Domen Krvina
ORCID: 0000-0002-2276-1156

The Growing Dictionary of the Slovenian Language (2014-) and Slovenian Neologisms: Study on Types of Data and Their Use

Slovenski jezik / Slovene Linguistic Studies 14/2022. 117–151.

DOI: https://doi.org/10.3986/sjsls.14.1.05

ISSN tiskane izdaje: 1408-2616, ISSN spletne izdaje: 1581-127

https://ojs.zrc-sazu.si/sjsls

Slovenski jezik – Slovene Linguistic Studies 14 (2022): 117–151

Domen Krvina (ORCID: 0000-0002-2276-1156)
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Slovenija
DOI: https://doi.org/10.3986/sjsls.14.1.05

The Growing Dictionary of the Slovenian Language (2014-) and Slovenian Neologisms: Study on Types of Data and Their Use

The article aims at presenting the methods of detection of Slovenian neologisms, used in the making of the *Growing Dictionary of the Slovenian Language*, accessible at the Fran portal <https://fran.si/>, which integrates various dictionaries into a single whole, form 2014 onwards. In the first year of compiling and for the following few years, the main source of the candidates was corpus Gigafida 1.0, built in 2013. Due to the corpus not being updated regularly (and unavailability of other appropriate sources), users' suggestions have taken over the main role. Users submit suggestions directly on the Fran portal. The corpus Gigafida and other (Janes, SlWaC) are still used for checking users' suggestions. Due to a high number of such suggestions and a growing demand for new lexical descriptions, their importance cannot be overlooked. The neologisms collected in the dictionary exhibit a number of characteristics, a brief overview of which is provided at the end of the study.

Keywords: Neologisms, Slovene, Growing Dictionary of the Slovenian Language, Data Detection, Corpora, Users' Propositions, Overview of Neologisms' Characteristics

Prispevek predstavlja metode zaznavanja slovenskih neologizmov, uporabljene pri izdelavi *Sprotnega slovarja slovenskega jezika*, ki je od leta 2014 dostopen slovarskem portalu Fran <https://fran.si/>. Ta združuje različne slovarje v eno celoto. V prvem letu nastajanja slovarja in nekaj naslednjih je bil glavni vir kandidatov za neologizme korpus Gigafida (zaključen leta 2013). Ker se ni redno posodabljal, drugi primerni viri pa tudi niso bili na voljo, so glavno vlogo prevzeli predlogi uporabnikov. Ti lahko svoje predloge oddajajo neposredno na portalu Fran. Korpusi Gigafida in drugi (Janes, SlWaC) ohranjajo vlogo gradiva za preverjanje

uporabniških predlogov. Zaradi velikega števila tovrstnih predlogov in velikega povpraševanja po novih leksikalnih opisih njihovega pomena ne le da ni mogoče zanemariti –postali so temelj opisa novejšega besedja. Kratek pregled njegovih temeljnih značilnosti je podan na koncu prispevka.

Ključne besede: novejše besedje, slovenščina, Sprotni slovar slovenskega jezika, gradivna zaznava, korpusi, predlogi uporabnikov, pregled značilnosti novejšega besedja

1 Background: transformation of Slovenian lexicography, the portal Fran and the rise of new type of dictionary in 2014

Neologisms constantly appear in language: they reflect developments in lifestyles, environment, perceptions of the world (ten Hacken 2020). In Slovene, the new lexis for the period 1991-2009 was comprehensively treated in the monograph *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri)* (Gložančev et al. 2009), mainly from a lexicological point of view, and lexicographically in the *Dictionary of New Slovenian Words* (2012). The neologisms presented in the dictionary spanned from 1991 to 2012 as the wordlist was compiled using the Nova beseda corpus in relation to the wordlist of the only (systematically compiled by a team of authors adhering to unified principles) monolingual general explanatory dictionary at the time – SSKJ: *Dictionary of the Slovenian Standard Language* (1970–1991).

In the following years Slovenian lexicography, after what could be called a preparatory decade, experienced some major shifts in its course, not unlike those that took place in English lexicography at the time of the COBUILD project (Sinclair et al. 1987), more than a decade before. Firstly, the corpus Gigafida 1.0, the first Slovene reference corpora to be fully equipped with formal POS tagging and at the same time accessible to the general public, built within the project *Sporazumevanje v slovenskem jeziku* <http://www.slovenscina.eu/>, was compiled in 2013. Secondly, that same year, three authors published a dictionary conceptualization plan proposing to compile a new, mainly corpus-driven explanatory dictionary, planned in different phases: from the first, computer-driven phase, whose

results would be only partially revised and would be available immediately, to the final phase with fully revised entries on various levels that are marked as completed (Krek et al. 2013). Thirdly, the first edition of SSKJ was updated and partially revised into SSKJ^2: *Dictionary of the Slovenian Standard Language, 2nd Edition* using the data from the corpus Gigafida 1.0.

These events set the stage for the following developments in the late 2014 and early 2015:

1. the emergence of the dictionary portal Fran <https://fran.si/> at the ZRC SAZU, Fran Ramovš Institute of the Slovenian Language;
2. the creation of the *Growing Dictionary of the Slovenian Language* and the publication of the first-year batch of entries;
3. the making of dictionary conceptualization plan for a completely new, corpus-based dictionary eSSKJ: *Dictionary of the Slovenian Standard Language, 3rd Edition*, which saw the publication of its first entries in 2016.

The main role of the portal Fran in 2014 was to bring together existing dictionaries and integrate them into a user-friendly and user-responsive website – by ensuring their transition into e-form by linking the data from various sources that are searchable through a single search engine (and results displayed from all the different sources all at once). The portal supports user-responsive interface. It enables general and highly advanced, targeted searches. Even when a dictionary is singled out by the user, the search is always performed against the entire background database – these results are shown separately from the main search in the navigation panel; see Figure 1 (Ahačič et al. 2015, Perdih 2018, 2020). The other important function of the portal was to serve as a platform on which completed batches of entries in new type of e-dictionaries could be published regularly, alongside with some (minor) changes to those new dictionaries on the level of microstructure, if necessary. These new-type dictionaries would be called *rastoči slovarji* ('growing' dictionaries).

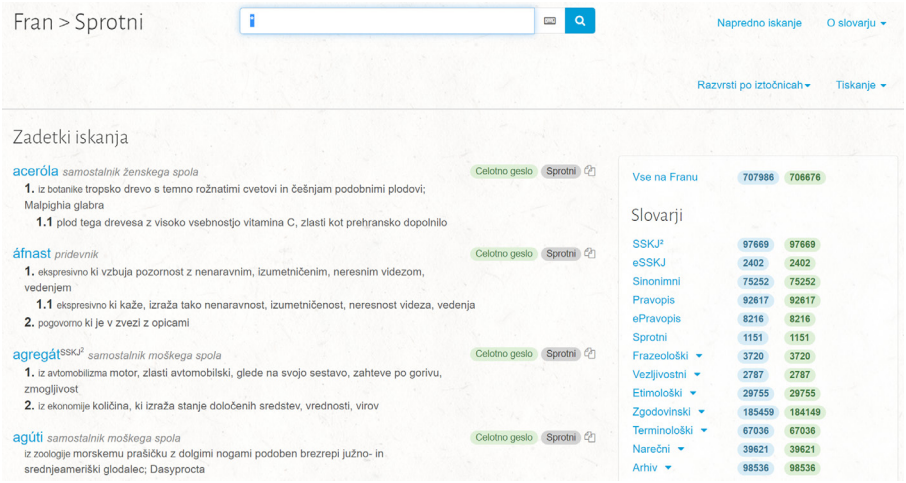

FIGURE 1: Portal Fran (*Growing Dictionary of the Slovenian Language*)

In October 2015, the portal adopted a policy of encouraging users to suggest 'missing' words and meanings as well as equivalents of loanwords as candidates for lexical description (FIGURES 2 and 3). Especially in the case of Slovenian equivalents of loanwords, Slovenian word-formation strategies (such as *sup*: *stojeska* a board for 'standing paddling', *plovček* 'sailing', 'rowing') would play a pivotal role. First seen as a part of user inclusion policy, this type of encouragement quickly turned out to be an extremely important source for propositions of neologisms, stemming directly from users' observations and answering their demand. These could be called **'neologisms from the users' point of view'**.

FIGURE 2: Portal Fran: suggesting new ('missing') words

FIGURE 3: Portal Fran: suggesting (and voting for) equivalents of loanwords

2 THE *GROWING DICTIONARY OF THE SLOVENIAN LANGUAGE*, COLLECTING POTENTIAL NEOLOGISMS AND STATE-OF-ART OF THEIR SOURCES

The *Growing Dictionary of the Slovenian Language*, which is the central point of our study, was the first one of a new type of dictionaries in the portal Fran – hence its name. Designed from the beginning as a

web dictionary, the *Growing Dictionary of the Slovenian Language* was one of the first to make good use of the adaptable environment of the portal Fran. As it was created literally at users request (and catering to their needs), editors decided that all the data should be presented as transparently and user-friendly as possible: no abbreviations (commonly used in linguistics and easily recognizable for linguists, but not necessarily for most other dictionary users) were to be used, hints to the dictionary content and structure were to be given in small grey frames on the right, the full list of all the word forms would be accessible by a simple click (FIGURE 4).

FIGURE 4: Growing Dictionary of the Slovenian Language: interface layout

The experience accumulated in the first two years of compiling the *Growing Dictionary of the Slovenian Language* was positive. This made the decision for the subsequent 'growing' dictionaries (*ePravopis*: *Slovenian Normative Guide* (2014–), eSSKJ: *Dictionary of the Slovenian Standard Language, 3rd Edition* (2016–) and NESSJ – *New Etymological Dictionary of Slovenian Language* (2017–)) to follow the same direction easier. It should be noted, however, that due to being the first of

'growing' kind, the *Growing Dictionary of the Slovenian Language* started more or less as the 'dictionary on the fly': apart from some basic principles of compilation (see below), most of its compiling criteria, especially in first years of compiling, would be dynamic rather than static. That was also linked to the circumstances regarding the availability of appropriate (corpora) resources and their (scarce or missing) updates. Therefore, dictionary compilation itself (as well as its assessment this paper is aiming at) could be seen as a certain experiment, particularly in the years prior to 2018-2019. The dynamic nature also applies to the dictionary's definition of 'neologism', which has been inclusive rather than exclusive, but more or less based on the three complementary approaches (see chapter 2.2) in dynamically changing proportions.

2.1 The Growing Dictionary of the Slovenian Language: main features and source limitations

The intention of the *Growing Dictionary of the Slovenian Language* was to continue the course of detecting and describing neologisms the *Dictionary of New Slovenian Word*s had started. The latter had defined a neologism in a somewhat straightforward way: the words (if the word already existed, also meanings – but this was rarer) not present in the SSKJ: *Dictionary of the Slovenian Standard Language* (1970–1991), but appearing in one of the first Slovene corpora Nova beseda, would qualify as candidates for dictionary description. Their frequency was of lesser importance, though given the scope of the corpus Nova beseda, it would be rather low in most cases. 6,000 such neologisms were described as dictionary entries (some contained several multi-world units), published in 2012.

For the present Slovenian state-of-art, it is important to note that there are no corpora of new Slovenian texts that are regularly updated. In late 2021, a project SLED (Spremljevalni korpus in spremljajoči podatkovni viri – SLED (ijs.si)), aimed at tracking neologisms, was announced – including a specialised corpus. However, its first version will not be available until late 2022. There are other specialised corpora of social media texts (Twitter, forums, blogs), such as Janes, built within

the same-name project in 2014-2018 (<http://nl.ijs.si/janes>; cf. Fišer et al. 2018), and corpora of web texts, such as SlWaC, built in 2011, and updated in 2014 (v. 2.1) using the web crawler SpiderLing (Erjavec, Ljubešić 2014). The main reference corpus Gigafida, published in 2013, saw a modest update of texts up to 2018 in 2019 (Gigafida 2.0). The past and especially present state-of-art, therefore, presented and still presents a substantial (but not insurmountable) obstacle to obtaining a completely corpus-driven candidate list of neologisms – which would contribute to its objectiveness.

As the goal of the *Growing Dictionary of the Slovenian Language* was to detect and analyse potential neologisms, it would make use of any appropriate resources at hand. At first, the Gigafida 1.0 corpus seemed sufficient (see below), but with no basic research of the newest lexis after 2013, its role could not be properly evaluated – at least not in a way it would remain the sole (major) source. With the number of users' propositions growing, the focus shifted to them, while Gigafida (and other corpora, as they became available) retained the role of sources used for checking such propositions. Due to the scarce (or non-existent) corpora updates, the – ever changing and expanding – web content came to the fore. With no widespread and readily available crawling tools for Slovenian (the one used in SlWac was the same as used for Czech), the dictionary would also not try to develop its own; partly because it would be time-consuming for a rather small dictionary outside the frames of general analysis of new lexis after 2013. Therefore, the option yet to be explored is a (semi-) automatic way of detecting neologisms in a process of comparing the content of all the available corpora against the expanding web content for the words not present in the corpora.

When decision was made in 2014 to start compiling the *Growing Dictionary of the Slovenian Language*, the first version (1.0) of the corpus Gigafida (2013) was the largest at hand and still relatively new. Therefore, it seemed feasible to retain the definition of neologism from the *Dictionary of New Slovenian Word*s: words (or, rarer, meanings) not present in the latter nor in the recently updated and partially revised SSKJ[2]: *Dictionary of the Slovenian Standard Language, 2nd Edition* but

appearing in Gigafida 1.0 would qualify as candidates for dictionary description. Taking into account the scope of a billion-word corpus Gigafida 1.0, additional limitations regarding the frequency and time of appearance were introduced: the frequency of corpus lemma should be below 1,000 (and above 500), the peak of occurrences in years 2009-2012 – the last three years covered in the corpus Gigafida 1.0. Thus, an additional frequency-time dimension (Slána 2017: 41) that corpus analysis allows for was provided – these could be called **'neologisms from the temporal point of view'**.

This procedure yielded some 500 candidates, out of which 224 (the majority of them with corpus frequency 700–500)[1] were chosen and then further processed all the way to the final dictionary entries. Among **various thematic fields** some stood out in particular – and would mostly continue to do so in the following years (cf. also Slána 2017: 42–43):[2]

a. **computing and technology**: *android*, *driftati* 'drive a car drifting', *inoks* 'stainless steel', *karbon* 'carbon used in bike frames', *kevlar* 'Kevlar', *multifunkcijski* 'multifunctional', *replikacija* 'replication', *večigralski* 'multi-player', *vtičnik* 'plugin';

b. **finances and economics**: *depozitarni* 'depository', *fiskalno* 'fiscally', *konsolidacija* 'consolidation', *prociklični* 'procyclic', *refinancirati* 'refinance', *volatilnost* 'volatility';

c. **medicine**: *artroskopija* 'arthroscopy', *epiduralni* 'epidural', *fibromialgija* 'fibromyalgia', *kandidiaza* 'candidiasis', *mirkocirkulacija* 'microcirculation', *obstruktiven* 'obstructive', *paradontalni* 'parodontal';[3]

[1] For further inclusion criteria see the chapter 2.3.

[2] Be aware that words listed above would qualify as neologisms in 2014, which may not be the case anymore. They will be probably sooner or later described also in general explanatory dictionaries, such as eSSKJ.

[3] In the fields of economics and especially medicine there is often a great deal of English-Slovene parallels both in form and meaning. For Russian-English comparison, see (Peredrienko and Istomina 2019).

d. **(healthy) food, leisure and lifestyle**: *falafel*, *gamber* 'prawn', *goji*, *makadamija* 'macadamia', *tahini*; *glamping*, *selfi* 'selfie', *selfness*, *skike*, *sup* 'SUP', *trimaran*.

Some users, accustomed to the *Dictionary of the Slovenian Standard Language*, which was both descriptive and normative, would still expect a dictionary to mark certain words for their 'foreign origin' – in the *Dictionary of New Slovenian Word*s this was done in cases the word retained the original written form from the donor language by applying the label cit. (lit. 'cited form'). Since the *Growing Dictionary of the Slovenian Language* was intended not to shy away from collecting many such words, it would not continue that tradition. The labels were to be used sparingly and 'loanword' would not automatically translate to 'colloquial', as this was often the case in earlier dictionaries, particularly in loanwords from German (the process of labelling was not straightforward; the fact of being borrowed, especially from German, would quite commonly point to a non-formal language layer, however).[4] Descriptiveness was the main goal and after two years users would embrace that fact – at least judging by their propositions, submitted (mainly) at the portal Fran.

In 2015, the total number of final dictionary entries was much lower (224 > 94),[5] although with some prominent additions, such as loanwords *bitcoin*, *karite* 'shea tree, butter', *overland*, *vloger*, *vlogerka* 'woman vlogger' etc.[6] This was mostly due to the fact that the initial supply of corpus candidates had been partially exhausted (note that until the modest above-mentioned update in 2019, the corpus remained virtually unchanged). Some uncertainties arose about how the potential neologisms with fewer than 500 occurrences should be treated: is this frequency still

[4] For further information on neologisms and purism in other European languages see (ten Hacken and Koliopoulou 2020), (Klosa-Kückelhaus and Wolfer 2020), (Marello 2020), (Panocová 2020).

[5] Partially also due to the decision taken at the Fran Ramovš Institut of the Slovenian Language to expand the smaller-scope 'growing' dictionaries by approximately 100 entries/units per year.

[6] The formation of feminine forms usually follows their neutral (grammatically 'male') counterparts rather quickly. For English-Slovene comparison and general information on gender of English loanwords in Slovene see (Stopar and Ilc 2019), (Sicherl 2019).

relevant in a corpus exceeding one billion tokens or not (provided the peak of occurrences occurs in final years still covered in corpus)? As it would turn out later when checking users' propositions, this frequency not only suffices – it is rather high: as the time passes, many potential neologisms may not be present in (non-updated) corpora at all. In 2015, the inflow of users' propositions was only gaining momentum to increase considerably in the following years and maintain the position of one of the most important methods of detecting neologisms.

2.2 The Growing Dictionary of the Slovenian Language: complementary approaches to collecting neologisms

As pointed out above, three main approaches have been developed and used complementarily, according to and in reaction to the available sources, in the *Growing Dictionary of the Slovenian Language* to collect potential neologisms:

a. **Straightforward data comparison** approach: the words (or meanings) not present in latest editions of explanatory dictionaries (if available, especially those of new words) but present in the latest version of corpora are very likely neologisms. This approach was used in the *Dictionary of New Slovenian Word*s and retained (especially for the first two-three years) in the *Growing Dictionary of the Slovenian Language*.

b. **Temporal corpus analysis** approach: the words with the peak of occurrences in the last years (data noise excluded) in each subsequent version of the corpus are potential neologisms for the time period covered in the corpus.

c. **Neologisms from the users' point of view**: words felt as 'new' by users themselves – according to their daily language use and observations.[7] Perhaps the most subjective of the three, but the subjectiveness is somewhat mitigated by the sheer number of

[7] Direct interaction with users via collecting and answering their questions concerning mainly everyday (and often not completely expected/systemic) language use is also the mainstay of Fran Ramovš Institute of the Slovenian Language Language Counselling.

such propositions coming from various users interested in various thematic fields.

The *Growing Dictionary of the Slovenian Language* first combined the approaches described above in the points a) and b). It found itself at a certain crossroads in the year 2015 – after the publication of the first yearly batch of entries. The upper half of words not present in available dictionaries but present in the corpus Gigafida 1.0 with frequency 1,000-500 and the peak of occurrences in the years 2009-2012 had been exhausted. Given the fact that the corpus Gigafida 1.0 had not received any update since 2013, 2015 was absolutely the last year in which 2009-2012 as a peak of occurrences seemed convincing enough for the temporal criteria (point b above) to be still applicable. Their typical (extremes at both ends are not taken into account) frequency plummeted from over 500 to 300. Fewer than 100 such words were processed all the way to the final dictionary entries – and it would be the last time corpus-only candidates made the vast majority of the final entries; see the line 'GF 1.0 (n/~ 500 initial)' in Figure 5. It became clear that new ways of collecting potential neologisms were to be actively sought out.

As mentioned, encouraging users to suggest 'missing' words and meanings (and equivalents of loanwords) as candidates for lexical description was first seen as a part of user inclusion policy – at the time no one could predict what an important source of collecting potential neologisms it would become. It should be noted that faced with the entire portal Fran content – from present-day to historical as well as terminological dictionaries in a unified electronic form – users had a powerful tool to compare entries which could serve as a kind of checkpoint: anything felt as 'new', but already described in one of the dictionaries or other manuals at portal Fran, would not qualify as such. Anything non-present anywhere at the portal Fran, however, identified as new – and, as it could eventually turn out, not present even in the latest (2.0), let alone the first (1.0) version of the corpus Gigafida – would have a high qualification as a potential 'new word' (neologism).

In 2015, however, user's propositions were few (7 were submitted)[8] and available only late in the year, and a number of other sources were selected in search of potential neologisms:

1. regular mail, telephone – usually alongside a linguistic question, answered by one of the researchers at the Institute;
2. the formalised way of answering such questions: Institute's Language Counselling site <https://svetovalnica.zrc-sazu.si/>, which is also integrated into the portal Fran;
3. systematic reading of new, mainly web texts of different genres which is done by students at Faculty of Arts in Ljubljana within their seminar work;
4. targeted reading of latest (news) web texts by paying special attention mainly to the fields which stood out in the first-year batch of entries (computing and technology, finances and economics, medicine, food, leisure and lifestyle); this is often done alongside the work on material for other growing dictionaries (eSSKJ, *ePravopis*);
5. external factors, such as projects which certain researchers from the Institute have taken part or interest in – e. g. *Janes*, alongside its proceedings.

These searches yielded some 20 candidates. As this was only a testing phase, they would not be processed further. The comparison with larger number of users' propositions was needed to better evaluate their position. These propositions came before long: 2016 saw an enormous increase in users' propositions submitted at the portal Fran (7 > 180).

Figure 5 shows how the proportions of neologism candidates from the approaches a)–c) have changed over time: from the domination of the straightforward data comparison along with temporal corpus analysis in 2014-2015 (the line 'GF 1.0 (n/~ 500 initial)' and the line 'published (entries)' as well as the line 'sum of the candidates' all follow the same curve) to the steep increase of the role of neologisms from the users' point of view (with temporal corpus analysis, when

[8] Among them was *sebek*, Slovene equivalent to *selfie* (2015), which would eventually make it to the final entries in 2018.

applicable, remaining an important part of entry processing) from 2016 onwards. While the line 'sum of the candidates' represents the sum of candidates from the initial approach a) + b) plus the all the candidates from the approach c) and other sources, the line 'sum of the propositions' unites only the latter: users' propositions + other sources, listed in above points 1-5. The content united under this line is shown in detail in FIGURE 6.

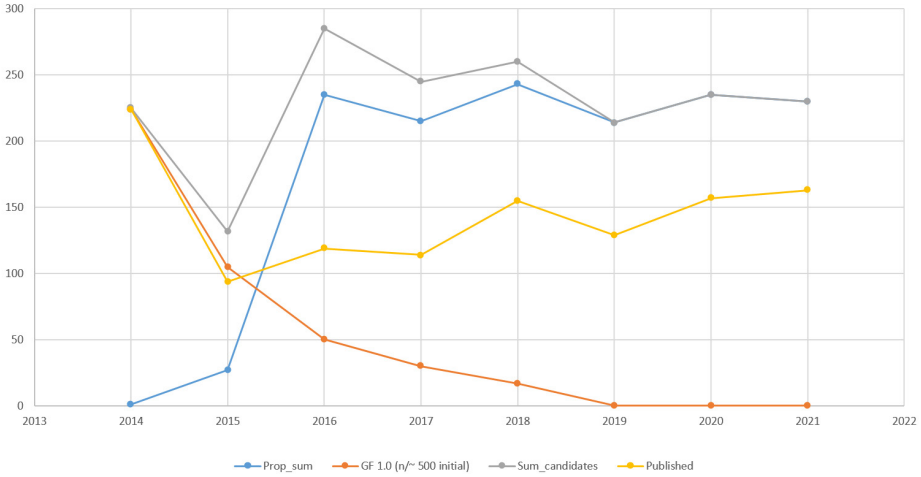

FIGURE 5: Data acquisition vs final entries

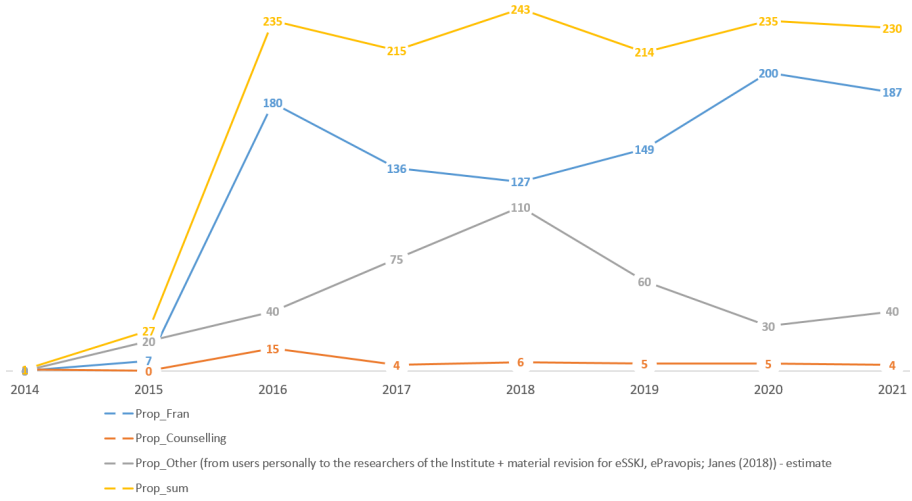

FIGURE 6: Proposition types

After 2015, a substantial number of propositions came from the sources listed in the above points 1-5, especially in the years 2016-2019. And an even larger number of users' propositions enabled their proper comparison with propositions from other sources (above points 1-5), which could not have been done in 2015. One of the other sources was the Language Counselling site, shown separately in Figure 6; most propositions were obtained either directly or indirectly from the questions related to either lexicology <https://svetovalnica.zrc-sazu.si/category/17/leksika>, <https://svetovalnica.zrc-sazu.si/category/19/pomenoslovje>, lexicography or word formation <https://svetovalnica.zrc-sazu.si/category/250/leksikografija>; <https://svetovalnica.zrc-sazu.si/category/25/besedotvorje>. The majority of propositions, however, stemmed from the process of compiling the dictionary eSSKJ and partially the normative guide *ePravopis*.

The 2018 was somewhat exceptional – the number of propositions from other sources, which had always been lower than those submitted at Fran by users, converged with the latter. This was mainly due to some researchers taking part or interest in the project *Janes* and its final proceedings, which was concluded in 2018. The project *Janes*, especially the corpus of social media (Twitter, forums, blogs) posts (<http://nl.ijs.si/janes/o-projektu/korpus-janes/>), contributed substantially to the content of the *Growing Dictionary of the Slovenian Language* – and not only in the 2018. Due to its specialized nature, this corpus cannot substitute the Gigafida corpus as an important tool in processing candidates (see the following chapter). However, together with web texts, the corpus proved very useful – particularly when the proposed candidates are nearly (frequency ≤ 8) or fully absent from the corpus Gigafida.

The combined use of three approaches (which also applies, to a certain degree, to the above point 3, done by students, and especially to point 4, with linguists taking role similar to that of general language users but with clear goal in mind) certainly allows for a greater degree of flexibility. The listed approaches are complementary – they help alleviate limitations that would arise when sticking disproportionally to only one of them (say, only corpus data without taking into account

user's observations or taking latter for granted without checking them thoroughly in corpora and other available sources). Thus, it makes sense that all of them should be used not only in collecting potential neologism candidates but also when processing them in the preparatory phase and then, if they pass the initial test, all the way to the final dictionary entries.

2.3 The Growing Dictionary of the Slovenian Language: dictionary inclusion criteria

What criteria must or should a neologism candidate fulfill to be included in the *Growing Dictionary of the Slovenian Language*? Reliance on corpora data alone was good enough only in 2014, when corpus Gigafida 1.0 was still relatively new – which allowed the frequency below 1000 and above 500 alongside the requirement for the peak occurrences in the 2009-2012 to function fair enough. After 2015 – a transient and in regards of inclusion criteria somewhat 'unsure' year (which resulted in the lowest number of entries published ever) –, 2016 saw a rise of number of users' propositions beyond expectations. The number of propositions from other sources (see the points 1-5 in the chapter 2.2) was substantial as well.

This required a careful consideration which neologism should be included immediately and which one should be put aside for possible inclusion later on. One could argue a big number of users' propositions alone is enough to lessen their subjectivity. Be it as it may, a decision was made they should, without any exceptions, undergo a process of verification in all the available corpora (not only Gigafida 1.0) and, if search yielded no results, also beyond corpora in web texts. From 2016-2017, web material and/or the corpus of web texts slWaC as well as corpus of academic texts KAS (in cases of determinologization), and from 2018 onwards also the corpus Janes, started being used much more frequently than before. The use of neologism candidates, along with frequency ≥ 10, in either of the listed corpora was preferred. However, should a candidate not be present in any of them, web texts still represented a sufficient last resort – although processing the data can hardly be as orderly as it is when doing it using corpora.

The non-included propositions were usually those not present in any of the corpora Gigafida, Janes or slWaC and at the same time barely present (or even absent) in the web texts. Meanwhile, the absence from the corpora alone – especially from Gigafida (1.0) and from the 2017 onwards – did not prevent the inclusion.

Non-included propositions are stored in the database, and they undergo a yearly check – when their presence becomes noticeable in various sources (at least in web texts), their inclusion can be reconsidered. When certain candidate is included, word formation also comes into play in the search for potential neologisms pertaining to parts of speech different from that of the proposed candidate – this is particularly true in Slovene, as well as other Slavic languages, which are known for their rich word formation. All word-formation candidates are subjected to the checking procedure described above; they are counted among 'other' propositions.

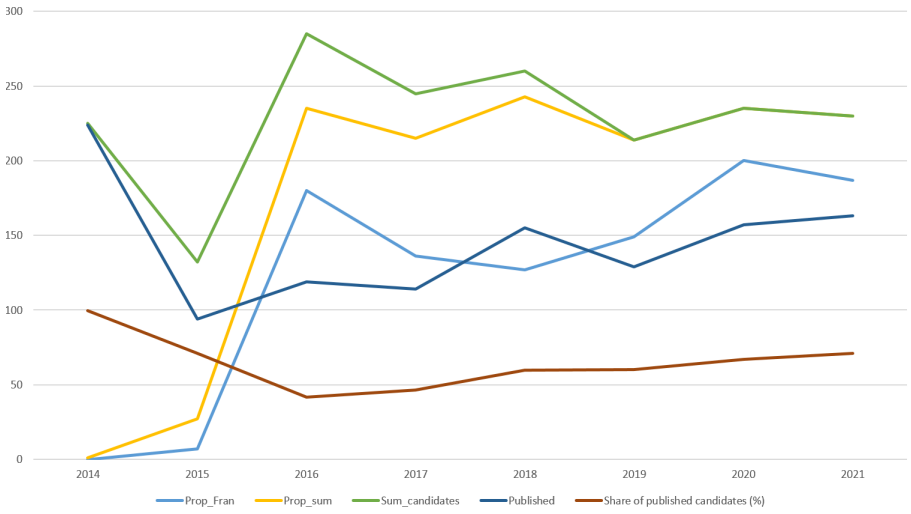

FIGURE 7: Neologism candidates vs published entries

As FIGURE 7 shows, from the total of all the candidates from all the sources (represented by the line Sum_candidates) – with exception of the first year when the corpus candidates were only available

– in average roughly about a half made it to the final entries each year. From the 2019, the line Prop_sum equals all candidates as the candidates from the initial Fran-Gigafida 1.0 alignment lost most of their initial relevance and stopped being used as source – the typical frequency of the lemmas in the corpus Gigafida 1.0 being also published dictionary entries was reduced from initial 500 in the 2014 to 40 in 2017. That, alongside a more streamlined process of checking users' and 'other' propositions (Prop_sum), caused the ratio between all the candidates and the published entries to begin a moderate, but steady rise towards 70%. For a process of checking the propositions to retain its relevance, a new specialized corpus, such as the announced SLED, is highly desired. The following chapter, which will serve as a kind of discussion entry point, will reveal the inefficiency of non-updated or scarcely updated reference corpora – such as it is the case in Slovenian – as the main source for neologism candidates not long after their compilation. Due to the sources being limited to the Slovenian corpora Gigafida 1.0 and 2.0, the results obtained apply only in regard to them, and thus cannot be generalized without taking into account the specific Slovenian situation described in the opening chapters. Further research and comparison of data from various languages may very well lead to different conclusions.

3 The feasibility of using (only reference) corpora for acquiring and/or processing the data

In regard to the corpus Gigafida 1.0, figures 8 and 9 show the absence or near absence – frequency ≤ 8, which makes appropriate processing of an entry relying solely on the limited corpus material very much inconvenient, impractical, if not outright undoable – of each year's (2014-2021) entries of the *Growing Dictionary of the Slovenian Language*; from 2019 also in the updated version of the corpus (2.0: <https://viri.cjvt.si/gigafida/System/About>). It has to be kept in mind, though, that due to the corpus Gigafida not being updated until 2019, most of the dictionary entries from the 2016 onwards stemmed from users' propositions. No large-scale analysis of the corpus Gigafida itself in terms of potential neologisms has been done. As shown in Figure 9, the update was rather modest

– not making a notable difference, at least as far as neologisms, described in the *Growing Dictionary of the Slovenian Language*, are concerned. There was some shift[9] from the complete (2019: 21 (1.0) > 11 (2.0); 2020: 50 > 44, 2021: 71 > 42) absence in 1.0 to near (2019: 46 (1.0) < 54 (2.0); 2020: 38 < 39, 36 < 50) absence in 2.0 – but the change was not substantial. Due to the initial Fran-Gigafida 1.0 alignment input, the absence or near absence from the corpus was nonexistent or negligible at first, but started gaining momentum with users' propositions and steady work on the material for both eSSKJ and *ePravopis* from 2016 onwards ('other' propositions). Starting with 2018, the sum of (nearly) absent entries represented at least a half of each year's entries, reaching up to 56-66% in 2020-2021 (both values show the impact of new corona lexis). If such trends continue, one could even argue that (near) absence from the corpus Gigafida 1.0 should become one of the criteria for inclusion of neologism candidates into dictionary entries, discussed in the chapter 2.3. Not something the *Growing Dictionary of the Slovenian Language* would seriously consider, of course.

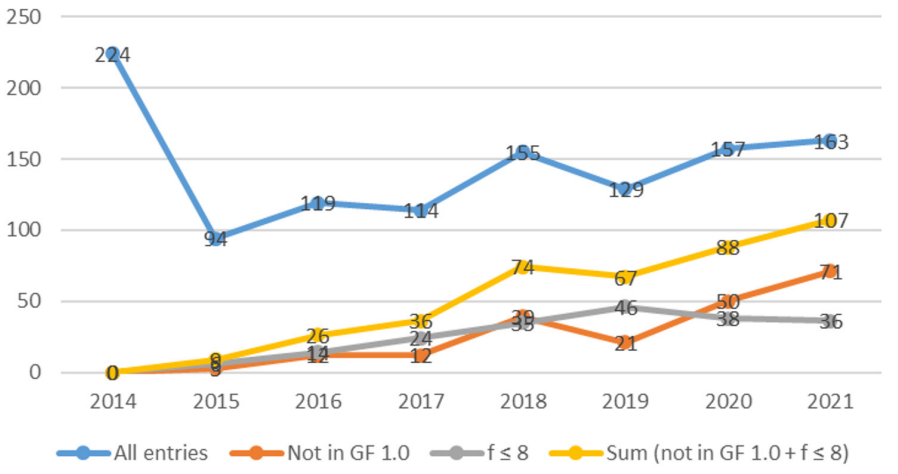

FIGURE 8: Presence of entries in corpus Gigafida 1.0

[9] With the corpus update there was also a shift in years of peak occurrences (2009-2012 → 2015-2018) to look for when processing neologism candidates.

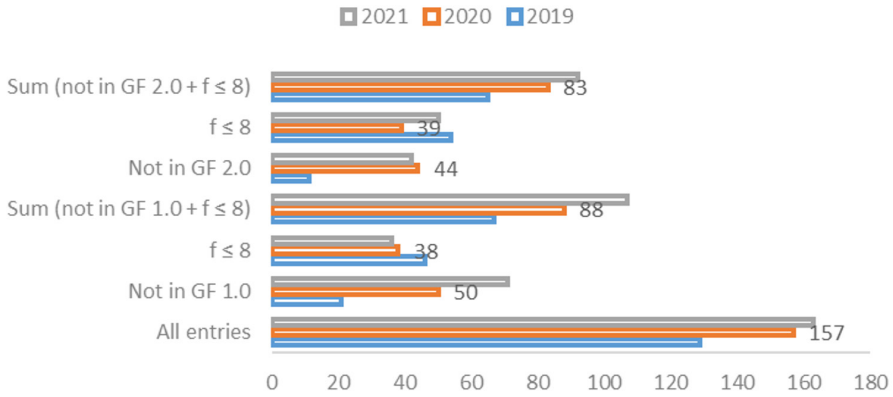

FIGURE 9: Presence of entries in corpus Gigafida 1.0 vs 2.0

The (nearly) absent entries are mainly, but not exclusively, loanwords (sometimes also in form of calques) and new derivatives, cf. (ten Hacken 2020), for example:

(2016) *helikopterski starši* 'helicopter parents',[10] *plačilomat* 'self-service payment machine', *vejpanje* 'vaping', *vejper* 'vaper'; *antropocen* 'Athropocene', *brezpilotnik* 'pilotless plane', *dismorfofobija* 'BDD', *camu camu*, *mangostin*, *zagonsko podjetje* 'start-up';

(2017) *emodži* 'emoji', *hipsterka* 'hipster woman', *ključnik* 'hashtag', *kriptovaluta* 'cryptocurrency', *pajkanje* 'web crawling', *zipline*; *antifeministka* 'antifeminist woman', *beachvolley*, *čustvenček* 'emoji', *hashtag*, *memorizirati* 'memorise', *netiketa* 'netiquette', *skrolati* 'to scroll', *smejko* 'emoji :-)', *tvitniti* 'to twit', *vitaminoza* 'vitaminosis';

(2018) *bestička* 'best she friend', *coworking*, *fixie*, *geolov* 'geocaching', *hejterka* 'hater woman', *influencer*, *influencerka*, *kretalec* 'user of sign language', *retvit*, *retvitati/-niti* 'to retwit/(pf)', *sebek* 'selfie', *selfiestick*, *supati* 'to sail on SUP', *tekstanje* 'texting', *vlogati* 'to publish on vlog'; *backpackerka* 'backpacker woman', *bestič* 'best friend', *chefinja* 'she chef', *klikanost* 'number of web clicks within an interval', *mikroplastika*

[10] Multi-word units are also included – one or more of them are listed under one of the components representing an entry. For further information on typology and treatment of multi-word lexical units in general monolingual explanatory Slavic dictionaries see (Perdih and Ledinek 2019).

'microplastic', *snorkljati* 'to snorkel', *sovrtičkar* 'kindergarten peer', *streamati* 'to stream', *trolati* 'to troll', *vstavljanka* 'toy with insertable parts', *youtubati* 'to publish on YouTube';

(2019) *časosled* 'timeline', *dojenčkati* 'to care for one's own baby', *fejmič* 'famous person', *hejtanje* 'hating', *jajcemat* 'self-service egg machine', *mikrozelenjava* 'microgreen', *odslediti* 'to unfollow', *prokrastinirati* 'to procrastinate', *risoroman* 'graphic novel', *spletinar* 'webinar', *vejpati* 'to vape'; *antidementiv* 'anti-dementia', *gentrificirati* 'to gentrify', *hendlanje* 'handling', *hrčkar* 'hoarder', *hrčkati* 'to hoard', *izsočiti* 'to extract juice', *jogistka* 'yogi woman', *kontrolfrik* 'control freak', *koruptibilen* 'corruptible', *nadkul* 'very cool', *napsihirati* 'to depress (pf)', *polajkati* 'to like on web (pf)', *polinkati* 'to link (pf)', *predtestirati* 'pretest', *rimoklepač* 'rapper', *shendlati* 'to manage', *takitos* 'taquito', *webinar*;

(2020) *alfakoronavirus*, *antikoronski* 'anti-corona', *brain freeze*, *halving*, *korona(čas, -humor, -kriza, -paket, -panika, ...)* 'corona-(time, humour, crisis, package, panic)', *plavajoča licenca* 'floating licence', *megapaket* 'mega-package', *odločbodajalec* 'decree-issuer', *ničti pacient* 'patient zero', *po(st)koronski* 'post-corona', *prekuževanje* 'infecting in order to build up immunity', *trikini*; *asimptomatično* 'asimptomatically', *bankster*, *brezsimptomen* 'asimptomatic', *brezstično* 'contactlessly', *hekaton* 'hackathon', *čredna imunost* 'herd immunity', *kohortna izolacija* 'cohort isolation', *megazakon* 'mega law-package', *novookužen* 'newly infected', *samoizolirati se* 'to impose self-quarantine';

(2021) *anticepilec*, *antivakser*, *anticepilski*, *antivakserski* (adj.) 'anti-vaxer', *antivakserka / proticepilka* 'woman anti-vaxer', *astroturfing*, *butaj* 'butai', *debelostnik* 'overweight person', *gerontocid*, *glinarjenje* 'working with clay', *hribarjenje* 'mountain hiking', *hudi* 'hoodie', *infodemija*, *instagramerka*, *kriptorudar* 'cryptocurrency miner', *kriptorudarjenje* 'cryptocurrency mining', *kriptorudarski* (adj.), *lockdown*, *nevrorazličnost* 'neurodivergence', *odrast* 'degrowth', *pokovidni/postkovidni* (adj.), *poobjavljati* 'retwit', *prebolelost* 'recovery from illness', predkoronski/predkovidni (adj.), *procepilec* 'provaxer', *procepilski* (adj.), *protiukrepni* (adj.) 'being/working against the measures', *razogljičiti* 'decarbonise', *senicid*, *tiktoker*, *tiktokerka* etc.

The (basic) meaning of many of these lexemes can be guessed even by a non-Slovenian speaker. They are certainly not 'exotic', yet none of them would make it to the entries relying only on the approaches a) + b) – it was users' propositions (approach c)) that proved crucial for their inclusion. The typical frequency of the entries present both in the dictionary and (as lemmas) in the corpus Gigafida 1.0, shown in Figure 10, further explains the diminishing role of the non-updated corpus as the reliable main source of neologism candidates without the aid provided by users' propositions. In 2014 and some following years the corpus played a very important role – when the substantial number of candidates at certain frequency was exhausted, the next effective number usually turned out to be at approximately half the previous frequency (500 > 300 > 160). Those were frequencies allowing for quite a comfortable analysis of data, typically using Sketch Engine, which would yield reliable collocations (common in the first two to three years, rare afterwards), distinct meanings etc.

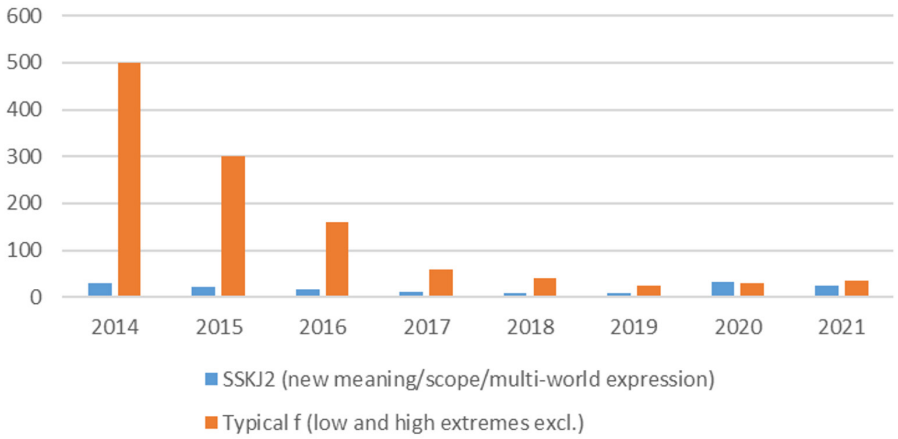

Figure 10: Typical frequency of entries also present as lemmas in corpus Gigafida 1.0

The change after the pivotal year 2016 was quite pronounced: the typical frequency of entries, also present as lemmas in the corpus Gigafida 1.0, was cut down to 60 (afterwards to even less).

The work turned from that resembling compilation of a general explanatory dictionary to 'trudging' through the material in search of examples which would reliably confirm the detected potential meanings.[11] With the ultimate goal of detecting (that was very much provided for by the users' propositions) and analysing potential neologisms at any cost, all the available corpora and (ever changing and expanding) web content came into play. One of the options yet to be explored is a potential (semi-)automatic way of detecting neologisms in a process of comparing the content of all the available corpora against expanding web content for the words not present in the corpora – with targeted search aimed mainly, but not exclusively, at the thematic fields standing out in the whole of entries of the *Growing Dictionary of the Slovenian Language* in the period 2014-2021. In this regard, the announced project (and a specialized corpus) SLED is also expected to prove extremely useful.

4 An overview of the basic characteristics of Slovenian neologisms in the period 2014–2021

This topic would require a study in a separate paper,[12] therefore only a brief introduction will be provided. Since 2012 the research of the neologisms has been limited to certain linguistic phenomena on various levels – such as word formation (Gložančev 2012), (Voršič 2015), (Štumberger 2015); semantics (Štumberger 2015a), (Zatorska 2016), (Fišer and Ljubešić 2018) – or varieties (Michelizza 2015), (Michelizza and Žagar-Karer 2018), (Zwitter Vitez and Fišer 2018). Most of them, except for those that study (colloquial) online language and are based mainly on the comparison of the corpus Janes with other corpora, are based on the material contained within the *Dictionary*

[11] Figure 10 also shows that the number of new meanings (or narrower/wider scope of a meaning or multi-word expressions) in lexis, already described in general explanatory dictionaries, such as SSKJ^2, is relatively low compared to the number of completely new words. Certain external events, such as present corona crisis, seem to augment that potential (see the years 2020-2021).

[12] A general overview of the topic, albeit with the focus on corona lexis (and its word-formation), is given in (Krvina 2021).

of New Slovenian Words. No comprehensive research on the newer material has been done yet, apart from a preliminary report, based on the entries of the *Growing Dictionary of the Slovenian Language*, as part of a lexicologically oriented project proposition. Therefore, for the time being only some basic insight into certain questions, which have arisen at different levels of linguistic description, can be provided.

Phonetical and morphological features:

a. Existence of variants, sometimes with different stylistic value, in both spelling and pronunciation (*dred/dread*, *selfi/selfie*, *snorkljati/šnorkljati*, *zero waste* [ˈziːrəu ˈuɛːist] : [ˈzeːrɔ ˈvɛːist]);

b. Types of nouns in which the accusative takes (also) animate forms, as is typical of many Slavic languages (*narediti selfi/selfij**a*** 'to make a selfie', *dobiti všeček/všečk**a*** 'to get a like');

   Types of words which also act as a type of adjective and form multi-word units in which the first element stays undeclined (*backpacker turist*, *korona kriza*) and their relation to their potential competing adjectival derivative (*backpacker**ski***, *koron**ski***).

Word-formation features:

a. Types of word derivatives from loanwords: verbs (*skrol**ati***, *sup**ati***, *tekst**ati***), their gerunds (*skrol**anje***, *sup**anje***, *tekst**anje***), animate agents (*sup**ar***, *sup**arka***); +/– existence of word-formation basis as an independent loanword (e. g. *skrol; *skrolati*, *skrolanje*);

b. Fully borrowed nouns (ending in *-er*) vs derivatives (with suffix *-ar*) from the verbs with the loanword as the basis (*vejper* : [sup-a-ti] *sup-ar*); occurrence of variants within the same word-formation basis (*youtuber* : [youtub-a-ti] *youtub-ar*);

c. Formation of verbs from loanwords as bases; the relationship between the suffix *-a-* and *-(iz)ira-* (*rent-**a**-ti*, *retvit-**a**-ti*, *stream-**a**-ti*, *vlog-**a**-ti* : *anonim-**izira**-ti*, *mentor-**ira**-ti*);

d. Types, frequency and ways of derivation of feminine forms from nouns in comparison to the neutral/masculine form

(*backpacker**ka***, *best**ica***. *chef**inja***, *influencer**ka***); the ratio of the respective suffixes *-ka*, *-ica*, *-inja*;

Semantic features:

a. Types and frequency of thematic fields predominantly contributing to the neologisms, mostly loanwords (computing and technology, finances and economics, medicine, (healthy) food, leisure and lifestyle);

b. Types and frequency of motivation for semantic shifts in the already existing words, usually via metaphor/metonymy or by expanding/narrowing/swapping the area of their use (*ambasador* 'of a country' : 'of an activity', *dopeči* 'to bake' 'to the end' : 'using special procedure in the shop', *sledilec* 'person following the track' : 'following the ideas, ideology; internet follower', *sodelo* 'cooperation in general' : 'a special type of cooperation – co-working', *vplivnež* 'influential in general' : 'influential in social media; influencer';

c. Types and frequency of the synonyms, especially in the relationship loanwords vs derivatives from the non-loanwords (*hashtag* : *ključnik*, *influencer* : *vplivnež*, *selfie* : *sebek*);

5 Further Discussion and Conclusions

The analysis was concerned with types of data and their use in the *Growing Dictionary of the Slovenian Language* (2014-), especially in regard to collecting potential neologisms (often called 'neologism candidates') and processing them in available corpora material and also beyond – in web texts, particularly when corpora analysis produces no results. In collecting and processing potential neologisms three complementary approaches, used in the *Growing Dictionary of the Slovenian Language*, were presented:

(a) straightforward data comparison (the words not present in latest editions of explanatory dictionaries but present in the latest version of the corpora);

(b) temporal corpus analysis (the words with the peak of occurrences in the last years in each subsequent version of the corpus);

(c) neologisms from the users' point of view (propositions submitted by users at the dictionary portal Fran).

The study has shown that the inclusion of users (even when initially viewed as a part of user inclusion policy rather than a way to obtain meaningful data) is an important part of methodology. Users' propositions can draw attention to neologisms or other meaningful phenomena that could remain nearly or completely undetected relying solely on (especially non-updated) corpus data – even that of a reference corpus such as Gigafida. Corpus use is still indispensable, particularly in general data processing: it represents the most systematic and statistically reliable way of analysis. A methodology of an individual dictionary should define the role and share of the corpus data according to the intended goals – and it should be dynamic rather than static.

As the growing type of dictionary has become commonplace in Slovenian lexicography in the last couple of years,[13] users are increasingly included in the compiling process in one way or another: usually via suggesting new entries, additions or corrections (this is the type of propositions the portal Fran encourages), sometimes also in the editing process itself (for *Collocations Dictionary of Modern Slovene* <https://viri.cjvt.si/sopomenke/eng/community>, <https://viri.cjvt.si/kolokacije/eng/about#>). Our study – based on the data obtained from the *Growing Dictionary of the Slovenian Language*, which from 2016 onwards heavily relies on users' propositions, and non-/scarcely updated corpus Gigafida – suggests dictionaries of (predominantly) neologisms in particular should try to provide a steady inflow of user's propositions (preferably in standardized electronic form allowing for easy processability and trackability) in

[13] Apart from the *Growing Dictionary of the Slovenian Language* there are eSSKJ (2016-), *ePravopis* (2014-) as well as the *Collocations Dictionary of Modern Slovene* (2018-) and *Thesaurus of Modern Slovene* (2018-). All of them are corpus-based or corpus-driven and use semi- (mainly in form of word sketches) or fully automated corpus data processing.

about 2-3 years after the start of compiling. This is especially true, if regularly updated corpora are not available. Such propositions can include those obtained from advanced users, such as other researchers, especially those working on material of a general explanatory dictionary. The candidates will most likely be the ones on the fringes of general lexis,[14] low in frequency (≤ 50) and possibly with occurrences mostly in the last years covered by the reference corpus – thus relying on the corpus data alone might not be able to provide a sufficient result.

Our study has also shown that the role a corpus, especially if it is not regularly updated, could play in detecting and processing neologisms may be well dependant on its 'age' – it seems to be much more efficient in a period not exceeding 3 years since the completion of the work on the corpus. After that, the corpus data alone, if not regularly updated (and even then, since a major overhaul of corpus data is rarely possible; cf. the case of Gigafida 1.0 vs 2.0), becomes less dependable, at least according to our study. The combined strategy – such as uniting approaches (a)-(c) in the *Growing Dictionary of the Slovenian Language* from 2015 onwards – can often be the most effective solution for a satisfactory degree of responsiveness. It also enables the advantages of one approach to mitigate the shortcomings of another. If a new or if old, completely overhauled, appropriate corpus appears, its share in detection and processing of the candidates is expected to rise according to its content – more so in specialised neologism corpora, such as (the first for the Slovenian, that is) recently announced SLED, whose role in description of potential neologisms only further research will be able to evaluate. For the peak of candidates' occurrences, the last years covered in the corpus are always preferred (with data noise excluded).

[14] The newest terminological lexis is also regularly monitored and Slovenian equivalents of foreign terms are suggested/evaluated, especially within the framework of Terminological counselling, which Fran Ramovš Institute of Slovenian Language provides as well. Due to a large quantity of specialized lexis entering general lexis via the process of determinologization, the results of Terminological counselling activity often prove valuable for the *Growing Dictionary of the Slovenian Language.*

Apart from using the most appropriate available corpora, the scope of analysis should be widened by including web texts and taking into account all the available statistics. To further increase chances of detection of possible candidates, a strategy of targeted reading of both corpora and web texts seems viable. Searching parameters, enabling also a potential (semi-)automated search, should be defined in advance: thematic fields (those standing out in the first years of compiling may serve as suggestion, others are not excluded) in general, text types, time interval (recent years) and derivative of frequency of a candidate's occurrence within it etc.

To sum up: the responsiveness of a (contemporary) dictionary should rest on a wide array of available data and approaches and try to combine them into an effective single whole, mainly by allowing for a flexibility in ratio of the individual approaches' shares according to a situation at hand. This is especially true of the dictionaries of neologisms since even at the beginning of their compilation most available corpora with content wide enough are likely to be somewhat old. Thus, the corpora efficiency in detecting and – due to the low frequencies – also processing neologisms will diminish over time. The share of users' propositions should preferably remain fairly high throughout the length of the process. For this to be possible, a kind of reference point should be made and maintained – such as the well-rounded portal Fran, the site of the Society for Danish Language and Literature or the site of the *Wielki słownik języka polskiego*, to name only a few. Targeted reading, complemented by possible development and implementation of (semi-)automated search and extraction of potential neologisms candidates – using parameters set in advance according to the already processed data (in the database of existing entries) –, should also not be neglected. Alongside recent and future developments in automated targeted search – despite a number of drawbacks (Kerremans et al. 2012), (Slána 2017), (Waszink 2019) – it may well become one of the most important sources of candidates.

As for the research on the characteristics of the neologisms, for the Slovenian only a brief overview is given at the end of the study. Even the preliminary research, based on the entries in the database of the

Growing Dictionary of the Slovenian Language, has revealed a number of subjects on various linguistic levels (phonetics, word-formation and semantics in particular) that are worth further, comprehensive research. Some subjects are shared between (related) languages: in Slavic languages for instance types of nouns in which the accusative takes (also) animate forms; types of words which can be written as one word, in which case they can be interpreted as a compound, or separately, in which case a multi-word units arise. For more effective research, it is preferable for the (dictionary) database to be structured in a standard processable format (e. g. TEI) – if the (dictionary and corpora) databases of different languages are to be compared.

All the above considered, the study of types of data and their use in the *Growing Dictionary of the Slovenian Language* has proven to be a worthwhile subject of study both in regard to detecting, collecting, processing (checking in corpora and beyond) and describing neologisms and examining their characteristics on various levels of linguistic description. Further research should focus on generalising the findings tied to the data of the *Growing Dictionary of the Slovenian Language* presented in this paper by comparing them to other (especially Slavic) languages in a wider scope and in greater detail.

References

Dictionaries

Krvina, Domen (ed.). *Sprotni slovar slovenskega jezika 2014–* [*Growing dictionary of the Slovenian Language 2014–*]. Available at: https://www.fran.si/132/sprotni-sprotni-slovar-slovenskega-jezika.

Slovar novejšega besedja slovenskega jezika [*Dictionary of New Slovenian Words*]. 2013. Available at: https://www.fran.si/131/snb-slovar-novejsega-besedja.

Slovar slovenskega knjižnega jezika [*Dictionary of the Slovenian Standard Language*]. Available at: https://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika.

Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja [*Dictionary of the Slovenian Standard Language, 2nd Edition*]. 2014. Available at: https://www.fran.si/133/sskj2-slovar-slovenskega-knjiznega-jezika-2.

eSSKJ: Slovar slovenskega knjižnega jezika 2016– [eSSKJ: *Dictionary of the Slovenian Standard Language, 3rd Edition*]. Available at: https://www.fran.si/201/esskj-slovar-slovenskega-knjiznega-jezika.

ePravopis: Slovar slovenskega pravopisa 2014– [*ePravopis – Slovenian Normative Guide*]. Available at: https://www.fran.si/135/epravopis-slovenski-pravopis.

Furlan, M. (ed.). *Novi etimološki slovar slovenskega jezika 2017–* [*New Etymological Dictionary of Slovenian Language*]. Available at: https://www.fran.si/207/nessj-novi-etimoloski-slovar-slovenskega-jezika.

Kolokacije 1.0: Kolokacijski slovar sodobne slovenščine [*Collocations Dictionary of Slovene*]. Available at: https://viri.cjvt.si/kolokacije.

Sopomenke 1.0: Slovar sopomenk sodobne slovenščine [*Thesaurus of Modern Slovene*]. Available at: https://viri.cjvt.si/sopomenke.

Society for Danish Language and Literature. Available at: https://dsl.dk/.

Wielki słownik języka polskiego. Available at: https://www.wsjp.pl/.

Corpora

Gigafida 1.0. Available at: http://www.gigafida.net/.

Gigafida 2.0: Korpus pisne standardne slovenščine. Available at: viri.cjvt.si/gigafida.

Janes. Available at: https://www.clarin.si/noske/run.cgi/corp_info?corpname=janes.

slWaC. Available at: https://www.clarin.si/noske/run.cgi/corp_info?corpname=slwac.

KAS. Available at: https://www.clarin.si/noske/run.cgi/corp_info?corpname=kas.

Language Counselling at ZRC SAZU. Available at: https://svetovalnica.zrc-sazu.si/.

Sporazumevanje v slovenskem jeziku. Available at: http://www.slovenscina.eu/.

Other literature

Ahačič, Kozma, Ledinek, Nina, Perdih, Andrej. 2015. Fran: The next generation Slovenian dictionary portal. In: K Gajdošová, A. Žáková (ed.). *Natural language processing, corpus linguistics, lexicography: proceedings*. Eighth International Conference: Bratislava. 21–22.

Erjavec, Tomaž, Lubešić, Nikola. 2014. The slWaC 2.0 corpus of the Slovene web. In: T. Erjavec. J. Žganec Gros (ed.). *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba*. Ljubljana: Institut Jožef Stefan. 19–24.

Fišer, Darja (ur.) 2018. *Viri, orodja in metode za analizo spletne slovenščine*.

Ljubljana: Znanstvena založba Filozofske fakultete. DOI: https://doi.org/10.4312/9789610600701

Fišer, Darja, Ljubešić, Nikola. 2018. Tviti kot leksikografski vir za analizo pomenskih premikov v slovenščini. *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete. 198–226.

Gložančev, Alenka, Jakopin, Primož, Micheliza Mija, Uršič, Lučka, Žele, Andreja. 2009. *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri)*. Ljubljana: Založba ZRC, ZRC SAZU.

Gložančev, Alenka. 2012**.** Novejša slovenska leksika v luči obravnave samostalniških zloženk v Slovenskem pravopisu 2001. *Pravopisna stikanja: razprave o pravopisnih vprašanjih*. Ljubljana: Založba ZRC. 125–139.

ten Hacken, Pius. 2020. Norms, New Words, and Empirical Reality. *International Journal of Lexicography* 33/2. 135–149. DOI: https://doi.org/10.1093/ijl/ecaa005

ten Hacken, Pius, Koliopoulou, Maria. 2020. Dictionaries, Neologisms, and Linguistic Purism. *International Journal of Lexicography* 33/2. 127–134. DOI: https://doi.org/10.1093/ijl/ecaa011

Kerremans, D. Stegmayr S., and Schmid H-J. 2012. The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. ' In Allan, K., Robinson, J. (eds), *Current methods in historical semantics*. De Gruyter Mouton. 59–96.

Klosa-Kückelhaus Annette, Wolfer Sascha. 2020. Considerations on the Acceptance of German neologisms from the 1990s. *International Journal of Lexicography*, 33/2:150–167. DOI: https://doi.org/10.1093/ijl/ecz033

Krek, Simon, Kosem, Iztok, Gantar, Polona. 2013. *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*. Accessed on 1–20 January 2022. Avaliable at: http://www.sssj.si/.

Krvina, Domen 2021. Sprotni slovar slovenskega jezika, covid-19 in z njim povezano (novejše) besedje. In: S. Ristić, I. Lazić Konjik, N. Ivanović (ed.). *Lexicography and lexicology in the light of current issues*. Beograd: Serbian language institute of SASA.

Marello, Carla. 2020. New Words and New Forms of Linguistic Purism in the 21st Century: The Italian Debate. *International Journal of Lexicography* 33/2. 168–186. DOI: https://doi.org/10.1093/ijl/ecz034

Michelizza, Mija. 2015. *Spletna besedila in jezik na spletu. Primer blogov in Wikipedije v slovenščini*. Ljubljana: Založba ZRC, ZRC SAZU.

Michelizza, Mija, Žagar Karer, Mojca. 2018. Internetna leksika v slovenščini. *Jezikoslovni zapiski* 24/1. 79–92.

Panocová, Renáta. 2020. Attitudes towards Anglicisms in Contemporary

Standard Slovak. *International Journal of Lexicography* 33/2. 187–202. DOI: https://doi.org/10.1093/ijl/ecaa006

Perdih, Andrej. 2018. Dictionary portal Fran: current state and future developments. In B. Niševa (ed.). *Slovanská lexikografie počátkem 21. století: sborník příspěvků z mezinárodní konference. Vyd. 1*. Praha: Slovanský ústav AV ČR, 57–65.

Perdih, Andrej. 2020. Portal Fran: od začetkov do danes. *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 46/2. 997–1018.

Perdih Andrej, Ledinek, Nina. 2019. Multi-word Lexical Units in General Monolingual Explanatory Dictionaries of Slavic languages. *Slovenski jezik / Slovene Linguistic Studies* 12. 113–134. DOI: https://ojs.zrc-sazu.si/sjsls/article/view/7629

Peredrienko, Tat'jana, Istomina, Ekaterina. 2019. Lexical Parallels in the Academic Vocabulary of Russian and English. *Slavistična revija* 67/4. 605–614. Available at: https://srl.si/ojs/srl/article/view/2019-4-1-5.

Sicherl, Eva. 2019. Določitev spola anglizmov v slovenščini. *Slavistična revija* 67/2. 343–352. Avaliable at: https://srl.si/ojs/srl/article/view/2019-2-1-22.

Sinclair, John McHardy. (ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. Collins ELT.

Slána, Jakub. 2017. K (polo)automatické excerpci neologismů. *Jazykovědné aktuality* 54/3-4. 34–46. Jazykovědné sdružení České republiky.

Stopar, Andrej, Ilc, Gašper 2019. Stilistična (ne)zaznamovanost moških in ženskih poimenovalnih parov za poklice v angleščini in slovenščini. *Slavistična revija*, 67/2. 333–342. Avaliable at: https://srl.si/ojs/srl/article/view/2019-2-1-21.

Štumberger, Saška. 2015a. Besedotvorje novejše slovenske leksike: medponskoobrazilne zloženke. *Zbornik prispevkov s simpozija Škrabčevi dnevi 8 (2013)*. Nova Gorica: Založba Univerze. 155–163.

Štumberger, Saška. 2015b. Leksikološka opredelitev novejše leksike in terminološka raba v slovenskem jezikoslovju. *Slavistična revija* 63/2. 249–259. Avaliable at: https://srl.si/ojs/srl/article/view/COBISS_ID-57985122.

Voršič, Ines. 2015. Tvorjenke s pomenom nosilnika lastnosti v novejšem slovenskem besedju. *Slavia Centralis* 8/1. 119–134.

Waszink, Vivien. 2019. Using Neoloog to detect and describe neologisms in online dictionaries. *Abstracts_IDS*. Instituut voor de Nederlandse Taal.

Zatorska, Agnieszka. 2016. Czasowniki w nowszej leksyce słoweńskiej. *Rozprawy komisji językowej* 62. 229–239.

Zwitter Vitez Ana, Fišer, Darja. 2018**.** Govorne prvine v nestandardni spletni slovenščini. *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete. 254–272.

Received January 2022, accepted March 2022.
Prejeto januarja 2022, sprejeto marca 2022.

Acknowledgments

The publication of article was made possible by programme *Slovenski jezik v sinhronem in diahronem razvoju* (P6-0038 (A)), which is financially supported by the Slovenian Research agency.

The author would like to thank Mitja Trojar for language editing and advice on terminological issues.

Summary

The Growing Dictionary of the Slovenian Language (2014-) and Slovenian Neologisms: Study on Types of Data and Their Use

The article examines types of data and their use in the *Growing Dictionary of the Slovenian Language*, which is integrated into Fran, a well-established dictionary portal for dictionaries and other language resources by the ZRC SAZU Fran Ramovš Institute of the Slovenian Language. The data in question is mainly input in the process of analysing and selecting data for dictionary entries; the dictionary is a so-called growing dictionary, which means that new entries are published every year. Most entries relate to neologisms; less commonly, there are new meanings of existing words. In the first year of compiling the dictionary (2014) and for the following few years, it was possible to rely on the Gigafida 1.0 corpus, built in 2013 (updated to 2.0 in 2019), for entry candidates; subsequently, with the "ageing" of the corpus, the main role has been assumed by user suggestions. Users can submit suggestion directly on the Fran portal ("suggest a new word"), which, with its extensiveness, serves as an important point of reference: if users feel that something is new, and it cannot be found on the Fran portal and is not an archaism, it is most likely a neologism. User suggestions are reviewed in all available resources (the Gigafida, Janes, SlWaC, KAS corpora, the web); the minimal criterion for inclusion in the dictionary is an adequate occurrence in web texts that is diverse enough in terms of sources and temporally recent. "Other" suggestions for candidates originate in the lexicographic work on other growing dictionaries (especially eSSKJ), in a seminar that is part of lexicology and lexicography lectures at the University of Ljubljana, Faculty of Arts, and partly in the Institute's Language Counselling service; these are far less numerous than user suggestions. About 50% of the total number of

all suggestions are included in the dictionary every year (those not included are re-analysed the following year); in 2020–2021, this share rose to nearly 70% through the appearance of COVID-related words. One of the main highlights of this analysis is that user engagement during the compilation of the dictionary is extremely important. Dictionaries consisting mostly of neologisms (new words), in particular, cannot rely only on corpus materials in detecting potential candidates for inclusion; if the corpus in question is a general (reference) one, it is outdated fairly quickly when it comes to neologisms. With Gigafida 1.0, which, put in comparison with SSKJ2 and SNB, was a major starting source for the *Growing Dictionary of the Slovenian Language*, the share of yearly entries that do not appear (or hardly appear) in the corpus ($f \leq 8$) rose from 0% in 2014 all the way to 66% in 2021 (exceeding 50% since 2018). The update of the corpus (2.0) in 2019 has improved the situation to some extent (instead of total absence, there is $f \leq 8$ presence), but not dramatically. A corpus that is being made within the SLED project is expected to be significantly more useful, and our analysis shows this work is justified. In terms of future research following up on the analysis in this article, it seems sensible to generalise the findings relating to the *Growing Dictionary of the Slovenian Language* in comparison with similar analyses of (dictionaries of) new words, especially in other Slavic languages.

Sprotni slovar slovenskega jezika (2014–) in slovensko novejše besedje: analiza tipov podatkov in njihove uporabe

Prispevek obravnava tipe podatkov in njihovo uporabo v *Sprotnem slovarju slovenskega jezika*, ki je integriran v uveljavljen slovarsko-jezikovni portal Fran Inštituta za slovenski jezik Frana Ramovša ZRC SAZU. Gre zlasti za vhodne podatke v procesu analize in izbora podatkov za slovarske iztočnice; slovar je t. i. rastoči slovar, kar pomeni, da se nove iztočnice objavljajo vsako leto. Med iztočnicami prevladujejo neologizmi, v manjši meri novi pomeni že obstoječih besed. V prvem letu nastajanja (2014) in še nekaj naslednjih se je bilo mogoče pri kandidatih za iztočnice nasloniti na leta 2013 končani korpus Gigafida 1.0 (posodobitev 2.0 2019), pozneje pa so s »staranjem« korpusa glavno vlogo prevzeli predlogi uporabnikov. Uporabniki predloge oddajajo neposredno na portalu Fran (»predlagaj novo besedo«), ki s svojo obsežnostjo služi kot pomembna primerjalna točka: kar uporabniki čutijo kot novo, pa tega ni na portalu Fran in ni arhaizem, je precej verjetno neologizem. Uporabniški predlogi so pregledani v vseh virih, ki so na voljo (korpusi Gigafida, Janes, SlWaC, KAS, splet), pri čemer minimum za uslovarjenje predstavlja zadostna, po virih dovolj pestra in časovno novejša pojavnost v spletnih besedilih. »Drugi« predlogi za kandidate prihajajo iz slovarskega dela za preostale rastoče slovarje (zlasti eSSKJ), seminarja v okviru predavanj

iz leksikologije in leksikografije na FF UL, delno tudi iz inštitutske Jezikovne svetovalnice; po številu jih je precej manj kot predlogov uporabnikov, skupaj pa tvorijo vsoto vseh predlogov. Letno je v povprečju uslovarjenih okoli 50 % te vsote (neuslovarjeni predlogi so znova analizirani naslednje leto); v letih 2020–2021 se je ta delež povzpel proti 70 %, k čemur je prispeval pojav koronabesedja. Kot enega glavnih poudarkov analize lahko izpostavimo, da je angažiranje uporabnikov v procesu nastajanja slovarja izjemno pomembno. Zlasti slovarji pretežno neologizmov (»novejšega besedja«) se, posebej pri zaznavi potencialnih kandidatov za uslovarjenje, ne morejo naslanjati zgolj na korpusno gradivo; če gre za splošni (referenčni) korpus, ta z vidika neologizmov precej hitro zastari. V primeru Gigafide 1.0, ki je pri *Sprotnem slovarju slovenskega jezika* ob sopostavitvi s SSKJ[2] in SNB predstavljala pomemben izhodiščni vir, se je delež vsakoletnih iztočnic, ki v korpusu (skoraj) niso prisotne ($f \leq 8$), od 0 % leta 2014 povečal vse do 66 % leta 2021 (in presegal 50 % od leta 2018 dalje). Posodobitev korpusa (2.0) v letu 2019 je stanje nekoliko izboljšala (namesto polne neprisotnosti prisotnost pod $f \leq 8$), vendar ne izrazito. Precej večjo uporabnost je pričakovati od korpusa v okviru projekta SLED, na utemeljenost katerega kaže tudi naša analiza. Kar se tiče prihodnjih raziskav kot nadgradnje analize v tem prispevku, se zdi smiselno ugotovitve, vezane na *Sprotni slovar slovenskega jezika*, posplošiti ob primerjavi s podobnimi analizami (slovarjev) novejšega besedja, zlasti v drugih slovanskih jezikih.