Novel Slovenian COVID-19 vocabulary from the perspective of naming possibilities and word formation

Senja Pollak[1], Ines Voršič[2], Boris Kern[3], Matej Ulčar[4]

[1] Jožef Stefan Institute, Ljubljana, Slovenia

[2] Faculty of Education and Faculty of Arts, University of Maribor, Slovenia

[3] ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Ljubljana Slovenia and University of Nova Gorica, School of Humanities, Nova Gorica, Slovenia

[4] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

E-mail: senja.pollak@ijs.si, ines.vorsic@um.si, boris.kern@zrc-sazu.si, matej.ulcar@fri.uni-lj.si

Abstract

We analyze a sample of novel Slovenian vocabulary related to COVID-19, focusing on naming possibilities and word-formation processes. We grouped previous descriptions of COVID-19 vocabulary and extended the list with a semi-automated selection based on embedding-based keyword expansion. In terms of naming possibilities, the analysis shows that a large majority of COVID-19 lexemes were created through derivation, showing high productivity and language vitality in Slovenian, and that a smaller number of examples are set phrases and neosemantisms, as well as explicit borrowings, whereas calques were not a productive strategy. From the point of view of the word-formation system, it is mainly possible to distinguish infix compounds, ordinary derivatives, and compositions. The most productive substructure is the root morpheme korona, which produces most of the infix subordinate compounds, but also higher-order adjectival derivatives (e.g., koronski 'corona') and compositions (e.g., protikoronski 'anti-corona'). Otherwise, infix subordinate compounds turn out to be the most productive word-formational type. The most productive derivatives are adjectival and nominal derivatives with the suffix -ost, and these are also the ones that show the most frequently confirmed combinatorics of the suffix -en- + -ost.

Keywords: COVID-19; embeddings; naming possibilities; word formation; formant combinatorics

1. Introduction

The COVID-19 pandemic has fundamentally altered our reality—and with it our linguistic reality. In this article, we extract and analyze a sample of novel Slovenian vocabulary related to COVID-19, focusing on naming possibilities and word-formation processes. Expansion of the lexicon due to pandemics is not specific to COVID-19. As summarized by Gustilo et al. (2021), the words *epidemic* and *pandemic* are related to the seventeenth-century plague in Europe, and *quarantine* was first used in the fourteenth century to describe the forty-day period during which ships were in isolation before landing during the Black Death.

This changed linguistic reality has been characterized in particular by the penetration of a large number of terms into common usage that have thus been subjected to determinologization, the emergence of neologisms, occasionalisms, and neosemantisms that have been subjected to a process of accelerated Slovenization, and the imposition of particular lexical variants.

Identifying and analyzing new vocabulary is of high importance from several aspects. First, from the lexicographic perspective, adding novel lexis is of crucial importance, both in terms of the synchronic description of a language, as well as from the point of view of the historical character of a special pandemic era. The goal of this study is not only to identify novel vocabulary, but especially to understand different naming possibilities and word-formation processes, which are signs of language vitality and are highly interesting from a linguistic perspective.

Our work groups existing descriptions of COVID-19 vocabulary and applies natural language processing methods to extend the dataset by extracting candidates for novel Slovenian vocabulary related to COVID-19. Specifically, we trained a fastText embedding model on a dataset of COVID-19 news articles from the initial period of the COVID-19 pandemic, and then, using seed words related to COVID-19 and keyword expansion via embedding of nearest neighbors, we extended this initial set. The resulting material is used for manual analysis of COVID-19 keywords in terms of naming possibilities and word-formation processes.

This article is structured as follows. Section 2 presents the background, including naming possibilities and word-formation processes in Slovenian, followed by related work in Section 3. Next, we introduce the methodology, including the natural language processing approach used for lexicon extraction, in Section 4. In Section 5, we present the findings and analysis from the perspective of naming possibilities and word-formation processes. Finally, the conclusions and plans for further work are presented in Section 6.

2. Background: naming possibilities and word formation in Slovenian

2.1 Naming possibilities

A naming typology for Slovenian lexemes was proposed by Ada Vidovič Muha (2013: 23–25) in her work *Slovensko leksikalno pomenoslovje* (Slovenian Lexical Semantics). Below, we present a slightly adapted naming typology. When a new denotatum or a need for a new denotation arises in a language, we can 1) search for naming possibilities in the language itself or 2) borrow from a foreign language. When looking for possibilities in the language itself, the possibilities are 1.1) a simplex or a set

phrase (e.g., *roka* 'hand', *osebna izkaznica* 'ID'), 1.2) a derivative (e.g., *nalivnik* 'fountain pen'), or 1.3) a neosemantic term (e.g., *hrošč* 'bug'). In the case of borrowings from a foreign language, this borrowing may be 2.1) disguised as a calque (e.g., *kolidž* 'college', *strežnik* 'server'), or 2.2) expressed (*powerpoint* 'Powerpoint (presentation)', *halloween* 'Halloween').

Further, Vidovič Muha's (2013) typology of naming possibilities in the language itself is further divided into two groups: simplex words (e.g., *roka* 'hand') and non-simplex words, which are further divided into set phrases, derivatives, and neosemantisms. In the case of acquisition from foreign languages, she distinguishes between disguised borrowings with groups of a) denotational and b) semantic calques (from classical languages, from a lingua franca, and from other languages), and explicit borrowing, with a) non-adapted, b) semi-adapted, and c) systemic acquisition. In citation, the word retains its form in the source language; in semi-quotation, there is a partial adaptation to the recipient language (especially in inflexion); and, in systemic acquisition, the word is completely integrated into the formal system of the recipient language.

2.2 Word formation in Slovenian

As a branch of linguistics, word formation is used to analyze the vitality of a language's lexicon and to chart the course of linguistic development. Moreover, word formation facilitates the formation of new words at two levels: linguistically described, predictably formative, and transformative processes as well as systemically unpredictable word-formation patterns. Modern Slovenian word-formation theory includes derivation by suffixation, derivation by prefixation, and compounding, among the traditional word-formation processes. Current systemic word formation is briefly presented by Plemenitaš, Stramljič Breznik & Voršič (2020), who conclude that in Slovenian suffixation is the most productive word-formation pattern, with more than 300 suffixes used for creating nouns, adjectives, and verbs. The majority of suffixes are used for the formation of nouns, which can be masculine, feminine, or neuter. Adjectives can be formed with approximately 70 affixes. Verbal word-formation, on the other hand, uses only 15 suffixes. Prefixation, including foreign prefixes, uses 14 nominal, 4 adjectival, 20 nominal and adjectival, and around 40 verbal prefixes (Toporišič, 2000: 142–234).

The most productive word-formational process in Slovenian is nominal suffixation. Research shows that the majority of Slovenian words are still formed through nominal suffixation (Stramljič Breznik, 2005). Nouns can be derived from verbs, adjectives, and other nouns through suffixation or prefixation. Denominal derivation of nouns also includes derivation via prepositional phrases. Verbs can be derived from nouns, adjectives, interjections, other verbs, and prepositional phrases through suffixation or prefixation. Prefixation in verbs typically involves deverbal derivation

from bases containing prepositional phrases or from prepositional verbs. Adjectives can be derived from nouns, verbs, adverbs, and other adjectives mainly through suffixation. Prefixation in adjectives typically involves denominal derivation via prepositional phrases. Adverbs can be derived from nouns, adjectives, verbs, other adverbs, and prepositional phrases. Adverbial derivation mainly uses suffixation, yielding the semantic categories of place, time, manner, and quantity.

There are also word-formation patterns that are unpredictable from a formative and transformative point of view; that is, non-systemic formations cannot be assigned to a syntactic stem or be unequivocally morphemized into a word-formation stem and an affix, given the unpredictability of the number and the fact that affixation words can be different parts of speech. Non-systemic formations are also distinguished from systemic formations by their function. Their central purpose is not a naming necessity or a lexical gap, but a striving for originality and attractiveness. As such, they deliberately break the laws of word formation. Among the non-systemic word-formation processes, the following are recognized in current Slovenian: blends, back formations, abbreviations, bicapitalizations, and graphoderivatives; that is, formations enriched with graphic elements (Voršič, 2010).

3. Related work

3.1 Studies of COVID-19 vocabulary in Slovenian

The COVID-19 pandemic period has had a profound impact on our lives, and it has certainly had a linguistic impact—and Slovenian is of course no exception. There are a multitude of linguistics articles on this impact for individual languages, but in this article we limit ourselves to a brief summary of works analyzing current COVID-19 vocabulary in Slovenian.

The lexicographic portal Fran.si, published by the Fran Ramovš Institute of the Slovenian Language, launched the subpage Fran.si, COVID-19 Version (7.1) in early April 2020. It "brings together the most important new and previously published dictionary entries and language advice, provided by the Language Advising Service, in the context of the COVID-19 epidemic and the novel coronavirus. In addition to current vocabulary and orthographic and terminological notes, the material also contains a thematic overview of the history and etymology of pandemic-related vocabulary, as well as contagion-related terms from Slovenian dialects." The most relevant part of the material, which was also included in this research, is new vocabulary from *Sprotni slovar slovenskega jezika* (Growing Dictionary of the Slovenian Language, GDSL).

This is a dictionary that contains living, newer words not yet registered in dictionaries, and at the same time contains the latest, emerging meanings of words

already registered. "The core of the dictionary consists of words not yet registered in dictionaries and whose use has been confirmed by corpus material in recent years, supplemented by suggestions from language users. Because these suggestions are usually relatively up to date, the glossary also contains words that are not (yet) present in current (time-limited) corpora of Slovenian, but whose use has been registered in other (especially digital) sources" (Krvina, 2023). The vocabulary described can be categorized into two groups: a) the lexicon already established whose use increased significantly or which acquired new meanings during the pandemic, and b) the lexicon that was newly created upon the emergence of the novel coronavirus and the associated pandemic. Among the words in the first group, the word *korona* should be mentioned first. This word was already introduced in *Slovar slovenskega knjižnega jezika* (Dictionary of the Slovenian Standard Language 2, SSKJ 2) with the meanings "1. *music* 'a semicircle with a dot, indicating an indefinite prolongation of a note or a pause [i.e., *fermata*]', and 2. *astronomy* 'a layer of the Sun's atmosphere, which passes into interplanetary space'." While the word *koronavirus* 'coronavirus', which came into Slavic languages as an integral English borrowing (Będkowska-Kopczyk & Łaziński, 2020), is already recorded in SSKJ 2 with the meaning of 'virus of the family *Coronaviridae'*, the word *korona* has acquired a new meaning from a dictionary point of view and is explained in GDSL (*Sprotni slovar slovenskega jezika*) as 'a coronavirus, in particular the highly contagious SARS-CoV-2, or a disease characterized by inflammation of the upper respiratory tract, in the severe form a pneumonia, caused by this virus'. It is also used adjectivally to mean 'that which is related to this virus or to the economic, social, or health consequences of an epidemic of the disease caused by this virus' (as a synonym of the derivative *koronski* 'corona'), and it can be considered a converse derivative in a word-forming sense. Otherwise, *korona* is one of the most productive bases for neologisms (see also Stramljič Breznik, 2021). The lexeme *koronavirus* also spread in phrasal usage during the pandemic as *novi koronavirus* 'a coronavirus, in particular the highly contagious SARS-CoV-2, or a disease characterized by inflammation of the upper respiratory tract, in the severe form a pneumonia, caused by this virus [i.e., novel coronavirus]'. Otherwise, there are two more synonymous terms for the disease; namely, *SARS-CoV-2* (severe acute respiratory syndrome coronavirus 2) and *COVID-19.* Both are also listed in GDSL.

In addition to the formations mentioned above, GDSL also records other words that increased in frequency or acquired a new or at least broader meaning during the coronavirus outbreak. These include *bolezen* 'illness', in the phrases *pridružena bolezen* and *spremljajoča bolezen* 'comorbidity, usually a chronic disease, which is already present in the patient at the onset, the start of treatment for another disease'); *distanca* 'distance', in the phrase *socialna distanca* '1. weak or less intense interaction between the usually dominant group and other groups of people due to personal, social, economic differences [i.e., social distance]; 2. avoiding non-necessary physical contact with others in order to prevent spreading of the virus'; *govorec* 'one that represents an authority, organization, or individual by presenting its views and

decisions in public [i.e., spokesperson]'; *helikopterski* 'helicopter' in the phrase *helikopterski denar* 'money granted by a state or a community of states in particular for individuals or companies in times of adversity or emergency to promote economic growth'; *izolacija* 'isolation', in the phrase *kohortna izolacija* 'isolation in which several patients with the same pathogen are placed in the same room'; *obhajilo* 'communion', in the phrase *duhovno obhajilo* 'a union with Christ without ingesting the consecrated host by focusing on him and longing for him'; *omejevanje* 'limitation', in the phrase *omejevanje socialnih stikov* 'avoidance of non-essential physical contact with others to prevent the spread of infection'; *pacient* 'patient', in the phrases *številka ena* and *ničti pacient* 'whoever in a particular area or country is the first to contract a communicable, usually highly contagious disease [i.e., patient zero]'; and *testirati* 'to carry out a procedure to determine the presence of a disease-causing agent or of a particular substance, active ingredient [i.e., to test]'. Among the formations already recorded in SSKJ 2 is *samoizolacija* 'isolation, closing oneself off from others, usually of one's own accord [i.e., self-isolation]'. The same formation with the meaning 'quarantine in the face of very probable infection, which individuals spend at home by order or choose to do so themselves because of their responsibility towards others' (GDSL) is cited in word-formation analysis as a higher-stage derivative of the verb *samoizolirati se* 'to self-isolate'. This, interestingly enough, is a verbal infix compound not previously recorded in the dictionary.

Another study on the coronavirus pandemic vocabulary was performed by Stramljič Breznik (2021). In her study, she analyzed the development of newly created words and "tried to evaluate the importance of word-formative processes involved in their coinage" (Stramljič Breznik, 2021: 321). The analysis includes new vocabulary found on the subpage Fran.si, COVID-19 Version (7.1), as well as random searches for lexis from various websites and "less formal social media" (Stramljič Breznik, 2021: 321). In the article, the author focuses on the compounds with the constituent *korona* 'corona', which are particularly problematic in Slovenian from an orthographic point of view—there is a dilemma as to whether to write them as one word or as two (e.g., *koronavirus* or *korona virus* 'corona virus'). The material the analysis is based on shows that in 74% of cases writing them together is chosen, which is quite surprising considering that the general tendency in the language for this type of compound is to write them separately. This was undoubtedly influenced by the advice of the Language Advising Service, which advocated writing these as compounds (Weiss & Dobrovoljc, 2020), published as early as March 2020. In this article, special attention is paid to word-formation productivity of new COVID-19 vocabulary and analysis of COVID-19 occasionalisms. COVID-19 vocabulary contains a large number of expressive lexemes, which the author explains by the fact that the epidemiological situation has brought with it a series of mental, social, and economic hardships that have also had a linguistic impact.

Last but not least, Voršič (2022) focused on ad-hoc formations with an expressive association with the various social consequences of the pandemic. Ad-hoc formations

are words that are formed through the most current word-formation processes and reflect the creative flexibility of a language. The study moves away from neologisms formed due to lexical gaps, focusing instead on ad-hoc formations with an expressive association with the various social consequences of the coronavirus pandemic (Voršič, 2022: 265).

3.2 COVID-19 and word formation

The closest research to this article is a corpus study by Gustilo et al. (2021), who focused on meanings and word-formation processes of pandemic-related lexemes across English varieties, also leveraging a news corpus. They identified COVID-19 terms in the News on the Web (NOW) English corpus and classified them as compounds or blends. Specifically, they differentiated between revitalized compounds, blends, and new formations.

Word formation related to COVID-19 neologisms was also addressed in Asif et al. (2020), who analyzed neologisms related to COVID-19 from various text sources, including social media, where in terms of word-formation processes the neologisms most frequently corresponded to compounds, abbreviations, acronyms, and blends. Compounding, blending, and affixation were the most frequent word-formation processes identified in a study focusing on COVID-19 by Akut (2020), and Al Salman and Haider (2021) identified compounding, clipping, blending, acronyms, and other dual word-formation processes in their study of lexemes from various online sources.

3.3 Natural language processing and COVID-19 vocabulary

Natural language processing methods have been used previously for analysis of COVID-19 texts and related vocabulary. For example, Lei et al. (2021) studied the emergence of COVID-19 neologisms in Chinese, based on data from the Baidu Index. They followed the dynamics of the usage of various words for COVID-19, grouped into five categories (official, stigmatizing, English abbreviations, etc.), during the first six months of the pandemic. Wang and Huang (2021) compared the usage of terms relating to contact prevention and social distancing in Chinese and English in two cities, Hong Kong and Guangzhou. They analyzed how cultural differences affected the evolution of social distancing terms during the COVID-19 pandemic in the two cities.

Another line of research focuses on named entity recognition. Truong et al. (2021) present a COVID-19 domain-specific dataset for Vietnamese, with annotated named entities, including epidemic-specific entity types, and they implement several baselines for this task. COVID-19 was also one of the corpora included in the Third Slavic Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages (Piskorski et al. 2021).

There are also works exploring semantic shifts related to COVID-19. Montariol et al. (2021) propose a novel semantic change detection method using contextual embedding cluster distribution comparison and apply it to a corpus of COVID-19 news. In another work, Kellert and Zaman (2022) introduced a novel dataset focusing on lexical change triggered by the COVID outbreak and compare various types of analyses capable of capturing linguistic change; namely, relative frequency analysis, *n*-gram analysis, lexical change analysis based on word embeddings, and topic modeling. They show that changes of word distributions in topics can provide insights into changes in words' pragmatic meanings.

4. Methodology

Our methodology consists of the following steps: training the embedding model on a news corpus, seed word selection related to COVID-19, expanded COVID-19 candidate vocabulary extraction via embedding nearest-neighbor extraction, vocabulary filtering, and manual selection of the final COVID-19 dataset used for fine-grained analysis in form of naming possibilities and word formation.

First, we selected words related to COVID-19 to be used either directly for analysis, or for embedding-based expansion. We used the list of COVID-19 vocabulary from GDSL (Krvina, 2014–). Next, we used the COVID-19 vocabulary from the CJVT Language Monitor (Kosem et al., 2021) and, third, the list of COVID-19 vocabulary of occasional words collected by Voršič (2022). The resulting joint list contains 186 unique keywords (or key phrases) for embedding-based expansion.

Next, we trained a fastText word embedding model (Bojanowski et al., 2017) on a large Slovenian corpus of 144,352 articles about COVID-19, described in detail in Pollak et al. (2021). The corpus, collected by the EventRegistry service (Leban et al., 2014), contains articles from news portals that contain at least one of the following keywords: *covid*, *koronavirus*, *sars-cov-2*, *covid19*, *covid-19*, *korona virus*, *koronavirusna*, or *koronavirusen* and cover the early pandemic period in Slovenia, between January 1st and December 31st, 2020. We selected a domain-specific corpus for output as closely related to COVID-19 as possible. The fastText model was chosen because of the size of our corpus and as it uses subword information, which makes it the most suitable for morphologically rich languages, and especially given the neologism detection and word-formation perspectives under investigation. Also, unlike the models from the BERT family (Devlin et al., 2019), the output tokens are not tokenized, which is important for our analysis.

For each word (or multiword expression) in the seed vocabulary, we first extracted its 200 nearest neighbors from the fastText embeddings model. Next, we lemmatized the tokens and used Levenshtein distance–based filtering to avoid extracting words that were too similar. We decided to perform lemmatization after the embedding-expansion step because we did not want to fully rely on lemmatization,

which can be unreliable when neologisms are in question. We filtered the extracted candidates by removing those that do not contain any letter of the Slovenian or English alphabet, as well as all words that were already included in the Slovenian lexicon Sloleks (Dobrovoljc et al., 2019) because we were only interested in novel vocabulary. In the end, we kept the most related neighbors for each seed word and grouped them in a joint list by removing duplicates.

Altogether, 4,947 lemmas were extracted. In our study, we analyzed 843 lemmas that occurred at least 5 times in our corpus. We selected the lemmas according to the criteria of direct relevance to COVID-19. As a result, 66 relevant lemmas were identified.

The word lists and categorisations are available at: https://kt.ijs.si/data/elex_covid.zip

5. Analysis and results

5.1 Analysis of naming possibilities

The analysis of naming possibilities included 149 lexemes. In addition to the 66 lexemes resulting from our embedding-based expansion process, 29 lexemes were added from GDSL by Krvina (2014–) and 54 from the COVID-19 vocabulary of occasional words collected by Voršič (2022).

We followed our typology presented in Section 2, distinguishing between: 1) a search for naming possibilities in the language itself, including 1.1) a simplex or a set phrase, 1.2) a derivative, 1.3) a neosemantic formation, when a lexeme acquires new meaning, and 2) borrowing from a foreign language, including 2.1) disguised borrowing or calques and 2.2) expressed borrowing; for a detailed description, see also Section 2. A schematic overview is presented in Figure 1.

The results show that 85.9% of the lexemes were created by word-formation processes (i.e., derivatives in Figure 1), showing high language productivity and language vitality. Among the naming possibilities derived from Slovenian, set phrases follow at 6% (e.g., *socialna distanca* 'social distance', *omejevanje socialnih stikov* 'limiting social contacts', and *pridružena bolezen* 'associated illness'), and neosemantisms account for 1.3% (e.g., *govorec* 'spokesperson' and *testirati* 'to test'). Explicit borrowings, on the other hand, account for 6.7% (e.g., *covid-19*, *korona* 'corona', *lockdown*, *webinar*), and we did not identify any calques in our examples.

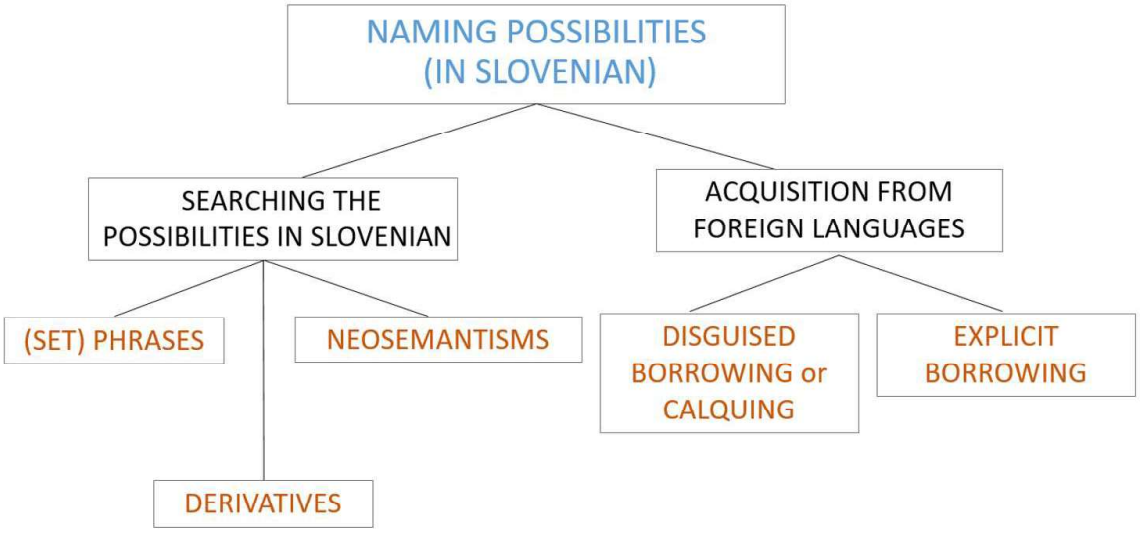

Figure 1: Naming possibilities in Slovenian (Vidovič Muha 2000, 2023)

5.2 Analysis of word-formation processes

5.2.1 Analysis of neologisms

A more detailed word-formation analysis included a total of 77 lexemes. 62 examples from our embedding-based extension method (out of the total list of 66 lexemes, four instances were not kept for analysis due to the fact that they were explicit borrowings from English and were not formed using word-formation processes in Slovenian) and 15 neologisms were included in the already confirmed GDSL, but were originally not kept in the embedding expansion results (because they did not appear above the selected frequency threshold).

The analysis of this dataset (see Table 1) shows that the most frequent are systemic formations (94.8%), out of which the most productive are interfixal compounds, followed by ordinary derivatives by suffixation and ordinary derivatives by prefixation, modificational derivatives by suffixation, derivatives from a prepositional phrase, coordinate interfixal-suffixal compounds, and subordinate interfixal-suffixal compounds. Compared to the systemic formations, the percentage of systemically unpredictable formations is much lower (5.2%), with abbreviations, blend words, and bicapitalizations.

Word-formation type	%	Example and gloss
Systemically predictable formations		
Interfixal compounds	41.56	*koronavirus* 'coronavirus (n.)'
Ordinary derivatives by suffixation	27.27	*koronavirusni* 'coronavirus (adj.)'
Ordinary derivatives by prefixation	15.58	*asimptomatski* 'asymptomatic'
Modificational derivatives by suffixation	3.90	*gripca* 'little flu (diminutive)'
Derivatives from a prepositional phrase	2.60	*brezkontakten* 'contactless'
Coordinate interfixal-suffixal compounds	2.60	*nosno-žrelni* 'nasopharyngeal'
Subordinate interfixal-suffixal compounds	1.30	*visokorizičen* 'high-risk'
Systemically unpredictable formations		
Abbreviations	2.60	*DSO < dom starejših občanov* 'retirement home'
Blends	1.30	*infodemija* 'infodemics'
Bicapitalizations	1.30	*OstaniDoma* 'StayHome'
Total	100	

Table 1: Categorization of examples by word-formation type.

Systemic derivatives

We first focus on systemic formations; that is, those that are formed in accordance with the word-formation rules of Slovenian. The systemic formations are categorized as compounds, derivatives by suffixation, and derivatives by prefixation.

Among the systemically predictable formations related to the coronavirus pandemic, the most frequent are nominal interfixal compounds; namely, those containing the prefix *korona-* in the first part. For example, in addition to the aforementioned borrowed word *koronavirus* 'coronavirus', these are neologisms of the type *koronačas* 'coronatime', *koronahumor* 'coronahumor', and *koronazakon* 'coronalaw'. Alternatively, the non-adapted term *covid-* is also productive for interfixal compounds, but, whereas compounds with the first component substituted are consistently written together in the material, compounds with the first borrowed component *covid-* in the first part can be written either together (e.g., *covidbolnišnica* 'covid hospital' and *covidoddelek* 'covid ward') or with a hyphen (e.g., *covid-redar* 'covid checker', *covid-pozitiven* 'covid-positive'). Among the noun interfix compounds in the collected material, there are also compounds with borrowed prefixoids (e.g., *alfakoronavirus* 'alphacoronavirus', *kiberkriminalec* 'cyber criminal') and unborrowed prefixoids (e.g., *samokarantena* 'self-quarantine'). Adjectival compounds are also confirmed (e.g., *novopotrjen* 'newly confirmed',

novookužen 'newly infected') and to a lesser extent verbal interfixal compounds; for example, *samoizolirati (se)*, *samoosamiti (se)* 'to self-isolate'. It is noticeable that among the more productive formations, especially those with the constituents *novo-* 'new(ly)' and *samo-* 'self' are the most productive. At the same time, among the compounds, there are those with both an acronymic (e.g., *PCR-metoda* 'PCR method', *PCR-test*) and a numeric (e.g., *10-dneven* 'ten-day', *14-dneven* 'fourteen-day') constituent in the first part. Among the subordinate interfixal-suffixal compounds, only the adjectival formation *visokorizičen* 'high-risk' is present. Coordinate interfixal-suffixal compounds are also rare; only the adjectival formations are attested, namely *nosno-žrelni* 'nasopharyngeal', *ustno-nosni* 'oral-nasal'. Certain formations, such as *videopovezava* 'vodeo connection', *14-dneven* 'fourteen-day', and *nosno-žrelni* 'nasopharyngeal', are not really neologisms because they were already in use before, but their frequency increased sharply during the pandemic.

Interfixal compounds are followed in frequency by ordinary derivatives. Here we can again mention the lexeme *korona* 'corona', which is productive of the adjectival derivative *koronski* 'corona(l)' and the nominal derivative *koronik* 'corona-positive person'. The lexeme *koronavirus* 'coronavirus' is the basis of the derivative *koronavirusni* 'coronavirus (adj.)'. Otherwise, adjectival formations are the most frequent among the derivatives (e.g., *prebolevniški* 'convalescent'), whereas derivatives from the noun stem *pandemija* 'pandemic'—for example, *pandemičen* 'pandemic (adj.)' and *pandemski* 'pandemic (adj.)'—are used synonymously. *The* derivative *samoizoliran* 'self-isolated' is derived from the verbal stem *samoizolirati* 'to self-isolate'. Among the noun derivatives, there are verb derivatives (e.g., *oksigenacija* 'oxygenation', *predihavanje* 'ventilation') and adjectival derivatives with the suffix *-ost* (e.g., *asimptomatičnost* 'asymptomaticity', *brezkontaktnost* 'the state of being contactless'). Another interesting derivative formation is the suffix *-izem* '-ism' (i.e., *starizem* 'ageism'). We also trace adverbial derivatives—in synonymous use, the derivatives of the adjectival stem are *asimptomatično* 'asymptomatically' and *asimptomatsko* 'asymptomatically'. Another set is derivatives from a prepositional phrase, for which only two formations with the native prefix *brez-* 'non-' or '-less' are found in the material; namely, the adjectival formations *brezstičen* 'contactless' and *brezkontakten* 'contactless', which are in a synonymous relationship. Among the modificational derivatives, the verbal formations *prekuževati* 'to develop herd immunity' and *predihavati* 'to ventilate' and the nominal diminutive form *gripca* 'little flu' appear. The prepositional phrase formations and the modifying derivatives thus prove less productive.

Finally, we also observe nominal and adjectival ordinary derivation by prefixation with borrowed prefixes (e.g., *superprenašalec* 'superspreader', *antitelesa* 'antibodies', *asimptomatičen* 'asymptomatic') and non-borrowed prefixes (e.g., *neinvaziven* 'non-invasive', *nekužen* 'non-contagious', *protikorona* 'anti-corona').

Non-systemic derivation

In the context of less frequent systemically unpredictable formations, the abbreviations *DSO* < *dom starejših občanov* 'retirement home' and *PKP* < *protikoronski paket* 'anti-corona package' have been confirmed. Next, we find bicapitalizations (e.g., *OstaniDoma* 'StayHome'), and blends (*infodemija* < *informacija* 'information' + *epidemija* 'epidemic' ('infodemic', an epidemic of false, misleading information).

5.2.2 Analysis of occasionalisms

Occasionalisms are words composed for a specific purpose, with low frequency, and they are only at the periphery of the language system. The ad hoc formations are words that are the manifestation of the most current word-formation processes and reflect the creative flexibility of language, which is always a reflection of society, what is happening in it, and the social changes that are taking place.

To complement the word-formation description of the COVID-19 vocabulary in Section 5.2.1, which is based on material either from the confirmed vocabulary from GDSL or corpus-grounded (extracted words appearing at least five times in our corpus), in this section we discuss occasionalisms, which are an important counterpart to the analysis of neologisms. These were used in the seed list part of our vocabulary consisting of occasionalisms by Voršič (2022), collected from various social media sources, or were found in the embedding-based results, but did not match the set frequency threshold. The categorization is based on the work by Voršič (2022).

Among the systemic derivatives, the most productive are verb derivatives from proper nouns: *beović-iti* 'to speak like Bojana Beović' ('to speak in such a way that you leave people in suspense' referring to Bojana Beović, who was the head of the Medical Chamber of Slovenia); similarly, *kaciniti* 'to speak like Kacin' ('to explain instructions in a mischievous and mildly threatening manner', referring to Jelko Kacin, who was the main governmental director of public relations). Such derivatives also give rise to higher-stage nominal derivatives; that is, *beovićenje* 'speaking like Beović', *kacinjenje* 'speaking like Kacin'.

In contrast to the results in Section 5.2.1, non-systemic derivatives are much more common in the context of ad hoc vocabulary. The essential characteristic of non-systemic formations is the indeterminacy of the syntactic stem and the impossibility of morphemization, but also the unpredictability of the number and different parts of speech of stem words that are merged into a neologism. The fact that most of the ad hoc formations are non-systemic formations is not surprising because these are words formed for stylistic effect and as a more attractive parallel to the existing lexemes. In the seed words by Voršič (2022), there are examples such as *kapitalizolacija** 'capitalisolation' < *kapital* 'capital' + *izolacija* 'isolation' ("you can go to work but you can't hang out with your friends"), or hashtags such as

#OstaniZdrav 'StayHealthy', where each component in the keyword is often capitalized, and so these formations could therefore also be defined as sets of bicapitalizations.

The high productivity among ad hoc formations is shown by blends. These are a more recent type of formation, formed by the compounding and back formation of two, or more rarely several, independent words that are expressively overlapping at a certain point. Sicherl and Žele (2018: 76) point out two basic conditions that have to be met to justify this type of formation; namely, 1) the overlap must be semantically recognizable, meaningful, and stylistically effective, and 2) the degree of back formation of individual subordinate words must be adapted to the pronunciation possibilities in the given language and determined by the creator. Stylistic effectiveness, wit, and jocularity are features also highlighted by Bugarski (2002: 217), who treats blend words as a distinctly sociolinguistically motivated word-formation process. Thus, blend words are formed on purpose to achieve a certain stylistic effect or with the intention of influencing (Sicherl & Žele, 2018: 82).

This is also reflected in the blend words analyzed, grouped into the following types: a) blend words in which the first part of the first sub-word and the whole of the second sub-word are joined; for example, *covinek** 'covidbend' < *covid* + *ovinek* 'bend' ('loosely avoiding the oncoming person while walking'); b) blend words, in which the whole of the first sub-word and the last part of the second sub-word are joined together; for example, *koronačitnice* 'coronavacation' < *korona* 'corona' + *počitnice* 'vacation' ('vacation in the time of corona'); c) blend words, in which the central part of the overlap is shared by the two base words and they overlap in this part; for example, *covidiot* < *covid* + *idiot* ('a person that ignores the measures'); d) blend words in which part of the second subword is inside the first: *opravljičilo** 'escusetale' < *opravičilo* 'excuse' + *pravljica* 'fairy tale' ('an obviously a made-up reason when the police stop you in the next municipality'; *natednovanje** 'strainweeking' < *nategovanje* 'straining' + *teden* 'week' ('which is every day for another 14 days and then we'll see'); and e) a special type of blend word structured by a mental association with current social conditions, already mentioned by Sicherl and Žele (2018): near homonyms (paronyms). The wide selection confirms that these were particularly common during the coronavirus pandemic: *dombola** 'homeraffle' < *dom* 'home' + *tombola* 'raffle' ('a raffle to see which parent can go for a walk alone and who stays at home with the children').

5.2.3 Word-formation combinatorics

Here we discuss the material from the perspective of formant combinatorics. Slovenian, like other Slavic languages, is characterized by a rich morphemic structure of words, which is the result of multi-stage formation; for example, from the adjective *star* 'old' the noun *starost* 'age' is formed in the first stage, from it the adjective

starosten 'age' in the second stage, from it the noun *starostnik* 'old man' in the third stage, and from it the possessive adjective *starostnikov* 'old man's', which is the fourth stage. This example demonstrates the associativity of the four suffixal forms: *-ost* + *-en* + *-ik* + *-ov*, and the associativity of the forms is to be understood as the ability of different word-forming elements to coexist in the context of a multi-stage formation, taking into account the meaning-formation aspect.

As shown by recent vocabulary referring to the outbreak of the coronavirus pandemic, the most productive word-forming type is the interfixal compound, with the constituent *korona* 'corona' in the first part proving to be particularly productive (e.g., *koronačas* 'corona time', *koronabedak* 'corona idiot', *koronazakon* 'corona law'). Otherwise, *korona* is productive for adjectival derivation (e.g., *koronski* 'corona(l)'), and compounding (e.g., *protikorona* 'anti-corona (n.)', *protikoronski* 'anti-corona (adj.)'). The second most frequent are common derivatives, which is not surprising because the word-formative derivation is the most common derivation process for Slovenian (Plemenitaš et al., 2020). Derivatives are also the word-formation type that most clearly demonstrates the sociability of affixes. This indicates, for example, the nominal compound *rizičn-ost* 'riskiness', derived from the adjective *rizič-en* 'risky'. In this case, it is the combinatorics of the suffix forms *-en* + *-ost*. The same combinatorics is also confirmed in the case of higher-stage derivatives from the noun *simptom* 'symptom': the adjective *simptomatičen* 'symptomatic' from the first-stage noun *simptom-atika* 'symptomatics' is productive in the material for the compound *asimptomatičen* 'asymptomatic', from which the two adjectives are derived, irrespective of the prefix *a-*, according to the pattern *simptomatičen* (adj.) → *simptomatičn-o* (adv.), or *simptomatičen* (adj.) → *simptomatičn-ost* (n.) forming the adverb *asimptomatičn-o* 'asymptomatically' and the noun *asimptomatičn-ost* 'asymptomaticness' at the same stage. As a synonym of the adjectival formations *simptomatičen* 'symptomatic', the adjective *simptomatski* 'symptomatic' appears, from which the adverb *simptomatsk-o* 'symptomatically' is formed, and in the material the adverb *asimptomatsk-o* 'asymptomatically' is derived as a higher-stage derivative from the compound *a-simptomatski* 'asymptomatic'. *Simptomatičn-o* 'symptomatic' is the adjective from which the adverb *simptomatsk-o* 'symptomatically' is formed. The companionability of the suffixal forms with the combinatorics *-en* + *-ost* is also expressed in the derivative *brezkontaktn-ost* 'contactlessness', which is a derivation of the derivative of a prepositional phrase: *kontakt* 'contact' → *brez-kontakt-en* 'contactless' → *brezkontaktn-ost* 'the state of being contactless', and in the verbal derivative from the prepositional phrase, productive for the formation of nouns with the meaning of properties: *pre-boleti* 'to recover (perfective verb)' → *prebol-el* 'recovered' → *prebolel-ost* 'recovery from disease; *pre-boleti* 'to recover (v. pf.)' → *prebol-eva*-ti 'to recover (v. impf.)' → *prebolev-en* 'recovering' → *prebolevn-ost* 'recovers from disease'. The adjectival derivative *prebolev-en* 'recovering' is also used to form the noun *prebolevn-ik* 'person that has recovered', which is a stem of the adjective *prebolevni-ški* 'relating to persons that have recovered'.

When observing word-formation combinatorics, we are mainly dealing with the combinatorics of two suffixal forms; namely, the adjectival suffix *-en* and the nominal suffix *-ost*. Only some of the formations show a combinatorics of three suffixes, namely *-atika* + *-en* + *-ost* and *-en* + *-ik* + *-ški*.

Among the ad hoc formations, only the verbal noun derivatives are confirmed within the systemic formation, which give rise to the verbal noun derivatives with the word-formative meaning of action, and they show the combinatorics of the suffixal forms *-iti* (e.g., *beović-iti* 'to speak like Beović') + *-enje* (e.g., *beovićenje* 'speaking like Beović') and *-ovati* (e.g., *krekovati* 'to speak like Milan Krek') + *-anje* (e.g., *krekovanje* 'speaking like Krek').

6. Conclusions and further work

This article analyzed COVID-19 Slovenian vocabulary from the perspective of naming possibilities and word formation, including formant combinatorics. We grouped various sources with COVID-19 vocabulary and used natural language processing techniques to expand this and acquire additional vocabulary. The results of our study have an impact on understanding various naming possibilities and word-formation processes in Slovenian, and, on the applied side, 41 newly identified words will be proposed for expansion of the current description of COVID-19 vocabulary in the Growing Dictionary of the Slovenian Language (ed. Krvina 2014–). The analysis shows that a large majority of lexemes were created through word-formation processes, whereas set phrases, neosemantisms, and explicit borrowings were much less frequent, and no calques were identified in the examples analyzed.

From the point of view of the word-formation system, systemic derivatives were the most frequent formation process, among them interfixal compounds, ordinary derivatives by suffixation, and derivatives by prefixation. The analysis also confirms the finding of Stramljič Breznik (2021) that the most productive substructure is the root morpheme *korona*, which produces most of the subordinate interfixal compounds, but also higher-stage adjectival derivatives by suffixation and by prefixation. In addition, we also analyzed occasionalisms, which are mainly non-systemic and employ the blending strategy. In terms of formant combinatorics, we found that the most productive derivatives are adjectival and nominal derivatives with the suffix *-ost*; these are also the ones that show the most frequently confirmed combinatorics of the suffix *-en- + -ost.*

Because this is only a preliminary study performed on a small sample and on the vocabulary extracted from existing resources and from the corpus from 2020, one must note that the sample might not be representative. The main goal was to detect various types of naming possibilities and word-formation processes. In the future, we plan to update the study using either the updated domain COVID-19 corpus, or by

adding material from the Slovenian Monitoring corpus (Kosem, 2022; Kosem et al., 2022). Whereas this study focused on single-word terms, further work also covering multiword expressions would be of interest.

7. Acknowledgements

This article was written as part of the project Formant Combinatorics in Slovenian (J6-3134) funded by the Slovenian Research and Innovation Agency (ARIS). We also acknowledge ARRS funding through the core programs Knowledge Technologies (P2-0103) and The Slovenian Language in Synchronic and Diachronic Development (P6-0038).

8. References

Akut, K. B. (2020). Morphological Analysis of the Neologisms during the COVID-19 Pandemic. *International Journal of English Language Studies*, 2(3), pp. 1–7. Available at: https://doi.org/10.32996/ijels.2020.2.3.11.

Al-Salman S., & Haider A.S. (2021). COVID-19 Trending Neologisms and Word Formation Processes in English. *Russian Journal of Linguistics*, 25(1), pp. 24–42. doi: 10.22363/2687-0088-2021-25-1-24-42

Asif, M., Zhiyong, D., Iram, A., & Nisar, M. (2021). Linguistic Analysis of Neologism Related to Coronavirus (COVID-19). *Social Sciences & Humanities Open* 4. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3608585.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.

Będkowska-Kopczyk, A. & Łaziński, M. (2020): COVID-19 Vocabulary in Slavic. In Marc L. Greenberg (ed). Encyclopedia of Slavic Languages and Linguistics Online. Available at: https://referenceworks.brillonline.com/browse/encyclopedia-of-slavic-languages-and-linguistics-online.

Bugarski, R. (2002). *Nova lica jezika*. Belgrade: Čigoja štampa.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 4171–4186. Minneapolis, MN: Association for Computational Linguistics.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., & Robnik Šikonja, M. (2019). *Morphological lexicon Sloleks 2.0, Slovenian Language Resource Repository CLARIN.SI*, ISSN 2820-4042. Available at: http://hdl.handle.net/11356/1230.

Fran, različica covid-19 (7.1), (2020). Available at: https://fran.si/o-portalu?page=Covid_19_2020.

Gustilo, L., Pura, C. M., & Biermeier T. (2021). Coronalexicon: Meanings and Word-Formation Processes of Pandemic-Related Lexemes across English Varieties. *International Journal of English Language Studies* 27(4), pp. 1–15. Available at: https://www.researchgate.net/publication/356999244_Coronalexicon_Meanings_and_Word-formation_Processes_of_Pandemic-related_Lexemes_across_English_Varieties.

Kosem, I. (2022). Trendi—A Monitor Corpus of Slovene. In *Proceedings of the XX EURALEX International Congress*. Mannheim, Germany, pp. 230–239. Mannheim: IDS-Verlag.

Kosem, I., Čibej, J., Dobrovoljc, K., & Ljubešić, N. (2022). Spremljevalni korpus Trendi: metode, vsebina in katalogizacija besedil. In *Proceedings of the Conference on Language Technologies and Digital Humanities,* Ljubljana, Slovenia, pp. 86–92. Ljubljana: Inštitut za novejšo zgodovino.

Kosem, I., Čibej, J., Gantar, P., Arhar Holdt, Š., Krek, S., Laskowski, C., Robnik Šikonja, M., Klemenc, B., Dobrovoljc, K., Gorjanc, V., Repar, A., & Ljubešić, N. (2021). *Sledilnik 1.0: Language Monitor.* Available at: viri.cjvt.si/sledilnik.

Krvina, D. (ed.). (2014–). *Sprotni slovar slovenskega jezika.* Available at: www.https://fran.si/iskanje?FilteredDictionaryIds=132&View=1&Query=%2A.

Krvina, D. (2023). *Sprotni slovar slovenskega jezika.: Uvod.* Ljubljana: Založba ZRC. Available at: https://fran.si/132/sprotni-sprotni-slovar-slovenskega-jezika/datoteke/Sprotni_Uvod.pdf

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014): Event Registry: Learning about World Events from News. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*, pp. 107–110. New York: Association for Computing Machinery.

Lei, S., Yang, R., & Huang, C.-R. (2021) Emergent Neologism: A Study of an Emerging Meaning with Competing Forms Based on the First Six Months of COVID-19. *Lingua,* 258: 103095. Available at: https://www.sciencedirect.com/science/article/pii/S002438412100067X.

Piskorski, J., Babych, B., Kancheva, Z., Kanishcheva, O., Lebedeva, M., Marcińczuk, M., Nakov, P., Osenova, P., Pivovarova, L., Pollak, S., Přibáň, P., Radev, I., Robnik-Sikonja, M., Starko, V., Steinberger, J., & Yangarber, R. (2021). Slav-NER: The 3rd Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pp. 122–133, Kyiv, Ukraine: Association for Computational Linguistics.

Plemenitaš, K., Voršič, I., & Stramljič Breznik, I. 2020. Derivational Networks in Slovene. In L. Körtvélyessy, A. Bagasheva, & P. Štekauer (eds.) *Derivational networks across languages*, pp. 105–113. Berlin: De Gruyter Mouton.

Pollak, S., Martinc, M., Pelicon, A., Ulčar, M., & Vezovnik, A. (2021). Covid-19 v slovenskih spletnih medijih: analiza s pomočjo računalniške obdelave jezika. In *Pandemična družba: Slovensko sociološko srečanje,* Ljubljana, pp. 260–268. Ljubljana: Slovensko sociološko društvo.

Sicherl, E., & Žele, A. (2018). Prekrivanke v slovenščini z vidika vpliva angleškega jezika. *Jezik in slovstvo,* 63(1), pp. 75–88. Available at: https://www.jezikinslovstvo.com/pdf.php?part=2018%7C1%7C75%E2%80%9388.

Stramljič Breznik, I. (2005). Kvantitativne lastnosti slovenskega tvorjenega besedja v poskusnem besednodružinskem slovarju za črko B 'Quantitative properties of Slovene complex words in the test dictionary of word families for entries starting in B'. *Slavistična revija* 53 (4): 505–520.

Stramljič Breznik, I. (2021). Pandemija koronavirusa – zunajjezikovni dejavnik jezikovne ustvarjalnosti. In P. Kowalski (ed.) *Słowotwórstwo w przestrzeni komunikacyjnej* (= *Prace Slawistyczne, Slavica 151*), pp. 307–322. Warsaw: Instytut Slawistyki Polskiej Akademii Nauk. Available at: https://ispan.waw.pl/ireteslaw/bitstream/handle/20.500.12528/1927/Pawe%c5%82_Kowalski_%28red%29_S%c5%82owotw%c3%b3rstwo_w_przestrzeni_komunikacyjnej.pdf?sequence=1&isAllowed=y.

Škerget, S. (2020). Novotvorjenke, nastale v času pandemije koronavirusa. *Liter jezika: literarno-jezikoslovna revija* 11: 14. Available at: http://literjezika.ff.um.si/?page_id=371.

Toporišič, J. (2000). *Slovenska slovnica*. Maribor: Obzorja.

Vičar, B. (2021). Koronabesedje. *Naši izviri: glasilo Občine Miklavž na Dravskem polju* 21(105), pp. 33–34.

Vidovič Muha, A. (2013). *Slovensko leksikalno pomenoslovje.* Ljubljana: Znanstvena založba Filozofske fakultete.

Voršič, I. (2010). Sistemske in nesistemske tvorjenke v tiskanih oglasih. *Jezikoslovni zapiski*, 16(1), pp. 107–120. Available at: http://www.dlib.si/details/URN:NBN:SI:DOC-5AGEWGAH.

Voršič, I. (2022). Priložnostne tvorjenke kot odraz dobe koronavirusa. In G. Nikolovski (ed.) *Slavistična prepletanja*, 3, pp. 253–266. Maribor: Univerza v Mariboru, Univerzitetna založba. Available at: https://press.um.si/index.php/ump/catalog/book/668.

Wang, X., & Huang, C. (2021). From Contact Prevention to Social Distancing: The Co-Evolution of Bilingual Neologisms and Public Health Campaigns in Two Cities in the Time of COVID-19. *SAGE Open*. Available at: https://journals.sagepub.com/doi/10.1177/21582440211031556.

Weiss, P., & Dobrovoljc, H. (2020). Kako pisati in sklanjati izraze “koronavirus” in bolezen “covid-19” ali “koronavirusna bolezen 2019”. *Jezikovna svetovalnica.* Available at: https://svetovalnica.zrc-sazu.si.