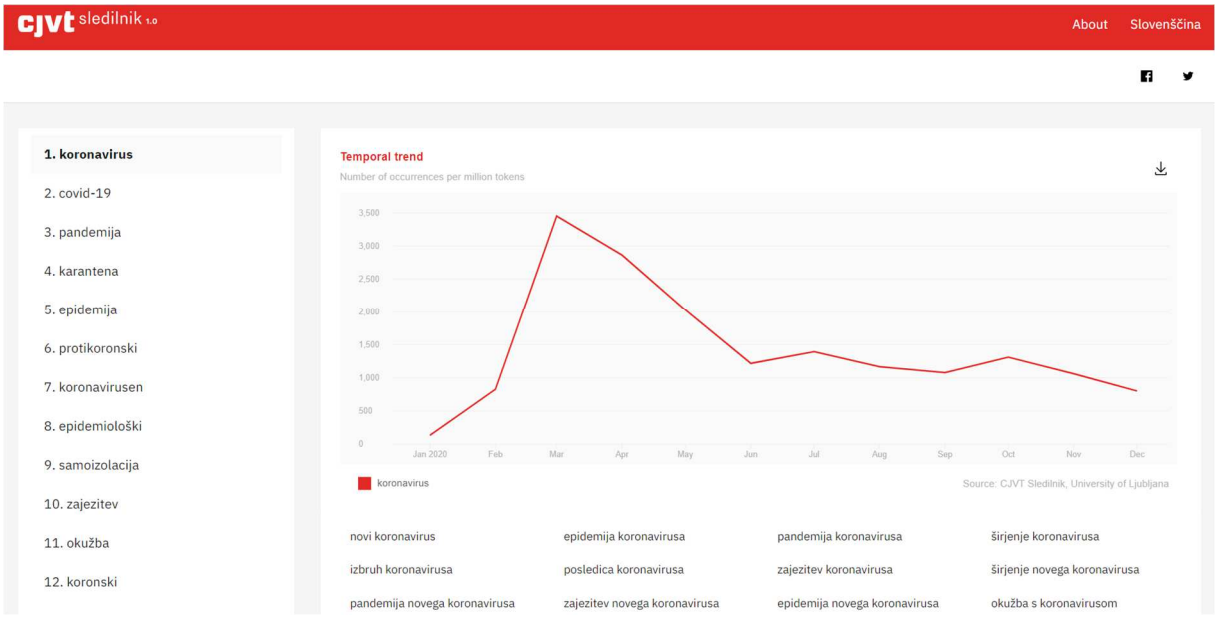

Figure 2: Line graph of the temporal trend for *koronavirus*.

Below the line graph, the most frequent n-grams featuring the word in question are listed. In this case, they contain expressions such as *novi koronavirus* ('novel coronavirus'), *izbruh koronavirusa* ('coronavirus outbreak'), *posledica koronavirusa* ('consequence of coronavirus'), *širjenje novega koronavirusa* ('spread of the novel coronavirus'), and so on.

The second section features temporal trends of word groups, i.e. groups of words that share a certain characteristic. At the end of March 2021, a total of 13 groups were available, for instance *Neologisms - February 2021* (containing salient words that first appeared in February 2021), *Words - February 2021* and *Words - January 2021* (salient words from January and February 2021, respectively), *Proper Nouns - January 2021* (prominent proper nouns from January 2021), and *Verbs - 2020* (salient verbs from 2020). Figure 3 shows the visualisation for *Words - February 2021* and features the list of available word groups on the left (the currently viewed word group is set in bold), a line graph with temporal trends of one or more salient words on the right (the first three are shown in the line graph by default; up to six can be visualised), and a clickable list of salient words below the graph. By selecting or unselecting words, the user can modify the line graph to visualise the relevant words. By clicking on the Download icon in the upper right corner of the line graph, the user can also export the line graph in .PNG format for further use.

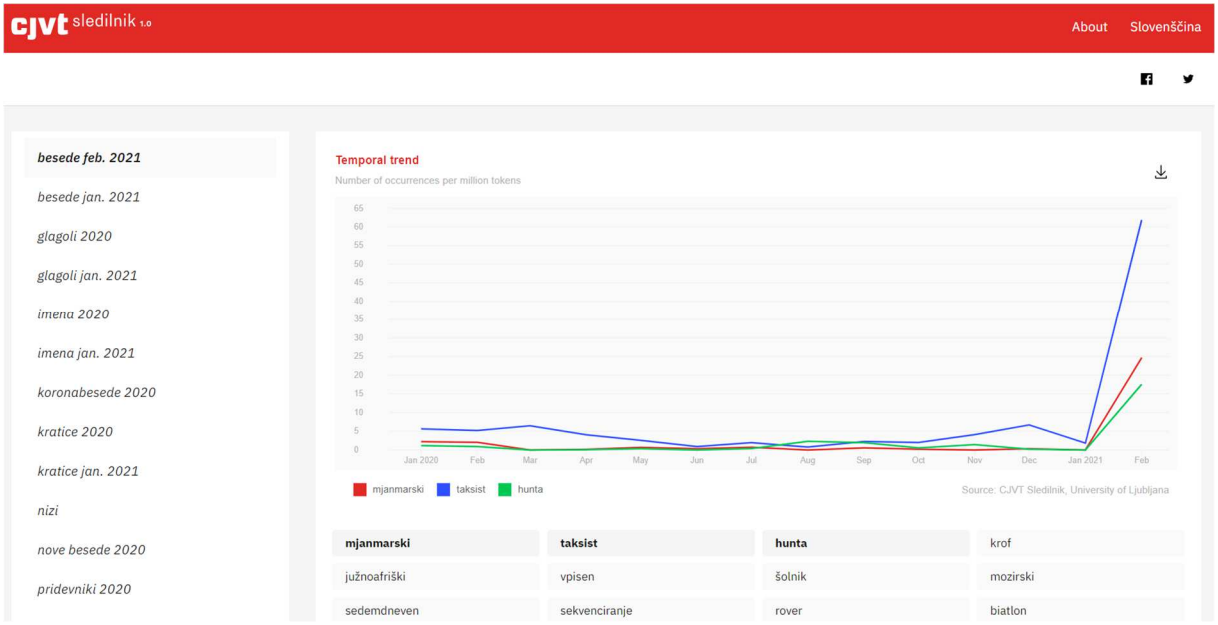

Figure 3: Line graph for the *Words - February 2021* word group.

The most salient words from February 2021 reflect most of the major events (both local and global) reported by Slovene media in that month, such as the coup d'etat in Myanmar (*mjanmarski* 'adjective, related to Myanmar', *hunta* 'junta'), seasonal holidays (*pusten* 'adjective, related to Mardi Gras', *krof* 'doughnut', *valentinovo* 'Valentine's Day'), the ongoing coronavirus epidemic (*južnoafriški* 'South African',

sekvenciranje 'sequencing', *cepljen* 'vaccinated'), political turmoil in the Slovene parliament (*nezaupnica* 'vote of no confidence'), sexual harassment revelations in Slovene society and subsequent changes to Slovene legislation regarding sexual violence (*nadlegovanje* 'harassment', *redefinicija* 'redefinition'), and NASA's rover mission to Mars (*rover* 'rover').

The third option is the neologism section, a special word group section which features salient words that are found in the compared corpus but have never appeared in the reference corpus. Shown in Figure 4 is the February 2021 neologism section, which features, for example *karanteval* (a lockdown version of a Mardi Gras parade; a portmanteau of *karantena* 'quarantine' and *karneval* 'carnival') and *astroturfing* (an English loanword which experienced a surge in use after a Slovene politician accidentally revealed their use of a fake Twitter profile to attack political opponents). Each neologism also features a sentence exemplifying its use, along with a link to the original article, its source and date of publication. In version 1.0, no line graph is provided for neologisms since the word has just entered language use and no trends are yet available.

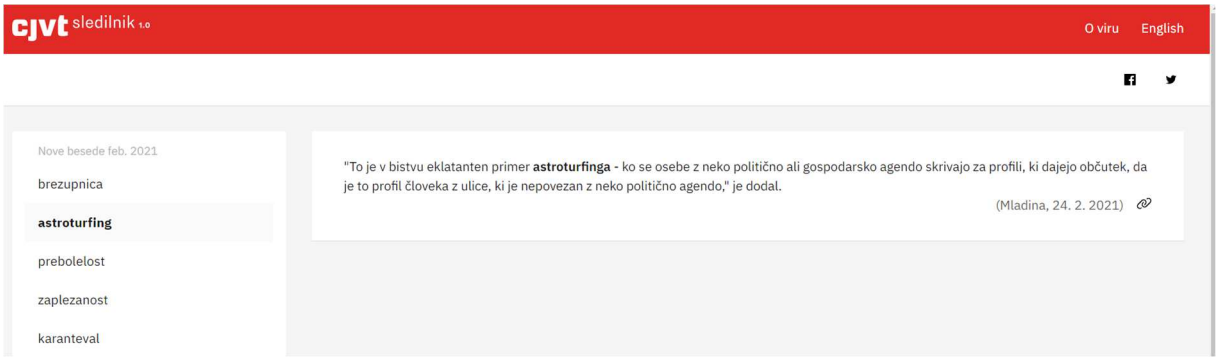

Figure 4: The neologism section (February 2021) of the Language Monitor 1.0.

The last section offers trend comparisons between words with data available in the Language Monitor 1.0. A total of 184 salient words were available for comparison by the end of March. The user can either select one of the preset comparisons (which have been prepared in advance) or generate a custom comparison by selecting up to six words from the list of available words (similar to the word group comparison, but this section allows for comparisons among all available words, not just within the relevant group). Figure 5 shows a preset comparison of the words *samoizolacija/samoosamitev* (both meaning 'self-isolation') and *izolacija/osamitev* ('isolation'). The trends show that the words *samoizolacija* (red) and *izolacija* (yellow) are both more frequent than their counterparts *samoosamitev* (blue) and *osamitev* (green).

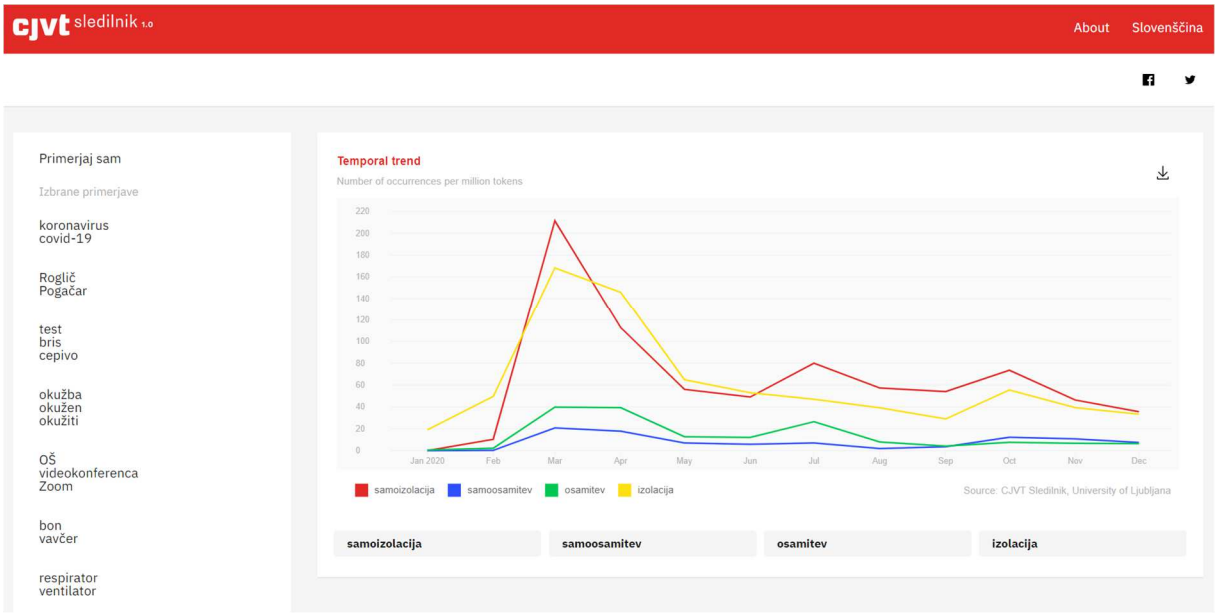

Figure 5: Trend comparisons in the Language Monitor 1.0.

4. Conclusions and future plans

The Language Monitor is a new addition to the infrastructure for contemporary Slovene, a resource that has made first strides towards consistent and constant monitoring of language change. Version 1.0 has focussed on presenting this information to the general public, using word lists in combination with different visual (line graphs) and interactive methods such as word groups and comparisons.

It was clear to us from the very beginning that the current methods of updating the Language Monitor were not sustainable nor desirable long-term, especially in view of the needs and wishes of researchers, lexicographers, and users. Considering the progress made in the area of lexical data extraction from Slovene corpora (e.g. Gantar et al., 2016) and the ongoing development of the Digital Dictionary Database for Slovenian (Klemenc et al., 2017; Kosem et al., forthcoming), which will consolidate different monolingual and bilingual lexical resources for Slovene, it is our aim to integrate the Language Monitor into this infrastructure.

Consequently, we have started preparing a pipeline that will extract various statistical information (e.g. raw frequency, number of different texts, source) on lemmas, collocations, multiword lexical units, etc., along with links to corpus examples, on a daily basis. In order to ensure data compatibility, the Gigafida 2.0 reference corpus for the years up to 2018 will need to be reprocessed with the same pipeline, using the latest versions of tools for morphosyntactic tagging, parsing and other annotation layers. This was not done for the Language Monitor 1.0, and we have already observed a number of issues caused by differences in lemmatisation and morphosyntactic tagging during

our manual analyses.

All the data extracted from the text using our pipeline will be fed into a relational database, which will store various information on different language phenomena in Slovene. Importantly, the database will hold the information on data from different types of corpora from different periods. Then, using a data warehouse solution, the information in the database will be analysed using different statistical methods (including Simple Maths, various association measures for collocations, etc.) and the results made available to lexicographers working on various lexical resources. Many of these calculations are already offered by corpus tools. However, lexicographers often need to take additional calculation steps during concordance analysis in order to obtain such information, and then make decisions based on it. It is our intention to use the data warehouse solution to provide lexicographers with alerts about significant changes in the usage of lexical items over time, or about important usage patterns in general (e.g. text type dispersion).

On the other hand, the database will directly feed the resources aimed at the general public, particularly the Language Monitor, which will offer users the possibility to not only observe but also explore the usage of words and collocations over time. Specifically, the ideas for the Language Monitor currently in preparation include implementing three methodologies: automatic extraction, manual analysis by linguists/lexicographers, and user involvement (crowdsourcing). In this manner, experts and users will work together in shaping the Language Monitor, and by feeding the results back into the database, their work will be of benefit to lexicographers and researchers.

5. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, *Language Resources and Technologies for Slovene*) and that the projects *Collocation as a basis for language description: semantic and temporal perspectives* (J6-8255) and *New grammar of modern standard Slovene: resources and methods* (J6-8256) were financially supported by the Slovenian Research Agency.

6. References

Ahlin, M., Lazar, B., Praznik, Z. & Snoj, J. (2014). Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. Izdali Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni inštitut Slovenske akademije znanosti in umetnosti, Inštitut za slovenski jezik Frana Ramovša. *Jezik in slovstvo*, 59 (4), pp. 121–127.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. In: J. Čibej, V. Gorjanc,

I. Kosem, S. Krek (eds.): Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. ISBN 978-961-06-0097-8). Ljubljana: Znanstvena založba Filozofske fakultete. 2018, pp. 401-410. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Barnhart, D.K. (1985). Prizes and pitfalls of computerized searching for new words for dictionaries. *Dictionaries* 7, pp. 253-260.

Cartier, E. (2019). Neoveille, web platform for finding and monitoring neologisms in monitor corpora. *Neologica*, 13, pp. 23–54.

Cook, P. (2010). *Exploiting Linguistic Knowledge to Infer Properties of Neologisms*. PhD Dissertation. Toronto: University of Toronto.

Cook, P., Rundell, M., Lau, J. H. & Baldwin, T. (2014). Applying a word-sense induction system to the automatic extraction of dictionary examples. In A. Abel et al. (eds.) *Proceedings of the XVI EURALEX International Congress. Bolzano, Italy: EURAC*, pp. 319–328.

Fišer, D. & Ljubešić, N. (2016). Detecting semantic shifts in Slovene Twitterese. In A. Horák, P. Rychlý & A. Rambousek (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pp. 1–8.

Freixa, J. & Torner, S. (2020). Beyond frequency: On the dictionarisation of new words in Spanish. *Dictionaries* 41(1), pp. 131-154.

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29(2), pp. 200–225.

Geeraerts, D. (2014). How words and vocabularies change. In J. Taylor (ed.) *The Oxford Handbook of the Word*.

Gulordava, K. & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 67–71.

Herman, O. & Kovář, V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 201*3. Brno: Tribun EU, pp. 79–85.

Heylen, K., Wielfaert, T., Speelman, D. & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157, pp. 153-72.

Ishikawa, S. (2006). When a word enters the dictionary: A data-based analysis of neologism. In *JACET Society of English Lexicography, English Lexicography in Japan*. Bunkyo-ku: Taishukan, pp. 39-52.

Janssen, M. (2008). NeoTrack: Un analyseur de néologismes en ligne. In M.T. Cabré, O. Domènech, R. Estopà & J. Freixa (eds.) *Proceedings of CINEO 2008*, pp. 1175-1188.

Kehoe, A. & Gee, M. (2009). Weaving Web data into a diachronic corpus patchwork. In A. Renouf & A. Kehoe (eds.) Corpus Linguistics: Refinements and Reassessments. Leiden: Brill, pp. 255-279.

Kerremans, D., Stegmayr, S., & Schmid, H. J. (2011). The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. In Allan & Robinson (eds) *Current Methods in Historical Semantics*, 73, pp. 59.

Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (eds.), *Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK, July 2009.* https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6-10, 2004*. Lorient: Universite de Bretagne – sud, pp. 105–116.

Klemenc, B., Robnik-Šikonja, M., Fürst, L., Bohak, C. & Krek, S. (2017). Technological Design of a State-of-the-art Digital Dictionary. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds) *Dictionary of modern Slovene: problems and solutions. 1st ed.* Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10-22.

Kosem I., Krek, S. & Gantar, P. (forthcoming). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. *Proceedings of EURALEX 2020, Volume II.*

Kosem, I, Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C A. (2018). Collocations dictionary of modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 989-997. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Kosem, I., Krek, S., Čibej, J., Gantar, P., Arhar Holdt, Š., Logar, N., Laskowski, C. A., Klemenc, B., Ljubešić N., Dobrovoljc, K., Gorjanc, V. & Pori, E. (2020). *The Orange workflow for observing collocation clusters ColEmbed 1.0*, Slovenian language resource repository CLARIN.SI. https://www.clarin.si/repository/xmlui/handle/11356/142.

Krek, S. (2014). Prva in druga izdaja SSKJ. *Slovenščina 2.0,* 2(2), pp. 114–160. Accessed on 11 April 2021. https://doi.org/10.4312/slo2.0.2014.2.114-160.

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. *Proceedings of the 12th Language Resources and Evaluation Conference" European Language Resources Association"*, pp. 3340-3345. https://www.aclweb.org/anthology/2020.lrec-1.409.

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

Metcalf, A. (2002). *Predicting New Words.* Boston, MA: Houghton Mifflin Company.

Nimb, S., Sørensen, N. H. & Lorentzen H. (2020). Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 8(2), pp. 112-138.

https://doi.org/10.4312/slo2.0.2020.2.112-138.

O'Donovan, R. & O'Neill, M. (2008). A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress* (Barcelona, 15-19 July 2008). Barcelona: IULA-UPF, pp. 571-579.

Pollak, S., Gantar, P. & Arhar Holdt, Š. (2019). What's New on the Internetz? Extraction and Lexical Categorisation of Collocations in Computer-Mediated Slovene, *International Journal of Lexicography*, 32 (2), pp. 184–206, https://doi.org/10.1093/ijl/ecy026.

Renouf, A. (1993), A Word in Time: first findings from dynamic corpus investigation. In J. Aarts, P. de Haan, & N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi, pp. 279-288.

Renouf, A. (2009). Corpus Linguistics beyond Google: the WebCorp Linguist's Search Engine. In R. Siemens & G. Shawver (eds.) *New Paths for Computing Humanists*, in Digital Studies / Le champ numérique Vol 1, No 1, the Society for Digital Humanities / Société pour l'étude des médias interactifs (SDH/SEMI).

Renouf, A. (2013). A finer definition of neology in English: The life-cycle of a word. In H. Hasselgård, J. Ebeling & S. Oksefjell Ebeling (eds.) *Corpus Perspectives on Patterns of Lexis* (Studies in Corpus Linguistics, 57), pp. 177-208.

Sagi, E., Kaufmann, S. & Clark, B. (2011). Tracing semantic change with latent semantic analysis. In K. Allan & J. A. Robinson (eds.) *Current Methods in Historical Semantics*. De Gruyter Mouton, Berlin, Germany.

Trampuš, M. & Novak, B. (2012). The Internals Of An Aggregated Web News Feed. *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).* http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf

Veale, T. (2006). Tracking the Lexical Zeitgeist with Wikipedia and WordNet. In *Proceedings of ECAI'2006, the 17th European Conference on Artificial Intelligence*.