Pedagogická fakulta
Faculty of Education

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých
Budějovicích Pedagogická fakulta
Katedra anglistiky

Bakalářská práce

Collocability of the most frequent nouns in COVID-19 forum threads

Kolokabilita nejfrekventovanějších substantiv na diskusních fórech ohledně COVID-19

Vypracoval: Martin Svoboda
Vedoucí práce: Mgr. Jaroslav Emmer

České Budějovice 2021

Prohlášení

Prohlašuji, že svoji bakalářskou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 8. července 2021.

Martin Svoboda

Poděkování

Na tomto místě bych rád poděkoval mému vedoucímu práce Mgr. Jaroslavu Emmerovi za jeho užitečné rady a připomínky, bez nichž by tato práce nemohla vzniknout.

Anotace

Tato bakalářská práce se zabývá kolokabilitou nejfrekventovanějších substantiv používaných účastníky na diskusních fórech ohledně tématu COVID-19. Internetová diskusní fóra jsou známá především různorodostí názorů jejich navštěvovatelů, kde každý z nich může mít svůj specifický styl vyjadřování, formování vět a také slovník slov, které často využívá. Teoretická část se zabývá především pojmy fraseologie a kolokace. V rámci kolokace je kromě obecného popisu pohlíženo i na jednotlivé druhy a typy. Závěr teoretické části se zabývá přiblížením pojmu kolokability. Praktická část obsahuje korpusový výzkum dat získaných pomocí skriptu z diskusního fóra Reddit. Tato data jsou podrobena analýze a je z nich vytvořen korpus s pomocí programu #LancsBox. Následně je v korpusu vytvořen frekvenční seznam substantiv a vypracován kolokační profil pro ta nejčastější substantiva. Práce zkoumá, jaká slovní spojení se nejčastěji vyskytují mezi neprofesionály na diskusních fórech ohledně aktuálního tématu COVID-19 a v jaké frekvenci se tak vyskytují.

Abstract

This Bachelor's thesis deals with the collocability of the most frequent nouns used by users in COVID-19 forum threads. Online discussion forums are well-known for the diversity of opinions of its users where every user may have their own specific style of expression, sentence formation, and also vocabulary of often used words. The theoretical part mostly concentrates on the terms "phraseology" and "collocations". Regarding collocations, besides general description a closer look at individual categories and types of collocations is also provided. The ending of the theoretical part concentrates on collocability and its concept. The practical part contains the corpus research of data gathered using a script from a Reddit forum. This data is subjected to analysis and is used in the creation of a corpus with the help of #LancsBox software. Subsequently, a frequency list of nouns is created and a collocational profile for the most frequent nouns is developed. The thesis analyses which collocations most often occur among non-professionals in discussion forums on the current topic of COVID-19 and at what frequency they occur.

Table of contents

I. INTRODUCTION 7
II. THEORETICAL PART 8
1. What is phraseology and idiomatics? 8
1.1. Phrasemes and idioms 9
2. What are collocations? 11
2.1. Examining collocations 12
2.2. Corpus linguistics and Corpus 13
2.3. Types of collocations 14
2.4. Selectional preferences 15
3. What is collocability? 17
III. PRACTICAL PART 18
1. Reddit.com 18
1.1. Method of research and data collection 19
1.2. Chosen subreddit 22
2. #LancsBox and used functions 23
2.1. The "Words" function 23
2.2. The "GraphColl" function 25
2.3. Associations measures 26
2.3.1. MI-score 26
2.3.2. T-Score 27
3. Analysis 28
3.1. Analysed posts 28
3.1.1. Post #1 28
3.1.2. Post #2 28
3.1.3. Post #3 29
3.1.4. Post #4 29
3.1.5. Post #5 30
3.1.6. Post #6 30
3.1.7. Post #7 30
3.1.8. Post #8 31
3.1.9. Post #9 31
3.1.10. Post #10 32
3.1.11. Post #11 32
3.1.12. Post #12 33
3.1.13. Post #13 33

3.1.14. Post #14 34
3.1.15. Post #15 34
3.2. Method of Analysis 35
4. Results 37
4.1. The word “People” 37
4.2. The word “Time” 37
4.3. The word “Mask” 40
4.4. The word “Home” 42
4.5. The word “Thing” 43
4.6. The word “Day” 45
4.7. The word “Virus” 47
4.8. The word “Lot” 49
4.9. The word “Case” 51
4.10. The word “Country” 52
IV. CONCLUSION 54
V. REFERENCES 55

I. INTRODUCTION

As of writing this, COVID-19 is still a pressing issue all around the world. It is practically impossible to avoid people speaking about it on TV, seeing people outside wearing masks or witnessing long conversations unfolding on social media and online forum threads.

On the one hand online forums are a great way to find like-minded individuals to discuss certain topics, search for useful information and tips left by fellow users or even help someone else with their problem or research. On the other hand, some online forums can also lead to a cesspool of misinformation, constant arguments, especially concerning politics, and even cyberbullying. Experiences may vary from person to person and by the forum one visits.

Even if it is not always positive, online forum threads can be a valuable source for analysis and interesting research. These types of sites are visited daily by millions of people from all around the world with different mindsets, interests, and language skills. However, forum threads also tend to vary greatly. While some may prefer news threads where they can discuss in the comments about the recent happenings in their country, some just like to browse funny pictures or just chat with others and relax.

One of, if not the most popular online forum, is Reddit. With more than 430 million monthly active users worldwide (Reddit, 2021) and being the sixth most popular social networking mobile app in the US (Statista, 2019), Reddit is almost starting to become a synonym for the word "forum". This popularity was most likely achieved by its sheer number of subreddits, which are a kind of subforum, there being more than 2.2 million different ones (Metrics For Reddit, 2020). The topics that these subreddits are dedicated to can vary from politics or current issues to cooking, fitness, beauty, or even true crime speculations.

Unsurprisingly, one of the more popular topics on Reddit currently is COVID-19, which took the world by storm. This worldwide issue spawned many subreddits. On some of them, users discuss COVID-19 in general, while on other discussions are held about COVID-19 on specific continents, in specific countries or even discussions between people who have suffered from it or are dealing with it right now. This thesis aims to observe the most frequent nouns used by COVID-19 subreddit threads participants and the collocates of these nouns.

II. THEORETICAL PART

1. What is phraseology and idiomatics?

Phraseology and idiomatics (PI henceforth) is a relatively young branch of linguistics (separate from lexicography because it has special needs for its application), and it only gradually comes to light where in language we may look for its units, idioms, and phrasemes. In language, they are, generally speaking, found in all branches of linguistics wherever combinations of units with meaning are involved. Moreover, PI has links to other non-linguistic disciplines, outside of language (Čermák, 2007).

With its origin in the Greek words of *phrasis* (“*way of speaking*”) and *-logia* (“*study of*”) PI is a scholarly approach to language which developed in the twentieth century (Knappe, 2004). After its tortuous beginnings in various countries, PI has become established as an independent scientific discipline and area of research. In part, this is also due to the European Association for Phraseology or Europhras, founded as late as the nineties, which nowadays organizes its own congresses and publishes their proceedings (Čermák, 2007).

Being a scientific discipline, PI is mostly concerned with the study of set or fixed expressions, such as idioms, phrasal verbs, and other types of multi-word lexical units (often collectively referred to as phrasemes), in which the component parts of the expression take on a meaning more specific than or otherwise not predictable from the sum of their meanings when used independently (Wikipedia, 2021).

Idiomatic and phraseological expression is a fixed combination of at least two words of any word class which does not have the same meaning when one of the words combines with a different element (the term for this being *non-compositionality*). The meaning of such a combination is figurative. Čermák (2009) said that a difference between a phrase and an idiom is in their formal and expressive point of view (phrase) and semantic and functional point of view (idiom) (Čermák, 2009).

1.1. Phrasemes and idioms

According to Čermák, phraseme is a unique combination of minimum two words, from which either of them does not work in the same way in a combination with a different word, or it occurs only in one expression. As the phraseme is fixed, it has a meaning as a whole and it is not possible to enter in any other elements (Čermák & Šulc, 2006). The elements of a phraseme can be either compatible or incompatible. A phraseme with compatible elements can have either idiomatic or literal meaning. Čermák explains this on Czech examples like "*bledá tvář*", which has both idiomatic and literal meanings (a white person in films about Indians/a face that is literally white) or "*dutá hlava*" which has only one idiomatic meaning (a fool). If one element would change, the meaning would be unrecognizable, e.g. "*dutá ruka*". (Čermák & Šulc, 2006)

An idiom is a "frozen expression" in which the meaning of the whole does not reflect the meanings of the component parts. For example: "*to kill two birds with one stone*" = achieving two things with one action; "*break a leg*" = wishing someone luck. It does not literally mean an act of killing any bird or the breaking of one's leg. (Benson et al., 1993) Some idioms can be difficult to tell apart, especially for a non-native speaker of a given language whose native language does not include a similar idiom that they can refer to.

As described in "Collocations in a Learner Corpus" (Nesselhauf, 2005), word combinations may also be divided into four groups:

1. Free combinations – the elements of combination are used in the literal sense, e.g. "*drink tea*" and substitution can happen within a semantic field.
2. Restricted collocations – at least one element is used in its literal meaning, the other one has non-literal meaning, e.g. "*perform a task*", and substitution is limited.
3. Figurative idioms – they have figurative meaning but have literal interpretation, e.g. "*U-turn*" – to change one's behaviour. Substitution is rarely possible.

4. Pure idioms – they have figurative meaning and do not have literal interpretation, e.g. “*blow the gaff*”. It is not possible to substitute the elements at all.

   (in Nesselhauf, 2005)

Phrasemes and idioms can be commonly seen in all languages with most of them being meant literally in their original use. As time passes, we will most likely encounter some brand-new idioms that have grown away from its original literal meaning and now may be associated with something entirely different.

2. What are collocations?

According to Futurelearn.com (2021) with the help of Macquarie University in Sydney, Australia, collocation is a group of two or more words that are almost always put together to create a specific meaning. Using a different combination of words sounds unnatural or awkward (Future Learn, 2021). These combinations "sound right" to native English speakers, who tend to use them regularly. For example, "*a fast train*" vs. "*a quick train*". While native English speakers associate the word "*fast*" with movement and the word "*quick*" with passage of time and can quickly tell which collocation is natural and which is not, non-native English speakers may have a problem telling the difference. This does not necessarily mean that the non-native speaker will not be understood, but the listener may have to concentrate harder on the speech, which may result in some communication problems or difficulties. Using the correct collocations can also be beneficial if a speaker wants their speech to contain more information in a shorter context (Barfield & Gyllstad, 2009).

Word pairings like these are very important and common among native speakers and unfortunately, there is no easy rule to learn. However, what might help with learning is that people tend to remember collocations more easily than individual words. Learning and remembering a collocation can be highly beneficial especially for learners since it can help them with further learning, if they can identify and spot a familiar collocation in a text and help them feel more confident about their language abilities (Nesselhauf, 2005).

According to Čermák (2006), collocations are also very important in the field of education where teachers can use textbooks and materials based on collocation studies to help their students sound more natural. In addition, he states that translators may also find some benefit in collocations; by consulting a dictionary one can find a more natural-sounding expression and make their translations more native-like. Up to a few decades ago English textbooks presented individual vocabulary as the most essential part of language in favour of collocations and their many variations which were overlooked (Barfield & Gyllstad, 2009).

2.1. Examining collocations

In theoretical terms, collocations can be defined as lexical relations between two or more words which tend to appear and co-occur within a few words of each other. In this broad sense, collocations may also take different shapes and forms. To get a better idea of the various levels at which the co-occurrence of words may be defined, we may examine the four types that Sinclair (1991) distinguished: *collocation, colligation, semantic preference, and semantic prosody* (Geeraerts, 2010). However, we will focus only on collocations because they are the main topic in this thesis.

As Geeraerts (2010) describes, the target word of a collocation is often called the *node,* and the co-occurring word the *collocate*. An often-used way of examining is to produce a concordance of a text or a set of texts, i.e. an alphabetical list of the words in those texts and their immediate context. The usual way of representing a concordance is the Key Word in Context index (or KWIC), although in this thesis its use is not as prevalent. It is mostly used as an optional way of examining the collocates of select nodes, their position around the node (right or left of the node), the distance between the collocates and the nodes and if the collocates are a part of the same sentence the node is in or not (Geeraerts, 2010).

The node of a collocation analysis may be a word form or a word, if lemmatization can be applied, i.e. if all the inflectional forms of a word are treated as instances of a single lexical unit. Nodes themselves may also be complex expressions or phrases. Although words like *a, the, is, are, by, from*, etc. (also called *stop words*) that are much less illuminating and carry less semantic value may negatively impact the result of collocational analyses, although there are ways to overcome this, such as the use of stop lists as filters or the use of different association measures that mostly filter such words out (Geeraerts, 2010). In this thesis, the MI-score association measure is used to mostly filter out stop words, although they are included in the results of the T-score association measure for illustration and data completeness.

2.2. Corpus linguistics and Corpus

Defined by some as “an area which focuses upon a set of procedures, or methods, for studying language” (McEnery, T. & Hardie, A., 2011), corpus linguistics is a term closely related to the study of collocations and is important for this thesis. While not being a separate branch of linguistics or a theory of language, it is used as a methodology to obtain and either quantitatively or qualitatively analyse language data. It can be applied to almost any area of language studies using authentic, naturally occurring language use as the object of the study (University of Helsinki, 2016).

One term also closely associated with corpus linguistics is the term “corpus” itself. Defined as “in linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database” (McArthur, 1992). One of the main purposes of a corpus is to verify a hypothesis about a language, for example to determine what variations can occur when using a particular sound, word, or syntactic construction. Corpora can also be used as a starting point of linguistic description (Crystal, 1991).

While it may sound to people unfamiliar with corpora that basically any kind of text can be a corpus or can be turned into a corpus, the reality is quite different. The text of a corpus must be representative with regard to the hypothesis, it has to have a defined size and its content is electronically stored, as it is easier to obtain information on frequencies, grammatical patterns, and collocations by means of a computer than manually. It should be also freely available, so the research results can be contrasted, compared, and repeated (University of Helsinki, 2016).

2.3. Types of collocations

According to Kaplan International Languages (2021), categorizing collocations can make it easier for people to learn them.

The first category they describe is Strong vs. Weak (or Lexical) Collocations. Strong collocations are those words that do not necessarily match with many other words. The connections are strong because there are not many other acceptable options to say the same thing. They point out that for example, the expression "turn on a light" is strong because most other synonyms will sound rather strange and unnatural, like "*start a light*" or "*activate a light*". Weak collocations are the exact opposite of this. They include words that have many other options. For example, the expression "*very interesting*" is commonly used, but the collocation is weak, so substitutes like "*extremely interesting*" or "really interesting" are also acceptable (Kaplan International, 2021).

The second category they describe is Grammatical Collocations. This is then further categorized into: Adverb collocations (adverb + adjective), Adjective collocations (adjective + noun), Noun collocations (noun + noun/verb) and Verb collocations (verb + noun/adverb) (Kaplan International, 2021).

Although Wei (1999) goes more into detail with Grammatical Collocations, he also describes a third collocational category in his work "Teaching Collocations for Productive Vocabulary Development". Concerning Grammatical Collocations, he divides them into two sub-categories, one being "Grammatical collocations that contain a preposition" and the other being "Grammatical collocations that involve a grammatical Structure". He then goes into more detail showing contrasting examples. As the third category, he decided to include Idiomatic expressions, saying that idiomatic expressions are the most fixed word combinations, where substitution of any of their components is virtually impossible, for example, "*kick the bucket*", "*play it by ear*", "*let one's hair down*", etc. (Wei, 1999).

2.4. Selectional preferences

Selectional preferences can be defined as the tendency of a predicate to favour particular arguments within a certain linguistic context and also as the tendency to reject other arguments that would result in conflicting or implausible meanings (Metheniti & Van de Cruys & Hathout, 2020).

Most predicates (such as verbs) have a strong tendency to favour certain arguments to others that even though syntactically correct, would more likely be judged as awkward or ill-formed by most native speakers. Consider the following examples:

1) *The cyclist is riding down the street.*

2) *The apple is riding down the fridge.*

The first example is semantically felicitous, as it is perfectly normal and possible for a cyclist to ride and a street to be ridden on. On the other hand, the second example is semantically infelicitous because both an apple and a fridge are inanimate objects that have no literal capability of motion. Although under very specific circumstances (animated film, cartoon, etc.) it is also possible for the second example to make perfect sense and be semantically felicitous (Metheniti & Van de Cruys & Hathout, 2020).

This act of preference of predicates is known as selectional preference. This phenomenon is very important within various natural language processing (NLP) applications and can be used as an additional knowledge source for various tasks involving natural language processing (Metheniti & Van de Cruys & Hathout, 2020).

When researching selectional preferences, we are often also interested in their strength. Selectional preference strength can be defined as the amount of information that a predicate tells us about the semantic class of its arguments. For example:

1) The verb "*eat*" tells us a lot about the semantic class of its direct objects (typically something that can be eaten or something with the ability to eat)

2) The verb "*be*" does not tell us much (as it can be used in a plethora of ways with multiple different meanings)

To what extent does the verb constrain its object can be determined by the difference in information between the distribution of expected semantic classes for any direct object and the distribution of expected semantic classes for this verb. The greater the difference, the more the verb is constraining its object (Stanford University, 2016).

3. What is collocability?

According to Čermák, collocability is the individual, formal and semantic combinability of language elements. This can also be explained as the ability of each language element to combine with another or others. It results from one or more of the element's collocational paradigms and (in regular combinations) is conditioned by its compatibility with them. Together with valency it is the main component of syntagmaticity of any language element. The individual realization of collocability produces a collocation (Čermák, 2007).

In his work Collocations, Collocability and Dictionary, he also claims that the whole collocational range (or collocability) of most words is and seems to be so large and unlimited that it is never given in full. Despite that, Čermák states that there is a select group of words that is evidently and strictly in its collocational capacity. This group has a very small list of collocates, which reverts the view adopted so far and suggests the possibility of viewing both the head and collocate as a single unit, identical, in many ways to idioms, compared to "*afraid*" (be afraid), "*afoul*" (run afoul), etc. (Čermák, 2006).

III. PRACTICAL PART

1. Reddit.com

In their own words: “Reddit is home to thousands of communities, endless conversation, and authentic human connection. Whether you are into breaking news, sports, TV fan theories, or a never-ending stream of the internet's cutest animals, there's a community on Reddit for you.” (Reddit Inc., 2021).

In other words, Reddit is a modern phenomenon among the younger generation. Being the 19th-most-visited website in the world, 7th-most-visited website in the US (as of April 2021) and having around 52 million daily active users, 100 thousand different communities and 50 billion daily views, speaks greatly of Reddit’s popularity (Reddit Inc., 2021). Basically, Reddit is a massive community-driven forum that consists of many other smaller forums called “subreddits”. Each of these subreddits typically has their own theme or subject matter, page and communities including regular users and moderators. These subreddits allow registered users to post various posts, that can consist of anything from questions, stories, or even links to media like videos or images. Naturally, content posted by users must follow not only overall rules given by Reddit itself but also subreddit-specific rules that can vary from subreddit to subreddit. If a user fails to follow the rules, either an automatic security bot or a moderator detects it and suspends the user. These suspensions can vary from short one- or two-hour bans to permanent bans from a subreddit or Reddit itself.

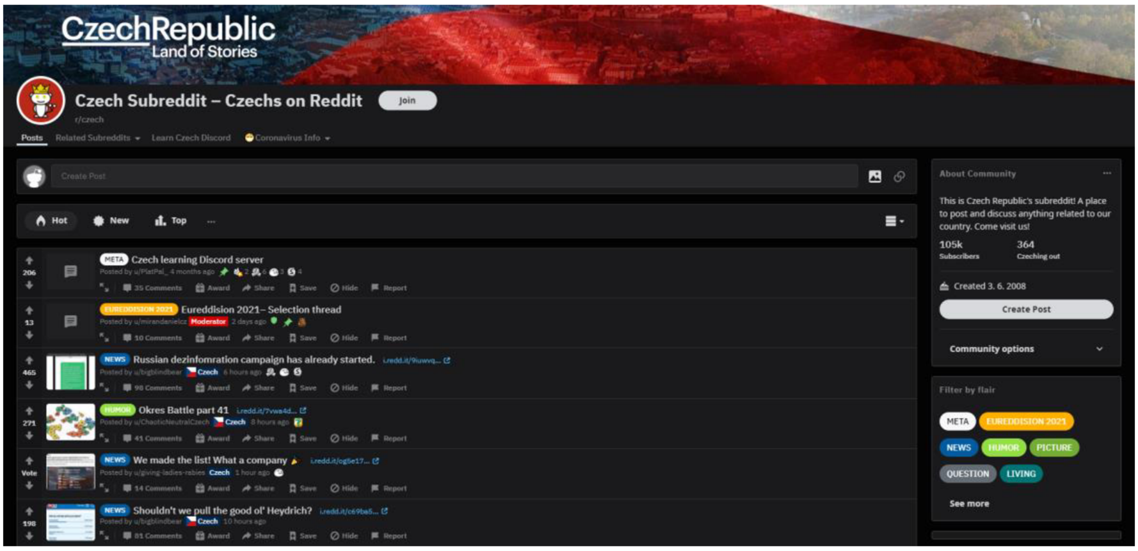
Users, also known as “Redditors”, can communicate with each other through different means. They can send each other private messages, talk through a direct chat function, or talk through comments and replies under posts. Although it is of no surprise that commenting and replying under posts seems to be the most common and widespread way of communication. And since not all comments tend to be nice and positive, there is also a voting system in place. Redditors can vote on posts or comments by using the “upvote” and “downvote” buttons. When a post or a comment receives a lot of upvotes, it rises to the top of the subreddit or even Reddit itself. On the contrary if a post or a comment receives a lot of downvotes, it sinks to the bottom of the subreddit or is even deleted or hidden from viewing. This system results in a constantly changing and volatile environment.

At the time of writing, there are over 2.75 million subreddits and about 1700 to 2200 new subreddits are created every single day. The number of subreddits created every month has nearly doubled ever since the COVID-19 pandemic has started (Metrics For Reddit, 2021). With its accessibility and sheer popularity that is at its peak, Reddit was the prime candidate for conducting research for this thesis. The level of language used will also most definitely vary from user to user and post to post, so this site will surely prove very resourceful in terms of analysis.

Figure 1: The logo of Reddit (Reddit.com).

1.1. Method of research and data collection

Since the COVID-19 pandemic began, numerous subreddits discussing the virus have sprouted up. Some subreddits discuss this issue on more of a global scale while others discuss in terms of countries or states. For this research, the former type of subreddit was chosen, specifically the r/Coronavirus subreddit (https://www.reddit.com/r/Coronavirus/). r/Coronavirus is a subreddit that discusses COVID-19 on a global and more general scale without any bias to a specific country or state, thus being great for observation of language use. Since users on this subreddit speak on a broader scale, it is more likely to find people who speak different levels of English. Most people who want to discuss happening in their own country or state will probably use a subreddit dedicated to that country or state. These subreddits will also most likely use that country's language. For example, the r/Czech subreddit (https://www.reddit.com/r/czech/), which despite featuring posts both in Czech and English language, sees interaction and comments mostly in Czech, by Czech-speaking users. For example, the r/Czech subreddit looks like this:



206 META Czech learning Discord server
Posted by u/PlatPal_ 4 months ago
35 Comments

13 EUREDDISION 2021 Eureddision 2021– Selection thread
Posted by u/mirandanielcz Moderator 2 days ago
10 Comments

465 NEWS Russian dezinfomration campaign has already started. i.redd.it/9iuwvq...
Posted by u/bigblindbear Czech 6 hours ago
98 Comments

271 HUMOR Okres Battle part 41 i.redd.it/7vwa4d...
Posted by u/ChaoticNeutralCzech Czech 8 hours ago
41 Comments

Vote NEWS We made the list! What a company i.redd.it/og5e17...
Posted by u/giving-ladies-rabies Czech 1 hour ago
14 Comments

198 NEWS Shouldn't we pull the good ol' Heydrich? i.redd.it/c69ba5...
Posted by u/bigblindbear Czech 10 hours ago
81 Comments

This is Czech Republic's subreddit! A place to post and discuss anything related to our country. Come visit us!

105k Subscribers
364 Czeching out

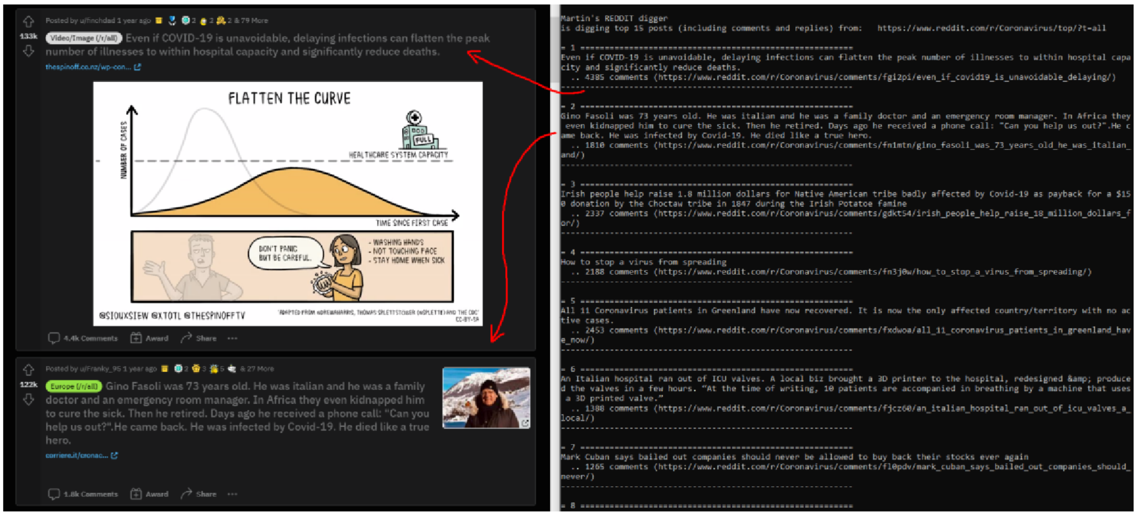
Created 3. 6. 2008

Figure 2: The frontpage of r/Czech subreddit (Reddit.com/r/Czech).

This is the "classic" layout of a subreddit page. The top of the page is occupied by the banner (which is optional) with the name of the subreddit right beneath it on the left. And right beneath that on the left, there are some post filtering options (Hot, New, Top, Rising) and the individual posts that have been posted on this subreddit and fall under those filters, while on the right there is a short "About community" article with a subscriber count and a number of currently online users, "Create post" button and some more useful filters.

To get data for the analysis, comments from the most popular or talked about posts were used. To find these posts, the subreddit first needed to be sorted by the top posts by selecting the "TOP" option on the front page. There is also an option that allows to further customize which top posts can be seen, by selecting if they must be from today, this week, month or even the top posts of all time on this subreddit. For this analysis, the "TOP" posts of "All time" were chosen.

The data in the form of comments were collected by executing a script written in the Python programming language. Simply said, after executing, this script goes to a given subreddit and converts the page into machine-readable format while collecting the text of select posts, including comments and replies to comments under these posts. The script then prints the collected data into a simple .TXT file based on defined rules.



Posted by u/finchdad 1 year ago 2 2 2 & 79 More

133k

Video/Image (/r/all) Even if COVID-19 is unavoidable, delaying infections can flatten the peak number of illnesses to within hospital capacity and significantly reduce deaths.

thespinoff.co.nz/wp-con...

FLATTEN THE CURVE
NUMBER OF CASES
FULL
HEALTHCARE SYSTEM CAPACITY
TIME SINCE FIRST CASE
DON'T PANIC BUT BE CAREFUL.
- WASHING HANDS
- NOT TOUCHING FACE
- STAY HOME WHEN SICK
@SIOUXSIEW @XTOTL @THESPINOFFTV
ADAPTED FROM @DREWAHARRIS, THOMAS SPLETTSTOESSER (@SPLETTE) AND THE CDC
CC-BY-SA

4.4k Comments Award Share ···

Posted by u/Franky_95 1 year ago 2 3 5 & 27 More

122k

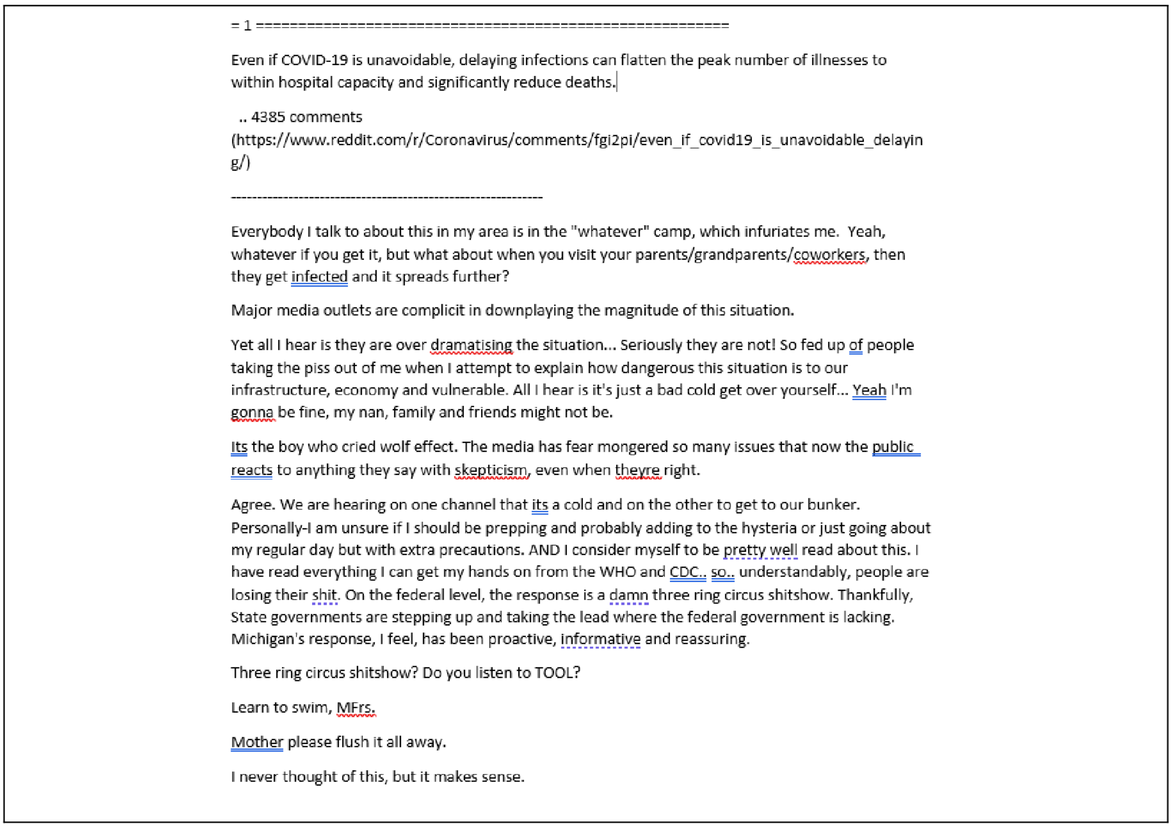
Europe (/r/all) Gino Fasoli was 73 years old. He was italian and he was a family doctor and an emergency room manager. In Africa they even kidnapped him to cure the sick. Then he retired. Days ago he received a phone call: "Can you help us out?".He came back. He was infected by Covid-19. He died like a true hero.

corriere.it/cronac...

1.8k Comments Award Share ···

```
Martin's REDDIT digger
is digging top 15 posts (including comments and replies) from:   https://www.reddit.com/r/Coronavirus/top/?t=all

= 1 ========================================================
Even if COVID-19 is unavoidable, delaying infections can flatten the peak number of illnesses to within hospital capa
city and significantly reduce deaths.
  .. 4385 comments (https://www.reddit.com/r/Coronavirus/comments/fgi2pi/even_if_covid19_is_unavoidable_delaying/)
------------------------------------------------------------

= 2 ========================================================
Gino Fasoli was 73 years old. He was italian and he was a family doctor and an emergency room manager. In Africa they
 even kidnapped him to cure the sick. Then he retired. Days ago he received a phone call: "Can you help us out?".He c
ame back. He was infected by Covid-19. He died like a true hero.
  .. 1810 comments (https://www.reddit.com/r/Coronavirus/comments/fn1mtn/gino_fasoli_was_73_years_old_he_was_italian_
and/)
------------------------------------------------------------

= 3 ========================================================
Irish people help raise 1.8 million dollars for Native American tribe badly affected by Covid-19 as payback for a $15
0 donation by the Choctaw tribe in 1847 during the Irish Potatoe famine
  .. 2337 comments (https://www.reddit.com/r/Coronavirus/comments/gdkt54/irish_people_help_raise_18_million_dollars_f
or/)
------------------------------------------------------------

= 4 ========================================================
How to stop a virus from spreading
  .. 2188 comments (https://www.reddit.com/r/Coronavirus/comments/fn3j0w/how_to_stop_a_virus_from_spreading/)
------------------------------------------------------------

= 5 ========================================================
All 11 Coronavirus patients in Greenland have now recovered. It is now the only affected country/territory with no ac
tive cases.
  .. 2453 comments (https://www.reddit.com/r/Coronavirus/comments/fxdwoa/all_11_coronavirus_patients_in_greenland_hav
e_now/)
------------------------------------------------------------

= 6 ========================================================
An Italian hospital ran out of ICU valves. A local biz brought a 3D printer to the hospital, redesigned &amp; produce
d the valves in a few hours. "At the time of writing, 10 patients are accompanied in breathing by a machine that uses
 a 3D printed valve."
  .. 1388 comments (https://www.reddit.com/r/Coronavirus/comments/fjcz60/an_italian_hospital_ran_out_of_icu_valves_a_
local/)
------------------------------------------------------------

= 7 ========================================================
Mark Cuban says bailed out companies should never be allowed to buy back their stocks ever again
  .. 1265 comments (https://www.reddit.com/r/Coronavirus/comments/fl0pdv/mark_cuban_says_bailed_out_companies_should_
never/)
------------------------------------------------------------

= 8 ========================================================
```

Figure 3: Example of how comments from Reddit look after executing the script.



= 1 ==

Even if COVID-19 is unavoidable, delaying infections can flatten the peak number of illnesses to within hospital capacity and significantly reduce deaths.

.. 4385 comments (https://www.reddit.com/r/Coronavirus/comments/fgi2pi/even_if_covid19_is_unavoidable_delaying/)

Everybody I talk to about this in my area is in the "whatever" camp, which infuriates me. Yeah, whatever if you get it, but what about when you visit your parents/grandparents/coworkers, then they get infected and it spreads further?

Major media outlets are complicit in downplaying the magnitude of this situation.

Yet all I hear is they are over dramatising the situation... Seriously they are not! So fed up of people taking the piss out of me when I attempt to explain how dangerous this situation is to our infrastructure, economy and vulnerable. All I hear is it's just a bad cold get over yourself... Yeah I'm gonna be fine, my nan, family and friends might not be.

Its the boy who cried wolf effect. The media has fear mongered so many issues that now the public reacts to anything they say with skepticism, even when theyre right.

Agree. We are hearing on one channel that its a cold and on the other to get to our bunker. Personally-I am unsure if I should be prepping and probably adding to the hysteria or just going about my regular day but with extra precautions. AND I consider myself to be pretty well read about this. I have read everything I can get my hands on from the WHO and CDC.. so.. understandably, people are losing their shit. On the federal level, the response is a damn three ring circus shitshow. Thankfully, State governments are stepping up and taking the lead where the federal government is lacking. Michigan's response, I feel, has been proactive, informative and reassuring.

Three ring circus shitshow? Do you listen to TOOL?

Learn to swim, MFrs.

Mother please flush it all away.

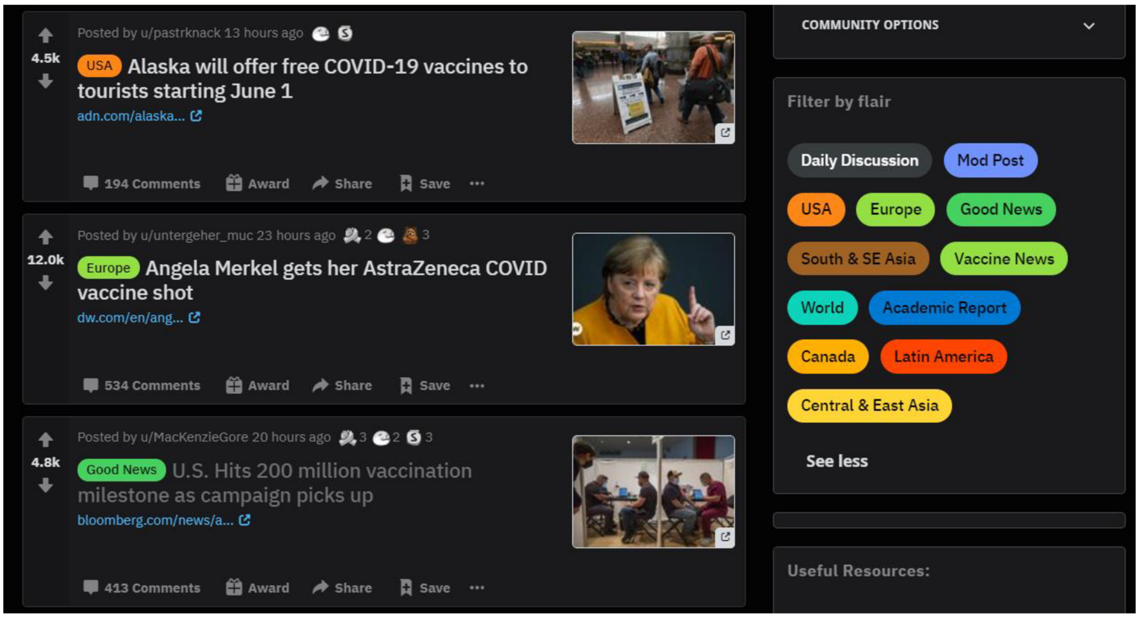
I never thought of this, but it makes sense.

Figure 4: Example of the output of the script shown in MS Word.

1.2. Chosen subreddit

https://www.reddit.com/r/Coronavirus/

The subreddit r/Coronavirus was chosen for this research because it is the largest subreddit concerning COVID-19 on Reddit. At the time of writing, it has around 2.4 million members with close to 7 thousand of them being online. Posts here are marked by their general theme (Good News, Academic Report, etc.) or by the part of the world they speak about (Europe, Latin America, World, etc.). While the number of members does not say anything about the levels of their English or the complexity of their posts and comments, I believe it will be sufficient enough to conduct a proper analysis.



4.5k
Posted by u/pastrknack 13 hours ago
USA Alaska will offer free COVID-19 vaccines to tourists starting June 1
adn.com/alaska...
194 Comments

12.0k
Posted by u/untergeher_muc 23 hours ago 2 3
Europe Angela Merkel gets her AstraZeneca COVID vaccine shot
dw.com/en/ang...
534 Comments

4.8k
Posted by u/MacKenzieGore 20 hours ago 3 2 3
Good News U.S. Hits 200 million vaccination milestone as campaign picks up
bloomberg.com/news/a...
413 Comments

Figure 5: The r/Coronavirus subreddit (Reddit.com/r/Coronavirus).

2. #LancsBox and used functions

#LancsBox is a free software package that was developed at Lancaster University specifically for the analysis of language data and corpora, making it a very useful tool for this research. Developed by a team of several talented people, #LancsBox's main features are, among many others, the ability to work with your own data or existing corpora, visualizing language data and said corpora, comparing multiple corpora, analyses of data in any language, automatic annotations of data for part-of-speech, and its ease of use (#LancsBox, 2021).

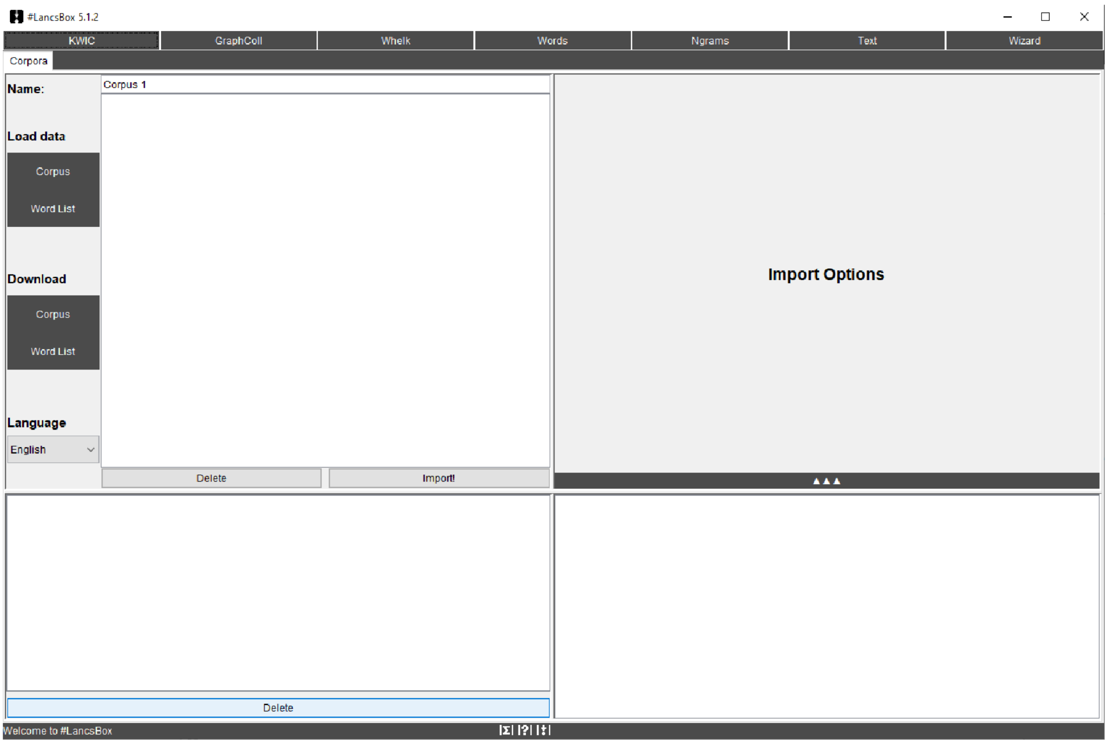
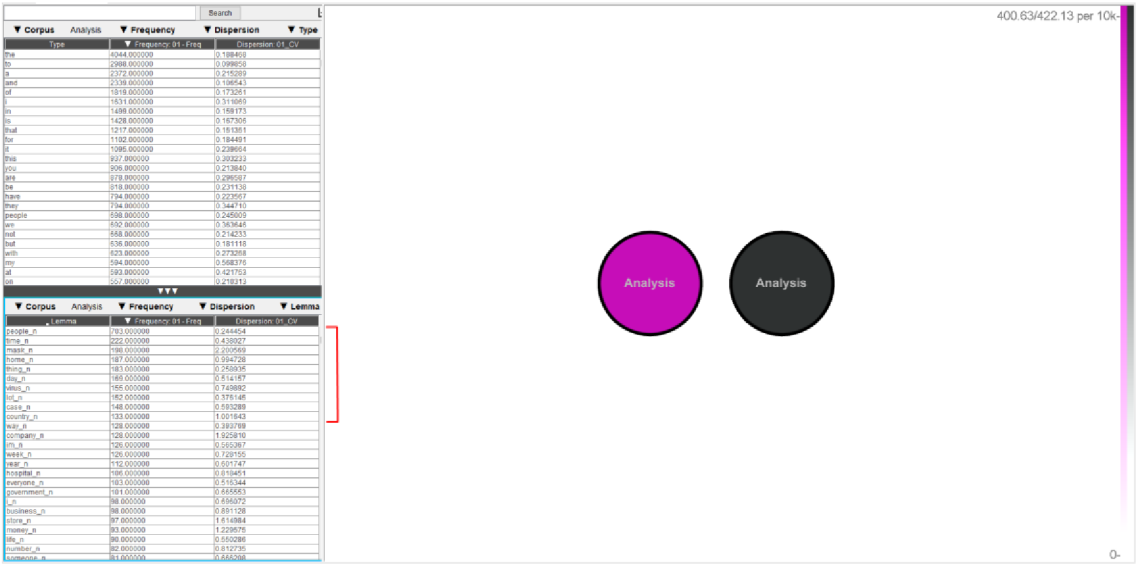


Figure 6: The interface of #LancsBox.

2.1. The "Words" function

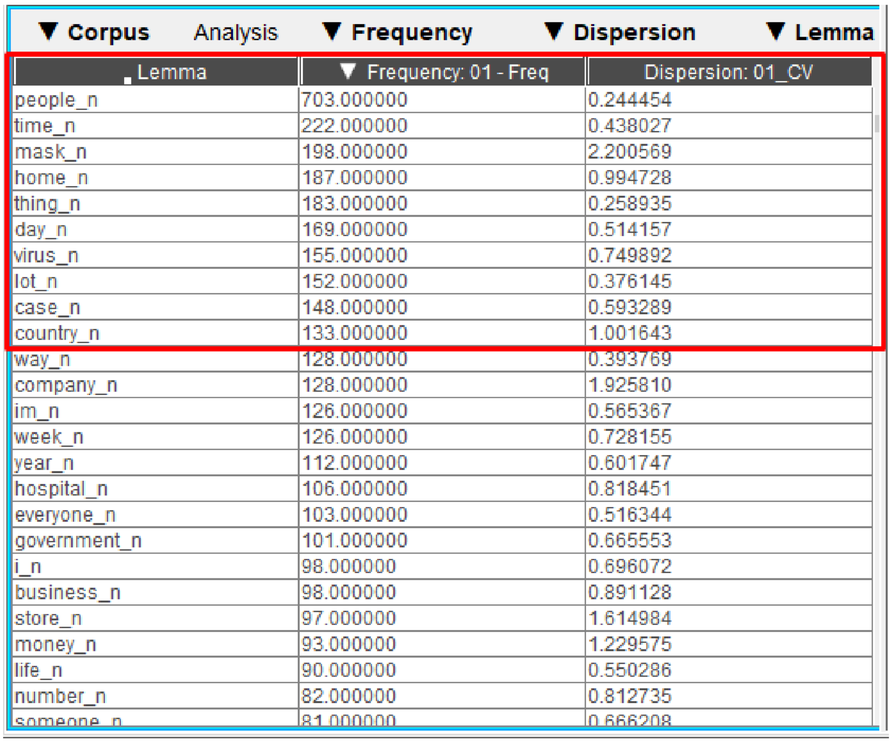
One of the two important functions of #LancsBox that was used is the "Words" function. This function allows users analyse the frequencies of types, lemmas, or POS categories. In addition, it also allows comparison of corpora using the "keywords" technique. In this thesis, this function was used to analyse the frequencies of types and lemmas in a corpus comprised of roughly fifteen Reddit posts with a total number of approximately 100 thousand words. The frequency list of lemmas was then filtered to show only nouns which were also sorted from the most frequent to the least frequent. Ten of the most frequent nouns were then chosen for further analysis.



Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
the	4044.000000	0.188468
to	2988.000000	0.099858
a	2372.000000	0.215289
and	2339.000000	0.106543
of	1819.000000	0.173261
i	1631.000000	0.311069
in	1499.000000	0.159173
is	1428.000000	0.167306
that	1217.000000	0.151351
for	1102.000000	0.184491
it	1095.000000	0.239664
this	937.000000	0.303233
you	906.000000	0.213840
are	878.000000	0.296587
be	818.000000	0.231138
have	794.000000	0.223567
they	794.000000	0.344710
people	698.000000	0.245009
we	692.000000	0.363646
not	668.000000	0.214233
but	636.000000	0.181118
with	623.000000	0.273258
my	594.000000	0.568376
at	593.000000	0.421753
on	557.000000	0.210313

Lemma	▼ Frequency: 01 - Freq	Dispersion: 01_CV
people_n	703.000000	0.244454
time_n	222.000000	0.438027
mask_n	198.000000	2.200569
home_n	187.000000	0.994728
thing_n	183.000000	0.258935
day_n	169.000000	0.514157
virus_n	155.000000	0.749892
lot_n	152.000000	0.376145
case_n	148.000000	0.593289
country_n	133.000000	1.001643
way_n	128.000000	0.393769
company_n	128.000000	1.925810
im_n	126.000000	0.565367
week_n	126.000000	0.728155
year_n	112.000000	0.601747
hospital_n	106.000000	0.818451
everyone_n	103.000000	0.516344
government_n	101.000000	0.665553
i_n	98.000000	0.696072
business_n	98.000000	0.891128
store_n	97.000000	1.614984
money_n	93.000000	1.229575
life_n	90.000000	0.550286
number_n	82.000000	0.812735
someone_n	81.000000	0.666208

Figure 7: The "Words" function.

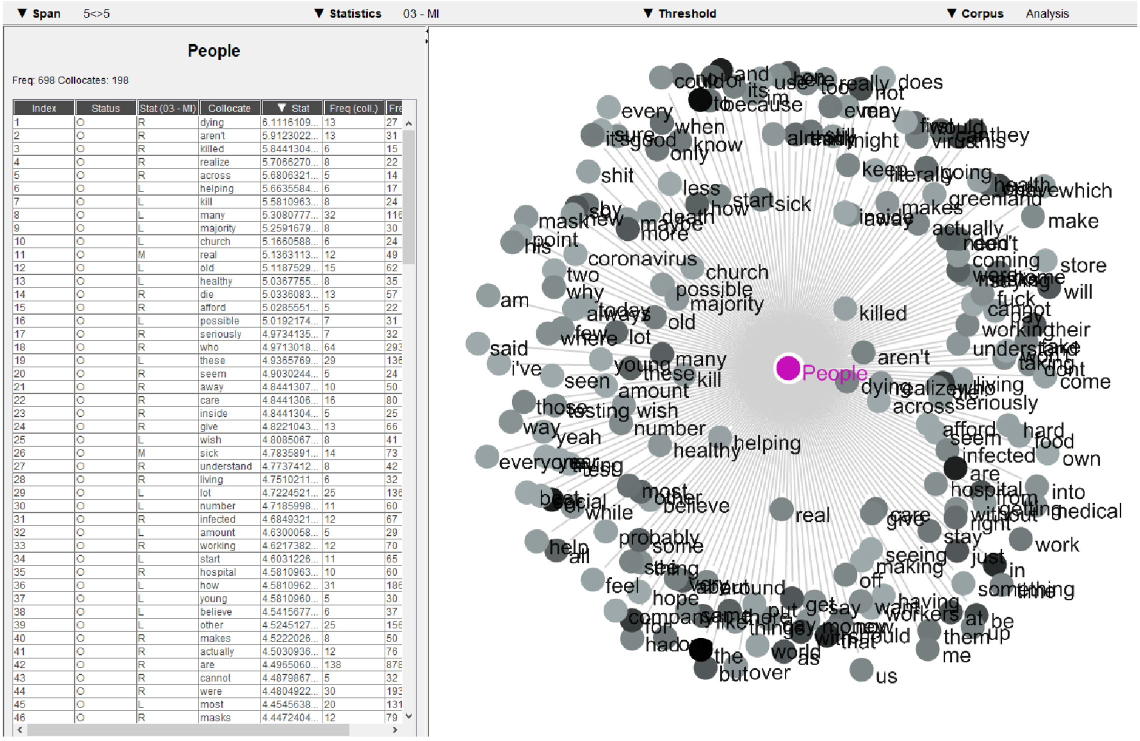

▼ Corpus
Analysis
▼ Frequency
▼ Dispersion
▼ Lemma

Lemma	▼ Frequency: 01 - Freq	Dispersion: 01_CV
people_n	703.000000	0.244454
time_n	222.000000	0.438027
mask_n	198.000000	2.200569
home_n	187.000000	0.994728
thing_n	183.000000	0.258935
day_n	169.000000	0.514157
virus_n	155.000000	0.749892
lot_n	152.000000	0.376145
case_n	148.000000	0.593289
country_n	133.000000	1.001643
way_n	128.000000	0.393769
company_n	128.000000	1.925810
im_n	126.000000	0.565367
week_n	126.000000	0.728155
year_n	112.000000	0.601747
hospital_n	106.000000	0.818451
everyone_n	103.000000	0.516344
government_n	101.000000	0.665553
i_n	98.000000	0.696072
business_n	98.000000	0.891128
store_n	97.000000	1.614984
money_n	93.000000	1.229575
life_n	90.000000	0.550286
number_n	82.000000	0.812735
someone_n	81.000000	0.666208

Figure 8: Top 10 most frequent nouns.

2.2. The “GraphColl” function

The second important function of #LancsBox that was used is the “GraphColl” function. This function allows users to identify and display collocations of words or phrases and to visualize them. In this thesis, ten of the most frequent nouns that were determined using the “Words” function of #LancsBox were subjected to analysis. Starting from the most frequent, these nouns were one by one inserted into the search bar. The Span, Statistics and Threshold parameters were also used to further specify the result of the analysis. The result was a graph depicting the analysed word along with its collocates that fall under the specific set of chosen parameters, their frequency, position, and collocation strength. This was then used in the creation of collocational profiles.



▼ Span 5<>5 ▼ Statistics 03 - MI ▼ Threshold ▼ Corpus Analysis

People

Freq: 698 Collocates: 198

Index	Status	Stat (03 - MI)	Collocate	▼ Stat	Freq (coll.)	Fre
1	○	R	dying	6.1116109...	13	27
2	○	R	aren't	5.9123022...	13	31
3	○	R	killed	5.8441304...	6	15
4	○	R	realize	5.7066270...	8	22
5	○	R	across	5.6806321...	5	14
6	○	L	helping	5.6635584...	6	17
7	○	L	kill	5.5810963...	8	24
8	○	L	many	5.3080777...	32	116
9	○	L	majority	5.2591679...	8	30
10	○	L	church	5.1660588...	6	24
11	○	M	real	5.1363113...	12	49
12	○	L	old	5.1187529...	15	62
13	○	L	healthy	5.0367755...	8	35
14	○	R	die	5.0336083...	13	57
15	○	R	afford	5.0285551...	5	22
16	○	L	possible	5.0192174...	7	31
17	○	R	seriously	4.9734135...	7	32
18	○	R	who	4.9713018...	64	293
19	○	L	these	4.9365769...	29	136
20	○	R	seem	4.9030244...	5	24
21	○	R	away	4.8441307...	10	50
22	○	R	care	4.8441306...	16	80
23	○	R	inside	4.8441304...	5	25
24	○	R	give	4.8221043...	13	66
25	○	L	wish	4.8085067...	8	41
26	○	M	sick	4.7835891...	14	73
27	○	R	understand	4.7737412...	8	42
28	○	R	living	4.7510211...	6	32
29	○	L	lot	4.7224521...	25	136
30	○	L	number	4.7185998...	11	60
31	○	R	infected	4.6849321...	12	67
32	○	L	amount	4.6300058...	5	29
33	○	R	working	4.6217382...	12	70
34	○	L	start	4.6031226...	11	65
35	○	R	hospital	4.5810963...	10	60
36	○	L	how	4.5810962...	31	186
37	○	L	young	4.5810960...	5	30
38	○	L	believe	4.5415677...	6	37
39	○	L	other	4.5245127...	25	156
40	○	R	makes	4.5222026...	8	50
41	○	R	actually	4.5030936...	12	76
42	○	R	are	4.4965060...	138	878
43	○	R	cannot	4.4879867...	5	32
44	○	R	were	4.4804922...	30	193
45	○	L	most	4.4545638...	20	131
46	○	R	masks	4.4472404...	12	79

Figure 9: The "GraphColl" function using the MI-score.

2.3. Associations measures

Association measures are mathematical procedures or formulas that are often used to find collocations on corpora. These measures have predominantly mathematical basis that stands on statistical testing of hypotheses, but there are also some otherwise motivated measures that include purely empirical ones unrelated to statistical relevance. Currently, dozens of association measures are known, the most used include *Dice, log-likelihood, MI-score, MI3, T-score* etc. Given the breadth and diversity of collocations from a linguistic and mathematical point of view, it is understandable that these individual measures may significantly differ from each other by the type of collocations, which they consider important.

Association measures typically deal with the frequency of the entire collocation, its individual members or parts, and the size of the corpus. This is then put into contingency tables and, based on them, the measures calculate the resulting numerical value according to a specified formula.

The resulting value for a given pair of words in the corpus expresses the degree of association between them, which can be negative in some measure, which shows a negative association, i.e. mutual "repulsion". The numerical values of one association measure are generally not comparable to the values of another association measure, but for comparison, numerical values are normally converted to rank in a list of collocations arranged according to the numerical values of the given measure (Český Národní Korpus, 2019).

The two main association measures used in this analysis are MI-score and T-score.

2.3.1. MI-score

In short, MI-score is an association measure that is used when searching for strong collocations with high relative frequency, and therefore rather exceptional or random.

An unwelcome feature of MI-score is the fact that it can be greatly influenced by the frequency of individual words. The highest values are more often than not reached by pairs of words with low frequency. For this reason, corpus managers like #LancsBox allow to set the lower frequency limit when calculating MI-score, thus completely avoiding the need to calculate the score for words below this limit.

MI-score values are mostly positive (negative values indicate mutual repulsion of units and are relatively sparse). The MI = 7 limit (for a 100 million corpus) is considered relevant, when it is justified to assume that it is a systemic collocation. This value depends on the size of the corpus and in the case of this analysis, the MI = 3 limit (for a 100 thousand word corpus) was chosen (Český Národní Korpus, 2019).

2.3.2. T-Score

T-score is based on the statistical method of testing hypotheses using the so-called t-test. It is also sometimes referred to as the degree of contrast. Unlike MI-score, T-score is sensitive to the frequency of the entire collocation. The results of this test regularly overestimate combinations of very frequent words (which are mostly grammatical words like articles and such) at the expense of less frequent lexical combinations.

The use of T-score is that in the case of collocations, we test whether the detected numbers of occurrences of individual words and their pairs correspond to the random distribution of words in the corpus. The higher the value of T-score, the less likely it is that it is a random distribution of words and, conversely, the more likely it is that it is a stronger, more stable combination of words, i.e. collocations. As mentioned before, this is the exact opposite of MI-score, results of which are mostly exceptional or random collocations with high relative frequency (Český Národní Korpus, 2019).

3. Analysis

For analysis, fifteen posts from the “TOP” category with the “All time” modifier were used to provide a representative sample. The “All time” modifier was chosen to single out the posts that have amassed the largest user engagement since the creation of this subreddit (the start of the pandemic). The number of comments under these posts ranged from 1265 to 11561 with a total combined number of 101 148 word tokens.

3.1. Analysed posts

The following sections contain information about analysed posts like their paraphrased content, some information about the comments under said posts, their total number as well as the number of upvotes said posts amassed and the general theme of the posts they were marked as.

3.1.1. Post #1

The first post was a sort of “motivational post” that urged people to help “flatten the curve” by taking COVID-19 seriously and delaying infections through proper hygiene and being careful around other people. Comments under this post were mixed, some blaming everything on the media and government while others tried staying rational and urging others to take the pandemic seriously. The post amassed a total of 4385 comments and around 133 thousand upvotes, making it the top post of all time on this subreddit. This post was marked as “Video/Image”, meaning that it does not follow a specific theme or a part of world, but it aims to create a discussion about the provided video or image.

3.1.2. Post #2

The second post was about Gino Fasoli, a 73-year-old Italian doctor and emergency room manager, who retired but amidst the COVID-19 pandemic volunteered to help the sick, got infected himself and sadly passed away. Comments under this post were very supportive, mostly mourning the death of Fasoli and others like him while some even shared their own stories of loss and criticized society. The post amassed a total of 1810 comments and around 122 thousand upvotes, making it the top second post of all time on this subreddit. This post was marked as “Europe”, meaning that the topic of this post is about or takes places in Europe and its countries.

3.1.3. Post #3

The third post was about the fact that Irish people helped raise nearly 2 million dollars for a Native American tribe badly affected by COVID-19 as a payback for a 150-dollar donation they received from this tribe during the Irish Potato famine. Comments under this post were extremely kind and supportive, most of them regarding how nice of a gesture it was and that their “faith in humanity was restored”. Some commenters even claimed that they are of Native American origin and that this donation really helped. The post amassed a total of 2337 comments and around 122 thousand upvotes, making it the top third post of all time on this subreddit. This post was marked as “Good news”, meaning that the topic is something good and positive that happened in the world.

3.1.4. Post #4

The fourth post involved instructions about how to stop a virus from spreading. It showed a gif that depicted a graph that further spread into more and more branches as it went on, mimicking the spread of a virus. Some of these branches were then greyed out because of people staying home, avoiding contact and similar actions. This post was met with mostly positive comments, some further stating the importance of social distancing and commending others for staying home and some were stating that disasters like this bring out the worst in humanity and criticizing society. The post amassed a total of 2188 comments and around 110 thousand upvotes, making it the top fourth post of all time on this subreddit. This post was marked as “Video/Image”.

3.1.5. Post #5

The fifth post was about the fact that all eleven COVID-19 patients in Greenland have recovered and that at the time of posting Greenland was the only affected country with no active cases. This post was met with some funny comments referring to a video game called Plague Inc. where people play as a virus and Greenland is notoriously hard to infect. Some people were even making fun of the virus and how people in Scandinavia and northern countries are more resistant to the virus. The post amassed a total of 2453 comments and around 106 thousand upvotes, making it the top fifth post of all time on this subreddit. This post was marked as "Good News".

3.1.6. Post #6

The sixth post was about an Italian business that helped a hospital that ran out of ICU valves by bringing a 3D printer to the hospital and printing new valves. At least 10 patients were then accompanied in breathing by machines that used these 3D printed valves. Comments under this post were positive, complimenting the Italian business for helping the hospital and possibly saving lives and some people even offered help, saying that they have a 3D printer and can print anything for anyone. These offers were met with even more positive responses. The post amassed a total of 1388 comments and around 101 thousand upvotes, making it the top sixth post of all time on this subreddit. This post was marked as "Good News".

3.1.7. Post #7

The seventh post was regarding the statement of Mark Cuban, who is an American billionaire and investor, who said that companies who received financial assistance bailed out of the market amidst the COVID-19 pandemic should not be allowed to buy back their stocks ever again. This post was met with mixed comments, some saying that such punishment would be too harsh for some companies while others said that this would at least teach some companies a lesson. There were even some comments in support of the companies who bailed out. Most of the comments were discussing politics and economics. The post amassed a total of 1265 comments and around 101 thousand upvotes, making it the top seventh post of all time on this subreddit. This post was marked as “Breaks Rule 3”, meaning that it violated the third rule of the subreddit (“Avoid reposting information”), but perhaps because of high user engagement was not removed.

3.1.8. Post #8

The eight post was about a statement made by the American and Southwest Airlines, that prohibited people who cannot wear a mask due to “medical” reasons, meaning people who just do not want to wear mask and use a fake medical reason to not wear it, from entering their planes. They instead told them to find alternate travel arrangements. Comments under this post were mixed, some in defence of people who do not want to wear masks, saying that it is their right, and that America is a free country, so they should be able to do what they want, others in defence of the Airlines and anti-COVID measures in general, saying that proper hygienic measures must be taken so the virus would not spread even more than it already has. The post amassed a total of 3734 comments and around 98 thousand upvotes, making it the top eighth post of all time on this subreddit. This post was marked as “USA”, meaning that the topic of this post is about or takes places in the USA and its states.

3.1.9. Post #9

The ninth post was about a twitter post made by the governor of New York, Andrew Cuomo, who said that they will not put a dollar figure on human life, public health strategy can be consistent with an economic one and that no one should be a victim for the sake of the stock market. Comments under this post were mostly in support of what Cuomo said, some praising him for what he said and believing it to be truly possible, others just accepting what he said with some criticism and remarks. Some commenters were also discussing politic and economic topics regarding the USA or their own countries. The post amassed a total of 3653 comments and around 95 thousand upvotes, making it the top ninth post of all time on this subreddit. This post was marked as “USA”.

3.1.10. Post #10

The tenth post quotes a Twitter post made by Joanna Killian who is Chief Executive of Surrey County council. In her Twitter post she urges people to not spread the virus and stay at home, says that it is not just a simple flu and people are actually dying. Comments under this post were mostly in agreement with Killian’s statement, discussing their everyday lives and how COVID impacted them while also praising essential workers for their hard work during these trying times. Some even shared their frustration with anti-COVID measures that were forced upon them. The post amassed a total of 3290 comments and around 93 thousand upvotes, making it the top tenth post of all time on this subreddit. This post was marked as “World”, meaning that the topic of this post either is from all over the world, or its contents can be applied to or are speaking of the world in general.

3.1.11. Post #11

The eleventh post was about the fact that one church in Kentucky had its Easter service filled to near capacity during quarantine measures. Local police force responded to this occasion by showing up in the parking lot and serving every one of the churchgoers with a ticket and ordered them into a 14-day quarantine. This post was also locked by a moderator so no new comments could be made because of overly political and uncivil comments. Comments under this post were mostly negative towards the churchgoers or just the church in general, some being angry at the churchgoers for purposely avoiding quarantine measures and possible aiding the spread of the virus, others expressing disappointment of the church’s disregard for quarantine measures and criticizing church in general, some even mocking church culture which inevitably spiralled into a highly political discussion. The post amassed a total of 3617 comments and around 92 thousand upvotes, making it the top eleventh post of all time on this subreddit. It was also the post with the most negative comments of the analysis. This post was marked as “USA”.

3.1.12. Post #12

The twelfth post was about at that time a 101-year-old man from Italy, born during the Spanish flu, who has successfully beaten COVID-19 and has fully recovered. This post was met with very positive comments, some being supportive of the man and praising his resilience towards the virus, others making funny remarks about the fact that he had survived the Spanish flu in the past and now survived COVID-19 in result making him practically immortal. A lot of the comments were also filled with video game references like him using a “god mode” cheat making him invincible or playing the game of life on “hardcore mode” and basically beating it. The post amassed a total of 1442 comments and around 91 thousand upvotes, making it the top twelfth post of all time on this subreddit. This post was marked as “Good News”.

3.1.13. Post #13

The thirteenth post was about a story from Vice News about Starbucks employees in the USA who were begging the company to shut down its stores because of coronavirus, saying that “Coffee is not essential” thus not making them essential workers. Even though the stores transitioned to a “to go” model to cope with the virus, employees still think that it is not enough to protect them, especially in severely afflicted areas or parts of cities. Comments under this post were mixed, some praising the employees for standing their ground a signing a petition, others criticizing the company for not valuing their employees enough and not respecting their wishes. Some were also in support of Starbucks, saying that even though the situation is not ideal, at least their employees can still work and have a stable income and that not all lines of work can provide financial stability amidst the pandemic. The post amassed a total of 6080 comments and around 90 thousand upvotes, making it the top thirteenth post of all time on this subreddit. This post was marked as “USA”.

3.1.14. Post #14

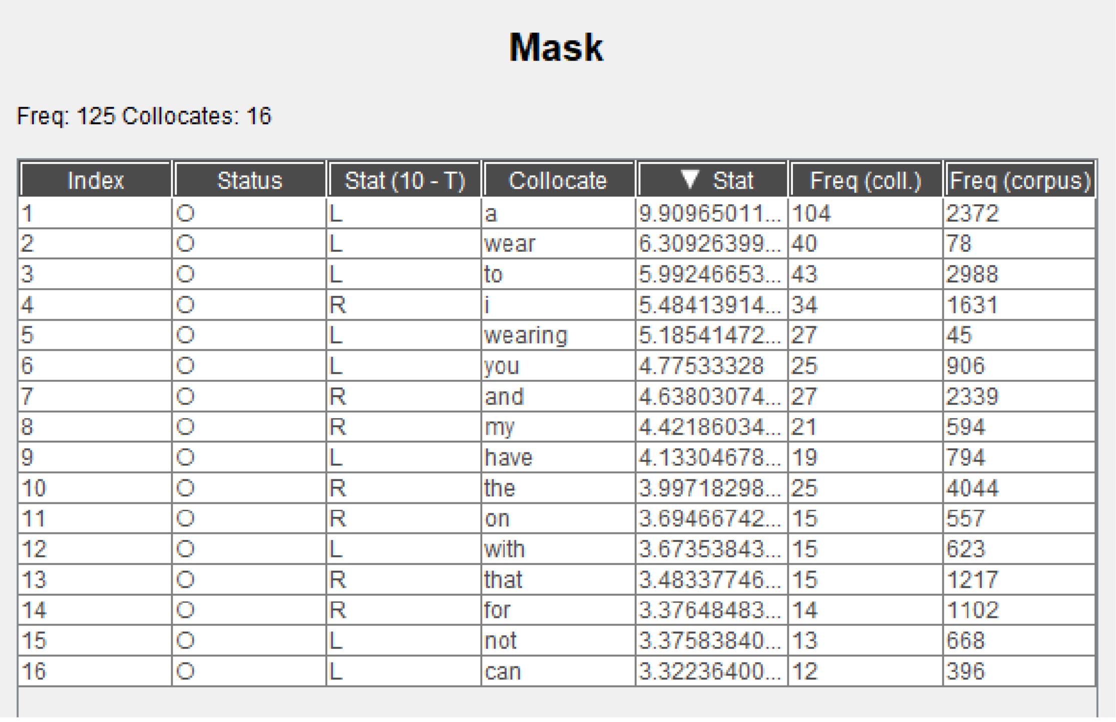
The fourteenth post was regarding a statement from Walmart titled “A Simple Step to Help Keep You Safe: Walmart and Sam’s Club Require Shoppers to Wear Face Covering”, in which Walmart basically says that to bring consistency across stores and clubs and prevent the spread of the virus, they would require customers to wear masks and that they would open only one entrance to the stores and station a health ambassador there. This post was met with mixed comments although most were in support of Walmart’s decision, some commending Walmart for doing what is right and helping to prevent the spread of the virus, others were criticizing people who do not already wear masks and people who refuse to wear masks. The latter posts were also split between serious critique of such people and making fun of such people, saying that there is about to be a whole new wave of funny videos and compilations of people refusing to comply to these new rules. The post amassed a total of 5620 comments and around 89 thousand upvotes, making it the top fourteenth post of all time on this subreddit. This post was marked as “World”.

3.1.15. Post #15

The fifteenth post was an “AMA” post which stands for “Ask Me Anything”. These posts are often made by some people who are either famous or had something interesting happen to them. This AMA was held by none other than Bill Gates (the founder of Microsoft) himself. Here Gates says that over the years he had a chance to study numerous diseases and that he and his foundation have committed over 100 million dollars to help with COVID-19 response around the world and 5 million dollars to support the state of Washington. He then urges other redditors to ask him anything about COVID-19 specifically or epidemics and pandemics in general. Comments under this post were mostly questions aimed at Gates or replies to these questions. These questions ranged from questions about COVID-19 and how to prevent its spread, general healthcare questions, questions about job security and proper safety measures against COVID-19 to questions about COVID-19 testing standards or even mental health. The post amassed a total of 11561 and around 87 thousand upvotes, making it the top fifteenth post of all time on this subreddit. It was also the post with the largest number of comments of the analysis, which makes sense as Bill Gates is a very popular and prominent figure in the tech industry. This post was marked as “AMA”, meaning that the author is willing to answer any questions in the comments about a certain topic the post is about.

3.2. Method of Analysis

As mentioned before, the “Words” and “GraphColl” functions were used to analyse the corpus. First the “Words” function was used determine the ten most frequent nouns from the corpus by selecting lemmas as the primary units and applying a “*_n” filter which narrowed the result to only nouns which were then ordered by frequency. Ten of the most frequent nouns were then noted and used in the “GraphColl” function. These nouns are: *people, time, mask, home, thing, day, virus, lot, case, country*. Using the “GraphColl” function, each of the ten nouns were analysed to find their collocates. Here two main association methods (types of statistics) were used, the MI-score and the T-score. MI-score was used to find the strongest collocates, that are often paired with select nouns, albeit these collocates may not always be very frequent. On the other hand, T-score was used mainly for illustration and the overall completeness of data, but was not particularly semantically significant, as the results of T-score comprise predominantly determiners, articles, and prepositions. Collocates analysed both by MI-score and T-score may also be further examined using the built-in “KWIC” function, that gathers all the appearances of select collocates in the corpus and displays them in the form of a short text, that can also be modified by changing the Context value to make the text shorter or longer. As the use of the “KWIC” function is purely optional, it is not described any further in this thesis.

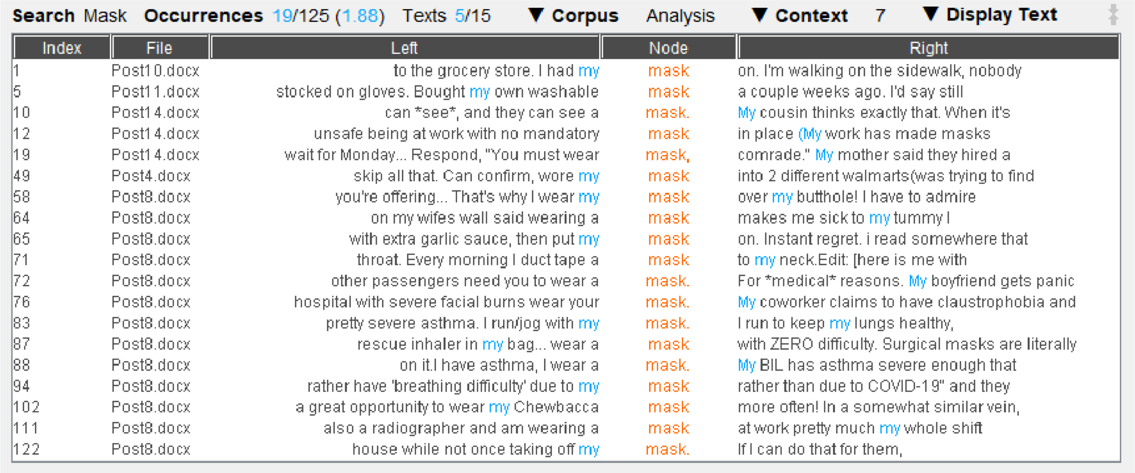


Mask

Freq: 125 Collocates: 16

Index	Status	Stat (10 - T)	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
1	○	L	a	9.90965011...	104	2372
2	○	L	wear	6.30926399...	40	78
3	○	L	to	5.99246653...	43	2988
4	○	R	i	5.48413914...	34	1631
5	○	L	wearing	5.18541472...	27	45
6	○	L	you	4.77533328	25	906
7	○	R	and	4.63803074...	27	2339
8	○	R	my	4.42186034...	21	594
9	○	L	have	4.13304678...	19	794
10	○	R	the	3.99718298...	25	4044
11	○	R	on	3.69466742...	15	557
12	○	L	with	3.67353843...	15	623
13	○	R	that	3.48337746...	15	1217
14	○	R	for	3.37648483...	14	1102
15	○	L	not	3.37583840...	13	668
16	○	L	can	3.32236400...	12	396

Figure 10: Results of T-score analysis of the noun "Mask".



Search Mask **Occurrences** 19/125 (1.88) Texts 5/15 ▼ **Corpus** Analysis ▼ **Context** 7 ▼ **Display Text**

Index	File	Left	Node	Right
1	Post10.docx	to the grocery store. I had my	mask	on. I'm walking on the sidewalk, nobody
5	Post11.docx	stocked on gloves. Bought my own washable	mask	a couple weeks ago. I'd say still
10	Post14.docx	can *see*, and they can see a	mask.	My cousin thinks exactly that. When it's
12	Post14.docx	unsafe being at work with no mandatory	mask	in place (My work has made masks
19	Post14.docx	wait for Monday... Respond, "You must wear	mask,	comrade." My mother said they hired a
49	Post4.docx	skip all that. Can confirm, wore my	mask	into 2 different walmarts(was trying to find
58	Post8.docx	you're offering... That's why I wear my	mask	over my butthole! I have to admire
64	Post8.docx	on my wifes wall said wearing a	mask	makes me sick to my tummy I
65	Post8.docx	with extra garlic sauce, then put my	mask	on. Instant regret. i read somewhere that
71	Post8.docx	throat. Every morning I duct tape a	mask	to my neck.Edit: [here is me with
72	Post8.docx	other passengers need you to wear a	mask.	For *medical* reasons. My boyfriend gets panic
76	Post8.docx	hospital with severe facial burns wear your	mask.	My coworker claims to have claustrophobia and
83	Post8.docx	pretty severe asthma. I run/jog with my	mask.	I run to keep my lungs healthy,
87	Post8.docx	rescue inhaler in my bag... wear a	mask	with ZERO difficulty. Surgical masks are literally
88	Post8.docx	on it.I have asthma, I wear a	mask.	My BIL has asthma severe enough that
94	Post8.docx	rather have 'breathing difficulty' due to my	mask	rather than due to COVID-19" and they
102	Post8.docx	a great opportunity to wear my Chewbacca	mask	more often! In a somewhat similar vein,
111	Post8.docx	also a radiographer and am wearing a	mask	at work pretty much my whole shift
122	Post8.docx	house while not once taking off my	mask.	If I can do that for them,

Figure 11: Optional examination of collocates using the "KWIC" function.

4. Results

The results were gathered using the Span of 5<>5, MI and T Statistics, default Threshold and default Type. The results are ordered by their respective scores.

4.1. The word "People"

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word "People" out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
dying	6.111611	13	27
aren't	5.912302	13	31
killed	5.844130	6	15
realize	5.706627	8	22
across	5.680632	5	14
helping	5.663558	6	17
kill	5.581096	8	24
many	5.308078	32	116
majority	5.259168	8	30
church	5.166059	6	24

Table 1: MI-score results for the word "People"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word "People" out of all the other collocates of the word "People".

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
the	14.918294	276	4044
to	13.649323	226	2988
of	11.301493	152	1819
are	11.226917	138	878
and	10.735154	146	2339
a	10.002476	131	2372
in	9.544955	111	1499
that	9.260447	102	1217
I	7.926025	84	1631
who	7.744977	64	293

Table 2: T-score results for the word "People"

The MI-score results for this word show that the strongest collocate was the word “dying” with MI-score of roughly 6.11 and with collocation frequency (number of times it was collocated specifically with the analysed word) of 13 out of 27 total appearances in the corpus, which translates to a relative frequency of roughly 48.15%. “Dying” + “People” was the most relatively frequent and strongest collocation because the mortality rate of COVID-19 is a highly monitored statistic in practically every country since it is in direct correlation with how successfully the country is coping or dealing with the pandemic. Overall, the MI-score results table seems to be comprised mostly of verbs, adjectives, and nouns, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word “the” with T-score of roughly 14.92 and collocation frequency of 276 out of 4044 total appearances in the corpus, which translates to an absolute frequency of roughly 6.82%. This is unsurprising as articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners, prepositions, and pronouns.

4.2. The word “Time”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Time” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
during	6.520064	9	52
first	6.119842	8	61
hard	5.991685	6	50
long	5.708187	7	71
same	5.704804	12	122
every	5.591147	8	88
off	5.143688	6	90
at	5.124093	39	593
use	4.963116	5	85
been	4.755123	11	216

Table 3: MI-score results for the word "Time”

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word “Time” out of all the other collocates of the word “Time”.

Collocate	***T-score***	***Freq (coll.)***	***Freq (corpus)***
the	9.397447	103	4044
a	8.020455	73	2372
to	7.430432	66	2988
at	6.065926	39	593
and	5.714246	41	2339
of	5.695704	39	1819
I	5.577104	37	1631
this	5.527909	34	937
in	5.346148	34	1499
for	5.097801	30	1102

Table 4: T-score results for the word "Time"

The MI-score results for this word show that the strongest collocate was the word “during” with MI-score of roughly 6.52 and with collocation frequency of 9 out of 52 total appearances in the corpus, which translates to a relative frequency of roughly 17.31%. “During” + “Time” was the most relatively frequent and strongest collocation because of people wanting to address, express or even educate other people on what to do and not to do during this time of crisis and how to live a semi-normal life during a global pandemic. In addition, the word “during” is derived from “-dure”, which is semantically closely related to time. Overall, the MI-score results table seems to be comprised mostly of verbs, adjectives, and even some prepositions, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word “the” with T-score of roughly 9.4 and collocation frequency of 103 out of 4044 total appearances in the corpus, which translates to an absolute frequency of roughly 2.55%. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners, prepositions, and pronouns.

4.3. The word “Mask”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Mask” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
wearing	8.918615	27	45
wear	8.692107	40	78
face	6.768056	5	37
cant	6.551244	5	43
due	6.459184	6	55
put	6.196149	5	55
she	5.900693	6	81
can't	5.848226	5	70
without	5.692107	5	78
a	5.144132	104	2372

Table 5: MI-score results for the word "Mask"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word “Mask” out of all the other collocates of the word “Mask”.

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
a	9.909650	104	2372
wear	6.309264	40	78
to	5.992467	43	2988
I	5.484139	34	1631
wearing	5.185415	27	45
you	4.775333	25	906
and	4.638031	27	2339
my	4.421860	21	594
have	4.133047	19	794
the	3.997183	25	4044

Table 6: T-score results for the word "Mask"

The MI-score results for this word show that the strongest collocate was the word “wearing” with MI-score of roughly 8.92 and with collocation frequency of 27 out of 45 total appearances in the corpus, which translates to a relative frequency of roughly 60%. “Wearing” + “Mask” was the most relatively frequent and strongest collocation because during the pandemic people in perhaps every country were forced to wear masks to help stop the spread of the virus. Users on r/Coronavirus mostly used this collocation to voice their opinions on the subject of wearing masks and why people should or should not wear them. In terms of MI-score the word "Wearing" is also closely followed by the word "Wear", which is the same lexeme but without the "-ing" suffix. It could also be assumed that the word "wear(ing)" will strongly prefer the subject "mask" in the context of COVID-19 as that is what is typically done with a mask to stop the spread of the virus. Overall, the MI-score results table seems to be comprised mostly of verbs, nouns, and even one article, but it is difficult to determine in some cases due to the lack of semantic context in the table. The table also contains the word "can't" twice due to a difference in spelling (with or without apostrophe). Both versions of the word were included because I was not quite sure how to combine their T-score values.

The T-score results for this word show that its most frequent collocate was the word "a" with T-score of roughly 9.91 and collocation frequency of 104 out of 2372 total appearances in the corpus, which translates to an absolute frequency of roughly 4.38%. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners, pronouns and even some verbs.

4.4. The word "Home"

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word "Home" out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
staying	8.058078	11	21
stay	7.947895	66	136
coffee	6.669035	5	25
working	6.321112	11	70
wish	6.218374	6	41
work	6.015211	30	236
at	5.988529	74	593
from	5.727138	38	365
sick	5.386101	6	73
go	5.378520	13	159

Table 7: MI-score results for the word "Home"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word "Home" out of all the other collocates of the word "Home".

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
at	8.466841	74	593
stay	8.091137	66	136
to	7.599645	69	2988
I	6.912247	54	1631
and	6.185104	47	2339
from	6.048041	38	365
the	5.947042	50	4044
work	5.392542	30	236
my	5.264080	30	594
you	4.954991	28	906

Table 8: T-score results for the word "Home"

The MI-score results for this word show that the strongest collocate was the word "staying" with MI-score of roughly 8.06 and with collocation frequency of 11 out of 21 total appearances in the corpus, which translates to a relative frequency of roughly 52.38%. “Staying” + “Home” was the most relatively frequent and strongest collocation due to the ubiquitous advice and recommendations from the government, media, and advertisements to stay at home as much as possible because of the pandemic. And since not all professions support working from home, this collocation was used in many comments discussing this topic. In terms of MI-score the word “Staying” is also closely followed by the word “Stay, which is the same lexeme but without the “-ing” suffix. Overall, the MI-score results table seems to be comprised mostly of verbs, nouns, and prepositions, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word “at” with T-score of roughly 8.47 and collocation frequency of 74 out of 593 total appearances in the corpus, which translates to an absolute frequency of roughly 12.48%. The word with the highest T-score is a preposition instead of an article like in previous T-score results tables. Overall, the T-score results table comprises mostly pronouns, prepositions and even some nouns or verbs.

4.5. The word “Thing”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Thing” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
whole	7.036818	5	40
same	6.565512	11	122
right	6.553002	16	179
its	5.621780	9	192
do	5.589734	16	349
only	5.137964	6	179
is	4.983302	43	1428
this	4.688474	23	937
about	4.644500	8	336
an	4.422108	6	294

Table 9: MI-score results for the word "Thing"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word “Thing” out of all the other collocates of the word “Thing”.

Collocate	***T-score***	***Freq (coll.)***	***Freq (corpus)***
the	8.031589	72	4044
is	6.350133	43	1428
to	5.702984	38	2988
a	4.965857	29	2372
this	4.609840	23	937
of	4.084937	20	1819
right	3.957400	16	179
do	3.916942	16	349
that	3.710367	16	1217
but	3.289325	12	636

Table 10: T-score results for the word "Thing"

The MI-score results for this word show that the strongest collocate was the word “whole” with MI-score of roughly 7.04 and with collocation frequency of 5 out of 40 total appearances in the corpus, which translates to a relative frequency of 12.5%. “Whole” + “Thing” was the most relatively frequent and strongest collocation due to people discussing the pandemic and overall happenings around COVID-19 and referring to it as the “whole thing” or “whole coronavirus thing” to generalize the matter or to speak in broader terms (although the collocation “the/this whole thing” is quite common even outside the context of COVID-19). And as this collocation was used only five times, this was its main use. Overall, the MI-score results table seems to be comprised mostly of adjectives, verbs, and some determiners, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word “the” with T-score of roughly 8.03 and collocation frequency of 72 out of 4044 total appearances in the corpus, which translates to an absolute frequency of roughly 1.78%. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners, pronouns, some prepositions and one adjective or adverb.

4.6. The word “Day”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Day” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
rainy	9.949269	6	6
every	7.659762	18	88
day	6.276844	8	102
other	6.248829	12	156
I'm	4.919522	6	196
still	4.913645	5	164
home	4.904875	6	198
one	4.695928	7	267
work	4.651588	6	236
going	4.556951	5	210

Table 11: MI-score results for the word "Day"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word “Day” out of all the other collocates of the word “Day”.

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
the	6.337782	48	4044
a	5.945191	40	2372
to	4.506278	26	2988
and	4.416036	24	2339
every	4.221660	18	88
in	4.133090	20	1499
of	3.676854	17	1819
on	3.449289	13	557
I	3.447012	15	1631
for	3.443745	14	1102

Table 12: T-score results for the word "Day"

The MI-score results for this word show that the strongest collocate was the word “rainy” with MI-score of roughly 9.95 and with collocation frequency of 6 out of 6 total appearances in the corpus, which translates to a relative frequency of exactly 100%. This means that in over 100 thousand words, whenever the word “rainy” appeared, it was always in collocation with the word “day”. The word “day” itself was also one of the results (e.g. “day in day out”, “day by day”). “Rainy” + “Day” was the most relatively frequent and strongest collocation not because its literal meaning (a day that it is raining on) but because of the economic term “Rainy Day fund”, which is basically a reserved amount of money to be used in times when regular income is disrupted or decreased, like in the case of the COVID-19 pandemic. In addition, the term “Rainy day” is also an idiom, which basically means “worse times” and can be seen even outside the context of COVID-19. Overall, the MI-score results table seems to be comprised mostly of adjectives, nouns, and some adverbs and verbs, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word “the” with T-score of roughly 6.34 and collocation frequency of 48 out of 4044 total appearances in the corpus, which translates to an absolute frequency of roughly 1.19%. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners, prepositions and one adjective and pronoun.

4.7. The word “Virus”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Virus” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
spreading	8.218937	7	16
antibodies	7.274079	5	22
spread	7.227158	11	50
virus	5.202129	8	148
being	5.062854	8	163
than	4.719092	7	181
any	4.711143	6	156
how	4.679779	7	186
this	4.584071	33	937
the	4.569567	141	4044

Table 13: MI-score results for the word "Virus"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word “Virus” out of all the other collocates of the word “Virus”.

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
the	11.374269	141	4044
to	6.677447	53	2988
this	5.505057	33	937
a	4.738399	29	2372
and	4.642446	28	2339
is	4.470969	24	1428
of	4.353776	24	1819
not	4.252808	20	668
that	3.821443	18	1217
for	3.730651	17	1102

Table 14: T-score results for the word "Virus"

The MI-score results for this word show that the strongest collocate was the word “spreading” with MI-score of roughly 8.21 and with collocation frequency of 7 out of 16 total appearances in the corpus, which translates to a relative frequency of

43.75%. The word "virus" itself was also surprisingly one of the results (e.g. when referring to the virus in consecutive sentences). "Spreading" + "Virus" was the most relatively frequent and strongest collocation because "the spread(ing) of the virus" or "virus spread(ing)" is a basic concept and a highly discussed topic in almost all forms of media from official government statements and advice on how to stop the spread of the virus with proper hygiene to general news articles and COVID-19 statistics. In terms of MI-score the word "Spreading" is also followed by the word "Spread", which is the same lexeme but without the "-ing" suffix. Overall, the MI-score results table seems to be comprised mostly of adjectives, nouns, and some verbs and determiners, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word "the" with T-score of roughly 11.37 and collocation frequency of 141 out of 4044 total appearances in the corpus, which translates to an absolute frequency of roughly 3.49%. The score difference between the first and the second word is also quite large which indicates how frequently was the word "the" used compared to the other options. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners and prepositions.

4.8. The word “Lot”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Lot” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
parking	9.418267	12	13
there's	6.337348	6	55
things	6.260726	9	87
there	5.570271	20	312
a	5.440798	139	2372
of	5.334172	99	1819
these	5.031245	6	136
more	5.025950	12	273
than	4.841254	7	181
people	4.730518	25	698

Table 15: MI-score results for the word "Lot"

These are the top ten collocates with the highest T-score (and absolute frequency), thus being the most frequent collocations with the word “Lot” out of all the other collocates of the word “Lot”.

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
a	11.518392	139	2372
of	9.703230	99	1819
the	5.551053	41	4044
people	4.811660	25	698
and	4.588850	27	2339
in	4.486167	24	1499
there	4.378013	20	312
are	4.324087	21	878
to	3.955262	23	2988
is	3.916913	19	1428

Table 16: T-score results for the word "Lot"

The MI-score results for this word show that the strongest collocate was the word “parking” with MI-score of roughly 9.42 and with collocation frequency of 12 out of 13 total appearances in the corpus, which translates to a relative frequency of

92.31%. This is yet another very high percentage meaning that almost every word "Parking" in the corpus was a collocate of the word "Lot". "Parking" + "Lot" was the most relatively frequent and strongest collocation mostly because of post #11, as over half of this collocation's occurrences came from this post alone. Otherwise, the word "parking" would likely not have such a high score. However, most uses of this collocation were to address and comment on how parking lots are full despite there being an ongoing pandemic. In addition, in the case of the word "Parking", the word "Lot" is used with a different and less frequent meaning of "something that is shared". As for the nine remaining collocates in the table, they are the most frequent meaning of "much/many". This is also precisely the reason, why the MI-score of the "Parking lot" collocation is so high (without the word "Parking", the word "Lot" appears very rarely with this specific meaning). Overall, the MI-score results table seems to be comprised mostly of nouns, and determiners and an adjective, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word "a" with T-score of roughly 11.52 and collocation frequency of 139 out of 2372 total appearances in the corpus, which translates to an absolute frequency of roughly 5.86%. The score difference between the first and the second word is also quite large which indicates how frequently was the word "the" used compared to the other options. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. Overall, the T-score results table comprises mostly determiners, verbs, prepositions, and a noun.

4.9. The word “Case”

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word “Case” out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
confirmed	9.504923	7	18
in	4.902671	24	1499
just	4.585795	5	389
this	4.455020	11	937
a	4.240567	24	2372
my	4.238137	6	594
with	4.169367	6	623
be	3.776499	6	818
we	3.754793	5	692
of	3.739003	13	1819

Table 17: MI-score results for the word "Case"

These are the only six collocates with their T-score (and absolute frequency), thus being the most frequent collocations with the word “Case”.

Collocate	*T-score*	*Freq (coll.)*	*Freq (corpus)*
in	4.735202	24	1499
a	4.639820	24	2372
the	3.862315	19	4044
and	3.407058	14	2339
of	3.335517	13	1819
this	3.165407	11	937

Table 18: T-score results for the word "Case"

The MI-score results for this word show that the strongest collocate was the word “confirmed” with MI-score of roughly 9.5 and with collocation frequency of 7 out of 18 total appearances in the corpus, which translates to a relative frequency of 38.89%. The score difference between the first and the second word is also quite large which indicates how frequently was the word “confirmed” used compared to the other options. “Confirmed” + “Case” was the most relatively frequent and strongest collocation because of numerous discussions surrounding COVID-19 statistics and the number of confirmed cases in each country. This was prevalent mostly during the first few months or even the first year of the pandemic, when numbers of confirmed cases were on the rise in perhaps every country, which in return gave people some idea on how their country is doing statistically compared to the rest of the world. Overall, the MI-score results table seems to be comprised mostly of prepositions, determiners, and an adjective, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word "in" with T-score of roughly 4.74 and collocation frequency of 24 out of 1499 total appearances in the corpus, which translates to an absolute frequency of roughly 1.6%. This is the second and last case, where the word with the highest T-score is a preposition instead of an article like in previous T-score results tables. The results also show that there were only six collocates that corresponded to the specific parameters used in the analysis. There were probably more collocates to this word, but their MI-score was lower than the required value of 3, so they were not included. Overall, the T-score results table comprises mostly determiners, and prepositions.

4.10. The word "Country"

These are the top ten collocates with the highest MI-score (and relative frequency), thus being the strongest and most frequently collocated with the word "Country" out of all the other words these were collocated with.

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
rest	7.666750	5	34
most	5.720790	5	131
from	4.920532	8	365
what	4.681857	6	323
this	4.367734	14	937
was	4.353943	7	473
of	4.188317	24	1819
their	4.173012	5	383
with	4.149196	8	623
the	4.094573	50	4044

Table 19: MI-score results for the word "Country"

These are the only eight collocates with their T-score (and absolute frequency), thus being the most frequent collocations with the word “Country”.

Collocate	***T-score***	***Freq (coll.)***	***Freq (corpus)***
the	6.657168	50	4044
a	4.967084	28	2372
of	4.630262	24	1819
to	4.344925	23	2988
is	3.741633	16	1428
this	3.560421	14	937
in	3.451718	14	1499
and	3.289244	14	2339

Table 20: T-score results for the word "Country"

The MI-score results for this word show that the strongest collocate was the word “rest” with MI-score of roughly 7.67 and with collocation frequency of 5 out of 34 total appearances in the corpus, which translates to a relative frequency of 14.71%. “Rest” + “Country” was the most relatively frequent and strongest collocation because when discussing the pandemic and how it affects certain people or groups, “the rest of the country” (in this case the most frequent collocation) also often comes into play, as people may be worried about the rest of their country or perhaps just want to compare their current situation with the rest of the population. Overall, the MI-score results table seems to be comprised mostly of determiners, prepositions, pronouns, and an adverb, but it is difficult to determine in some cases due to the lack of semantic context in the table.

The T-score results for this word show that its most frequent collocate was the word “the” with T-score of roughly 6.66 and collocation frequency of 50 out of 4044 total appearances in the corpus, which translates to an absolute frequency of roughly 1.24%. This is unsurprising as stated previously, because articles (both definite and indefinite) are very commonly found near nouns but carry no real semantic meaning. The results also show that there were only six collocates that corresponded to the specific parameters used in the analysis. There were probably more collocates to this word, but their MI-score was lower than the required value of 3, so they were not included. Overall, the T-score results table comprises mostly determiners, prepositions, and a verb.

IV. CONCLUSION

The aim of this thesis was to create a corpus and subsequently with its help create collocational profiles for the most frequently used nouns in COVID-19 forum threads to find out, what were their most frequent collocates. To gather the required data sample to create a corpus, a total of fifteen forum posts and their comments from the website Reddit.com were selected and their content “downloaded” using a script and imported into #LancsBox. A corpus was then created and used in the creation of collocational profiles for the most frequently used nouns.

First before creating the collocational profiles, a frequency list of nouns needed to be created using the “Words” function of #LancsBox. This list was used to determine which nouns were the most frequent. Ten of the most frequent nouns from the frequency list were then chosen and had two sets of collocational profiles (first using MI-score, second using T-score) created for them using the “GraphColl” function of #LancsBox.

The first set of collocational profiles (MI-score) showed some surprising results like in the case of the word “Case” where its most frequent collocate, the word “Confirmed”, was almost twice as frequent as the next word after that because of the pandemic and discussions about confirmed COVID-19 cases. Other profiles also showed how even some common words, like “People” or “Home”, can have their collocates influenced under special circumstances like the ongoing pandemic.

The results of the second set of collocational profiles (T-score) were however quite unsurprising because except for two cases, either the article “the” or “a” were the most frequent collocates. This is to be expected as T-score deals with absolute frequency and nouns are often accompanied by articles. The two times, when this was not the case, were with the words “Case” and “Home”, where the most frequent collocates were prepositions “in” and “at” respectively, because “in case” and “at home” are idioms (which is why both prepositions also have a relatively high MI-score). This perfectly shows us that in case of these words, the prepositions “in” and “at” are not just grammatical.

Overall, the research can be considered successful since we clearly found out what the most frequent nouns were and created their collocational profiles, the results of which showed how usage of certain words and collocations on online forum threads can change in response to a worldwide event such as the COVID-19 pandemic.

List of tables and figures

Table 1: MI-score results for the word "People" 37

Table 2: T-score results for the word "People" 37

Table 3: MI-score results for the word "Time" 38

Table 4: T-score results for the word "Time" 39

Table 5: MI-score results for the word "Mask" 40

Table 6: T-score results for the word "Mask" 40

Table 7: MI-score results for the word "Home" 42

Table 8: T-score results for the word "Home" 42

Table 9: MI-score results for the word "Thing" 43

Table 10: T-score results for the word "Thing" 44

Table 11: MI-score results for the word "Day" 45

Table 12: T-score results for the word "Day" 45

Table 13: MI-score results for the word "Virus" 47

Table 14: T-score results for the word "Virus" 47

Table 15: MI-score results for the word "Lot" 49

Table 16: T-score results for the word "Lot" 49

Table 17: MI-score results for the word "Case" 51

Table 18: T-score results for the word "Case" 51

Table 19: MI-score results for the word "Country" 52

Table 20: T-score results for the word "Country" 53

Figure 1: The logo of Reddit (Reddit.com). 19

Figure 2: The frontpage of r/Czech subreddit (Reddit.com/r/Czech). 20

Figure 3: Example of how comments from Reddit look after executing the script. 21

Figure 4: Example of the output of the script shown in MS Word. 21

Figure 5: The r/Coronavirus subreddit (Reddit.com/r/Coronavirus) 22

Figure 6: The interface of #LancsBox. 23

Figure 7: The "Words" function. 24

Figure 8: Top 10 most frequent nouns. 24

Figure 9: The "GraphColl" function using the MI-score. 25

Figure 10: Results of T-score analysis of the noun "Mask". 36

Figure 11: Optional examination of collocates using the “KWIC” function. 36

V. REFERENCES

Primary sources:

BARFIELD, A., & GYLLSTAD, H. (Eds.). (2009). *Researching Collocations in Another Language: Multiple Interpretations*. Palgrave Macmillan. ISBN 9780230245327.

BENSON, M., BENSON, E., ILSON, R. (Eds.). (1993). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam, Philadelphia: John Benjamins Publishing Company. ISBN 9780915027804.

CRYSTAL, David. *A Dictionary of Linguistics and Phonetics*. 3. B. Blackwell, 1991. ISBN 9780631178712.

ČERMÁK, František a Michal ŠULC. *Kolokace*. 2. Praha: Nakladatelství Lidové noviny, 2006. ISBN 80-7106-863-2.

ČERMÁK, František. *Collocations, Collocability and Dictionary*. Turín: Edizioni dell'Orso, 2006. ISBN 88-7694-918-6.

ČERMÁK, František. *Frazeologie a idiomatika - česká a obecná*. 1. Praha: Karolinum, 2007. ISBN 978-80-246-1371-0.

ČERMÁK, František a Jiří HRONEK. *Slovní české frazeologie a idiomatiky*. Praha: Leda, 2009. ISBN 978-80-7335-219-6.

GEERAERTS, Dirk. *Theories of Lexical Semantics*. New York: Oxford University Press, 2010. ISBN 978-0-19-870030-2.

KNAPPE, Gabriele. *Idioms and Fixed Expressions in English Language Study Before 1800: A Contribution to English Historical Phraseology*. Michiganská univerzita: P. Lang, 2004. ISBN 9780820473451.

McENERY, T. & HARDIE, A., 2011 - McEnery, T., & Hardie, A. (2011). *What is corpus linguistics?* In Corpus Linguistics: Method, Theory and Practice (Cambridge Textbooks in Linguistics, pp. 1-24). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511981395.002.

NESSELHAUF, N. (2005). *Collocations in a Learner Corpus*. Amsterdam, Philadelphia: John Benjamins Publishing Company. ISBN 9789027222855.

SINCLAIR, John a Les SINCLAIR. *Corpus, Concordance, Collocation*. 2. Kalifornská univerzita: Oxford University Press, 1991. ISBN 9780820473451.

Internet sources:

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). *#LancsBox v. 5.x.* [software]. Dostupné z: http://corpora.lancs.ac.uk/lancsbox.

CVRČEK, Václav a Olga RICHTEROVÁ. *Asociační (kolokační) míry*. Pojmy:asociacni_miry [online]. Příručka ČNK, 2021 [cit. 2021-6-28]. Dostupné z: http://wiki.korpus.cz/doku.php?id=pojmy:asociacni_miry&rev=1554815423

GOMEZ, Cristobal. *Collecting collocations*, Kaplan International Languages [online]. Kaplan International Languages, 2021 [cit. 2021-6-28]. Dostupné z: https://www.kaplaninternational.com/blog/learning-languages/eng/collecting-collocations-speak-like-a-native

Homepage - Reddit [online]. Reddit [cit. 2021-6-28].
Dostupné z: https://www.redditinc.com

McARTHUR, T. (1992). *The Oxford Companion to the English Language*. [online] Oxford University Press. [cit. 2021-6-28].
Dostupné z: https://www.cambridge.org/core/journals/language-in-society/article/abs/tom-mcarthur-ed-the-oxford-companion-to-the-english-language-oxford-new-york-oxford-university-press-1992-pp-xxix-1184/4D4806A81E2909AD4CCF2BBC36561966

METHENITI, Eleni, Tim VAN DE CRUYS a Nabil HATHOUT. *How Relevant Are Selectional Preferences for Transformer-based Language Models?* ACL Anthology [online]. Barcelona: International Committee on Computational Linguistics, 2020 [cit. 2021-6-28].
Dostupné z: https://www.aclweb.org/anthology/2020.coling-main.109

Metrics For Reddit [online]. 2021 [cit. 2021-6-28].
Dostupné z: https://frontpagemetrics.com

Phraseology [online]. Wikipedia, The Free Encyclopedia, 2021 [cit. 2021-6-28].
Dostupné z: https://en.wikipedia.org/w/index.php?title=Phraseology

Reddit [Online]. Reddit [cit. 2021-6-28].
Dostupné z: https://www.reddit.com

Reddit - r/Coronavirus [online]. Reddit [cit. 2021-6-28].
Dostupné z: https://www.reddit.com/r/Coronavirus/

Reddit - r/Czech [online]. Reddit [cit. 2021-6-28].
Dostupné z: https://www.reddit.com/r/czech/

Selectional Restrictions and Preferences [online]. Stanford University [cit. 2021-6-28].
Dostupné z: https://web.stanford.edu/~jurafsky/slp3/slides/22_select.pdf

TANKOVSKA, H. *Reddit - Statistics & Facts*. Statista [online]. 2021 [cit. 2021-6-28].
Dostupné z: https://www.statista.com/topics/5672/reddit/

WEI, Yong. *Teaching Collocations for Productive Vocabulary Development*. [online] ERIC - Institute of Education Sciences, 1999. [cit. 2021-6-28]. Dostupné z: https://eric.ed.gov/?id=ED457690

What are collocations? [online]. Future Learn, Macquarie University, 2021 [cit. 2021-6-28].
Dostupné z: https://www.futurelearn.com/info/courses/improve-ielts-speaking/0/steps/98854

What is a corpus, what is corpus linguistics? [online]. University of Helsinki, 2016 [cit. 2021-6-28].
Dostupné z: https://www.futurelearn.com/info/courses/improve-ielts-speaking/0/steps/98854