Filozofická fakulta Univerzity Palackého

Keywords of Coronavirus Pandemic in Czech Online Newspapers

(Bakalářská práce)

2021

František Špaček

Filozofická fakulta Univerzity Palackého

Katedra anglistiky a amerikanistiky

Keywords of Coronavirus Pandemic in Czech Online Newspapers
(Bakalářská práce)

Autor: František Špaček

Studijní obor: Anglická filologie a obecná lingvistika a teorie komunikace

Vedoucí práce: Mgr. Michaela Martinková, Ph.D.

Počet stran: 57

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a uvedl úplný seznam citované a použité literatury.

V Olomouci dne   Podpis

I would like to thank my supervisor Mgr. Michaela Martinková, Ph.D. for her insight and valuable advice regarding this thesis.

V Olomouci dne   František Špaček

Abstract

The attention that the coronavirus pandemic gained from the media allowed for the creation of special-purpose corpora due to a high number of online articles mentioning the coronavirus. This thesis focuses on developing a method for downloading relevant articles, creating corpora from the data collected from the first and second wave of the pandemic, and comparing these corpora. The first and second wave corpora are compared against each other from a lexical point of view.

Keywords

coronavirus corpus, covid corpus, exploratory research, coronavirus keywords, covid keywords, Czech coronavirus keywords, Czech covid keywords

Anotace

Mediální pozornost, kterou koronavirová pandemie získala, umožnila díky vysokému počtu online článků zmiňujících koronavirus vytvoření specializovaných korpusů. Tato diplomová práce se zaměřuje na vývoj metody pro stahování relevantních článků, vytváření korpusů z dat nasbíraných z první a druhé vlny pandemie a porovnání těchto korpusů. Korpusy z první a druhé vlny jsou porovnány mezi sebou z hlediska lexikálního.

Klíčová slova

koronavirový korpus, covidový korpus, průzkum, koronavirová klíčová slova, covidová klíčová slova, česká koronavirová klíčová slova, česká covidová klíčová slova

Contents

1 Introduction

This thesis aims to analyze and compare keywords from articles from three Czech online newspapers – Lidovky.cz, idnes.cz and Novinky.cz – written during the first and second wave of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The keyword analysis will be performed on corpora made from these articles by using Sketch Engine (Kilgarriff et al. 2014), and the results for the first and second wave will be compared.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1] is the virus responsible for the highly infectious illness that first occurred in late 2019, and the situation was declared a global pandemic by the World Health Organization on 11 March 2020. With its emergence, newspaper publishers all over the world started publishing articles concerning the pandemic situation. The first confirmed cases of coronavirus in the Czech Republic were reported on 1 March 2020, and there has been a great number of articles written since then, making corpus analysis possible.

Sketch Engine is an online software for managing and creating corpora that offers useful tools for text analysis. These tools include keyword extraction, which is used to define the main topic of the selected corpus by comparing it to a reference corpus and analyzing the differences between these two.

Although Sketch Engine offers the option to create corpora from web, it often returns unsatisfactory results that include irrelevant data for this research, such as comments sections below the articles or links to other articles. Therefore, a custom-made web-scraping script was developed for extracting only the relevant data from relevant articles from the newspaper websites mentioned above. The criteria for deciding what is and what is not a relevant article will be discussed later. The script was made using the Python programming language and it is programmed to perform all necessary text pre-processing tasks to get data in a format which can be used in Sketch Engine. Data from each article is saved as a plain text file and then these files are divided into two sets representing the first and second wave of the coronavirus.

To reiterate, the main goals of this thesis are to develop a method for automatic data gathering from newspaper web articles, create special-purpose corpora and compare these corpora in terms of their keywords to identify lexical differences in the description of the two waves.

[1] The name “coronavirus” will be used in the rest of the thesis to describe Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

2 Literature review

This section reviews the literature related to keyword analysis and corpus linguistics.

2.1 Keywords, keyness and other measures

In corpus linguistics, a keyword is defined as “a term for a word that is statistically characteristic of a text or set of texts” (Culpeper and Demmen 2015, 90), with keyness being the measure of how “key” the given keyword is. Keyness is a textual quality, which means that different words might be “key” in different texts, and different texts might have different keywords (Scott and Tribble 2006, 56).

2.2 Requirements for keyword analysis

For analyzing keywords, one first needs to meet some requirements. These requirements are:

1. a focus corpus,
2. a reference corpus,
3. corpus annotation tools (such as lemmatizer and part of speech tagger),
4. a metric for calculating keyness and identifying keywords. (Kilgarriff et al. 2014, 25-26)

2.2.1 Focus corpus and reference corpus

Focus corpus is a corpus that is the subject of analysis. Reference corpus is a corpus that the focus corpus is compared with. Although the choice of the reference corpus is important, “the size of reference corpus is not very important in making a keyword list” (Xiao and McEnery 2005, 70). According to Culpeper and Demmen (2015, 97), there is not a clear consensus of what is a sufficient reference corpus size, but they also mention that “reference corpora are typically the same size as the target corpus, or very much larger.” The size of a reference corpus does not seem to be very important concerning keywords, but the content of a reference corpus does (Scott, 2009).

2.3 Calculating keyness

Since keyness is a statistical measure, there are several different methods for calculating it. Sketch Engine uses a rather, as the name would suggest, simple method for calculating keyness called Simple maths developed by Kilgarriff (2009). Simple maths uses the following formula:

$$\frac{fpm_{focus} + n}{fpm_{ref} + n}$$

“where fpm_{focus} is the normalized (per million) frequency of the word in the focus corpus, fpm_{ref} is the normalized (per million) frequency of the word in the reference corpus, n is so-called smoothing parameter” (Sketch Engine 2020, www.sketchengine.eu).

The smoothing parameter is there mainly for cases when there are no occurrences of the given word in the reference corpus – in that case there would be division by zero, which is impossible. The simplest solution is to add one (Kilgarriff 2009, 2), which is also the default value in Sketch Engine. Choosing a different value for n would put focus on different parts of the vocabulary, with higher values focusing on more common words and lower values focusing on rarer words (Kilgarriff 2009, 3).

2.4 Statistical significance

Another important measure in keyword analysis is statistical significance which is a measure that says whether a difference in frequencies of a given phenomenon in different samples is significant. Although many programs used for keyword analysis come with built-in significance testing, Sketch Engine uses only the already mentioned Simple maths model that does not work with statistical significance. For this reason, an online calculator (Český národní korpus, n.d.) of statistical significance was used to test the significance of selected keywords.

While statistical significance “reveals whether we have a sufficient amount of data,” it does not reveal if the observed differences are relevant and carry some useful information for description (Fidler and Cvrček 2015, 227). When ranking keywords by statistical significance, it is possible to get misleading results, as the metrics place words with less different relative frequencies higher than words with more different relative frequencies if the difference in total frequency is larger. This was illustrated by Fidler and Cvrček on a model scenario, see tables below.

Table (1): Model scenario with word A

	Corpus 1	Corpus 2
fq(A)	130	100
N	100,000	100,000
RelFq(A)	0.013	0.010

Table (2): Model scenario with word B

	Corpus 1	Corpus 2
fq(A)	1100	1000
N	100,000	100,000
RelFq(A)	0.011	0.010

In the case of the word A, there is a 30% difference in the frequency of the word between the two corpora, while in the case of the word B, the difference is only 10%. In both cases, the difference is significant, but log-likelihood test indicates that the difference for the word B is more significant. Despite the relative frequencies of the word B being closer than of the word A, ranking these words by statistical significance would rank the word B higher than the word A (Fidler and Cvrček 2015, 227-228).

2.5 Effect size

There is also a third important measure concerning keywords, namely effect size. Effect size is a measure showing how relevant a difference between two phenomena is (Pojmy:din - Příručka ČNK, n.d.). Even though the words *significance* and *relevance* might sound somewhat similar in this context, they cover different concepts and fill different needs. As was mentioned in the previous chapter, significance does not reveal if the observed differences have some descriptive value. If there is enough data, even a small change might be considered to be significant, but that does not mean it is relevant (Pojmy:din - Příručka ČNK, n.d.). For these reasons, there is the effect size.

The effect size measure is also part of the online calculator (Český národní korpus, n.d.) mentioned earlier. It uses the difference index (DIN) proposed by Fidler and Cvrček (2015) and uses the following formula:

$$DIN = 100 \times \frac{fpm_{focus} - fpm_{ref}}{fpm_{focus} + fpm_{ref}}$$

The resulting value always ranges from -100 to 100 and is interpreted in the following way:

Table (3): Values of DIN (Fidler and Cvrček 2015, 230)

–100	The word is present only in the reference corpus and not in the focus corpus
0	The word occurs equally often in the focus corpus and reference corpus (with
100	The word is present only in the focus corpus

Furthermore, Fidler and Cvrček (2015, 230-231) write that DIN was made to be immune to the problem when a word is present only in the focus corpus (Simple maths solved this with the smoothing parameter, see chapter 2.3) and it gives the same value (DIN = 100) for all keywords found only in the focus corpus. They admit that this is not very helpful in deciding whether the absent words are relevant, but it lets the researcher know that keywords like these need special attention.

3 Methodology

As was already mentioned in the introduction, one of the goals is to develop a method for collecting articles from online newspapers. The creation of the article-collecting script is described in section 3.1. Another goal is to then extract keywords from these articles and to compare the keywords from the first wave with the keywords from the second wave to observe if there was any language change. Sections 0-3.9 are focused on the keyword extraction and data processing. The last step is to compare the data, which is done in section 4.

Many examples of Czech words appear in the following sections – except for proper nouns, all Czech words and words with different spelling in English are translated and glossed. The words in Czech are followed by their POS, the English words include the POS only when the POS is not clear from the isolated word. As most of the example words are nouns, only POS other than nouns are included in the gloss.

3.1 The Collected Corpora

Table (1Table (4) below shows the number of tokens (punctuation, numbers, abbreviations or anything else between two spaces) and words (tokens starting with a letter) as well as the number of files (i.e., individual articles) that make up the corpora. It also includes the individual sections of the newspapers that were used for data collection, and the time span that the corpora cover.

Table (4): Information about the collected corpora

Corpus	Tokens	Words	Files	Time span	Sections included
IDNES_1	2,208,699	1,861,209	3,645	1.3.2020-30.6.2020	Zprávy, Domácí, Věda (News, Home, Science)
IDNES_2	2,193,541	1,850,076	3,462	1.8.2020-28.2.2021	
LIDOVKY_1	887,743	749,205	1,623	1.3.2020-30.6.2020	Domov, Svět (Home, World)
LIDOVKY_2	655,468	554,491	1,312	1.8.2020-28.2.2021	
NOVINKY_1	1,521,257	1 260,854	3,475	1.3.2020-30.6.2020	Koronavirus
NOVINKY_2	1,454,787	1 202,215	2,987	1.8.2020-28.2.2021	

As can be seen from the table, all corpora covering the second wave are smaller despite covering a longer time span, which is rather unexpected.

The LIDOVKY corpora are considerably smaller than the other corpora, only about a half the size of the other ones. One possible explanation is that since Lidovky.cz is just an extended online version of a printed newspaper, one could argue that the newspaper tries to maintain some level of prestige and does not report on so many things and events as the other two newspapers that are purely online. However, that is only a mere speculation and deciding whether the newspapers are doing tabloid journalism or not goes beyond the scope of this thesis. The following section describes where and how the data were collected.

3.1.1 Data collection

A custom-made script for each newspaper server was made using Python programing language. As of March 1, 2021, Python is listed as the most used programming language in the world (PYPL index, n.d.) which makes finding resources and guides online rather easy. Python also offers a wide range of libraries, which are collections of resources used in software development that let the programmer use pieces of already existing code rather than making everything from scratch on their own, thus saving time.

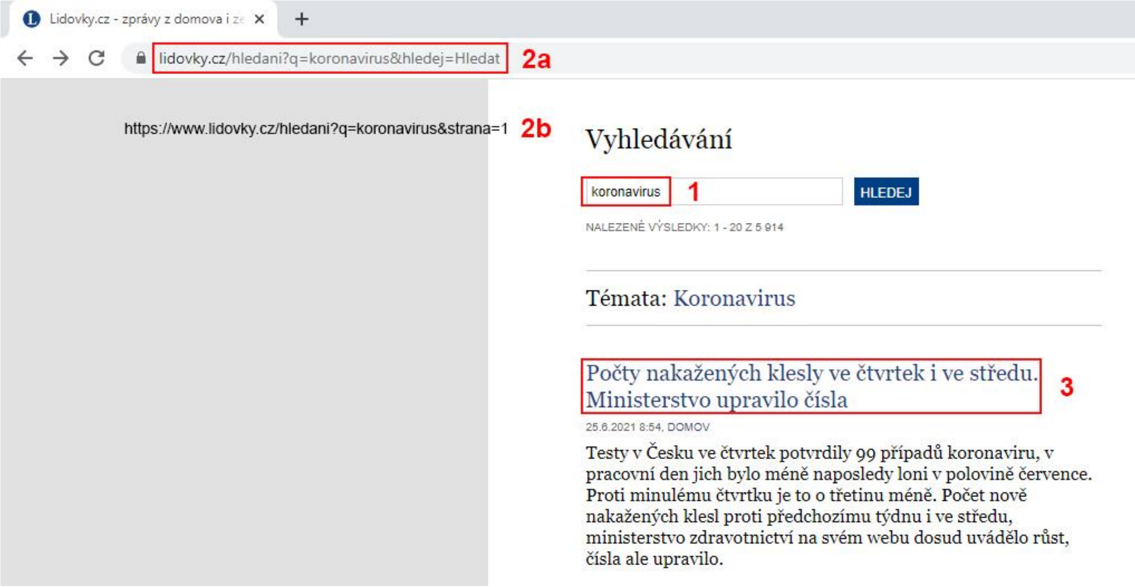
Although it would quite likely be possible to make a universal script that would work for all the website servers used in this thesis, it was not the main point of the thesis, and it would also most likely require much deeper programming knowledge and a lot of time. Therefore, each newspaper server has its own script. The scripts are fundamentally the same, but every website had its specific problems that needed to be solved. These will be described in the following section.

3.1.2 Lidovky.cz and idnes.cz

Lidovky.cz and idnes.cz are operated by the same company, which is reflected in their websites; they are very similar. Therefore, the respective scripts are mostly the same.

First, the user needs to decide on what keyword to use for searching the articles on the newspaper website. In this case the keyword “koronavirus” (number 1 in Figure 1) is used for both websites. Then the user needs to slightly modify the URL of the search results (number 2a in Figure 1) and this modified URL (number 2b in Figure 1) serves as the first input. Both the original and modified URL lead to the same page, but the modified URL allows to go through all the search results by simply changing the number in the URL, which is what the script does. Apart from dividing the articles by their publication date the script will do all of the work on its own from this point.

Figure 1: keyword used for searching articles, search results, and URL



lidovky.cz/hledani?q=koronavirus&hledej=Hledat

https://www.lidovky.cz/hledani?q=koronavirus&strana=1

Vyhledávání

koronavirus

NALEZENÉ VÝSLEDKY: 1 - 20 Z 5 914

Témata: Koronavirus

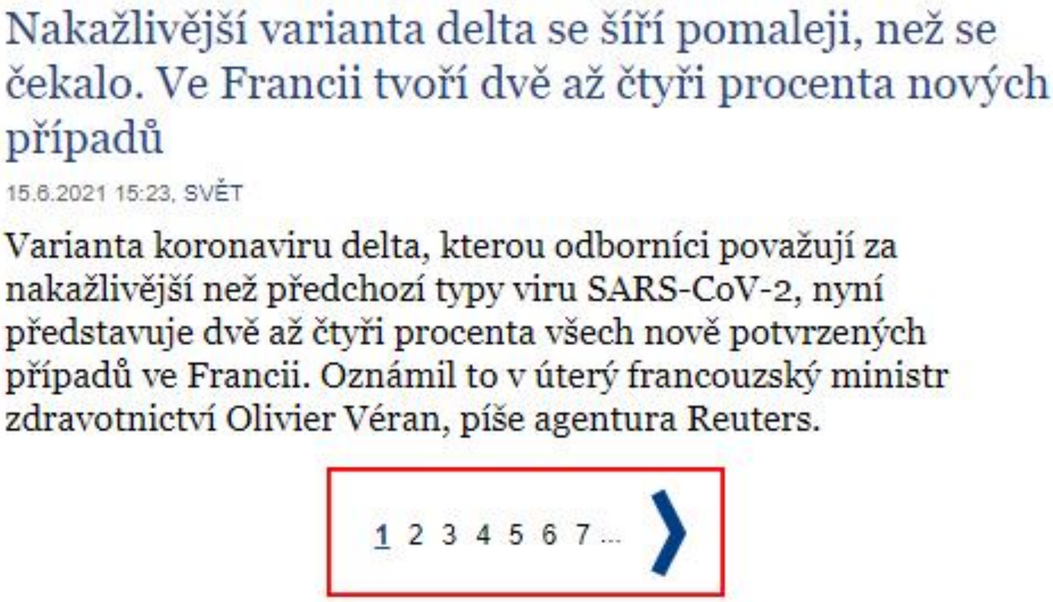
Počty nakažených klesly ve čtvrtek i ve středu. Ministerstvo upravilo čísla

25.6.2021 8:54, DOMOV

Testy v Česku ve čtvrtek potvrdily 99 případů koronaviru, v pracovní den jich bylo méně naposledy loni v polovině července. Proti minulému čtvrtku je to o třetinu méně. Počet nově nakažených klesl proti předchozímu týdnu i ve středu, ministerstvo zdravotnictví na svém webu dosud uvádělo růst, čísla ale upravilo.

Both Lidovky.cz and idnes.cz present their search results in the form of a numbered list with 20 articles per page. The script searches the webpage for the URLs leading to these pages (Figure 2) with other search results and saves all of them. Then it opens the URL with page number higher by 2 than the current page number (i.e., it goes from page 1 to page 3, then page 5 and so on) and saves the new URLs (Figure 3).

Figure 2: pages of search results to be saved

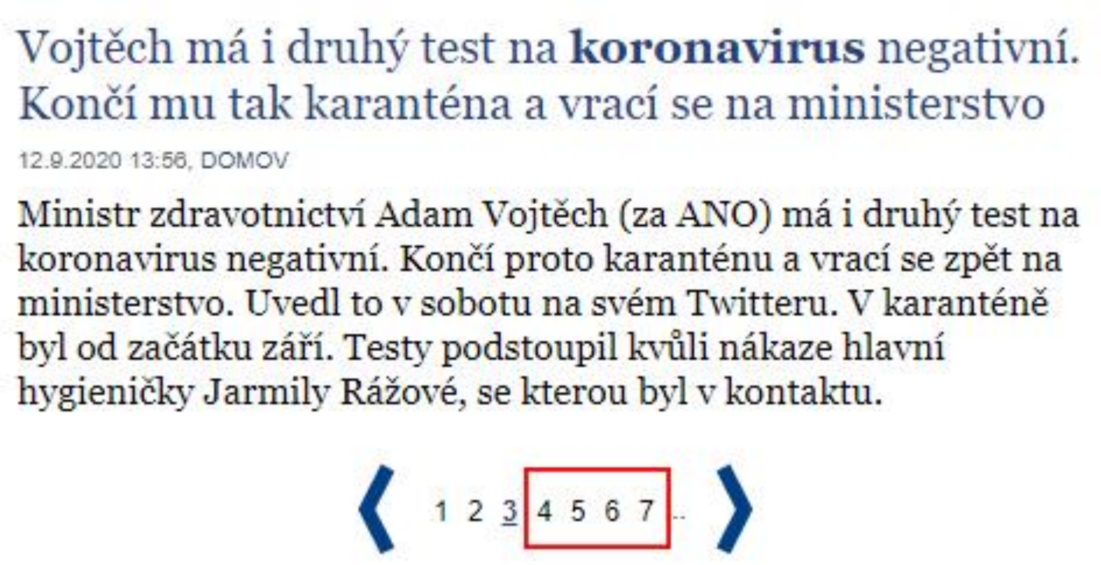


Nakažlivější varianta delta se šíří pomaleji, než se čekalo. Ve Francii tvoří dvě až čtyři procenta nových případů

15.6.2021 15:23, SVĚT

Varianta koronaviru delta, kterou odborníci považují za nakažlivější než předchozí typy viru SARS-CoV-2, nyní představuje dvě až čtyři procenta všech nově potvrzených případů ve Francii. Oznámil to v úterý francouzský ministr zdravotnictví Olivier Véran, píše agentura Reuters.

Figure 3: pages of search results to be saved after opnening a new URL



Vojtěch má i druhý test na **koronavirus** negativní. Končí mu tak karanténa a vrací se na ministerstvo

12.9.2020 13:56, DOMOV

Ministr zdravotnictví Adam Vojtěch (za ANO) má i druhý test na koronavirus negativní. Končí proto karanténu a vrací se zpět na ministerstvo. Uvedl to v sobotu na svém Twitteru. V karanténě byl od začátku září. Testy podstoupil kvůli nákaze hlavní hygieničky Jarmily Rážové, se kterou byl v kontaktu.

In the next step, the script opens all the URLs saved in the previous step, finds all the URLs leading to the individual articles (number 3 in Figure 1) and again saves all of them.

The initial search found all the articles containing the selected keyword, but it returned way too many results and articles from unwanted sections such as “Sport” were included. Therefore, in the following step, the saved article URLs were filtered. Each article URL contains the section the article was published in, which made the filtering rather easy. Articles from sections “Domácí” and “Svět” were kept in case of lidovky.cz and sections “Zprávy”, “Domácí” and “Věda” were kept for idnes.cz, which has articles divided into more sections than Lidovky.cz. As a result, more sections were selected for keeping. It is assumed that articles from sections such as “Sport” are not concerned with the coronavirus specifically, but rather with the impact that the coronavirus had on sports, which is not the primary focus of this thesis. Paid articles from all sections were filtered out in both cases.

In this step, the script opens the individual article URLs and script saves the whole HTML code from the webpage into a separate file with a unique name. Because file names cannot contain certain special characters, such as forward slashes used in URLs, the unique file name is created with a hash function that takes the article URL and creates a string of 64 characters, making a unique file name for each article. This step is not completely necessary, but it makes the whole script run faster, because webpages such as these have some sort of bot protection, which temporarily interrupts the connection to the server after too many connection attempts were made in a short period of time. Because of this protection, after loading each article from the website, there is a six second pause to avoid interrupting the connection. It was expected that the script would be run more than once, and this is where the saved HTML code comes into play – the script uses the hashing function on each article URL and checks if a file with that specific name exists, and if it does, it loads the article HTML from the file rather than the webpage instead, avoiding the six second pause. With thousands of articles being processed each time the script is run, this step saves a lot of time.

After the HTML code is saved, the script filters out unwanted elements that are present in the article itself, such as twitter posts, polls, or maps. After filtering these out, the article text is found and saved into a plain text file whose name is made from the article's timestamp, making division of the texts later on possible.

3.1.3 Novinky.cz

The fundamentals of the script for this newspaper server remain the same, but the webpage has one major difference making the functionality of the script different and suboptimal.

While Lidovky.cz and idnes.cz presented search results in a numbered list with a limited number of articles per page, novinky.cz webpage presents its articles in a form of a continuous list, meaning there is a button at the bottom of the list that only loads 20 new articles and makes the list longer. Each time the new articles are loaded all article URLs are saved. After 100 articles were loaded and the button to load more articles pressed, 20 articles from the top of the list are removed and 20 new articles are added at the bottom, which creates a moving window of 80 articles. Since all article URLs are saved, these 80 articles from the moving window would be present more than once in the final list of article URLs. For that reason, they are filtered out each time new URLs are saved, making this script run considerably slower than the other two.

Novinky.cz webpage has a dedicated coronavirus section, therefore no search of articles by a keyword and result filtration was necessary as was the case above, although some articles might belong into a section that would be filtered out in the case of the other scripts. The steps including saving the whole HTML code, removing unwanted elements, and making a filename from the article's timestamp remained the same.

However, an issue with this newspaper server occurred later in the process of data collection. It turned out that this website has a variable HTML code, meaning that it changes over time. The script uses names of certain HTML elements to find specific parts of the webpage (such as the article URLs or article texts), but since the names are changed after some time, it renders the script useless after a few weeks. Although it would be possible to make an updated version every time the script is needed, it would be highly impractical. Even after updating it there remains another issue – the updated version of the script would not work with the older saved HTML codes, therefore, the user would need to delete the files with HTML codes and then access each article from the website rather than from the saved HTML, thus eliminating the option to skip the 6 second pause needed after each article. The most recent update was done on March 1, 2021 to collect data until the end of February 2021.

3.2 Errors in text pre-processing

During the development of the script, only small numbers of the articles were downloaded to save time while debugging, during which the script was run after each adjustment to see if it functions correctly. After each iteration, the texts were either examined directly or uploaded into Sketch Engine to perform a keyword extraction on a small sample to see if there were any more adjustments to be made.

After examining the data, one could see that there were some errors in the texts. These errors can be divided into two major groups, each of them will be discussed in the following chapters.

3.2.1 Lowercase letters followed by uppercase letters

One of the most frequent issues that occurred during the text pre-processing was lowercase letters followed by uppercase letters. Since the script saves the text from the articles into plain text file format, which does not support text formatting, the individual paragraphs from the articles were joined into a single block of text. Some of the text files contained sentences with words that had a lowercase letter followed by an uppercase letter – a result of joining the last sentence of a paragraph with the first sentence of the following paragraph.

Considering that this issue concerns only the first and last sentences of paragraph, and these sentences need to have a full stop missing on top of that for the error to occur, only a small number of sentences are affected. Despite this fact, this issue was decided to be resolved because it was easy to fix – a step that adds a space between every lowercase letter followed by an uppercase letter was added to the script.

However, this fix had some unforeseen consequences as different issues emerged after implementing it. It turned out that many companies have a lowercase letter followed by an uppercase letter in their name, and the solution mentioned above affected these words as well. One example of this would be the name of the company *BioNTech* that was divided into words *bio* and *ntech*. Although not ideal, one can assume that this will not affect the results in a major way. The word *ntech* is nonsensical and it is safe to assume that it will not appear anywhere else, therefore, the total number of occurrences of the word *ntech* should represent the total number of occurrences of the whole company name. This is similar for many other company names and abbreviations appearing in the data. In any case, this is something that would be worth improving in future versions of the script.

3.2.2 Foreign words from social media posts

Although it is not uncommon for non-Czech words to appear in the keyword list, as will be seen during keyword analysis, some words stood out. Some of the articles cited social media posts in English or Russian/Ukrainian, and the contents of these posts were included in the saved files. Letters of Cyrillic alphabet were easy to distinguish from the rest of the keywords, and the unwanted English words were common words, such as pronouns or prepositions. Since all the texts are in Czech, it was safe to assume that words like these are not supposed to be keywords. Even though these words do not affect the results in a major way, they should not be present in the final keyword list and filtering out the social media posts that contained them would be another useful improvement for future version of the script.

3.3 Dividing the data

Since the focus of this thesis is to compare the language of the first and second wave of the coronavirus pandemic, the collected data from each server needed to be divided. Even though one could argue, after looking at the statistics of infected people (Onemocnění aktuálně, n.d.), that there were more than two coronavirus waves, the collected data was divided only into two parts for simplicity and to have two larger sets of data for each server rather than having more smaller sets. That way, a more reliable analysis can be performed, because a small size of a corpus may have a negative effect on its usefulness (Weisser 2016, 31). However, it is important to note that with a more specific corpus, the requirements on its size are less important, which makes it possible to have a smaller corpus (Weisser 2016, 31).

March 1, 2020 was selected as the start of the first wave, as it was the day when the first three coronavirus cases were confirmed in the Czech Republic (Aktuální inormace o COVID-19, 2020). The end of the first wave was not decided by some specific event but was chosen somewhat arbitrarily while looking at the graph of infected people (Onemocnění aktuálně, n.d.) to be June 30, 2020, as the number of infected people was quite steadily low. August 1, 2020 was chosen as the first day of the second wave, again somewhat arbitrarily, but it was around that time when the number of infected people began to steadily grow. The end of the second wave was decided to be February 28, 2021. Although there will without a doubt be more data to collect, the data collection had to be ended to proceed with the analysis.

There is a one-month break between the first and the second wave to have a clear division between the waves rather than having continuous sets of data.

3.4 Processing the data in Sketch Engine

After the collection and division of data was done, the data was uploaded into Sketch Engine. Two corpora, one from the articles from the first wave, and one from the articles from the second wave, were created for each newspaper server, making six corpora in total. All of the processing, including lemmatization and Part of speech (POS) tagging, was done by the Sketch Engine software. However, automatic data processing is not without error, therefore, some manual correction was done later on in the process. The specific corrections are discussed in section 3.7.

3.4.1 Errors in annotation

The two most important parts of corpus annotation in this paper are POS tagging and lemmatization. POS tagging is deciding whether a word is a noun or verb etc., and lemmatization is merging inflected word forms into one lemma, or dictionary form, to allow for their analysis as a single item. POS tagging is important for lemmatization, and lemmas are the word forms used for keyword extraction. An example of why correct tagging is important would be the word *září*, which in this form can be either a noun (September), or a verb (third person singular of the verb *zářit* – to shine). There were several cases of the noun form being tagged as the verb form instead, leading to incorrect lemmatization of the given word.

Although McEnery writes that some of the tools' accuracy rates are really high for languages like English or Spanish, with error rates at around 3 per cent (2012, 8), which is rather low, Sketch Engine does not specify the method and its accuracy of their annotation tools, so one cannot just assume such a high rate of accuracy. On top of that, the language in question is Czech which has more word inflection than English, making annotation more difficult (Fonseca 2019).

Another issue with lemmatization is foreign words and names. With the corpora settings set for the Czech language, the software consistently struggled with lemmatizing non-Czech, but even some Czech words correctly. It was most commonly names of companies and foreign people. With these words there was not an issue with tagging homonyms incorrectly, but rather with the assignment of the correct lemma. An example of a word like this would be *Pfizer* or *koronavirus* (coronavirus). Word forms occurring with the word *koronavirus* (coronavirus) include: *koronavir, koronavirma, koronavirem,* and *koronaviro* (word forms taken from Table (8)).

The last problem was with some Czech verbs, for example *očkovat*$_{\text{V}}$ (to vaccinate), specifically not lemmatizing some forms of the verb correctly and not converting them to infinitive form, which resulted in having both *očkovat*$_{\text{V}}$ (to vaccinate, infinitive) and *očkován*$_{\text{V}}$ (vaccinated, masculine passive) in the list of keywords.

3.5 Sketch Engine settings, significance testing, and difference index

3.5.1 Sketch Engine settings

Section 2.3 focused on calculating keyness and mentioned that Sketch Engine uses the so-called smoothing parameter. It is not exactly clear what should be the right value of this parameter, however, for the purposes of analysis in this thesis, the default value of $n = 1$ seemed to be suitable. Results with this value show words from both sides of the spectrum, but leaning towards the rarer side, and therefore showing many words that are not present in the reference corpus.

3.5.2 Significance testing

The online calculator mentioned in section 2.4 offers three different significance tests – Chi-square test, Fisher's exact test and Log-likelihood. The Chi-square test is the default one and it was used for testing in this paper. Although it is said that the Chi-square test does not work well with small samples (Pojmy:chi2 - Příručka ČNK, n.d.), changing the significance test did not seem to change the results. The user can also change the p-value, but it was left on the default value of 0.05.

3.5.3 Difference index

For the reason mentioned in section 2.5, the DIN measure is not used to rank the keywords, as it would place all the words not found in the reference corpus above all the other words, which would be an extreme change to the ranking, see Table (5) and Table (6) below for an example. Freq is frequency, F is focus corpus, R is reference corpus, FPM is frequency per million, SCORE is keyness value from Sketch Engine and DIN is difference index value.

Table (5): Keywords ranked by SCORE value

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN
1	antigenní$_{\text{ADJ}}$ *antigen*$_{\text{ADJ}}$	750	0	341.9	0	342.9	100
2	blatný	1,371	4	625	1.8	222.7	99.42
3	pfizer	491	1	223.8	0.5	154.8	99.60
4	lockdown	324	0	147.7	0	148.7	100

Table (6): Keywords ranked by DIN value

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN
1	antigenní$_{\text{ADJ}}$ *antigen*$_{\text{ADJ}}$	750	0	341.9	0	342.9	100
2	lockdown	324	0	147.7	0	148.7	100
3	pes *dog*	303	0	138.1	0	139.1	100
4	ntech	256	0	116.7	0	117.7	100

Table (5) shows keywords ranked by the SCORE values, while Table (6) shows keywords ranked by the DIN values. Two words are found in both tables, from which one changed its rank, and one remained the same, which does not seem to be an extreme change. To prove the extremity of the change, Table (7) with former top words follows.

Table (7): Former top keywords

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN
160	pfizer	491	1	223.8	0.5	154.8	99.60
161	blatný	1,371	4	625	1.8	222.7	99.42

The keywords *Pfizer* and *Blatný* that were formerly in the top four keywords are now at rank 160 and 161, respectively. That is quite a radical change to say the least. One way to approach this would be to reject all the words where DIN = 100, but that would mean also rejecting some important words, as all the words from a sample as little as the top four keywords (Table (6)) are related to the pandemic.

Furthermore, using both SCORE and DIN values to rank the keywords had no effect on the rankings. Therefore, DIN is used only as a secondary measure to see how prominent a keyword is. The wiki page of the Czech national corpus also states that for texts up to 20,000 words DIN value in the range 75-100 is a marker for a possibly prominent and relevant unit (Pojmy:din - Příručka ČNK, n.d.). It is not clear whether "texts up to 20,000 words" mean each subcorpus present in a corpus, or a whole corpus. Since the corpora created for this thesis do not contain any subcorpora, it would be wise to assume that "texts" mean whole corpora. However, as DIN is not extensively used yet (Fidler and Cvrček 2015, 226), no information about texts above 20,000 words was found. For this reason, DIN value of 75 will be used as a threshold for highly prominent units.

3.6 Choosing the right reference corpus

Using a general-purpose corpus as the reference corpus for the purposes of this thesis would not yield very interesting results, as the aim is to compare the first and second wave of the coronavirus pandemic. One can assume that comparing the special-purpose corpora used in this thesis with a general-purpose corpus would give very similar results for both waves, as one can expect some shared vocabulary between the first and second wave. This assumption can be easily verified by comparing keyword lists of all the created corpora made using a general-purpose corpus as a reference corpus and looking at the results.

In Table (8) below, there are top 15 keywords for each corpus. The size of these corpora ranges from 650 000 to 2 400 000 tokens, the reference corpus is csTenTen2017 (Suchomel 2018) – a Czech corpus consisting of internet texts whose size is over 12,5 billion tokens. No metrics such as the keyness score are included in this table as it serves only as an illustration of how similar the results would be if a general-purpose corpus was used as the reference corpus. The keywords here are also not glossed for the same reason.

Table (8): Top 15 keywords with csTenTen2017 as the reference corpus

IDNES_1	IDNES_2	LIDOVKY_1	LIDOVKY_2	NOVINKY_1	NOVINKY_2
koronavir*	koronavir*	koronavir*	koronavir*	koronavir*	koronavir*
covid-19**	covid-19*	covid-19**	koronavirus*	covid-19**	covid-19**
koronavirus*	covid	koronavirus*	covid-19**	koronavirus*	koronavirma*
karanténa	koronavirus*	koronavirma*	covid	koronavirma*	covid
pandemie	covidu-19**	nakažený	covidu-19**	nakažený	nakažený
koronavirma*	pandemie	koronavirem*	covidem-19**	koronavirem*	covidu-19**
nakažený	nakažený	pandemie	nakažený	karanténa	koronavirus*
rouška	covidem-19**	karanténa	pandemie	pandemie	covidem-19**
koronavirem*	rouška	covidem-19**	koronavirma*	koronaviro*	pandemie
koronavirový	antigenní	nákaza	koronavirem*	rouška	sars-co
respirátor	karanténa	koronavirový	koronavirový	respirátor	koronavirem*
koronaviro*	respirátor	koronaviro*	karanténa	sars-co	v-2
nákaza	koronavirový	rouška	pendler	nákaza	vakcína
epidemiolog	covidový	sars-co	antigenní	v-2	rouška
prymula	prymula	respirátor	sars-co	koronavirový	karanténa

* These words were incorrectly lemmatized and should be a single lemma. This is discussed more thoroughly in section 3.7.
** Same case as before.

Although it certainly would not be impossible to create a viable keyword list with csTenTen2017 as the reference corpus, using a special-purpose corpus seems to be the better way. For this reason, the first wave corpus is compared with the second wave corpus and vice versa, which results in a keyword list that does not include the shared portion of the vocabulary, but only the differences. The resulting keyword lists are shown in Table (9) below.

Table (9): Top 15 keywords with the second respective corpus as the reference

IDNES_1	IDNES_2	LIDOVKY_1	LIDOVKY_2	NOVINKY_1	NOVINKY_2
stoltenberg	antigenní	cuoma	blatný	zotavený	blatný
dluhopis	blatný	litovel	antigenní	hypotéka	antigenní
cuoma	lockdown	ruslan	peso	litovel	lockdown
letadlový	peso	cummings	pfizer	uničův	ntech
hydroxy-chlorochin	pfizer	uničův	norek	pevninský	pfizer
repatriační	ntech	hcq	sputnik	chlorochin	sputnik
mlékárna	naočkovat	výletní	bio	splatnost	peso
nájemník	pfizra	litoměřice	lockdown	splátka	naočkovat
pevninský	zeneca	macao	skóre	autosalon	skóre
an-124	skóre	salomon	ntech	hyunda	pfizra
vojvodství	očkován	raab	index	ušít	zenek
amnestie	očkovat	herbst	moderna	splácení	očkovat
námořník	sputnik	slobodník	zpěv	nošovice	ivermektin
ruslan	očkovaný	olomoucko	zmutovaný	darkov	vakcinace
velikonoce	stříkačka	dluhopis	říjnový	benátky	naočkovaný

Comparing Table (8) and Table (9), it is clear that the tables are very different. While a large portion of the keywords in the first table was the same, the keywords extracted from each couple of corpora in the second table are very different. The only similarities in the second table are found between the first corpora from each couple and the second corpora from each couple, which is an expected result as these corpora cover the same time span.

It is worth noting that the word *koronavirus* (coronavirus) does not appear as a keyword in Table (9). This is due to the fact that the word *koronavirus* (coronavirus) was used as the keyword for searching the articles. One can then expect that most of the articles contain this keyword and that there is not a significant difference in the absolute frequency of this word in the corpora from Table (9). Therefore, it does not appear as a keyword in Table (9), even though it does in Table (8).

3.7 Manual corrections and cleaning up the data

Section 3.2 focused on errors in the text pre-processing, i.e., preparing the texts for Sketch Engine. Correcting those kinds of errors was just a matter of tediously going through the portion of keywords that was going to be analyzed and either identifying words not belonging there and deleting them or finding the other half of a word that was originally one word but was divided into more parts (see chapter 3.2.1). These errors could be fixed by improving the script, and for that reason this chapter will talk only about errors in lemmatization, in other words errors made by Sketch Engine. As was mentioned above, the automatic corpus annotation is not perfect, but it works well enough to make manual corrections possible if needed (McEnery 2012, 8), because cleaning up a few mistakes made by a software is considerably more realistic than annotating the whole corpus alone. The annotation software struggles with foreign words, but also with names and especially with names of companies that are often coinages. This would probably not be the case for English corpora, but since Czech uses more inflection than English, and these words have an unusual form, the software has a hard time deciding what the correct lemma is supposed to be. This is illustrated in Table (10) below.

Table (10): Examples of wrong annotation – multiple forms of Pfizer in top 300 keywords from IDNES_2 corpus

Rank	Word	Focus Frequency	Reference Frequency	SCORE
5	pfizer	222	0	102.2
8	pfizra	255	1	80.7
177	pfizero	14	0	7.4

One can see how much the annotation software struggles with non-Czech words as the word was incorrectly lemmatized more frequently than correctly.

However, in cases such as these, it was quite easy to recognize what the correct lemma is supposed to be and fixing it therefore was not an issue – it was only a matter of finding all the individual word forms, adding up their frequencies in the focus and reference corpus, and calculating the keyness score again using the formula from chapter 2.3. Table (11) below shows the same lemma after corrections.

Table (11): Lemma Pfizer in IDNES_2 after corrections

Rank	Word	Focus Frequency	Reference Frequency	SCORE
4	pfizer	491	1	154.8

In this case, the correction moved the lemma Pfizer one rank higher in the keyword list, from rank 5 to rank 4, which is a notable change. Also, the total frequency more than doubled compared to the correct lemma from the previous table and it gives a more accurate reflection of what the newspapers wrote about.

However, only words that seemed relevant to the pandemic were treated this way. There were some words that were left uncorrected simply because they seemed unimportant – mostly names of people and geographical locations. This is discussed in more detail in section 3.9.1.

In a few cases, there were some names of companies consisting of two words in the keyword lists. Since all words separated by a space are individual tokens, Sketch Engine has no way of detecting that some words may form a single term.[2] One example of a company like this would be *Eli Lilly*. When multi-word terms such as these were found, the two parts were merged and treated as a single keyword.

3.8 Keyword list comparison

For purposes of the analysis, a sample of top 300 keywords from each corpus was taken. From these 300 keywords, some were incorrectly lemmatized and subsequently merged together, so the final keyword lists slightly vary in their length.

Table (12): Final keyword lists

Corpus	KW list length	Words not found in the other corpus	Tokens
IDNES_1	294	154	2,208,699
IDNES_2	286	155	2,193,541
LIDOVKY_1	299	273	887,743
LIDOVKY_2	289	225	655,468
NOVINKY_1	298	192	1,521,257
NOVINKY_2	280	167	1,454,787

Table (12) shows the final length of the sample keyword lists. The first sample from each couple is consistently longer. This is most likely due to more proper nouns (company names etc.) that were incorrectly lemmatized and then merged, resulting in a shorter list, appearing more in the second wave of the pandemic. Also, the first sample from each couple has more keywords that are not found in the second sample. The IDNES samples also have the most shared words from all the samples, which is probably due to the IDNES corpora being the largest – there simply might not be enough data in the LIDOVKY and NOVINKY corpora for some words to appear in the top 300 keywords. It is possible that if the corpora for these servers were larger, there would also be more shared keywords between their first and second wave corpora.

3.9 Keyword exclusion and inclusion

As was mentioned in Section 0, top 300 keywords from each corpus were taken to perform the analysis on. Since the goal of the analysis is to find changes in language use in relation with the coronavirus pandemic, not all of these keywords were subjected to analysis. The following sections describe which words were excluded from and which were included in the final keyword lists, respectively.

[2] It is important to note that Sketch Engine does offer multi-word term extraction which would probably reveal these terms, but that is not the aim of this thesis.

3.9.1 Excluded keywords

It is quite common for proper nouns in general to appear in keyword lists (Scott and Tribble 2006, 55-72), and Section 3.7 already mentioned that names of people and names of geographical locations seemed unimportant. There were many names of ministers and other politicians found in the keyword lists, and even though these names reflect what was happening during the pandemic quite well and they carry some descriptive value, they are not related to the pandemic directly.

It is a similar case with the names of geographical locations. Furthermore, it would be difficult to stay unbiased in deciding which words to keep and which ones not to. For example, the names of the cities *Litovel* and *Uničov* appeared in each corpus for the first wave, and they were both ranked high. They are both significant and highly prominent keywords due to media attention – both cities were locked down almost completely due to a high number of infected people, therefore an argument for keeping these keywords could be made here. However, knowing that these cities were in some way "special" is a part of a wider context. It is conceivable that there could be more "special" cities such as these two in the keywords lists but they would be excluded simply due to a lack of information related to them on my side. One way to avoid such a bias would be to keep all the names of geographical locations, but due to their overall low value for the goal of this thesis it was decided to exclude them completely.

Not all proper nouns were excluded, however. Names of vaccine manufacturers and names of medical drugs are two exceptions to this. See Section 3.9.2 for description of included keywords.

There were also thematically unrelated and common words that were excluded. These words include words such as *dluhopis* (financial bond), *gang*, *islámský*$_{\text{ADJ}}$ (Islamic), *rezidenční*$_{\text{ADJ}}$ (residential), *námořnictvo* (navy), *hotovost* (cash), *tržnice* (market), *chléb* (bread) or *ovoce* (fruit). Although there is without a doubt a reason why these words appeared in the keyword lists, and they most likely are related to the pandemic in one way or another, it is not an obvious relation but rather an obscure one. Analyzing these obscure relations goes beyond the scope of this thesis, and therefore these words were excluded.

To summarize, there are three major groups of excluded keywords – names of people, geographical locations, and thematically unrelated common words.

3.9.2 Included keywords

The method for deciding which keywords to include was to look for keywords that are related to the pandemic. Although Section 3.9.1 talked about proper nouns not being included, there are two exceptions – names of companies that produce vaccines or other medical supplies, and names of medical drugs, for example *Pfizer* or *hydroxychlorochin* (hydroxychloroquine). As was already mentioned in section 3.7, names of companies are often coinages, which means they are rather easy to spot among all the other words. It is a similar case with names of medical drugs – they are either coinages or medical terms that are very different from Czech words, which makes them easy to find just as well.

Another kind of included keywords is neologisms. These are words from the second wave corpora that are not found in the first wave corpora, and keywords from the first wave corpora that are not found in the csTenTen2017 corpus[3]. Some of these keywords might not be neologisms in the true sense, but they are in the context of these corpora – they are simply words that were not used before and then they became quite common. Table (13) shows two examples of such words.

Table (13): Occurences of example neologisms *lockdown* and *antigenní* across corpora

	lockdown		antigenní$_{\text{ADJ}}$ *antigen*$_{\text{ADJ}}$	
Corpus	Frequency	FPM	Frequency	FPM
IDNES_1	0	0	0	0
IDNES_2	324	147.7	750	341.9
LIDOVKY_1	0	0	0	0
LIDOVKY_2	72	109.8	258	393.6
NOVINKY_1	0	0	0	0
NOVINKY_2	340	233.7	479	329.3

The table shows the words *lockdown* and *antigenní*$_{\text{ADJ}}$ (antigen$_{\text{ADJ}}$) with their frequencies in each of the collected corpora. The FPM (frequency per million) column shows that these words were not used at all during the first wave of the pandemic, they were used only during the second wave, indicating that these words are neologisms.

Keywords related to vaccines make up the next group of included words. For example, there are many forms of the word *očkovat*$_{\text{V}}$ (to vaccinate) in the keyword lists – *naočkovat*$_{\text{V}}$ (to vaccinate, lit. on-vacciante), *vyočkovat*$_{\text{V}}$ and *proočkovat*$_{\text{V}}$ to name a few.

Another group of keywords consists of words related to face masks and respirators. These are words such as *celoobličejový*$_{\text{ADJ}}$ (full-face$_{\text{ADJ}}$) or *nanorespirátor* (nano respirator).

The following group of keywords concerns the virus itself. Words in this group are for example *mutace* (mutation) or *zmutnovaný*$_{\text{ADJ}}$ (mutated).

[3] Names of companies and medical drugs were excluded from this comparison because keywords from this category are mostly proper nouns.

Keywords concerning coronavirus testing and tracing make up the next group of words. There are words such as *sebereportování* (self-reporting) or *samotest* (self-test) in this category.

The last group of included keywords are medical expressions and terms and words derived from them, for example *imunolog* (immunologist) or *epidemický*$_{\text{ADJ}}$ (epidemic$_{\text{ADJ}}$).

There is some overlap between the aforementioned groups of keywords. For example, the keyword *antigenní*$_{\text{ADJ}}$ (antigen$_{\text{ADJ}}$) was grouped as a neologism, but it could just as well be put in the group of keywords related to coronavirus testing or the group with medical expressions and terms. However, having these overlaps is not an issue. The goal was to come up with a simple categorization of keywords to make the keyword samples smaller and to filter out words that are unimportant. Having precisely defined groups of words would be of no help in this matter.

To sum up the categorization – the keywords were divided into seven groups:

1. names of companies and medical drugs,
2. neologisms,
3. vaccines,
4. face masks and respirators,
5. the virus itself,
6. coronavirus testing,
7. medical expressions and terms.

4 Analysis

After filtering out the keywords and dividing the rest of the keywords into the categories described in Section 3.9.2, the keywords were compared in two ways. The first comparison was done for each server separately, comparing the first and second wave. The second comparison was a cross-server comparison, looking for general similarities and patterns in the language changes. Table (14) shows the numbers of keywords for each corpus and coronavirus wave after dividing them into the categories. Names of some categories were abbreviated, but the order remains the same as in 3.9.2.

Table (14): Keywords divided into categories

	C+D	NEO	VAX	Masks	Virus	Tests	Ex+T	Total
IDNES_1	4	2	0	2	0	0	2	8
IDNES_2	11	3	15	1	5	3	6	44
LIDOVKY_1	8	0	0	0	0	0	3	11
LIDOVKY_2	11	3	12	0	6	4	2	38
NOVINKY_1	5	0	0	1	0	0	3	9
NOVINKY_2	12	3	17	2	4	6	6	50

Even this rather general table shows some patterns already. Names of companies and medical drugs make up the majority of the keywords in the first wave corpora. Also, no first wave corpus has keywords related to vaccines, the virus or testing – this is most likely reflecting the events of the pandemic, as vaccines were still being developed during the first wave and they became a major topic only later on. The same applies to testing.

Concerning the keywords related to the virus, there are often words such as *zmutovaný*$_{\text{ADJ}}$ (mutated), and it makes sense that such words appeared only later on, as evolution takes some time to take place.

Section 3.9.2 mentioned that neologisms are keywords from the second wave not found in the first wave, and keywords from the first wave not found in the csTenTen2017 corpus. There are two keywords from the first wave – *rouškovník* (face mask tree) and *rouškomat* (face mask vending machine) – and one keyword from the second wave – *rouškovné* (a one-off compensation for face mask expenses for people receiving a pension) that could be considered even true neologisms, as they are not present in the csTenTen2017 corpus. However, a more thorough analysis of these keywords would be needed to completely confirm this claim.

There are also overall more keywords in each of the second wave corpora, which is probably reflecting the higher severity of the pandemic during the second wave, as the numbers of infected people were considerably higher. This can be seen from the graph of total number of infected people (Onemocnění aktuálně, n.d.).

4.1 Single server comparison

For the single server comparison, keywords within each category were compared against each other. Section 4 already talked about the numbers of keywords within each category, this section will examine these keywords more closely.

4.1.1 Idnes.cz

Table (15) and Table (16) show the names of companies and medical drugs which appeared as keywords during the pandemic. The column headings are repeated here for convenience – Freq again stands for absolute frequency, F stands for focus corpus, R for reference corpus, FPM is frequency per million, SCORE is keyness value from Sketch Engine, DIN is difference index, and Chi2 test shows values of significance testing. Some values in the Chi2 test column are written in exponential notation due to their length. All keywords from the IDNES corpora passed the Chi2 test as significant.

Table (15): Names of companies and medical drugs in IDNES_1

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
2	favipiravir	43	0	19.47	0	20.47	100	6.364e-11
3	hydroxychlorochin *hydroxychloroquine*	82	2	37.13	0.91	19.94	95.21	3.399e-18
15	chlorochin *chloroquine*	36	0	16.30	0	17.30	100	2.24e-09
134	paracetamol	24	2	10.87	0.83	6.50	84.52	1.731e-05

Table (16): Names of companies and medical drugs in IDNES_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
3	pfizer	491	1	223.84	0.45	154.77	99.60	6.895e-109
6	biontech	286	0	130.38	0	131.38	100	1.36e-64
8	astrazeneca	249	2	113.52	0.91	60.10	98.42	3.56e-55
11	sputnik	115	1	47.50	0.45	33.38	98.11	2.357e-26
15	moderna	248	8	102.43	3.62	22.38	93.17	3.171e-51
23	regeneron	40	0	16.52	0	17.52	100	2.204e-10
26	novavax	37	0	15.28	0	16.28	100	1.036e-09
89	ivermektin *ivermectin*	20	0	8.26	0	9.26	100	7.204e-06
105	bamlanivimab	18	0	8.21	0	9.21	100	2.069e-05
120	eli lilly	17	0	7.02	0	8.02	100	3.514e-05
146	sinopharm	15	0	6.20	0	7.20	100	0.0001018

Apart from the keyword *hydroxychlorochin* (hydroxychloroquine) in Table (15), all words are above the threshold of 75 points for the DIN value, meaning they are highly prominent and relevant. And even *hydroxychlorochin* (hydroxychloroquine) is below the threshold only by less than 4 points, which means it is still rather prominent.

All the keywords in the IDNES_1 list are referring to medical drugs and no names of companies appear, while the IDNES_2 list includes mostly names of vaccines or vaccine manufacturers. Also, IDNES_2 includes more than twice as many keywords in this category than IDNES_1.

Table (17) and Table (18) show neologisms. Table (17) shows neologisms from the IDNES_1 sample, meaning these keywords had the csTenTen2017 as the reference corpus. The rank of the keyword does not apply here, as these keywords were looked for specifically and individually.

Table (17): Neologisms in IDNES_1 with the csTenTen2017 as the reference corpus

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
NA	roúškomat *face mask vending machine*	15	0	6.79	0	7.8	100	0
NA	rouškovník *face mask tree*	14	0	6.34	0.41	7.3	100	0

Table (18): Neologisms in IDNES_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
1	antigenní$_{\mathrm{ADJ}}$ *antigen*$_{\mathrm{ADJ}}$	750	0	341.91	0	342.91	100	2.809e-166
4	lockdown	324	0	147.71	0	148.71	100	6.262e-73
209	rouškovné *a one-off compensation for face mask expenses for people receiving a pension*	12	0	5.47	0	6.47	100	0.0005088

All of the keywords are highly relevant and above the 75-point threshold for the DIN value this time. Section 4 already mentioned that the keywords *rouškovník* (face mask tree), *rouškomat* (face mask vending machine), and *rouškovné* (a one-off compensation for face mask expenses for people receiving a pension) could be also considered true neologisms. Considering the rest of the IDNES_2 keywords, the words *antigenní*$_{\mathrm{ADJ}}$ (antigen$_{\mathrm{ADJ}}$) and *lockdown* started to appear only during the second wave. Even though both of these keywords appear in the csTenTen2017 corpus, the word *antigenní*$_{\mathrm{ADJ}}$ (antigen$_{\mathrm{ADJ}}$) does not appear with the word *test*, and the word *lockdown* appears mostly as a sports term (lockdown corner) or a technological term (IP lockdown). *Lockdown* with the meaning “restriction of free movement of people” occurs in the context of school shootings and rather infrequently.

Table (19) shows keywords related to vaccines in the IDNES_2 corpus. There were no keywords in this category for the IDNES_1 corpus.

Table (19): Keywords related to vaccines in IDNES_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
7	naočkovat$_{V}$ *lit. on-vaccinate*	289	1	131.75	0.45	91.38	99.32	1.341e-64
10	očkovat$_{V}$ *to vaccinate*	1025	23	467.28	10.41	41.03	95.64	6.676e-212
12	očkovaný$_{ADJ}$ *vaccinated*	300	8	136.77	3.62	29.81	94.84	1.325e-62
14	očkovací$_{ADJ}$ *vaccination*$_{ADJ}$	854	30	389.32	13.58	26.77	93.26	2.524e-170
19	očkování *vaccination*	2119	104	966.02	47.09	20.11	90.70	0
21	naočkovaný$_{ADJ}$ *lit. on-vaccinated*	60	1	27.35	0.45	19.52	96.74	3.425e-14
30	vakcinační$_{ADJ}$ *vaccination*$_{ADJ}$	102	4	46.50	1.81	16.90	92.50	1.245e-21
56	vakcinace *vaccination*	239	17	108.96	7.70	12.64	86.80	4.13e-44
79	proočkovat$_{V}$ *lit. through-vaccinate*	63	4	28.72	1.81	10.57	88.14	4.612e-13
119	přeočkování *re-vaccination*	17	0	7.75	0	8.75	100	3.514e-05
125	naočkování *lit.* *on-vaccination*	34	2	15.50	0.91	8.66	88.96	8.601e-08
128	vakcína *vaccine*	3530	415	1609.27	187.89	8.52	79.09	0
155	vakcinolog *vaccinologist*	29	2	13.22	0.91	7.46	87.17	1.124e-06
161	vyočkovat$_{V}$ *lit. vaccinate out*	14	0	6.38	0	7.38	100	0.0001736
232	proočkování *lit. through-vaccination*	23	2	10.49	0.91	6.03	84.10	2.473e-05

These keywords are mostly different forms of “vaccine” or “to vaccinate” with various prefixes that slightly change the meaning. There are also two etymologically different keywords – *vakcinace* (vaccination) and *očkování* (vaccination), and words derived from them. The noun *vakcína* (vaccine) is more frequent than the noun *očkování* (vaccination), but concerning verbs, there are only different forms of the verb *očkovat*$_{\mathrm{V}}$ (to vaccinate), verb *vakcinovat*$_{\mathrm{V}}$ (to vaccinate) does not appear in the top 300 keywords. It seems that *vakcína* (vaccine) is the more preferred noun, while *očkovat*$_{\mathrm{V}}$ (to vaccinate) is the preferred verb.

The various forms of the verb *očkovat*$_{\mathrm{V}}$ (to vaccinate) are – *naočkovat*$_{\mathrm{V}}$ (to vaccinate, lit. on-vaccinate)*, očkovat*$_{\mathrm{V}}$ (to vaccinate)*, proočkovat*$_{\mathrm{V}}$ (to vaccinate with intention of vaccinating a large group of people, lit. through-vaccinate)*, vyočkovat*$_{\mathrm{V}}$ (to vaccinate with intention of using up a set amount of vaccine, lit. vaccinate out). There is also the noun *přeočkování* (re-vaccination) which does not have its verbal counterpart present in the keyword list.

The next group of keywords is words related to face masks and respirators. These are shown in Table (20) and Table (21).

Table (20): Keywords related to face masks and respirators in IDNES_1

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
115	celooblíčejový$_{\mathrm{ADJ}}$ *full-face*$_{\mathrm{ADJ}}$	12	0	5.43	0	6.43	100	0.000556
165	ffp3	119	19	53.88	8.66	5.68	72.30	4.358e-16

Table (21): Keywords related to face masks and respirators in IDNES_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
277	nanorespirátor *nano respirator*	10	0	4.56	0	5.56	100	0.001508

Overall, there are not many keywords in this category. The word *celooblíčejový*$_{\mathrm{ADJ}}$ (full-face$_{\mathrm{ADJ}}$) appears only during the first wave, while *nanorespirátor* (nano respirator) appears only during the second wave. The respirator class *FFP3* is present in both corpora but is a keyword only in IDNES_1. However, it is below the 75-point DIN value threshold, but again, by less than 4 points.

The following group of keywords in Table (22) is related to the virus itself. Once more, there is no data for this group of keywords in IDNES_1.

Table (22): Keywords related to the virus itself in IDNES_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
22	covidový$_{ADJ}$ *covid*$_{ADJ}$	583	28	265.78	12.68	19.51	90.89	1.694e-112
29	postcovidový$_{ADJ}$ *post-covid*$_{ADJ}$	35	0	15.96	0	16.96	100	2.912e-09
35	mutace *mutation*	815	49	371.55	22.19	16.07	88.73	6.869e-151
36	zmutovaný$_{ADJ}$ *mutated*	33	0	15.04	0	16.04	100	8.195e-09
117	covid	1654	187	754.03	84.67	8.81	79.81	1.647e-258

It is an interesting observation that words *covidový*$_{ADJ}$ (covid$_{ADJ}$) and *covid*$_{N}$ appear in both corpora (although only in IDNES_2 as a keyword), but *postcovidový*$_{ADJ}$ (post-covid$_{ADJ}$) appears only in IDNES_2. This is similar to the keywords *mutace* (mutation), occuring during both waves, and *zmutovaný*$_{ADJ}$ (mutated), occuring only during the second wave.

Table (23) is showing keywords related to coronavirus testing. There are again no keywords from this category in the IDNES_1 corpus.

Table (23): Keywords related to coronavirus testing in IDNES_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
76	kloktací$_{ADJ}$ *gargling*$_{ADJ}$	21	0	9.57	0	10.57	100	4.258e-06
243	samotest *self-test*	11	0	5.01	0	6.01	100	0.0008745
248	samotestování *self-testing*	11	0	5.01	0	6.01	100	0.0008745

The keywords are *kloktací*$_{ADJ}$ (gargling$_{ADJ}$), *samotest* (self-test), and *samotestování* (self-testing). All of them are found only during the second wave.

Table (24) and Table (25) are showing medical expressions and terms.

Table (24): Medical expressions and terms in IDNES_1

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
77	uzdravený$_{\text{ADJ}}$ *healed*	144	18	65.20	8.21	7.19	77.64	6.463e-23

Table (25): Medical expressions and terms in IDNES_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
91	epidemický$_{\text{ADJ}}$ *epidemic*$_{\text{ADJ}}$	280	26	127.65	11.77	10.07	83.11	3.708e-48
95	prodělání *suffering an illness in the past*	48	3	21.88	1.36	9.70	88.31	2.517e-10
111	reinfekce *reinfection*	27	1	12.31	0.45	9.16	92.90	8.147e-07
157	epidemicky$_{\text{ADV}}$ *epidemically*	36	3	16.41	1.36	7.38	84.71	1.122e-07
196	protiepidemický$_{\text{ADJ}}$ *anti-epidemic*$_{\text{ADJ}}$	570	85	259.85	38.48	6.59	74.20	7.927e-81
199	imunolog *immunologist*	122	17	55.62	7.70	6.51	75.69	3.66e-19

There is only one keyword in this category for IDNES_1 – *uzdravený*$_{\text{ADJ}}$ (healed). There are no keywords meaning “healed” in the IDNES_2 corpus, which is an interesting finding. It is safe to assume that people did not stop healing from the coronavirus during the second wave, it is rather caused by the media that stopped reporting on the numbers of healed people. The IDNES_2 medical expressions and terms are more concerned with the epidemic and the virus, with keywords such as *epidemický*$_{\text{ADJ}}$ (epidemic$_{\text{ADJ}}$) and *protiepidemický*$_{\text{ADJ}}$ (anti-epidemic$_{\text{ADJ}}$) or *prodělání* (suffering an illness in the past) and *reinfekce* (reinfection). Except for the keyword *protiepidemický*$_{\text{ADJ}}$ (anti-epidemic$_{\text{ADJ}}$), all medical expressions and terms from IDNES_2 corpus are above the DIN value threshold.

4.1.2 Novinky.cz

All keywords from NOVINKY corpora passed the Chi2 test as significant. Table (26) and Table (27) show names of companies and medical drugs.

Table (26): Names of companies and medical drugs in NOVINKY_1

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
6	chlorochin *chloroquine*	40	0	26.29	0	27.29	100	6.215e-10
54	respilon	18	0	11.83	0	12.83	100	3.34e-05
82	ibuprofen	27	1	17.75	0.69	11.11	92.54	1.614e-06
119	hydroxychlorochin *hydroxychloroquine*	67	5	44.04	3.44	10.15	85.52	1.086e-12
136	medicago	13	0	8.55	0	9.55	100	0.000422

Table (27): Names of companies and medical drugs in NOVINKY_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
4	sputnik	199	0	136.79	0	137.79	100	3.556e-47
5	pfizer	639	4	439.24	2.63	121.30	98.81	9.746e-145
7	biontech	389	3	267.39	1.97	90.31	98.54	1.675e-88
10	ivermektin *ivermectin*	85	0	58.43	0	59.43	100	4.182e-21
11	astrazeneca	435	7	299.01	4.60	53.56	96.97	2.121e-96
28	eli lilly	28	0	19.25	0	20.25	100	6.265e-08
47	bamlanivimab	20	0	13.75	0	14.75	100	4.804e-06
48	covax	20	0	13.75	0	14.75	100	4.804e-06
53	regeneron	47	2	32.31	1.31	14.39	92.18	4.5e-11
114	moderna	254	29	174.60	19.06	8.75	80.31	4.774e-43
151	diana biotechnologies	10	0	6.87	0	7.87	100	0.001222
222	sinopharm	50	6	34.37	3.94	7.15	79.41	1.462e-09

There are again more than twice as many keywords appearing during the second wave than the first. With the exception of *respilon* and *medicago*, which are company names, are all keywords in NOVINKY_1 referring to medical drugs, while the majority of keywords in NOVINKY_2 is referring to companies with the only two exceptions being *ivermektin* (ivermectin) and *bamlanivimab.*

Table (28) shows neologisms in the NOVINKY_2 corpus. There were no neologisms found in NOVINKY_1.

Table (28): Neologisms in NOVINKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
2	antigenní$_{\text{ADJ}}$ *antigen*$_{\text{ADJ}}$	479	0	329.26	0	330.26	100	5.857e-111
3	lockdown	340	0	233.71	0	234.71	100	2.588e-79
235	antigenový$_{\text{ADJ}}$ *antigen*$_{\text{ADJ}}$	8	0	5.50	0	6.50	100	0.003824

All three keywords appear only during the second wave, there is not a single occurrence during the first wave. There are two forms of the adjective "antigen" – *antigenní*$_{\text{ADJ}}$ and *antigenový*$_{\text{ADJ}}$, from which the former one is clearly the preferred form.

Table (29) shows keywords related to vaccines in the NOVINKY_2 corpus. There were no keywords from this category in NOVINKY_1.

Table (29): Keywords related to vaccines in NOVINKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
8	naočkovat$_{\text{V}}$ *lit. on-vaccinate*	389	3	267.39	1.97	90.31	98.54	1.675e-88
12	očkovat$_{\text{V}}$ *to vaccinate*	957	18	657.83	11.83	51.34	96.47	4.536e-208
13	vakcinace *vaccination*	200	3	137.48	1.97	46.59	97.17	1.896e-45
14	naočkovaný$_{\text{ADJ}}$ *lit. on-vaccinated*	109	1	74.93	0.66	45.81	98.26	5.992e-26
15	vakcinační$_{\text{ADJ}}$ *vaccination*$_{\text{ADJ}}$	77	1	52.93	0.66	32.54	97.55	1.31e-18
17	očkování *vaccination*	1767	67	1214.61	44.04	26.99	93.00	0
29	proočkování *lit. through-vaccination*	28	0	19.25	0	20.25	100	6.265e-08
33	proočkovat$_{\text{V}}$ *lit. through-vaccinate*	64	2	43.99	1.31	19.44	94.20	5.489e-15
36	naočkování *lit. on-vaccination*	25	0	17.18	0	18.18	100	3.171e-07
65	očkovaný$_{\text{ADJ}}$ *vaccinated*	232	18	159.47	11.83	12.51	86.19	7.108e-44
77	vyočkovat$_{\text{V}}$ *lit. vaccinate out*	14	0	9.62	0	10.62	100	0.0001301

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
89	očkovací$_{\text{ADJ}}$ *vaccination*$_{\text{ADJ}}$	814	84	559.53	55.22	9.97	82.04	2.258e-138
113	proočkovanost *lit. the rate of through-vaccination*	45	4	30.93	2.63	8.80	84.33	1.798e-09
164	vyočkovaný$_{\text{ADJ}}$ *lit. vaccinated out*	10	0	6.87	0	7.87	100	0.001222
182	proočkovaný$_{\text{ADJ}}$ *lit. through-vaccinated*	23	2	15.81	1.31	7.26	84.65	1.612e-05
225	vakcinologický$_{\text{ADJ}}$ *vaccinological*	22	2	15.12	1.31	6.97	84.00	2.753e-05
276	vakcína *vaccine*	4194	694	2882.90	456.20	6.31	72.68	0

There are again the two etymologically different keywords – *vakcinace* (vaccination) and *očkování* (vaccination), and words derived from them. The verbs derived from the verb *očkovat*$_{\text{V}}$ (to vaccinate) include – *naočkovat*$_{\text{V}}$ (to vaccinate, lit. on-vaccinate), *očkovat*$_{\text{V}}$ (to vaccinate), *proočkovat*$_{\text{V}}$ (to vaccinate with intention of vaccinating a large group of people, lit. through-vaccinate), and *vyočkovat*$_{\text{V}}$ (to vaccinate with intention of using up a set amount of vaccine, lit. vaccinate out). There are also nominal counterparts to these verbs – *očkování* (vaccination), *proočkování* (vaccination with intention of vaccinating a large group of people, lit. through-vaccination), *naočkování* (vaccination, lit. on-vaccination), and *proočkovanost* (the rate of people vaccinated, lit. the rate of through-vaccination), and adjectives such as *naočkovaný*$_{\text{ADJ}}$ (vaccinated, lit. on-vaccinated), *vakcinační* (vaccination), *očkovaný* (vaccinated), *očkovací* (vaccination), *vyočkovaný* (the feature of having used up a set amount of vaccine, lit. vaccinated out), *proočkovaný* (the feature of groups of people of having been vaccinated, lit. through-vaccinated), and *vakcinologický*$_{\text{ADJ}}$ (vaccinological). The noun *vakcína* (vaccine) is again more frequent than *očkování* (vaccine), but it is less key and also slightly below the DIN value threshold. There again is no verbal counterpart to *vakcína* (vaccine), all the verbs are different forms of *očkovat*$_{\text{V}}$ (to vaccinate).

Table (30) and Table (31) show keywords related to face masks and respirators.

Table (30): Keywords related to face masks and respirators in NOVINKY_1

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
122	maska *mask*	283	26	186.03	17.87	9.91	82.47	5.809e-46

Table (31): Keywords related to face masks and respirators in NOVINKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
60	kn95	18	0	12.37	0	13.37	100	1.435e-05
166	nanorouška *nano face mask*	32	3	22.00	1.97	7.74	83.54	4.771e-07

There are not many keywords in this category – the only keyword appearing in NOVINKY_1 is *maska* (mask), and the two keywords in NOVINKY_2 are *KN95*, only appearing in NOVINKY_2, and *nanorouška* (nano face mask).

Table (32) shows keywords related the virus itself. There were not any keywords in this category in NOVINKY_1.

Table (32): Keywords related to the virus itself in NOVINKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
25	covidový$_{\text{ADJ}}$ *covid*$_{\text{ADJ}}$	284	12	195.22	7.89	22.08	92.23	5.143e-59
76	mutace *mutation*	930	86	639.27	56.53	11.13	83.75	6.153e-163
98	proticovidový$_{\text{ADJ}}$ *anti-covid*$_{\text{ADJ}}$	13	0	8.94	0	9.94	100	0.0002269
133	covid	1159	141	796.68	92.69	8.51	79.16	1.405e-185

There are again keywords *covidový*$_{\text{ADJ}}$ (covid$_{\text{ADJ}}$) and *covid*. The other two are *mutace* (mutation) and *proticovidový*$_{\text{ADJ}}$ (anti-covid$_{\text{ADJ}}$), which appears only in NOVINKY_2.

Table (33) shows keywords related to coronavirus testing. Once again, no keywords from this category are found in NOVINKY_1.

Table (33): Keywords related to coronavirus testing in NOVINKY_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
79	samoodběr *self-sampling*	14	0	9.62	0	10.62	100	0.0001301
107	samotest *self-test*	12	0	8.25	0	9.25	100	0.0003966
148	sebetrasování *self-tracing*	10	0	6.87	0	7.87	100	0.001222
154	samotestování *self-testing*	10	0	6.87	0	7.87	100	0.001222
186	pcr	333	47	228.90	30.90	7.21	76.22	1.358e-51
261	kloktací$_{\text{ADJ}}$ *gargling*$_{\text{ADJ}}$	8	0	5.50	0	6.50	100	0.003824

Except for *PCR*, all keywords are found only in NOVINKY_2. One keyword here is related to tracing – *sebetrasování* (self-tracing), the rest are related to testing – *samoodběr* (self-sampling), *samotest* (self-test), *samotestování* (self-testing), and *kloktací*$_{\text{ADJ}}$ (gargling$_{\text{ADJ}}$).

Table (34) and Table (35) show medical expressions and terms.

Table (34): Medical expressions and terms in NOVINKY_1

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
1	zotavený$_{\text{ADJ}}$ *recovered*	222	0	145.93	0	146.93	100	4.292e-48
228	ebola	59	6	38.78	4.12	7.76	80.78	1.601e-10
295	dezinfikován$_{\text{V}}$ *disinfected*	31	3	20.38	2.06	6.98	81.62	2.969e-06

Table (35): Medical expressions and terms in NOVINKY_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
30	epidemický$_{\text{ADJ}}$ *epidemic*$_{\text{ADJ}}$	200	9	137.48	5.92	20.02	91.75	9.267e-42
66	protiepidemický$_{\text{ADJ}}$ *anti-epidemic*$_{\text{ADJ}}$	483	39	332.01	25.64	12.50	85.66	1.488e-88
170	epidemicky$_{\text{ADV}}$ *epidemically*	17	1	11.69	0.66	7.65	89.35	0.0001102

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
174	pandemický$_{\text{ADJ}}$ *pandemic*$_{\text{ADJ}}$	140	18	96.23	11.83	7.58	78.10	1.698e-23
229	reinfekce *reinfection*	21	2	14.44	1.31	6.67	83.31	4.703e-05
262	autoprotilátka *autoantibody*	8	0	5.50	0	6.50	100	0.003824

There are a few keywords in NOVINKY_1 – *zotavený*$_{\text{ADJ}}$ (recovered), appearing only in NOVINKY_1, then *ebola* and *dezinfikován*$_{\text{V}}$ (disinfected). NOVINKY_2 contains twice as many medical expressions and terms. Half of them mentions an epidemic – *epidemický*$_{\text{ADJ}}$ (epidemic$_{\text{ADJ}}$), *protiepidemický*$_{\text{ADJ}}$ (anti-epidemic$_{\text{ADJ}}$), and *epidemicky*$_{\text{ADV}}$ (epidemically), but only one mentions a pandemic – *pandemický*$_{\text{ADJ}}$ (pandemic$_{\text{ADJ}}$). The keyword *autoprotilátka* (autoantibody) is present only in NOVINKY_2.

4.1.3 Lidovky.cz

All keywords in LIDOVKY corpora are statistically significant and above the 75-point threshold for DIN value. Table (36) and Table (37) show names of companies and medical drugs.

Table (36): Names of companies and medical drugs in LIDOVKY_1

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
62	ritonavir	14	0	15.77	0	16.77	100	0.001304
74	lopinavir	13	0	14.64	0	15.64	100	0.001947
89	antimalarikum *antimalarial drug*	12	0	13.52	0	14.52	100	0.002914
106	favipiravir	11	0	12.39	0	13.39	100	0.004373
172	ibuprofen	9	0	10.14	0	11.14	100	0.009942
215	antivirotikum *antiviral drug*	23	1	25.91	1.53	10.65	88.88	0.0001467
228	chlorochin *chloroquine*	8	0	9.01	0	10.01	100	0.01508
275	hydroxychlorochin *hydroxychloroquine*	8	0	9.01	0	10.01	100	0.01508

Table (37): Names of companies and medical drugs in LIDOVKY_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
3	pfizer	200	0	305.13	0	306.13	100	7.194e-61
7	sputnik	68	0	103.74	0	104.74	100	8.237e-22
8	biontech	131	1	199.86	1.13	94.46	98.88	9.35e-40
11	moderna	93	1	141.88	1.13	67.19	98.42	1.661e-28
33	regeneron	18	0	27.46	0	28.46	100	7.913e-07
42	astrazeneca	112	5	170.87	5.63	25.91	93.62	2.205e-31
64	sinopharm	13	0	19.83	0	20.83	100	2.716e-05
114	dexametazon *dexamethazone*	9	0	13.73	0	14.73	100	0.0004806
132	sinovac	8	0	12.21	0	13.21	100	0.000996
183	bamlanivimab	7	0	10.68	0	11.68	100	0.002077
202	covax	7	0	10.68	0	11.68	100	0.002077

LIDOVKY_1 contains only names of medical drugs, none of which is mentioned in LIDOVKY_2 except for *antivirotikum* (antiviral drug). It is the opposite with LIDOVKY_2 – almost all keywords are names of companies with the only two exceptions being *dexametazon* (dexamethasone) and *bamlanivimab*. All keywords are either exclusive to their respective corpus, or extremely rare in the reference corpus.

Table (38) shows neologisms in LIDOVKY_2. There were no keywords from this category in LIDOVKY_1.

Table (38): Neologisms in LIDOVKY_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
2	antigenní$_{\text{ADJ}}$ *antigen*$_{\text{ADJ}}$	258	0	393.61	0	394.61	100	5.487e-78
6	lockdown	72	0	109.85	0	110.85	100	5.335e-23
127	odmítač *rejector*	8	0	12.21	0	13.21	100	0.000996

The usual keywords *antigenní*$_{\text{ADJ}}$ (antigen$_{\text{ADJ}}$) and *lockdown* are present, but there is also a new keyword *odmítač* (rejector) that does not appear anywhere else.

Table (39) shows keywords related to vaccines in LIDOVKY_2. Once again, there were not any data from this group of keywords in LIDOVKY_1.

Table (39): Keywords related to vaccines in LIDOVKY_2

Rank	Word	Freq(F)	Freq(R)	FPM(F)	FPM(R)	SCORE	DIN	Chi2 test
15	vakcinační$_{\text{ADJ}}$ *vaccination*$_{\text{ADJ}}$	29	0	44.24	0	45.24	100	3.677e-10
26	očkovat$_{\text{V}}$ *to vaccinate*$_{\text{V}}$	248	9	378.36	10.14	34.06	94.78	9.667e-69
47	očkovaný$_{\text{ADJ}}$ *vaccinated*	69	3	105.27	3.38	24.27	93.78	5.195e-20
48	naočkovat$_{\text{V}}$ *lit. on-vaccinate*	68	3	103.74	3.38	23.92	93.69	1.027e-19
62	proočkovat$_{\text{V}}$ *lit. through-vaccinate*	13	0	19.83	0	20.83	100	2.716e-05
98	vakcinace *vaccination*	89	7	135.78	7.89	15.39	89.02	2.337e-23
122	naočkovaný$_{\text{ADJ}}$ *lit. on-vaccinated*	19	1	28.99	1.13	14.10	92.52	2.012e-06
123	očkování *vaccination*	605	58	923.00	65.33	13.93	86.78	1.757e-142
153	proočkování *lit. through-vaccination*	8	0	12.21	0	13.21	100	0.000996
210	očkovací$_{\text{ADJ}}$ *vaccination*$_{\text{ADJ}}$	281	33	428.70	37.17	11.26	84.04	9.433e-64
234	přeočkování *re-vaccination*	6	0	9.15	0	10.15	100	0.004363
250	naočkování *lit. on-vaccination*	6	0	9.15	0	10.15	100	0.004363

There are words derived from the word *vakcína* (vaccine), such as *vakcinační*$_{\text{ADJ}}$ (vaccination$_{\text{ADJ}}$) or *vakcinace* (vaccination), but the noun *vakcína* (vaccine) itself does not appear in the LIDOVKY_2 sample as a keyword. There are also other keywords that are derived from *očkovat*$_{\text{V}}$ (to vaccinate). The verbs here are *očkovat*$_{\text{V}}$ (to vaccinate), *naočkovat*$_{\text{V}}$ (to vaccinate, lit. on-vaccinate), and *proočkovat*$_{\text{V}}$ (to vaccinate with intention of vaccinating a large group of people, lit. through-vaccinate). The nouns derived from these verbs are *očkování* (vaccination), *proočkování* (vaccination with intention of vaccinating a large group of people, lit. through-vaccination), *přeočkování* (re-vaccination), and *naočkování* (vaccination, lit. on-vaccination). The adjectives include *očkovaný*$_{\text{ADJ}}$ (vaccinated), *naočkovaný*$_{\text{ADJ}}$ (vaccinated, lit. on-vaccinated), and *očkovací*$_{\text{ADJ}}$ (vaccination$_{\text{ADJ}}$).

The LIDOVKY corpora do not contain any keywords related to face masks and respirators. The next group of keywords is related to the virus itself. These keywords are shown in Table (40). LIDOVKY_1 does not contain any keywords from this category.

Table (40): Keywords related to the virus itself in LIDOVKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
13	zmutovaný$_{\text{ADJ}}$ *mutated*	33	0	50.35	0	51.35	100	2.302e-11
25	mutace *mutation*	343	12	523.29	13.52	36.11	94.96	1.175e-94
27	covidní$_{\text{ADJ}}$ *covid*$_{\text{ADJ}}$	21	0	32.04	0	33.04	100	9.655e-08
59	covidový$_{\text{ADJ}}$ *covid*$_{\text{ADJ}}$	122	7	186.13	7.89	21.06	91.87	5.007e-33
164	covid	628	72	958.09	81.10	11.68	84.39	4.028e-141
206	proticovidový$_{\text{ADJ}}$ *anti-covid*$_{\text{ADJ}}$	7	0	10.68	0	11.68	100	0.002077

The keywords here are *zmutovaný*$_{\text{ADJ}}$ (mutated), not appearing in LIDOVKY_1, and *mutace* (mutation). There are also two forms of the adjective covid – *covidní*$_{\text{ADJ}}$ and *covidový*$_{\text{ADJ}}$ – with the former appearing only in LIDOVKY_2, but the latter being the more frequent form. The last two keywords are *covid* and *proticovidový*$_{\text{ADJ}}$ (anti-covid$_{\text{ADJ}}$), which is not found in LIDOVKY_1.

Table (41) shows keywords related to coronavirus testing in LIDOVKY_2. Again, no keywords from this category were found in LIDOVKY_1.

Table (41): Keywords related to coronavirus testing in LIDOVKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
46	kloktací$_{\text{ADJ}}$ *gargling*$_{\text{ADJ}}$	16	0	24.41	0	25.41	100	3.238e-06
138	sebereportování *self-reporting*	8	0	12.21	0	13.21	100	0.000996
182	rt-pcr	7	0	10.68	0	11.68	100	0.002077
186	gargtest	7	0	10.68	0	11.68	100	0.002077

The keywords are *kloktací*$_{\mathrm{ADJ}}$ (gargling$_{\mathrm{ADJ}}$), *sebereportování* (self-reporting), *rt-pcr*, and *gargtest*, which is a non-Czech word. All the keywords here are not present in LIDOVKY_1.

Medical expressions and terms are shown in Table (42) and Table (43).

Table (42): Medical expressions and terms in LIDOVKY_1

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
65	medicinální$_{\mathrm{ADJ}}$ *medicinal*	14	0	15.77	0	16.77	100	0.001304
279	antibakteriální$_{\mathrm{ADJ}}$ *antibacterial*	8	0	9.01	0	10.01	100	0.01508
280	uzdravený$_{\mathrm{ADJ}}$ *healed*	89	6	100.25	9.15	9.97	83.27	1.003e-12

Table (43): Medical expressions and terms in LIDOVKY_2

Rank	**Word**	**Freq(F)**	**Freq(R)**	**FPM(F)**	**FPM(R)**	**SCORE**	**DIN**	**Chi2 test**
37	reinfekce *reinfection*	17	0	25.94	0	26.94	100	1.6e-06
158	epidemicky$_{\mathrm{ADV}}$ *epidemically*	26	2	39.67	2.25	12.50	89.25	6.909e-08

The keywords in LIDOVKY_1 include *medicinální*$_{\mathrm{ADJ}}$ (medicinal) and *antibakteriální*$_{\mathrm{ADJ}}$ (antibacterial), both found only in LIDOVKY_1, and *uzdravený*$_{\mathrm{ADJ}}$ (healed). Keywords from LIDOVKY_2 include *reinfekce* (reinfection), found only in LIDOVKY_2, and *epidemicky*$_{\mathrm{ADV}}$ (epidemically).

The sections above analyzed and compared keywords for each server separately. It was shown that there are some patterns occurring – keywords related to vaccines and keywords related to testing are appearing only in the second wave, first wave corpora contain more names of drugs while second wave corpora contain more names of companies, or that there are some words appearing exclusively during the second wave. The next section compares these keywords across all the servers.

4.2 Cross-server comparison

When comparing the newspaper servers with each other, one can see that there are some patterns in the data. Using the same order of categories and looking at the keywords within each category, it is clear that within the category of names of companies and medical drugs, there were more drugs than companies in the first wave corpora. These include some nonprescription drugs such as *ibuprofen* and *paracetamol* or specialized drugs such as *ritonavir* and *hydroxychlorochin* (hydroxychloroquine). These keywords are either rare or not present in the second wave corpora. This is shown in Table (44) below.

Table (44): Cross-server comparison of names of companies and drugs in the first wave corpora

IDNES_1			NOVINKY_1			LIDOVKY_1		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
favipiravir	43	0	chlorochin *chloroquine*	40	0	ritonavir	14	0
hydroxychlorochin *hydroxychloroquine*	82	2	respilon	18	0	lopinavir	13	0
chlorochin *chloroquine*	27	0	ibuprofen	27	1	antimalarikum *antimalarial drug*	12	0
paracetamol	24	2	hydroxychlorochin *hydroxychloroquine*	67	5	favipiravir	11	0
			medicago	13	0	ibuprofen	9	0
						antivirotikum *antiviral drug*	23	1
						chlorochin *chloroquine*	8	0
						hydroxychlorochin *hydroxychloroquine*	8	0

Considering the second wave, there were mostly names of companies such as vaccine manufacturers *Pfizer, BioNTech,* and *AstraZeneca*. The majority of the keywords here are shared between all corpora and the data shows that all of these keywords started appearing mostly during the second wave. From the three newspapers, Novinky.cz mentions the Russian vaccine *Sputnik* the most and it is also the top keyword in this category, while idnes.cz and Lidovky.cz have *Pfizer* as the top keyword. See Table (45) for complete comparison.

Table (45): Cross-server comparison of names of companies and drugs in the second wave corpora

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
pfizer	491	1	sputnik	199	0	pfizer	200	0
biontech	286	0	pfizer	639	4	sputnik	68	0
astrazeneca	249	2	biontech	389	3	biontech	131	1
sputnik	115	1	ivermektin *ivermectin*	85	0	moderna	93	1
moderna	248	8	astrazeneca	435	7	regeneron	18	0
regeneron	40	0	eli lilly	28	0	astrazeneca	112	5
novavax	37	0	bamlanivimab	20	0	sinopharm	13	0
ivermektin *ivermectin*	20	0	covax	20	0	dexametazon *dexamethazone*	9	0
bamlanivimab	18	0	regeneron	47	2	sinovac	8	0
eli lilly	17	0	moderna	254	29	bamlanivimab	7	0
sinopharm	15	0	diana biotechnologies	10	0	covax	7	0
			sinopharm	50	6			

Concerning the neologisms, only IDNES_1 contains some neologisms from all the first wave corpora, and these keywords are *rouškovník* (face mask tree) and *rouškomat* (face mask vending machine). Table (46) illustrates the second wave corpora, which all have two words in common – *antigenní*$_{\text{ADJ}}$ (antigen$_{\text{ADJ}}$) and *lockdown*, which is a non-Czech word. Each of the second wave corpora also has one exclusive keyword – *rouškovné* (a one-off compensation for face mask expenses for people receiving a pension) in IDNES_2, *antigenový*$_{\text{ADJ}}$ (antigen$_{\text{ADJ}}$) in NOVINKY_2, and *odmítač* (rejector) in LIDOVKY_2. None from the just mentioned words appear in the first wave corpora.

Table (46): Cross-server comparison of neologisms in the second wave corpora

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
antigenní$_{ADJ}$ *antigen*$_{ADJ}$	750	0	antigenní$_{ADJ}$ *antigen*$_{ADJ}$	479	0	antigenní$_{ADJ}$ *antigen*$_{ADJ}$	258	0
lockdown	324	0	lockdown	340	0	lockdown	72	0
rouškovné *a one-off compensation for face mask expenses for people receiving a pension*	12	0	antigenový$_{ADJ}$ *antigen*$_{ADJ}$	8	0	odmítač *rejector*	8	0

There are no keywords related to vaccines found in the first wave corpora, and most of the second wave keywords related to vaccines are found in each of the corpora. Even though many of these keywords did occur during the first wave, they were not key, and they were much more frequent during the second wave. Table (47) shows the full comparison of the second wave keywords.

Table (47): Cross-server comparison of keywords related to vaccines in the second wave corpora

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
naočkovat$_{V}$ *lit. on-vaccinate*	289	1	naočkovat$_{V}$ *lit. on-vaccinate*	389	3	vakcinační$_{ADJ}$ *vaccination*$_{ADJ}$	29	0
očkovat$_{V}$ *to vaccinate*	1025	23	očkovat$_{V}$ *to vaccinate*	957	18	očkovat$_{V}$ *to vaccinate*	248	9
očkovaný$_{ADJ}$ *vaccinated*	300	8	vakcinace *vaccination*	200	3	očkovaný$_{ADJ}$ *vaccinated*	69	3
očkovací$_{ADJ}$ *vaccination*$_{ADJ}$	854	30	naočkovaný$_{ADJ}$ *lit. on-vaccinated*	109	1	naočkovat$_{V}$ *lit. on-vaccinate*	68	3
očkování *vaccination*	2119	104	vakcinační$_{ADJ}$ *vaccination*$_{ADJ}$	77	1	proočkovat$_{V}$ *lit. through-vaccinate*	13	0
naočkovaný$_{ADJ}$ *lit. on-accinated*	60	1	očkování *vaccination*	1767	67	vakcinace *vaccination*	89	7
vakcinační$_{ADJ}$ *vaccination*$_{ADJ}$	102	4	proočkování *lit. through-vaccination*	28	0	naočkovaný$_{ADJ}$ *lit. on-vaccinated*	19	1

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
vakcinace *vaccination*	239	17	proočkovat$_V$ *lit. through-vaccinate*	64	2	očkování *vaccination*	605	58
proočkovat$_V$ *lit. through-vaccinate*	63	4	naočkování *lit. on-vaccination*	25	0	proočkování *lit. through-vaccination*	8	0
přeočkování *re-vaccination*	17	0	očkovaný$_{ADJ}$ *vaccinated*	232	18	očkovací$_{ADJ}$ *vaccination*$_{ADJ}$	281	33
naočkování *lit. on-vaccination*	34	2	vyočkovat$_V$ *lit. vaccinate out*	14	0	přeočkování *re-vaccination*	6	0
vakcína *vaccine*	3530	415	očkovací$_{ADJ}$ *vaccination*$_{ADJ}$	814	84	naočkování *lit. on-vaccination*	6	0
vakcinolog *vaccinologist*	29	2	proočkovanost *lit. the rate of through-vaccination*	45	4			
vyočkovat$_V$ *lit. vaccinate out*	14	0	vyočkovaný *lit. vaccinated out*	10	0			
proočkování *lit. through-vaccination*	23	2	proočkovaný *lit. through-vaccinated*	23	2			
			vakcinologický$_{ADJ}$ *vaccinological*	22	2			
			vakcína *vaccine*	4194	694			

Considering keywords related to face masks and respirators, there are no such keywords in the LIDOVKY corpora. There are, however, keywords in the IDNES and NOVINKY corpora, which are shown in Table (48) and Table (49), and all of them are exclusive to their corpus. There are not any meaningful differences between the first and second wave in these keywords, though.

Table (48): Cross-server comparison of keywords related to face masks in the first wave corpora

IDNES_1			NOVINKY_1		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
celoobličejový$_{ADJ}$ *full-face*$_{ADJ}$	12	0	maska *mask*	283	26
ffp3	119	19			

Table (49): Cross-server comparison of keywords related to face masks in the first wave corpora

IDNES_2			NOVINKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
nanorespirátor *nano respirator*	10	0	kn95	18	0
			nanorouška *nano face mask*	32	3

Looking at the keywords related to the virus itself, there are again no keywords from the first wave. Table (50) shows the keywords from the second wave corpora. All of these corpora include the keywords *covidový*$_{ADJ}$ (covid$_{ADJ}$), *covid*, and *mutace* (mutation), which are also present in the first wave corpora, but not as keywords. There are also keywords exclusive to the second wave – *postcovidový*$_{ADJ}$ (post-covid$_{ADJ}$), *zmutovaný*$_{ADJ}$ (mutated), *proticovidový*$_{ADJ}$ (anti-covid$_{ADJ}$), and *covidní*$_{ADJ}$ (covid$_{ADJ}$).

Table (50): Cross-server comparison of keywords related to the virus in the second wave corpora

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
covidový$_{ADJ}$ *covid*$_{ADJ}$	583	28	covidový$_{ADJ}$ *covid*$_{ADJ}$	284	12	zmutovaný$_{ADJ}$ *mutated*	33	0
postcovidový$_{ADJ}$ *post-covid*$_{ADJ}$	35	0	mutace *mutation*	930	86	mutace *mutation*	343	12
mutace *mutation*	815	49	proticovidový$_{ADJ}$ *anti-covid*$_{ADJ}$	13	0	covidní$_{ADJ}$ *covid*$_{ADJ}$	21	0
zmutovaný$_{ADJ}$ *mutated*	33	0	covid	1159	141	covidový$_{ADJ}$ *covid*$_{ADJ}$	122	7
covid	1654	187				covid	628	72
						proticovidový$_{ADJ}$ *anti-covid*$_{ADJ}$	7	0

Keywords related to coronavirus testing are another case of keywords that are found only during the second wave. The only keyword present in a first wave corpus is *pcr*, the rest of the keywords does not occur during that time at all. Then there is only one keyword shared between all the corpora – *kloktací* (gargling), and there are several keywords including the prefix *samo-* or *sebe-* (self-). LIDOVKY_2 also includes another keyword of English origin – *gargtest*. Table (51) below contains the full comparison.

Table (51): Cross-server comparison of keywords related to testing in the second wave corpora

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
kloktací$_{ADJ}$ *gargling*$_{ADJ}$	21	0	samoodběr *self-sampling*	14	0	kloktací$_{ADJ}$ *gargling*$_{ADJ}$	16	0
samotest *self-test*	11	0	samotest *self-test*	12	0	sebereportování *self-reporting*	8	0
samotestování *self-testing*	11	0	sebetrasování *self-tracing*	10	0	rt-pcr	7	0
			samotestování *self-testing*	10	0	gargtest	7	0
			pcr	333	47			
			kloktací$_{ADJ}$ *gargling*$_{ADJ}$	8	0			

There are not many medical expressions and terms appearing during the first wave. IDNES_1 and LIDOVKY_1 contain the word *uzdravený*$_{ADJ}$ (healed), while NOVINKY_1 contains the word *zotavený*$_{ADJ}$ (recovered). All the other keywords are not shared between the corpora. See Table (52) for all the first wave key medical expressions and terms. With the second wave medical expressions and terms shown in Table (53), there are two keywords shared among the corpora – *reinfekce* (reinfection) and *epidemicky*$_{ADV}$ (epidemically). Despite the coronavirus situation fitting the definition of a pandemic, all three second wave corpora mention an epidemic, but only NOVINKY_2 contains a keyword referring to a pandemic – *pandemický*$_{ADJ}$ (pandemic$_{ADJ}$).

Table (52): Cross-server comparison of medical expressions and terms in thefirst wave corpora

IDNES_1			NOVINKY_1			LIDOVKY_1		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
uzdravený$_{ADJ}$ *healed*	144	18	zotavený$_{ADJ}$ *recovered*	222	0	medicinální$_{ADJ}$ *medicinal*	14	0
			ebola	59	6	antibakteriální$_{ADJ}$ *antibacterial*	8	0
			dezinfikován$_{V}$ *disinfected*	31	3	uzdravený$_{ADJ}$ *healed*	89	6

Table (53): Cross-server comparison of medical expressions and terms in the second wave corpora

IDNES_2			NOVINKY_2			LIDOVKY_2		
Word	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**	**Word**	**Freq(F)**	**Freq(R)**
epidemický$_{ADJ}$ *epidemic*$_{ADJ}$	280	26	epidemický$_{ADJ}$ *epidemic*$_{ADJ}$	200	9	reinfekce *reinfection*	17	0
prodělání *suffering an illness in the past*	48	3	protiepidemický$_{ADJ}$ *anti-epidemic*$_{ADJ}$	483	39	epidemicky$_{ADV}$ *epidemically*	26	2
reinfekce *reinfection*	27	1	epidemicky$_{ADV}$ *epidemically*	17	1			
epidemicky$_{ADV}$ *epidemically*	36	3	pandemický$_{ADJ}$ *pandemic*$_{ADJ}$	140	18			
protiepidemický$_{ADJ}$ *anti-epidemic*$_{ADJ}$	570	85	reinfekce *reinfection*	21	2			
imunolog *immunologist*	122	17	autoprotilátka *autoantibody*	8	0			

These sections showed the already mentioned patterns in keywords more clearly. The data shows that for example keywords related to face masks had little to no difference between the first and the second wave, which is indicated by the overall low number of keywords. Vaccines, on the other hand, were the largest group of keywords, although most of the keywords in this group were words derived from *očkovat*$_{V}$ (to vaccinate)/*očkování* (vaccination). It was also shown that neologisms are the same for all the servers for the most part, with one keyword being different for each server. The following section attempts to interpret why some of these patterns occurred.

4.3 Interpreting the data

The data shows that keywords are a good reflection of a situation. One can observe this simply from the number of keywords that occur in each category – there are usually considerably lower numbers of relevant keywords during the first wave than the second (see Table (14) for the overview of the numbers of keywords). This is likely due to the coronavirus crisis not being severe during the first wave as there was simply not much to report on, but as the number of infected people grew, so did the media attention and consequently the number of keywords appearing.

At first, people were using nonprescription drugs and doctors were prescribing drugs used to treat similar illnesses similar to coronavirus. The coronavirus vaccine became a major topic only later on, and that is why names of vaccine manufacturers and keywords related to vaccines appear as keywords only during the second wave.

Looking at neologisms, the two most frequent keywords were *antigenní*$_{\text{ADJ}}$ (antigen$_{\text{ADJ}}$) and *lockdown*, with both of them appearing only during the second wave. There seems to be no Czech single-word equivalent to the word *lockdown* (Zajímavé dotazy - Ústav pro jazyk český, n.d.), and newspapers simply adopted the English term for convenience. People, however, showed some witty coinage in other areas with keywords such as *rouškovník* (face mask tree)*, rouškomat* (face mask vending machine) and *rouškovné* (a one-off compensation for face mask expenses for people receiving a pension). Furthermore, the whole coronavirus crisis was probably made worse due to a wide spread of fake news and hoaxes, which resulted in some people not believing the severity of the situation, and the keyword *odmítač* (rejector) reflects that.

The keywords related to face masks reflect government orders quite well, as covering one's face with a face mask of a specified class was mandatory. Furthermore, this portion of the data shows that there is not almost any difference between the first and the second wave. This can mean two things – either there were not many keywords in this category in general, or the way face masks were talked about did not change between the first and second wave, resulting in only a few keywords appearing. Other than that, this portion of the data does not reveal much.

It is difficult to explain with certainty the lack of keywords related to the virus itself in the first wave, but it is likely due to the already mentioned lower media attention at that time. Concerning the second wave, however, there are a few interesting observations to make. Even though the word *mutace* (mutation) occurs in the first wave corpora, it is key in the second wave corpora as compared to the first. However, the keyword *zmutovaný*$_{\text{ADJ}}$ (mutated) does not occur in the first wave corpora at all, which is quite likely due the time needed for the virus to evolve and mutate. INDES_2 also contains the keyword *postcovidový*$_{\text{ADJ}}$ (post-covid$_{\text{ADJ}}$). At first sight, it seems rather strange that such a keyword appears when the situation was still quite extreme during the time when the data was collected. However, after looking at the words following the keyword *postcovidový*$_{\text{ADJ}}$ (post-covid$_{\text{ADJ}}$), one can see that the expressions this keyword appears in are *postcovidový syndrom* (post-covid syndrome), *postcovidové centrum* (post-covid center), or *postcovidová péče* (post-covid care), and not *postcovidová doba* (post-covid age), as one might have thought.

The lack of keywords related to coronavirus testing during the first wave can be explained with more certainty than was the case with the keywords related to the virus – there simply were not so many infected people during that time, therefore there was not a need for extensive testing. On the other hand, the keywords from the second wave, such as *kloktací*$_{\text{ADJ}}$ (gargling$_{\text{ADJ}}$) or *gargtest* show the effort to develop non-invasive testing methods, while the prefixes *sebe*- and *samo*- (self-) show the effort to decrease human contact in some situations when testing was needed.

With keywords such as *dezinfikován*$_{\text{V}}$ (disinfected) or *antibakteriální*$_{\text{ADJ}}$ (antibacterial), the medical expressions and terms appearing during the first wave show some effort of decreasing the chance of spreading the disease, while the keywords *zotavený*$_{\text{ADJ}}$ (recovered) and *uzdravený*$_{\text{ADJ}}$ (healed) show that media were trying to inform about a positive aspect of the coronavirus statistics. The reason why these words do not appear as keywords during the second wave is perhaps that the newspapers stopped reporting on the numbers of people that recovered, and not that people would stop recovering. The keyword *ebola* referring to a different infection than the coronavirus appeared likely due to some similarities between these two viruses that allowed for a faster vaccine development due to the research done on the *ebola* virus. The second wave key medical expressions and terms show that the risk of being infected with the coronavirus for the second time was relevant only during the second wave, which is indicated by the keyword *reinfekce* (reinfection). Other than that, this group of keywords does not show much except for that the three selected newspapers were referring to the pandemic as an epidemic.

5 Conclusion

This thesis aimed to collect articles related to the coronavirus pandemic from online newspapers and to create corpora from these articles to allow for comparison of keywords between the first and the second wave of the pandemic. The data was collected with a custom-made script for each newspaper and the corpora were created with Sketch Engine.

Sketch Engine was also used for keyword extraction. The extracted keywords were manually corrected to get rid of errors made by the software and then divided into thematical groups. Keywords within these groups from the first wave and the second wave were then compared against each other to find differences between the waves.

It was found that the keywords reflect the events of the pandemic rather accurately, but also less obvious information when comparing the first and second wave corpora was found – namely occurrence of newly coined and borrowed words and lexical change.

Other research that could be done on this sample of data is multi-word keyword analysis, collocation analysis or analysis of conceptual metaphors related to the virus. It would also be possible to perform the same analysis on a number of keywords larger than top 300 or collecting data from more newspaper servers and repeating the analysis from this thesis.

6 Resumé

S rostoucí mediální pozorností koronavirové krize začal růst i počet článků zmiňujících koronavirus, což umožnilo provést korpusový průzkum těchto článků. Za účelem takového průzkumu byl použit online software Sketch Engine, který umožňuje vytváření vlastních korpusů a také nabízí řadu nástrojů k jejich zkoumání.

I když Sketch Engine nabízí možnost automatického stahování článků z internetu, takto stažené články často obsahovaly i data nechtěná pro průzkum, jako například diskuze pod články. Z tohoto důvody byly v rámci této práce vyvinuty tři skripty, z nichž každý stahoval relevantní články z jiného webu. Přestože tyto programy nejsou bezchybné a jejich funkčnost lze v některých oblastech zlepšit, osvědčily se jako adekvátní metoda pro získávání dat.

Nasbíraná data byla rozdělena na články z první a druhé vlny koronaviru a z takto rozdělených dat byly vytvořeny korpusy. Tímto způsobem vzniklo šest korpusů, dva pro každý server – idnes.cz, Lidovky.cz a Novinky.cz.

Pro analýzu klíčových slov byl také použit Sketch Engine, který však ve svých výpočtech používá metodu Simple maths, která nepoužívá testy signifikance. Za účelem ověřování statistické signifikance byla použita korpusová kalkulačka (Český národní korpus, n.d.), která také vypočítává tzv. difference index (DIN), což je hodnota určující relevanci daného klíčového slova.

Jako referenční korpus pro analýzu klíčových slov vždy posloužil druhý korpus z dané dvojice, korpus z první vlny tedy měl korpus z druhé vlny jako referenční a naopak. Použití obecného korpusu se v rámci této práce jevilo jako nedostačující pro srovnání obou vln.

Klíčová slova byla rozdělena do tematických skupin a poté byla slova z první vlny koronaviru z dané skupiny porovnána s klíčovými slovy ze stejné skupiny z druhé vlny. Tímto způsobem byly objeveny rozdíly v mezi jazykem první a druhé vlny. Bylo zjištěno, že vlivem koronavirové krize se v češtině začaly používat výpůjčky z angličtiny. Dále také bylo zjištěno, že se objevilo několik neologismů a že klíčová slova velmi dobře reflektují průběh dané situace.

Vyvinuté metody pro stahování článků je možné využít v případě potřeby stahování článků o koronaviru pro účely provedení podobného průzkumu na větším vzorku dat. Lze je však využít i pro stahování článků z daných webů za účely vytvoření jakéhokoliv jiného specializovaného korpusu. Již nasbíraná data lze použít i v případě budoucího výzkumu koronavirové krize.

7 Works cited

Aktuální informace o COVID-19. 2020. "V České Republice Jsou První Tři Potvrzené Případy Nákazy Koronavirem." Accessed March 1, 2021. https://koronavirus.mzcr.cz/v-ceske-republice-jsou-prvni-tri-potvrzene-pripady-nakazy-koronavirem/.

Culpeper, Jonathan, and Jane Demmen. 2015. "Keywords." *The Cambridge Handbook of English Corpus Linguistics.* Edited by Douglas Biber, and Randi Reppen. 90-105. Cambridge: Cambridge University Press.

Český národní korpus. n.d. "Calc: Korpusová Kalkulačka." Accessed June 12, 2021. https://www.korpus.cz/calc/.

Fidler, Masako, and Václav Cvrček. 2015. "A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis." *Journal of Slavic Linguistics* 23, no. 2: 197–239. https://doi.org/10.1353/jsl.2015.0018.

Fonseca, Erick. 2019. "State-of-the-Art Multilingual Lemmatization." Medium. Towards Data Science Accessed March 23, 2021. https://towardsdatascience.com/state-of-the-art-multilingual-lemmatization-f303e8ff1a8.

Kilgarriff, Adam. 2009. "Simple maths for keywords." In *Proceedings of Corpus Linguistics Conference CL2009,* edited by Michaela Mahlberg, Victorina González-Díaz, and Catherine Smith. Liverpool: University of Liverpool. https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years On." *Lexicography* 1, no. 1: 7–36. https://doi.org/10.1007/s40607-014-0009-9. http://www.sketchengine.eu/

McEnery, Tony. 2012. "Corpus Linguistics." *Oxford Handbooks Online.* https://doi.org/10.1093/oxfordhb/9780199276349.013.0024.

Onemocnění aktuálně. n.d. "COVID-19: Onemocnění aktuálně MZČR." Accessed March 1, 2021. https://onemocneni-aktualne.mzcr.cz/covid-19

Pojmy:chi2 - Příručka ČNK. n.d. "Test Chi Kvadrát (χ2 Test)." Accessed June 12, 2021. https://wiki.korpus.cz/doku.php/pojmy:chi2

Pojmy:din - Příručka ČNK. n.d."DIN." Accessed June 12, 2021. https://wiki.korpus.cz/doku.php/pojmy:din

PYPL index. n.d. "PYPL PopularitY of Programming Language Index." Accessed March 1, 2021. https://pypl.github.io/PYPL.html.

Scott, Mike. 2009. "In Search of a Bad Reference Corpus." *What's in a Word-list? Investigating Word Frequency and Keyword Extraction.* Edited by Dawn Archer. 79-92. Oxford: Ashgate.

Scott, Mike, and Christopher Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia: J. Benjamins.

Sketch Engine. 2020. "Simple Maths with Keywords and Terms." Accessed March 3, 2021. https://www.sketchengine.eu/documentation/simple-maths/

Suchomel, Vít. 2018. “csTenTen17, a Recent Czech Web Corpus.” In *Proceedings of Twelfth Workshop on Recent Advances in Slavonic Natural Language Processing*, edited by Aleš Horák, Pavel Rychlý, and Adam Rambousek. 111-123. Brno: Tribun EU.

Weisser, Martin. 2016. *Practical corpus linguistics: an introduction to corpus-based language analysis*. Malden, Massachusetts, USA: Wiley-Blackwell.

Xiao, Zhonghua, and Anthony McEnery. 2005. “Two Approaches to Genre Analysis.” *Journal of English Linguistics* 33, no. 1: 62-82. https://doi.org/10.1177/0075424204273957.

Zajímavé dotazy - Ústav pro jazyk český. n.d. “Zajímavé Dotazy.” Accessed July 22, 2021. https://ujc.avcr.cz/jazykova-poradna/zajimave-dotazy/201022-lockdown.html.