MERGING PROFESSIONAL AND COLLABORATIVE LEXICOGRAPHY: THE CASE OF CZECH NEOLOGY

Michal Škrabal
Institute of the Czech National Corpus, Charles University (michal.skrabal@ff.cuni.cz)

Martin Kavka
(kavka.martin@gmail.com)

Abstract
This paper aims to relate two linguistic phenomena: neology (along with sources for its study) and collaborative lexicography. A pair of case studies is presented concerning two thematically defined groups of recent Czech neologisms: those abusing the Czech ex-president V. Havel’s name and those reflecting the Covid-19 pandemic. An initial dataset was provided by the user-generated content web dictionary of non-standard Czech Cestina 2.0 and the Neomat neology database, fostered by professional linguists. The objective data from a monitor corpus of Czech is used in contrast with the initial dataset and thereby leads to some open questions, especially with regards to the extent to which amateur and professional, two branches of lexicography, can inspire and enrich each other.

Key words: user-generated content, collaborative lexicography, crowdsourcing, neologisms, Czech, Vaclav Havel, COVID-19

1. Introduction
By its very nature, neology is probably the worst lexicographically processed part of the lexicon. It poses a significant problem for lexicographers, starting with the question: which newly registered word will actually come into use and which will sink into oblivion. Earlier dictionaries of neologisms, mostly on paper, were often published with considerable delay and recorded words that were no longer used at the time. Today, this handicap can be easily overcome by regularly updated electronic dictionaries. Linguists register new words in specialized neological databases; in addition, there is also bottom-up lexicography (Carr 1997), built on voluntary contributors willing to take part in proverbial harmless drudgery and generate various collaborative dictionary projects, such as Wiktionary, Urban
Dictionary, Macmillan’s Open Dictionary, Wordnik, and many others. There is no indication that these users perceive their lack of professional qualification as an obstacle to participation in these collaborative works. Professional lexicographers can significantly benefit from the efforts made by amateur colleagues by applying their theoretical expertise and should strive to find the desired synergy between these two approaches rather than look askance at laymen contributors. In other, Michael Rundell’s (2017: 1), words: ‘Crowdsourcing – in its various forms – should be seen as an opportunity rather than as a threat or diversion.’

In this article, we will focus on the phenomenon of collaborative lexicography in the Czech environment, which is promisingly developing similarly to other countries (see, among others, Fuertes-Olivera 2009; Lew 2011; Hanks 2012; Meyer and Gurevych 2012; Creese 2013; Kosem, Gantar and Krek 2013; Lew 2014; Čibej, Fišer and Kosem 2015; Rundell 2017). To find out how viable and fecund it is and how much ‘the wisdom of crowds’ (Surowiecki 2005) can be used by professional lexicographers, we will focus on neology or the processing of neologisms, respectively. It seems that neology is an attractive topic within so-called folk linguistics (Niedzielski and Preston 2003), and neologisms generally enjoy a great deal of attention among language speakers. Besides, neologisms are repeatedly mentioned (Rundell 2012: 80; Lew 2014: 25) as an area of lexicon where amateurs can contribute most significantly, along with slang, regional varieties, and technical terminology.

In Section 2, we give a general presentation on the user-generated content web dictionary of non-standard Czech Čeština 2.0 and its offshoot in book form, as well as on the Neomat neology database, fostered by professional linguists. The new monitor corpus of Czech called Online is also presented. In Section 3, two case studies concern relatively large groups of recent Czech neologisms: those exploiting Czech ex-president Václav Havel’s name (3.1) and those reflecting the actual Covid-19 pandemic (3.2). These examples were chosen deliberately, as they differ in terms of the time of their coining and thus also the degree of their dynamics and entrenchment in contemporary Czech. On the contrary, they have in common the fact that both play a quite prominent role in contemporary Czech public discourse. We also outline a simple typology of neologism formations for these lexical sets. Subsequently, we contrast procedures and results of collaborative lexicography with objective data from the Online corpus to formulate some open questions (Section 4), especially in the sense of how these two branches of lexicography – amateur and professional – can inspire and enrich each other. Moreover, we consider the advantages and disadvantages of engaging amateur lexicographers in a dictionary-compiling process.

2. Data sources

2.1. Čeština 2.0

The Čeština 2.0 (‘Czech 2.0,’ hereinafter: C20) web portal is a collaborative user-generated content (UGC) project that has been running publicly since January 2009. Its creator, freelance journalist and copywriter Martin Kavka, founded it as the ‘Czech Urban Dictionary’ to gather ‘brand new words as well as slang and vernacular expressions and other interesting words from all areas of the mother tongue,’ as he states on the webpage. The initial search for new, hitherto unheard and original-sounding words in the public sphere (and their recording in a notebook) soon developed into a specialized website. First
it was filled by the author himself and a handful of his friends and family members, but, over time, the ranks of contributors grew, as did the length of the list of entries.\(^3\) Currently, the C20 dictionary contains more than 20 thousand words; in the last year, it grew by more than 4,000 entries, an average of 135 entries per week, while about 80 of these are edited and published. Individual entries are written by users (of which there are thousands, though only tens of them are actually active\(^4\)), and Kavka, who still oversees the project, edits, and, over time, publishes them.

In October 2018, a selection of entries from the online dictionary was published in the book format under the title *Hacknutačestina* (‘Czech Hacked,’ hereinafter: CH – Kavka and Škrabal 2018). It can be seen as a sequel to older dictionaries of non-standard Czech (Ouředník 1988; Obrátl 1999-2000; Hugo et al. 2006) that are quite popular among Czech readers. Knowing that a book, a manufactured article, would be – unlike the web dictionary – unchangeable and irreparable, it was necessary to engage a professional lexicographer in the project. In particular, it was essential to adapt the source material, edited up to this point by an amateur, to specific professional standards. This was one of the biggest challenges of the project: we did not want to smooth out the peculiar ‘handwriting’ of individual contributors and get rid of the charming tinge of amateur lexicography; on the other hand, it was necessary to establish at least elementary lexicographic principles and stick to them during the compilation process.

The book itself contains more than 3,000 entries, which represent over a quarter of the then online volume (about 11,000 entries). The selection criteria were not strictly defined but rather subjective. The aim was to include all categories of words so that vernacular, slang, loanwords, neologisms, and others appeared in the dictionary alongside frequent manifestations of linguistic creativity and humour. These entries are assumed to represent the proverbial ‘lexical chronicle of the present-day,’ describing trends and situations in politics, society, and sports that have arisen over the last ten years. Because they mainly capture temporary and ephemeral realities, they are mostly nonce words that manifest a particular contributor’s idiolect and/or a proof of his or her linguistic creativity. The use of such ad hoc expressions is then understandably very limited: they are rather puns and wordplay jokes with a very short life span.

The book has received considerable attention among both the general public and the experts (four reprints, 19,000 copies had been printed by the end of 2019). But what we consider to be especially positive is the fact that it attracted readers’ attention not only to the Czech lexicon itself but also to dictionaries in general.\(^5\) In addition to the further growth of active contributors, an offspring series of short videos for the internet television channel *Stream.cz* appeared (almost 600 words were presented in 32 parts), followed in 2020 by a podcast devoted to professional and lay slang terms and even a board game. A sequel to CH could be released in the future, albeit with some slight conceptual changes (see Section 4).

### 2.2. Neomat

Our other data source is an online\(^6\) database of Czech neology, named Neomat (hereinafter: N). We perceive it as an, in a way, ‘orderly’ supplement to C20 for our purposes, under the umbrella of an official institution (Czech Language Institute, within the Czech Academy of Sciences). Today’s version of the database, revised in 2005, is conceived more broadly than theretofore, as an archive of lexical dynamics (registering new verbal valencies
or shifts in the stylistic evaluation of words, among others), and is open to both the lay and professional public. Although there is the possibility of providing feedback, it does not primarily assume an active, contributing role for its users. The main burden in filling the database thus falls on professional excerptors. Clearly, the project is far from collaborative lexicography, at least compared to C20.

The archive is constantly being extended and updated at weekly intervals; at the end of October 2020, it had 345,372 entries (tokens, not types). The semantic content is not the focus of this archive, and therefore entries do not include a dictionary definition, with infrequent exceptions.

2.3. Corpus Online
C20 and N, which do not include any frequency information, are complemented by corpus data. Regularly updated corpora – ideally, on a daily basis – are most suitable for this paper’s purpose. In fact, such a corpus is newly available within the Czech National Corpus infrastructure; it is called Online⁷ (hereinafter: O), and it allows to observe the dynamics of the Czech lexicon, including neologisms. Its strength lies in two major points. Firstly, none of the numerous synchronic corpora of Czech provides such an up-to-date language. Secondly, it is not limited to the simply searchable web, as are others (the TenTen or Aranea corpora families), but may get beyond the public domain. As of October 22th, 2020, it has amassed more than 6.4 billion tokens and was structured into several subcorpora, representing individual media types. The acquisition of texts is targeted: O focuses on the dynamic web content – be it professional (such as most of the media portals) or user-generated (such as social networks, discussions, and forums). However, some text types representing a rather formal language (especially commercial presentations of companies) are not downloaded at all. The corpus includes internet discussions (under online articles) and hobby discussion forums, as well as the content of social networks (Facebook, Twitter). The opposite of this private communication is news websites (both traditional newspapers and magazines and leisure journalism). With respect to neology, the latter media seem to reflect a higher degree of social acceptability and entrenchment of new words and, conversely, a lower degree of personal involvement and expressiveness. Whereas social networks, discussions, and forums mirror naturally spoken Czech much better.

3. Case studies
Our goal in this section is to describe the specifics of two thematically defined groups of neologisms. We deliberately chose different examples in terms of the time of their coining and thus also the degree of their dynamics (see Figures 1 and 2) and entrenchment in contemporary Czech. While the first group refers to the legacy of an already deceased statesman and has been gaining ground in Czech for several years, the second group is brand new and is thriving even as we write this paper. Each group also differs in terms of the number of its members: it is evident that the Covid-19-related words cover a much larger spectrum of reality and will therefore provide richer material.

3.1. Case study #1: Havlophobes vs. Havlophiles
The first case study notes the traces of former Czech president Václav Havel⁸ in the current Czech lexicon. These traces are robust: even though Havel has been dead for a decade, his
Figure 1. Frequency of usage of Havel-related words in O. Occasional peaks can be explained by some regularly recurring anniversaries, for example, of the Velvet Revolution (November 17th) and Jan Palach’s Week which led to the revolution in 1989 (January 15th-21st) or Havel’s birthday (October 5th).

Figure 2. Frequency of usage of Covid-related words in selected online media.\textsuperscript{22}
name has given rise to a large number of neologisms, most of which have negative connotations – opponents of Havel’s philosophy and ethos use them to depreciate Havel’s sympathizers. Often contempt for Havel is signified by a lower-case ‘h’ in his name.

Some of these words – a total of 13 entries – were already recorded in the dictionary of Czech neologisms (Martincová et al. 2004: 148). Furthermore, other Havel-related words can be found in both C20 and N. As of October 22nd, 2020, C20 had a total of 14 entries containing the string havel/havl,9 plus additional 7 entries directly linked to Havel, and there were 65 relevant neologisms in N. However, there were only 6 entries (out of 80, that is, 7.5 %) shared by these two lists. Together with two extra words from Martincová et al. (2004) that were found neither in C20 nor in N, we gained exactly 82 entries for our initial dataset, which can be compared with frequency data in O.

A significant majority of neologisms (77%) are documented in O, but often only in negligible frequencies: 52 words (63%) have an ipm less than 0.1, while only three words have an ipm higher than 1.0: havloid ‘[derogatorily] a supporter of the first president of the Czech Republic V. Havel’ (6.41 ipm); havloidni ‘of, relating to, or associated with havloid*’ (3.65); havlista ‘a fanatic admirer of V. Havel’ (2.84). These often serve as a base for further derivation; for example, havloid is a base for the adjective havloidní as well as the nouns havloidismus and havloidstatanista. However, besides these three derivatives from our dataset, a survey of O offers a far more diverse inventory of neologisms and nonce words, the most recurring of which are the nouns havloidka and havloidísmus, the adjective havloidský, the adverb havloidně, and the compounds chazarohavloidní, havloidiot, and posthavloidně. The ipm’s of all these derivatives range from 0.1 to 0.01. Other words (no less than 600 different word forms)11 do not even exceed 0.01. Similarly, a plethora of neologisms is derived from pravdolaška r (‘a person who professes the legacy of […] V. Havel (based on his famous statement: ‘Truth and love must prevail over lies and hatred.’), with a liberal worldview, sometimes acting as a pathetic defender of democracy’ < pravda ‘truth,’ láška ‘love’) including its several synonyms pravdoláskista, pravdoláskovec and pravdoláškovec, the female form pravdoláškařka, the noun derives pravdoláška, pravdoláškařství, pravdoláškismus, pravdoláškovství, the adjective derives pravdolášský, pravdoláškovní, pravdoláškový, and the adverbial derivate pravdoláškovský. Except for two words: pravdoláška (0.58 ipm) and pravdoláškový (0.19), none of them has an ipm higher than 0.06. It is obvious that, from a lexicographer’s perspective, none of these words would normally get into a list of entries of any dictionary except those specializing in neologisms.

The dynamics of the use of this lexical set is indicated in Figure 1.

The actors of the Havel discourse can be divided into two rivalrous camps, according to their attitude towards the ex-president, and are named variously, largely derogatorily. Havel’s critics seem to be ‘louder,’ concentrating chiefly on social networks or anti-systemic media (while mainstream media refer to Havel seldom and in a neutral way) and trolling online discussions. Remarkably, there are far fewer of these words for a group of Havel’s opponents: in our dataset, we find only the word lženávístník (‘opposite of a pravdoláškař; a hater of ex-president Václav Havel’) and its rare synonym lženávista, and the somewhat neutral adjective compound antihavlovský ‘anti-Havel*,’ or its Czech calque variant protihavlovský. Thus, Havel’s supporters seem to be more benevolent towards their ideological adversaries and to make do with already coined words when referring to them. At the same time, they often identify with pejorative sobriquets (in the same way decadents
once adopted the originally mocking denomination for them), thereby as if blunting their edge. Paradoxically, a given word may have ambiguous semantic prosody, depending on who uses it – whether it is a Havel’s opponent or a supporter.

3.2. Case study #2: As many face masks you possess, as many times you are a human being

A sad truth of the times: whoever does not write about Covid-19 these days need not bother writing at all. The unprecedented pandemic was probably reflected in every language, including Czech, especially in the area of the lexicon. New words, multi-word units, and even idioms and proverbs arose (and then disappeared, often unrecorded) spontaneously and rapidly and flooded mass media and the internet. With a bit of exaggeration, we might say that it would be possible to compile a special dictionary of only these words. Apparently, they are often motivated psychotherapeutically: to transform something unknown and hostile hiding behind official, medical terms into a common language understandable to laypeople with the desired effect: to make life in forced quarantine more bearable. Even numerous words trivializing and mocking the whole situation have occurred, reflecting the Czech people’s grim humour. Most neologisms have an expressive nature (číu virus ‘greetings at the time of the coronavirus epidemic;’ covnivál ‘a person who is too keen about twaddle and hoaxes regarding coronavirus’ < bounivá́l ‘dung beetle’) and clearly manifest a speaker’s attitude (koronapičuš 1. unbalanced person, influenced by a media craze around the coronavirus, with unlimited belief in it and eager to make unsubstantiated and unfounded decisions; 2. author of mammoth-scale measures against coronavirus; 3. coronavirus’ < pičuš ‘fucker’).

The rise of neologisms was also recorded in C20. The very first Covid-19 related word is from January 26th, 2020: skorovirus (skoro ‘almost’) ‘unconfirmed case of coronavirus from China.’ Since then, the number of neologisms has grown daily; as of November 22th, it numbered no less than 711 entries. N is only slightly less affluent than C20: it stored 707 Covid-related entries. The common intersection of both datasets, however, is surprisingly small: 129 out of 1292 (10%). This ratio would decrease slightly (along with a decrease in the number of lemmas) if we did not consider purely spelling variants of the same lemmas; yet, there would still be a significant incompatibility between the two lists of entries: only every tenth word is shared. This result (along with that from the previous case study), which is somewhat startling, demonstrates at least two things: (1) different results are obtained when applying different approaches to the acquisition of neologisms (crowdsourcing, collaborative lexicography × controlled excerpt performed by professionals), and (2) difficulties in capturing this unstable and fluctuating part of the lexicon. This is exactly why the complementarity of these different approaches and data sources is desirable. Only by combining these sources and contrasting them with frequency data from a monitoring corpus we will get a more transparent overview of a) the behaviour of a given lexical set in the current language and b) candidates for inclusion in the official dictionary of Czech, which, coincidentally, is emerging at the moment (Akademický… 2012nn.).

The motivation for designating new Covid-19-related concepts in Czech is multifarious, and the diversity of word-forming patterns, as well as language creativity, can be well examined. First, we set apart words that already exist in Czech and have gained a new meaning during the pandemic; there are only 21 of them (1.6%). A few examples are based on either phonological or morphological proximity of both old and new meanings (nádržka ‘small reservoir’ > ‘face mask’ < na ‘on’ + držka ‘gob;’ koroner ‘coroner’ > ‘a person
infected with coronavirus; korýš ‘crustacean’ > ‘coronavirus’) or conceptual closeness (náhubek ‘muzzle’ > ‘face mask; nechráněný styk ‘unprotected contact’ > ‘violation of coronavirus quarantine by meeting without a face mask’).

As regards truly new lexemes, the simplest way is to create a compound with the first component korona- or its truncated version koro- (altogether approximately 42% of new entries). There is a difference again between N preferring longer forms and C20 with shorter ones, especially if the second part of a compound starts with n-: koronánákaza × koronákaza ‘the coronavirus epidemic’), often both variants referring to the same concept compete with each other: koronadovená ‘forced home-office due to the spread of a new coronavirus’ × korodovená × korolená ‘forced leave due to the coronavirus epidemic.’ Obviously, these compounds are too long for everyday communication, and users tend to shorten them or replace them with blends. This group of words is highly productive, as combinatorial limits are not completely obvious here: the initial segment koro(na)- can be combined with both concrete and abstract nouns, with words of domestic and foreign origin, it can form all content words.17 In fact, this type of formation is so straightforward and widespread that it can actually give rise to a new prefixoid, such as euro-, bio-, ex- or others.

The truncated version of the given segment can also be inserted inside actual words, as a few examples suggest: ekoronomika ‘economy severely affected by the coronavirus pandemic’ (< ekonomika ‘economy’); hypokoronadr ‘a person who constantly thinks he has the coronavirus or will soon catch it’ (< hypochondriad ‘valetudinarian’); velikoronoce ‘Easter during the coronavirus epidemic, when national quarantine was declared.’ However, this type partly overlaps with another group (of mostly non-compounds) where there is only a minor modification of an existing word: koronténa ‘quarantine for people suspected of having caught the coronavirus’ (karanténa ‘quarantine’); mateřišnˇka ‘face protection during the coronavirus pandemic that was sewn by one’s mother’ (< mateřidouška ‘Thymus,’ mater ‘maternal’); maskurbace ‘touching a respirator or a face mask too often’ (< masturbace ‘masturbation’); syndrom vymotení ‘state in which one has fully identified with the harsh Covid-19 regulations and continues to live by them […]’. Often the comic effect of combining a foreign and domestic element is desired: korokdák 1. hoax, nonsensical statement about coronavirus; 2. ill-considered proposal for measures to mitigate the effects of the coronavirus epidemic’ (< kdák ‘cackle’). Users are even aware of the metalinguistic nature of some neologisms: covid dokonavˇý, ‘crowned condition in which coronavirus disease has definitely disappeared; opposite of covid nedokonavˇý (< vid dokonavˇý ‘perfective aspect’); covid nedokonavˇý ‘condition in which the coronavirus disease Covid-19 is still present on planet Earth; opposite of covid dokonavˇý (< vid nedokonavˇý ‘imperfective aspect’).

Quite popular is blending, the combining of two words: koax ‘hoax regarding coronavirus; nákazník ‘store visitor potentially infected with coronavirus’ (< nákaza ‘infection’ + zákazník ‘customer’); zoomestr ‘summer semester of the academic year 2019/20’ (and other words derived from the Zoom application: zoombik < zombie, zoomčastnít se ‘attend an event held online’ < zúčastnit se ‘take part’). For more details on older blends in N, see Filipcová 2016.

There are only a few loanwords without any modifications, such as covid/korona free, social distance, early openers, virus stories, or Coronagate. This proves that such elements seem unnatural in highly inflected Czech, and a speaker tries to adapt them naturally (it
might even be a challenge for him or her) to the Czech phonetic, morphological, and grammatical system. Lockdown is a perfect example: it fits into the target language smoothly, being inflected easily as a hard-stem masculine, and even serves as a root word for subsequent derivation: the adjective lockdownový, the aspectual verbs lockdownovat, lockdownout, zalockdownovat, the diminutives lockdownek, lockdowniček or numerous compounds such as pololockdown ‘semi-,’ pseudolockdown, skorolockdown ‘almost-,’ samolockdown ‘auto-,’ protilockdownový ‘anti-‘, fulllockdownový and many others, albeit with negligible frequencies. The smoothness of the adaptation process indicates why the potential Czech equivalents recorded in C20 (zdravora and zarach) are not found in O at all, and the phonetic transcription lokdaun rarely (0.23 ipm) – they are felt to be superfluous.

Other examples of secondary word-formation from previous neologisms include cověan ‘a person who is easily manipulated by bigwigs abusing panic during a coronavirus epidemic to achieve their political goals’ (< covid + ověan ‘citizen who acts like a sheep,’ coined at C20 in 2009 as a blend of Czech words ové ‘sheep’ and oběan ‘citizen’) or digitální domád ‘a person working at a home-office during forced quarantine,’ modifying the original digitální nomád (doma ‘at home’) ‘a person who travels the world while working remotely for his clients, all he or she needs is a computer and an internet connection’ from 2015.

New multi-word units appear, too, of course. In this respect, the approach to lemmatization differs for both sources: while N is based on one-word lemmatization and collocations are available in sublemma only via advanced settings, the C20 list of entries commonly includes multi-word lemmas (in our dataset, a total of 68 cases out of 711, which is 9.6%). Newly formed combinations of older words (škola v pyžamu ‘distance school teaching during coronavirus epidemic,’ lit. ‘school in pyjamas;’ žíznivé okno ‘temporary pub set up after the closure of pubs due to the spread of coronavirus,’ lit. ‘thirsty window’) predominate, only some use neologisms (na coviděnou ‘farewell at the time of the Covid-19 disease;’ koronovirová opona ‘closed state border due to a coronavirus pandemic’ < železná opona ‘Iron Curtain’). Older phrases and idioms are exploited in an original and playful way: dát si dvacet ‘have an hour free and sew twenty face masks during that time’ (< ‘have forty winks’); šlápnout do pedálu ‘sit down at a sewing machine and sew face masks on it all day’ (< ‘pump the pedals’); na adama ‘without a mask, that is, with an exposed face [. . .]’18 (< ‘skinny dipping’). Even topical modifications of older proverbs are emerging: Rouška kvapná, málo platná ‘A hastily made face mask makes waste’ (< Práce kvapná. . . ‘Haste makes waste’) or Kolik roušek mäs, tolikrát jsi clověkem ‘The number of face masks you possess, the number of times you are human’ (< Kolik jazykářečí umíš… ‘The number of languages you know…’) to name just a few (cf. also Šemelík 2020: 5).

As for the keywords that serve as the derivational basis for neologisms, they are most often official medical terms, that is, koronavirus, Covid-19, or just the general designation virus (or vir in spoken Czech). Together, these three types cover up to 61.8% of the neologisms from our dataset. Furthermore, a face mask is involved very frequently, which is not surprising, given the enthusiasm with which the Czech people began to make protective equipment at home after the government failed to provide them in spring 2020. Besides the official term for a face mask (rouška or ústenka), literally, dozens of new words emerged not only for the mask itself but also for its various types (ropuška ‘used, unwashed face mask’ < ropucha ‘toad’ + rouška; šourka ‘coronavirus face mask sewn from old boxer shorts or briefs’ < šourek ‘scrotum;’ trikiny ‘swimsuit comprising a bikini and a face mask from the same material; see also koronakiny, trojdílné plavky [three-piece swimsuit]).
people (not) wearing them (rouškař ‘a person wearing a face mask;’ bezrouškař ‘a person violating the obligation to wear a face mask during coronavirus epidemic;’ rouškarián ‘a person who has a strong belief in the benefits of face masks against the spread of coronavirus and therefore wears it constantly; the opposite is bezrouškarián’) or other related concepts (rouškije ‘selfie in a face mask;’ rouškiss ‘kiss through a face mask’). The Czech Republic has become a ‘face mask power,’ which is well reflected in several neologisms: roušpublika ‘obligatorily face-masked country’ (< rouška + republika ‘republic’); rouškistán; prorouškovanost ‘degree of voluntary wearing of face masks even in situations where their wearing is not obligatory.’ To a much lesser extent, this also concerns other protective equipment: respír/respoš ‘respirator;’ respirozek ‘person wearing a respirator on his forehead’ (< respirator + jednorozec ‘unicorn’); sněhulák ‘special suit for paramedics testing people for coronavirus’ (lit. ‘snowman’).

Other keywords include karanténa ‘quarantine’ (tyranténa ‘restriction of personal freedom during the coronavirus crisis;’ Karande ‘date which respects coronavirus quarantine and thus a safe distance; ideally in a virtual and contactless mode’ (< randa ‘date’)), pandemie ‘pandemic’ infodemie ‘disseminating an excessive amount of information about a problem, often unverified and misleading;’ pandemáček ‘child conceived at the time of the coronavirus pandemic’); sometimes it is a combination of two keywords (koropanda ‘Covid-19 pandemic’ [< panda ‘panda bear’]; pandavirus ‘Chinese coronavirus (Covid-19 disease’)}. Numerous neologisms originated from the name of the Czech epidemiologist (and the short-term Minister of Health) Roman Prymula, who became the symbol of the fight against the pandemic: prymulex ‘set of government measures for coronavirus;’ prymulká ‘home-made face mask of poor quality (e.g., from pyjamas), worn because of duty during the coronavirus epidemic;’ deprimulovaný ‘depressed by measures stemming from the head of the chief epidemiologist Roman Prymula’). Other people were reflected sporadically in our dataset, like Prime Minister Andrej Babiš:19 na babiše (lit. ‘in Babiš’s manner’) ‘[to wear] a mask with strings untied.’ Surprisingly, a word inspired by the name of the former Minister of Health Adam Vojtěch appeared only scarcely (5 entries in C20).

The last issue we want to touch in our pre-corpus analysis is the number of senses. Most neologisms are monosemous, as we found only 45 cases of polysemous words in the C20 sub-dataset (6.3%). Most of them have two senses, like koronovaný ‘1. infected by the coronavirus, 2. infected by the brain-washed media hysteria over the Chinese coronavirus […]’ (< korunovaný ‘crowned’), in isolated cases, three or more senses can be found.20 This could prove semantic ambivalence of neologisms, as their meanings are far from definite but are still being formed and discussed. Yet, an overwhelming majority of neologisms pass out before gaining another sense.

It is a sad paradox that the term ‘coronavirus’ is by no means new (it goes back to 1968), but until 2020 its use has been limited to the professional domain of language. Therefore, it is logically registered, albeit scarcely, by the older corpora of contemporary Czech: 30 hits, 0.25 ipm in SYN2015 (Křen et al. 2015). However, unlike the previous case study, we would look in vain for its-related neologisms from our dataset in these corpora. Regularly updated web corpora21 are the best choice. Figure 2 shows a dramatic increase in the frequency of selected Czech words (koronavirus or covid and their derivatives) in the initial stage of the coronavirus crisis, yet a much more stable course later on despite the second wave in autumn 2020. Online news websites, be it mainstream or tabloids, refer to
the pandemic significantly more often than social media, which is the opposite of the previous case study.

Not only the mere frequencies are dynamic and fluctuating, but also collocates of some neologisms may change over time. If, for example, we compare collocation profiles of these expressions from the first and second wave (March 25th vs. October 25th, 2020), the results may be quite surprising. Out of the 40 top collocates for either period (attribute: lemma, window span: -3 to 3, sorted according to logDice), there are only 5 (7%) in common: boj ‘fight,’ epidemie ‘epidemic,’ onemocnění ‘disease,’ pacient ‘patient,’ and šíření ‘spread.’ This overlap, however minimal, represents the invariant core which clearly refers to the semantics of the word (‘a disease that is spreading among patients until it becomes an epidemic and must be fought against’), while the variables reflect the temporal and/or spatial specifics (for example, banální ‘banal,’ celoplošný ‘nationwide,’ Charles, mikroskopický ‘microscopical,’ odhalený ‘revealed,’ popírač ‘denier,’ princ ‘prince,’ rychlotest ‘rapid test,’ frequent are the numbers referring to statistics of infected/dead patients).

Table 1 gives a brief outline about Covid-19-related lexemes (not necessarily neologisms) in terms of their saliency within selected media types. Data from the SYN2015 corpus is added for a comparison with a pre-pandemic situation.

The steep rise in the frequency of most words is not at all surprising, and many of them are undoubtedly hot candidates for the Word of the Year 2020. It is much more interesting to note the unequal distribution within different media types: newspapers, regardless of their degree of reliability, provide information on the topic virtually without interruption (and their ipm’s in the table exceed the average level for almost every word) and discussions under articles are – from a lexical perspective – just their extension. The rich Covid-19 discourse also takes place on anti-establishment websites, which are an alternative to the official media for a specific part of society. On the contrary, the topic is discussed to a much lesser extent in the private sphere. Various hobby forums are thematically predefined, and the current discourse infiltrates them far less and rather furtively. The low ipm values may surprise among users of social networks, although they can probably be attributed to the preference for more colloquial variants of official words (for koronavirus: koronák, koronač, korona, and others, including non-diacritical variants in the online environment).

3.3. Conclusion

If we are to summarize our findings from both case studies briefly, it is necessary to underline the following points:

The Czech language has coped with the flood of neologisms during the Covid-19 crisis, proving its flexibility and openness to foreign elements (which some speakers consider undesirable). The Czechs have struggled with a difficult situation by using creativity and humour, albeit sometimes at the cost of being politically incorrect.

On the other hand, the Havel discourse is already firmly entrenched in the Czech environment and has been ongoing at least since Havel’s death in 2011, with some regularly recurring peaks around the anniversaries connected with the ex-president’s life. Havel’s objectors, concentrated mostly on social networks, internet discussions, and anti-establishment media, seem to have the upper hand in it, as most neologisms have markedly defamatory traits, although Havel’s sympathizers often identify with these unfavourable apppellations.
Both lexical sets also differ in their extent (1292 vs. 82 in our datasets). Thanks to the prefixoid korona- (less also covid-), the first group is de facto an open class, continually growing, while the growth of the second one is limited to ad hoc created nonce words with petty ipm’s. There are similar limitations regarding the typology of these neologisms too. In the pandemic discourse, almost any word-forming strategies available in Czech are being used: derivation, compounding (including blending), shortening, borrowing (including calquing), creating multi-word units, or figurative use. On the other hand, Havel-related words are typically simple derivatives or, to a lesser extent, also compounds. The original datasets obtained from various sources were significantly different, having only every thirteenth or tenth, respectively, word in common. Therefore, if we want to

\[ \begin{array}{cccccccc}
\text{main-stream} & \text{tabloids} & \text{online} & \text{hobby} & \text{anti-system} & \text{Facebook} & \text{Online} & \text{SYN2015} \\
\text{media} & \text{discussions} & \text{forums} & \text{webpages} & \text{now} & \text{(total)} & \\
\text{coron.}^* & 4.9 & 6.7 & 94.8 & 16.4 & 56.0 & 56.5 & 42.9 & 0.3 \\
\text{covid.}^* & 827.1 & 889.8 & 906.5 & 116.1 & 906.2 & 488.2 & 614.4 & 0 \\
\text{dezinformace} & 17.8 & 3.3 & 18.2 & 2.7 & 61.0 & 16.4 & 17.7 & 0.7 \\
\text{‘disinformation’} & 297.2 & 316.9 & 179.2 & 19.0 & 415.3 & 88.6 & 141.3 & 8.4 \\
\text{epidemie} & 406.2 & 833.9 & 313.9 & 80.7 & 272.1 & 174.1 & 272.7 & 1.8 \\
\text{‘epidemic’} & 1331.7 & 6195.1 & 614.1 & 61.9 & 1243.0 & 285.3 & 687.1 & 0.3 \\
\text{karanténa} & 28.6 & 10.5 & 8.3 & 2.7 & 24.4 & 10.2 & 13.9 & 0 \\
\text{‘quarantine’} & 26.7 & 4.7 & 4.0 & 0.7 & 7.8 & 4.6 & 8.8 & 0.3 \\
\text{koronavirus} & 17.7 & 10.0 & 44.2 & 6.4 & 25.6 & 16.7 & 30.4 & 0 \\
\text{koronakrise} & 425.0 & 24.5 & 151.6 & 31.6 & 297.9 & 109.0 & 164.2 & 135.4 \\
\text{‘coronacrisis’} & 927.3 & 1196.1 & 658.8 & 57.6 & 591.2 & 299.4 & 460.9 & 71.0 \\
\text{kurzarbeit} & 22.7 & 11.6 & 22.2 & 0.7 & 13.2 & 3.4 & 9.8 & 0 \\
\text{lock(-)down} & 70.0 & 43.7 & 12.5 & 0.7 & 47.4 & 35.4 & 62.2 & 0.5 \\
\text{pacient ‘patient’} & 225.2 & 205.9 & 99.7 & 73.1 & 178.3 & 64.3 & 103.3 & 86.3 \\
\text{opatřeni} & 536.5 & 385.6 & 1590.4 & 208.0 & 392.4 & 667.1 & 709.0 & 4.4 \\
\text{‘measure[s]’} & 113.4 & 78.1 & 182.4 & 14.9 & 414.9 & 62.2 & 87.6 & 5.8 \\
\end{array} \]
examine the group of neologisms in question with a sufficiently empirical approach, only a combined study of multiple sources would provide adequately reliable information. In other words, amateur contributors help to minimize gaps in a given lexical subsystem, drawing attention to neologisms that professional lexicographers might have omitted.

4. Outlook and discussion

Among other things, the possible sequel to CH mentioned earlier (2.1) leads us to consider merging methods of collaborative and professional lexicography. As its creators, we are looking for a way to present the sequel in a fresh way, especially if the book form is preserved. There are certainly many possible solutions, of which contrasting user-generated content with objective data represented by language corpora (O in our case) appeals to us the most. How specifically to do this depends on the degree of data-drivenness.

The list of entries, conceived purely subjectively in CH, should be newly compared to corpus data, and a certain frequency limit should be set for entries to be included in the dictionary. Of course, the microstructure of the entry should also change (see Figure 3): it will be expanded with frequency data, statistics according to various media types, the first occurrence in the corpus along with the date of entry in C20, as well as examples provided by contributors and newly compared with real examples from the corpus. Alongside the lemma, thumbs up/down mirror the other users’ rating and the popularity of the given entry. The icon of a magnifying glass refers to the word being found in O, the crossed-out glass to an entry recorded only in C20. Larger space will be needed for these updated entries, which will logically reduce their number if a book is to be an outcome.

Undoubtedly, it is difficult to predict which of the deluge of new words will become permanently entrenched in the language (as a matter of fact, only a very few will24) and which will not. The lexicon as the most dynamic level of language – especially in turbulent times – is a mycelium for linguistic change. Nevertheless, it is good that new words are being recorded to an unprecedented extent, not only by linguists (N) but also by laymen (C20). Even if they do not get into the official dictionaries, they contribute to a more complex and apt description of contemporary language, thus capturing not only the literary form of the Czech language but also its non-standard, yet commonly spoken variant. This is especially important if we consider the actual post-diglossic situation in Czech (Bermel 2014). What is especially gratifying is that this initiative arises ‘from below,’ from amateurs who voluntarily record neologisms – for it implies that they do care about their mother tongue. After all, there is nothing new under the sun: what would the Oxford English Dictionary be without thousands of anonymous contributors?

We are aware of the disadvantages and limits of amateur lexicographers (cf. Gao 2012: 427-430, Hanks 2012: 77-82 or Lew 2014): due to a lack of scientific training, they write their definitions instinctively and subjectively, they come up with unnatural example sentences and even create their own words. Moreover, they are selective when choosing which word they will contribute to the dictionary with – originality, wittiness, and intense emotions are pursued, whereas ‘boring’ words (read: those overused by mass media) are often neglected. However, these objections can be muted by the intervention of a professional lexicographer as a dictionary’s editor who can alter problematic entries in terms of wording, examples, et cetera, or supplement missing entries, seemingly unattractive ones for contributors.
Amateurs’ straightforwardness can be, on the contrary, counted as an advantage too. By creating their own entries, they may show us their idea of user-friendly written entry and how a dictionary should look like from their perspective (cf. Meyer and Gurevych 2012: 291). Their engagement can also reveal more general notions of language itself among the public, as the metalinguistic role of language gains prominence. This, along with the unquestionable merit of zero or minimum costs, helps to outweigh the drawbacks. The synergy between lay-approaches and theory-based lexicography, supported by corpus frequency data, contributes to a richer description of (not only) neologisms.

Similarly, inspiration in the opposite direction may be desirable too: the quality of collaborative web dictionaries will certainly improve with the adoption of specific lexico- graphic standards that professionals can provide, as an upgrading process of the C20 raw material into CH suggests. More strict rules adopted in CH should subsequently affect
editing rules for C20, and with more reliable editors (than just the one at present), the whole project may become semi-professionalised. Theoretically, it is also possible to supplement C20 with links to appropriate corpora and/or N, where users would get more information about using the word in its natural context. Although content would no longer be exclusively user-generated, it would remain user-oriented.

Thus, the possibilities are open and include theoretical questions, which may have received considerable attention in Czech linguistics, but still with little use of corpus methods and data. Related to our article, one of the many potential research questions is this: which of the words posted at C20 are actually used by Czech speakers – and to the contrary, which words are actually used but yet not reflected at all by crowdsourcing, and why would that be so? It will be no less interesting to see how firmly the words from the two groups described above have become entrenched in Czech and to what extent reference materials will reflect them in the future. However, this requires a sufficient time interval as well as regularly updated corpus data.

Notes

1. Attitudes to neologisms, especially to loanwords (and especially from English), can differ within the public, from favorable to neutral to negative; cf. the recently published study regarding Slovak (Panocová 2020). A specific context common for both Czech and Slovak should be mentioned here, viz (from our perspective unfortunate) the long-term influence of the prescriptive orientation of linguistics, the residues of which persist up to now. This can also be a strong source of motivation for laymen: to be allowed to contribute their piece of knowledge without a language expert’s testimonial.

2. https://cestina20.cz/

3. Increase of the list of entries in the last 5 years: 2015 – 5,801 words; 2016 – 7,078 words; 2017 – 9,605 words; 2018 – 12,200 words; 2019 – 16,250 words. In the first half of this year (until June 7th), no less than 2,500 words were added, most often in connection with the COVID-19 pandemic (nearly 600 words).

4. We know practically nothing about the C20 contributors, except for their name or nickname and IP address, as the contribution process is entirely anonymous. In this regard, we find the following studies inspiring: Müller-Spitzer, Wolfer and Koplenig 2015; Wolfer and Müller-Spitzer 2016; Sköldberg and Wenner 2020.

5. Cf. a recently published popular book about lexicography (Lišková and Šemelík 2019).

6. http://neologismy.cz/

7. In fact, there are two Online corpora: 1) ONLINE_NOW (Cvrček and Procházka 2020a) contains data from the current month and six previous months; it is updated daily, and 2) ONLINE_ARCHIVE (Cvrček and Procházka 2020b) contains data from February 2017 to the month in which ONLINE_NOW begins; it is always updated at the beginning of the month. The corpora do not overlap in the covered periods, so for a search in the entire time range, it is enough to merge the query results from both corpora (as we did in our case studies) and further manual adjustments to remove the overlap are not necessary.

8. Born 1936, died 2011; the president of Czechoslovakia 1989-1992; the president of the Czech Republic 1993-2003. A general view of the frequency of the collocation
'Václav Havel' over time can be seen in the Word at Glance application (https://www.korpus.cz/slovo-v-kostce/search/cs/V%C3%A1clav%20Havel– section Frequency of the word during time). Two spikes relate to the end of Havel’s presidency in 2003 and his death in 2011.

9. Havl- is an alternated (e > 0) stem of a nominative singular form Havel.

10. Whenever possible, the definitions come from the contributors themselves; we quote them to illustrate the ‘peculiar handwriting’ and ‘charming tinge of amateur lexicography’ mentioned in subsection 2.1. Definitions marked by a star (*) are provided by another source, as the relevant entry is missing in C20.

11. In such sparsely documented cases, the lemmatization is a major problem for inflective languages, along with the non-diacritic version of spelling (pravdolaskarsky instead of pravdolaskarský). The results may also be skewed by frequent duplication, and some form of deduplication of O will be necessary.

12. Probably the harshest word is hadrychia´da ‘period of repression associated with sanctions and denouncing during the coronavirus epidemic, e.g., for not wearing face masks in public,’ blending the Czech words hadry, ‘togs’ and heydrichia´da ‘period of heavy repression carried out by the German Nazi occupiers during World War II in the Protectorate of Bohemia and Moravia after the death of Deputy Reich Protector Reinhard Heydrich*,’ during which more than 1,400 Czechoslovak citizens lost their lives. After all, Heydrich himself is said to have called the Czechs ‘laughing beasts’ (smějìcí se bestie).

13. Cf. skoronavirus ‘influenza so strong that you get the suspicion that you have caught the coronavirus’ from January 28th, 2020.

14. We searched in all parts of an C20 entry (lemma, definition, example of usage) by keywords, or these character strings, respectively: [ck]oron?ा?, [ck]ovid, vir[uo], kar- ant, pandem, prymul, respir, roušek. Subsequently, the lists of found entries were manually checked and cleaned, as not all of the words referred to the coronavirus. In N, we utilised the special Covid-19 filter.

15. Cf. a list of 746 entries for German Covid-19-related neologisms at https://www.owid.de/docs/neo/listen/corona.jsp#. (Accessed October 26th, 2020). For an overview of relevant neologisms in six Slavic languages (Russian, Ukrainian, Polish, Czech, Croatian, Slovene) see Będkowska-Kopczyk and Łaziński 2020.

16. In addition to the slight phonetic distinction (karantenizovat × karanténizovat ‘quarantine a larger number of people suspected of the coronavirus disease’), sometimes there is also the choice of using either the original (corona-, coro-, covid-) or the adapted spelling of new words (korona-, koro-, kovid-). N clearly prefers the first option, C20 the second one (koronakrize × koronakrize ‘the coronavirus crisis,’ cov-idový × kovidový ‘relating to a coronavirus epidemic or covid-19 disease; also in the variant covidový’). For example, there are 29 entries starting with corona- in N, but in C20, there is only one, which, furthermore, is used parodically: coronavirus ‘hangover from drinking the Mexican beer Corona.’ It can be therefore assumed that C20 contributors are less afraid of adapting loanwords to the Czech spelling.

17. Nevertheless, nouns clearly prevail (93.9%), followed by adjectives (3.5%), idioms/MWUs (1.5%), verbs (0.9%), and adverbs (0.2%). For the whole dataset, statistics is similar: nouns (81.5%), adjectives (9.6 %), idioms/MWUs (4.9%), verbs (3.7 %), and adverbs (0.3%)
18. This idiom even has its feminine version: *na evu* (literally ‘on Eve’).

19. Otherwise, Babiš is, without a doubt, the most influential Czech public official in terms of the Czech lexicon. There are more than 150 (!) entries in C20 regarding him (cf. Havel with 21 entries).

20. In this regard, *korunavirus*, with its six senses, keeps a record: ‘1. condition where you have the last few crowns left, which is manifested by clear symptoms: weariness, weakening, thirst and lack of appetite; 2. very low work ethic due to low salary; 3. weakening of the open Czech economy, caused by the deteriorating state of the economy in Germany or the EU/world, respectively; 4. weakening of the Czech koruna [CZK] against the euro; 5. trying to make the most of the coronavirus pandemic and fear of it; 6. virus that president Miloš Zeman allegedly had during the ceremonial unlocking of the crown jewels […].’

21. Such as Mark Davies’ The Coronavirus Corpus (at https://www.english-corpora.org/corona/Otherwise) or The Oxford English Corpus (available via the Sketch Engine interface), to name at least two examples.

22. The peak on the Facebook curve from the beginning of February is due to technical reasons (on February 7th, a two orders of magnitude lower volume of data was retrieved compared to other days).

23. These five collocations from C20 all refer to the Covid-19 virus: *čínská pomsta* (lit. ‘Chinese vengeance’), *čínská chřipka* (lit. ‘Chinese flu’), *čínská rýmčka* (lit. ‘Chinese sniffle’), *čínus* (< Čína ‘China’ + virus) and wuchanvir.

24. The team of an emerging monolingual dictionary of Czech (*Akademický… 2012nn.*) processed, as a matter of priority, entries of the most frequent pandemic-related words and published them online as a special list of entries (http://www.slovnikcestiny.cz/covid19.php). Only a few of them are actual neologisms, not yet recorded in any Czech dictionary: *covid-19/covid* and *koronavirový, koronavirus/koronavir,* most words have been integrated into the Czech lexicon for a long time: *epidemie* ‘epidemic,’ *imunita* ‘immunity,’ *izolace* ‘isolation,’ *karanténa* ‘quarantine,’ *pandemie* ‘pandemic,’ *respirátor* ‘respirator,’ *rouska* ‘face mask,’ *smrtnost* ‘lethality,’ *úmrtnost* ‘mortality.’ Cf. similar lists of entries for Slovene (Fran, različnica covid-19: https://fran.si/o-portalu?page=Covid_19_2020) or Croatian (*Pojmovnik koronavirusa: https://jezik.hr/koronavirus*). Prestigious online dictionaries that are updated on a regular basis had to deal with the influx of new pandemic-related lexemes during the last months. For example, Merriam-Webster’s last update in April 2020 numbered 535 new words and phrases, among them also *self-isolate, physical distancing, contactless, WFH* [‘working from home’], *PPE* [‘personal protective equipment’], *forehead thermometer, intensivist, epidemic curve, immune surveillance, community immunity, herd immunity,* or *remdesivir*. In March 2020, a special unscheduled update was made, and entries *COVID-19, coronavirus, SARS-CoV, SARS-CoV-2, MERS-CoV, nCoV, index case, index patient, patient zero, contact tracing, community spread, super-spreader, social distancing,* and *self-quarantine* were added. Cf. updates to the Oxford English dictionary that can be found at https://public.oed.com/updates/.

25. Rare exceptions ought to be named here: *Lisková 2018, Sláma 2017, 2019.*
Acknowledgements

This study was supported by the programme Progres Q08 “Czech National Corpus” implemented at the Faculty of Arts, Charles University and by the European Regional Development Fund-Project “Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World” (No. CZ.02.1.01/0.0/0.0/16_019/0000734). We also thank both anonymous reviewers for their careful reading of our manuscript and their insightful comments and suggestions. Special thanks to our colleagues Václav Cvrček and Jan Kocek for their help with data processing and visualisation.

References

A. Dictionaries

Akademický slovník současné češtiny. 2012nn. Praha: Ústav pro jazyk český AV ČR, Accessed on 28 October. http://slovnikcestiny.cz/uvod.php.

Hugo, J., M. Fidlerová, K. Adámková, Z. Juráňková (eds). 2006. Slovník nespisovné češtiny. Praha: Maxdorf.

Kavka, M. and M. Škrabal (eds). 2018. Hackmutá čestina. Brno: Jan Melvil Publishing.

Martincová, O., V. Mejstrík, I. Janovec, J. Mravinacová, Z. Opavská, A. Rangelova, J. Světlá, P. Šmidová, Z. Tichá and I. Tintěrová. 2004. Nová slova v češtině 2. Slovník neologizmů. Praha: Academia.

Obrázil, K. J. 1999–2000. Kryptadia. Príspěvky ke studiu pohlavního života našeho lidu. Praha: Paseka.

Ouředník, P. 1988. Smírbusch jazyka českého. Paris: Edice K.

B. Other literature

Bedkowska-Kopczyk, A. and M. Łaziński. 2020. ‘Covid-19 Pandemic Reflected in Slavic.’ In Greenberg, M. L., L. A. Grenoble, S. Dickey, M. U. Fidler, R. Genis, M. Łaziński, A. Peti-Stantić, B. Wiemer and N. V. Zorixina-Nilsson (eds), Encyclopedia of Slavic Languages and Linguistics Online. Leiden: Brill. Accessed on 17 January 2021. https://referenceworks.brillonline.com/browse/encyclopedia-of-slavic-languages-and-linguistics-online.

Bermel, N. 2014. ‘Czech Diglossia: Dismantling or Dissolution?’ In Árokay, J., J. Gvozdanović and D. Miyajima (eds), Divided Languages? Diglossia, Translation and the Rise of Modernity in Japan, China and the Slavic World. Dordrecht: Springer, 21–37.

Carr, M. 1997. ‘Internet Dictionaries and Lexicography.’ International Journal of Lexicography 10.3: 209–230.

Creese, S. 2013. ‘Exploring the Relationship between Language Change and Dictionary: Compilation in the Age of the Collaborative Dictionary.’ In Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (eds), Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, Tallinn, Estonia. Ljubljana / Tallinn: Trojina, Institute for applied Slovene studies / Eesti keele instituut, 392–406.

Čibej, J., D. Fiser and I. Kosem. 2015. ‘The Role of Crowdsourcing in Lexicography.’ In L., Kosem, M. Jakubiček, J. Kallas and S. Krek (eds), Electronic lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2013 Conference. Ljubljana / Brighton: Trojina, Institute for Applied Slovene Studies / Lexical Computing, 70–83.

Filáčková, S. 2016. ‘Blendy v elektronickém neologickém archivu Neomat.’ Jazykovědné aktuality 53.3–4: 100–105.
Fuertes-Olivera, P. A. 2009. ‘The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-language Internet Dictionary.’ In Bergenholtz, H., S. Nielsen and S. Tarp (eds), Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow. Bern: Peter Lang, 99–134.

Gao, Y. 2012. ‘Online English Dictionaries: Friend or Foe?’ In Fjeld, R. V. and J. M. Torjusen (eds), Proceedings of the 15th EURALEX International Congress 7-11 August 2012. Oslo: University of Oslo, 422–433.

Hanks, P. 2012. ‘Corpus Evidence and Electronic Lexicography.’ In Granger, S. and M. Paquot (eds), Electronic lexicography. Oxford: Oxford University Press, 57–82.

Kosem, I., P. Gantar and S. Krek. 2013. ‘Automation of Lexicographic Work: An Opportunity for Both Lexicographers and Crowd-sourcing.’ In I., Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (eds), Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, Tallinn, Estonia. Ljubljana / Tallinn: Trojina, Institute for applied Slovene studies / Eesti keele instituut, 32–48.

Lew, R. 2011. ‘Online Dictionaries of English.’ In P. A., Fuertes-Olivera and H. Bergenholtz (eds), E-lexicography. The Internet, Digital Initiatives and Lexicography. London / New York: Continuum, 230–250.

Lew, R. 2014. ‘User-generated Content (UGC) in Online English Dictionaries.’ OPAL – Online publizierte Arbeiten zur Linguistik 2014.4: 8–26.

Lišková, M. 2018. Neologismus v linguistickém, mediálním a běžném obrazu světa. PhD thesis, Praha: Univerzita Karlova, Filozofická fakulta.

Lišková, M. and M. Šemelík (eds). 2019. Jak se piší slovníky aneb Lexikografie pro každého. Praha: Nakladatelství Lidové noviny.

Meyer, C. M. and I. Gurevych. 2012. ‘Wiktionary: A New rival for Expert-built Lexicons? Exploring the Possibilities of Collaborative Lexicography.’ In Granger, S. and M. Paquot (eds), Electronic Lexicography. Oxford: Oxford University Press, 259–291.

Müller-Spitzer, C., S. Wolfer and A. Koplenig. 2015. ‘Observing Online Dictionary Users: Studies Using Wiktionary Log Files.’ International Journal of Lexicography 28.1: 1–26.

Niedzielski, N. A. and D. R. Preston. 2003. Folk Linguistics. Berlin – New York: Mouton de Gruyter.

Panocová, R. 2020. ‘Attitudes Towards Anglicisms in Contemporary Standard Slovak.’ International Journal of Lexicography 33.2: 187–202.

Rundell, M. 2012. ‘It Works in Practice but Will It Work in Theory? The Uneasy Relationship between Lexicography and Matters Theoretical.’ In Fjeld, R. V. and J. M. Torjusen (eds), Proceedings of the 15th EURALEX International Congress 7-11 August 2012. Oslo: University of Oslo, 47–92.

Rundell, M. 2017. ‘Dictionaries and Crowdsourcing, Wikis and User-generated Content.’ In Hanks, P. and G.-M. de Schryver (eds), International Handbook of Modern Lexis and Lexicography. Berlin / Heidelberg: Springer, 1–16.

Sköldberg, E. and L. Wenner. 2020. ‘Folkmun.se. A Study of a User-generated Dictionary of Swedish.’ International Journal of Lexicography 33.1: 1–16.

Sláma, J. 2017. ‘K (polo)automatické excerpcí neologismů.’ Jazykovědné aktuality 54.3-4: 34–46

Sláma, J. 2019. ‘K efektivitě manuální a poloaautomatické excerpte neologismů.’ Naše řeč 102.1-2: 64–75.

Surowiecki, J. 2005. The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. New York: Doubleday.

Šemelík, M. 2020. ‘Korosloví.’ Pátek Lidových novin 22 May 2020: 5.

Wolfer, S. and C. Müller-Spitzer. 2016. ‘How Many People Constitute a Crowd and What Do They Do? Quantitative Analyses of Revisions in the English and German Wiktionary Editions.’ Lexikos 26: 347–371.
C. Corpora and other electronic sources

Cvrček, V. and P. Procházka. 2020a. ONLINE_NOW: monitorovací korpus internetové češtiny. Praha: Institute of the Czech National Corpus. Accessed on 28 October 2020. http://www.korpus.cz.

Cvrček, V. and P. Procházka. 2020b. ONLINE_ARCHIVE: monitorovací korpus internetové češtiny. Praha: Institute of the Czech National Corpus. Accessed on 28 October 2020. http://www.korpus.cz.

Čeština 2.0. Accessed on 28 October 2020. http://cestina20.cz.

Databáze excerpčního materiálu Neomat. 2015. Verze 3.0. Praha: Oddělení současné lexikologie a lexicografie Ústavu pro jazyk český AV ČR. Accessed on 28 October 2020. http://neologismy.cz.

Křen, M., V. Cvrček, T. Čapka, A. Čermáková, M. Hnátková, L. Chlumská, T. Jelínek, D. Kováříková, V. Petkevič, P. Procházka, H. Skoumalová, M. Škrabal, P. Trunček, P. Vondříčka and A. Zasina. 2015. SYN2015: reprezentativní korpus psané češtiny. Praha: Institute of the Czech National Corpus. Accessed on 28 October 2020. http://www.korpus.cz.