Athens Journal of Technology & Engineering 2022, 9: 1-29
https://doi.org/10.30958/ajte.X-Y-Z

Using Digital Humanities for Understanding COVID-19: Lessons from Digital History about Earlier Coronavirus Pandemic

By Tado Jurić[*]

At the time of the COVID-19 epidemic, it is useful to look at what lessons (digital) history can give us about the past pandemics and dealing with them. We show that the Google Ngram Viewer (GNV) can discover hidden patterns in history (of pandemics). Our study is searching for evidence that the COVID-19 is not a unique phenomenon in human history. By using the approach of Digital Humanities, we are testing the hypothesis that the flu-like illness that caused loss of taste and smell in the late 19th century (Russian flu) was caused by a coronavirus. This approach could give hints on how the COVID-19 might develop in the following years. The objective was to calculate the ratio of increasing to decreasing trends in the changes in frequencies of the selected words representing symptoms of the Russian flu and COVID-19. The primary methodological concept of our approach is to analyse the ratio of increasing to decreasing trends in the changes in frequencies of the selected words representing symptoms of the Russian flu and COVID-19 with the Google NGram analytical tool. Initially, keywords were chosen that are specific and common for the Russian flu and COVID-19. We show the graphic display on the Y-axis what percentage of words in the selected corpus of books (collective memory) over the years (X-axis) make up the word. To standardise the data, we requested the data from 1800 to 2019 in English, German and Russian (to 2012) book corpora and focused on the ten years before, during and after the outbreak of the Russian flu. We compared this frequency index with "non-epidemic periods" to test the model's analytical potential and prove the significance of the results. The limitation of this study is that it is difficult to formulate a hypothesis for a microbiological aetiology of a pandemic that occurred 133 years ago based on symptoms. There are indications that COVID-19 is not unique because the Russian flu (1889-1891) might be a coronavirus infection. The most crucial observation of similarities between the Russian flu pandemic and COVID-19 is the loss of smell and taste (anosmia and ageusia). Results show that all the three analysed book corpora (including newspapers and magazines) show the increase in the mention of the symptoms "loss of smell" and "loss of taste" during the Russian flu (1889-1891), which are today undoubtedly proven to be key symptoms of COVID-19. The mention of symptoms and the pandemic-related words fell sharply after the pandemic stopped. According to our analysis of historical records with the approach of GNV, 1) the 'natural' length of a pandemic is two to five years; 2) the pandemic stops on its own; 3) the viruses weaken over time; 4) the so-called "herd immunity" is not necessary to stop the pandemic; 5) Our approach has shown that a significant crisis does not need to occur after the COVID-19 pandemic. According to our study, GNV clearly shows the influence that social changes

[*]Assistant Professor, Catholic University of Croatia, Croatia.

have on word frequency. The results of this study open a discussion on the usefulness of the Google Ngram insights possibilities into past socio-cultural development, i.e. epidemics and pandemics that can serve as lessons for today. However, this method has severe limitations and can be useful only under cautious handling and testing. Despite the numerous indications, we are aware that this thesis still cannot be confirmed and that it requires further historical and medical research.

Keywords: *Google Ngram, big data, epidemic, COVID-19, Russian flu, digital Humanities*

Introduction[1]

An answer to the future development of the COVID-19 pandemic is of high importance for all societies and countries worldwide. By messages from the media and official reports, we know that they are unreliable and that epidemiological predictions are uncertain. Because medical evidence and epidemiological estimates cannot answer this question, looking at history's lessons can be helpful.

Studying past pandemics shows that elements relevant to the COVID-19 pandemic are repeated and that the measures that we undertake today are precisely the same as what they did in Spanish flu and partially in Russian flu – social distancing, wearing masks, quarantining, travel restrictions (King 2021). But just as individuals forget the past, so do societies (Halbwachs 1992). This paper shows that Digital Humanities approaches might be used to track historical epidemics and renew knowledge from the past.

According to Brüssow (2021) and Van Ranst (King 2021), the Russian flu might have been a coronavirus infection. Due to the limitations, it is impossible to have medical evidence for this thesis. Therefore, we set the hypothesis that the tools of Digital Humanities, especially Google Books Ngram Viewer (GNV), can help find the clinical data from the historical reports.

Our goal in this paper is to analyse the epidemiological literature on the development of the Russian flu pandemic (1889) for hints on how the COVID-19 might develop in the following years and compare the similarities. The historical record of past pandemics might thus provide us with the so-called "retrodictions" (Brüssow 2021) on possible future scenarios for the COVID-19 pandemic.

According to Van Ranst, the first coronavirus was transmitted from bovines to humans. According to this thesis, what we are experiencing today has already been experienced in the late 19th century (King 2021). To find evidence, we have analysed the indices by the clinical data from the historical reports from the Google corpus of digitised books that includes 15 million books (12% of all books ever published). We asked ourselves, especially if the COVID-19 pandemic is a unique

[1]The study was released as a pre-print version on 6 February 2022 during the peak of the pandemic: Jurić T (2022) *Using digital humanities for understanding COVID-19: lessons from digital history about earlier coronavirus pandemic*. MedRxiv, https://www.medrxiv.org/content/10.1101/2022.02.02.22270333v1.full-text. DOI: https://doi.org/10.1101/2022.02.02.22270333.

occurrence in humanity, whether it will disappear or become endemic, and the future consequences.

According to our study, the GNV clearly shows the influence that social changes have on word frequency. The relationship between values fostered in a society and its language is close (Brüssow and Brüssow 2021). Our basic assumption is that when culture and language are linked, one should impact the other. Furthermore, it has been recently shown that during seasonal influenza epidemics, users of Google are more likely to engage in influenza-related searches and that this signature of influenza epidemics corresponds well with the results of CDC surveillance (Jurić 2021b). We, therefore, reasoned that the Big Data and Digital Humanities approaches might be used to track historical epidemics and give us answers to some questions that would otherwise remain unanswered.

Big Data in Digital Humanities

The expression "Big Data" has been spreading since 2011. The term is used in academia, industry and the media, but it is not even today precisely clear what it means. Is it an object of study, a method, a group of technologies or a discipline (Rojas Castro 2017)? The definitions combine two essential ideas: storing a large volume of data and analysing this data quantitatively and visually to find patterns, establish laws, and predict conduct (Ward and Barker 2013). The classic definition of "Big Data" is a formula - the three "Vs": Volume, Velocity (data that is constantly generated) and Variety (texts, images, sounds) (Ward and Barker 2013).

According to Oza, Digital Humanities is "a broad field of research and scholarly activity covering the use of digital methods by arts and humanities researchers and how the arts and humanities offer distinctive insights into the major social and cultural issues raised by the development of digital technologies" (Oza 2020). Work in this field is methodological and interdisciplinary in scope, involving multiple skills, disciplines, and areas of expertise with the investigation, analysis, synthesis and presentation of data electronically (Oza 2020). According to Burdick et al. (2012) "Digital Humanities is less a unified field than an array of convergent practices exploring a universe in which print is no longer the primary medium in which knowledge is produced and disseminated" (Burdick et al. 2012).

Big Data is widely used today in digital culture as a promising method for deriving new understanding from massive aggregations of information. The ability to collect a vast amount of data from text, images, and media and to analyse it using computerised algorithms creates endless opportunities in many areas (Ophir 2016). "Big data" methodologies bring new potential not just for medicine and business analytics but also for humanities research and social sciences. Latour believes that big data can resolve the gap between the micro and the macro in sociology, the unexplained relations between macro-social phenomena and the individuals taking part in that phenomena (Latour 2014).

In the humanities, one can only speak of Big Data in connection with the technologies associated with this phenomenon, such as data mining, stylometry or natural language processing (Rojas Castro 2017). It is crucial to differentiate

between "data", "raw material", and "information". According to Castro, more than the finished product, what matters in the Digital Humanities is the creative process when a phenomenon is "modelled". The aim is to gain new knowledge and meanings by generating an external object that represents it (Rojas Castro 2017). Humanistic disciplines such as history, philosophy, and philology are characterised by a specific object of study and a method that seeks to understand particular, unusual and even unique cases through text commentary. According to Castro, Big Data in humanities will unquestionably affect certain clichés about the Humanities and their classic objects of study (Rojas Castro 2017). Although the tool may help develop specific theories concerning socio-cultural phenomena, many researchers claim that the data obtained with Google Books Ngram Viewer is not reliable enough to confirm these theories (Zięba 2018) (see Limitations).

Google NGram Viewer

Reading small collections of carefully chosen works enables scholars to make robust inferences about trends in human thought. However, this approach rarely allows precise measurement of the underlying phenomena. According to Michel et al. (2011a), computational analysis of the digitalised corpus of books enables us to observe cultural trends and subject them to quantitative investigation. This new field, *Culturomics*, extends the boundaries of scientific inquiry to a wide array of new phenomena (Lieberson and Horwich 2008).

One of the tools that serve Digital Humanities is GNV.[2] This tool has been created on top of Google Books, the largest digitised collection of books. GNV is creating a graphical representation of the frequency of occurrence of search terms over the years in a selected corpus of digitised books (Michel et al. 2011a). It contains a corpus of over 15 million digitised books and over 600 billion words in 2022. It is actually the world's largest archive - which is also available online and for free. Google states that its team, together with Cultural Observatory, Harvard University, Encyclopaedia Britannica and the American Heritage Dictionary, have digitised over 15% of all books that have ever been published from over 40 university libraries (such as the University of Michigan and the New York Public Library) and individual publishers.[3] In 2004, Google began by scanning books (OCR). The first version in 2009 had six million books; in 2012, the second version incorporated eight million books (Lin et al. 2012), and the 2019 version had over 15 million books. Due to the wide scale of digitally archived texts, these corpora are not limited to specific genres. It includes all sorts of literature, ranging from academic publications to biographies and novels (Chumtong and Kaldewey 2017). The collection contains books dating back to as early as 1473 and texts in 478 languages (Michel et al. 2011b). Of the 15 million books scanned, the country of publication is known for 91.5%, authors for 92.1%, publication dates for 95.1%, and the language for 98.6%. The OCR quality is generally higher for the languages

[2]Google NGram View: https://books.google.com/ngrams.
[3]Ibid.

that use a Latin alphabet (English, French, Spanish, and German), and more books are available (Michel et al. 2011b).

The new version of GNV from 2019 is characterised by improved optical character recognition (OCR) and better underlying library and publisher metadata (Younes and Reips 2019). Google estimates that over 98% of words are correctly digitised for modern English books (Michel et al. 2011a). The GNV does offer differentiation by language. Subcorpora exist for eight languages, with the English corpus being the biggest, containing more than 350 billion words. The corpus covers a period from 1500 until 2008. However, Michel et al. (2011b) point out that search inquiries between 1800 and 2000 will deliver the highest data density and quality. The problem is that smaller language communities are not included.

Compared to other big data sets, the GNV enables fast and easy access to this pool of information (Chumtong and Kaldewey 2017). Next to a regular search field for the term or phrase of interest, the online tool offers filtering options for the period, the language, the degree of smoothing that affects how the graphs of the search result are displayed, and a case insensitive option. It is also possible to search for more than one term or phrase for direct comparison (Chumtong and Kaldewey 2017). Next to avoid overwhelming the diagram in any given year, the graph will only show books with the term(s) if there are more than 40 occurrences. To deal with the problem presented by the increase in published books over time, the results are normalised by the number published each year (University of London n.d.).

Without a normalisation, it would be impossible to compare the frequency of a specific n-gram over time, as the number of books published in 1500 is not equal to the number of books published in 2010.[4] The viewer, therefore, displays a percentage of the number of occurrences, where the percentage is calculated out of the total number of books published in a given year. Clicking on a point in the plotted graph shows the rate of occurrences for that year (Ophir 2016). The data generated by specific inquiries can then be exported as a list and processed with alternative software packages (for example, "R"), particularly with spreadsheet applications (Chumtong and Kaldewey 2017).

GNV can be used as a tool for discovering hidden patterns of conceptual trends, trends in knowledge, the relative importance of concepts etc. (Kratzer 2019). The main challenge for Digital Humanities will be to take patterns discovered by digital analysis and discern correlations to historical events, to explain patterns by historical forces, causes and relations (Ophir 2016).

How to Use Google NGram Viewer in Digital Humanities

The GNV calculates how often a certain n-gram appears in the selected corpus of a given year relative to the total number of n-grams (Michel et al. 2011a). In computational linguistics, an Ngram is a contiguous sequence of *n* items from a given sequence of text[5], and the items can be phonemes, syllables, letters or

[4]Ibid.

[5]Google NGram View: https://books.google.com/ngrams.

words. The GNV database supports n-gram sequences of up to five elements (Ophir 2016). For example, "I" is a 1-gram and "I am" is a 2-grams - this means that if the researcher searches for one word (unigram), he will get the percentage of this word to all the other words found in the corpus of books for a specific year (Kratzer 2019). If the researcher entered more than one word or phrase, each one is represented by a colour-coded line to contrast with the other search terms. This is similar to Google Trends (see Jurić 2021a), except the search covers a longer period (Karch 2021).

The researcher can modify searches by time frame, degree of detail and corpus type, including several different languages as mentioned. As well as verbs and nouns, scholars can also search for adjectives, adverbs, pronouns, determiners, prepositions and more, using the tags listed on this helpful page of tips. Google estimates the accuracy of this tagging at 95% (Kratzer 2019).

A few features of the GNV may appeal to users who want to dig a little deeper into phrase usage: *wildcard search*, *inflexion search*, *case insensitive search*, *part-of-speech tags* and *n-gram compositions* (Kratzer 2019). For comparisons of several n-grams, it is possible to combine or separate two expressions and divide or multiply expressions to compare n-grams of different frequencies or to isolate frequencies of one n-gram in relation to another. Adding a "+" operator between n-grams allows the researcher to combine multiple frequencies into one. Adding the operator "-" between n-grams allows the subtraction of frequencies from the right from the frequencies from the left and thus enables the measurement of frequency connectivity (Michel et al. 2011a).

Adding the "/" operator between n-grams allows isolating the movement of one frequency to another. Adding the operator "*" between n-grams multiplies the frequency on the left by the frequency with the selected value, that is, by the given number. It allows a comparison of two distinctly different frequencies. Adding the ":" operator between n-grams uses the n-gram on the left and the corpus on the right, and compares n-grams in different corpora (Michel et al. 2011a).

Representation of words in multiple grammatical categories can be achieved by adding the code "_INF" as a suffix to the word's root. Example: "book_INF" generates the appearance of words such as "books", "booking", "booked" for viewing in a single graphical display.[6] GNV offers the option to tag words in search, such as "_NOUN_" (noun), "_VERB_" (verb), "_ADJ_" (adjective), "_ADV_" (adverb). These labels can serve as part of a word or make up the word itself. By entering the operator "=>," it is possible to show the relationship between words and their connection in a sentence.

There is also a case-insensitive option - displaying words written in lowercase, uppercase only, or a combination of words. If smoothing factor "1" is selected as the smoothing level, it means that the data are shown for - for example, 1990 will be the average of the raw data for 1990 summed with one value on each side (previous and future years) and divided by the number year (data for 1989 + data for 1990 + data for 1991) (Kardaš 2020). GNV does not make the search result available for further processing. Even though it is possible to download the

[6]Google NGram View: https://books.google.com/ngrams.

raw data, this option only addresses extensive scale analyses that require technical resources and advanced know-how in computer science.[7] However, there is a pragmatic way of extracting data from the HTML source code shown by Chumtong and Kaldewey (2018).

The primary method used by GNgram is text mining. It is a method for gathering structured information from unstructured text and discovering meaningful relationships (Berry 2012). Text mining has significant potential for academic application (Berry 2012) to 1) develop new hypotheses, 2) systematic reviewing of literature, and 3) testing of hypotheses. Documents can be mined to confirm or deny an existing hypothesis. In many cases, this might be the first opportunity to test an established belief about something (Berry 2012).

Text mining enables the identification of patterns and relationships within a large body of texts that would otherwise be extremely difficult or time-consuming to discover. Therefore, it is a method that can speed up research and allow us to pose new questions or test the old ones. One of the merits of this tool is that it enables the socio-cultural researcher to spend more time analysing data than on their collection, which is usually very time-consuming (Zięba 2018).

According to Zięba (2018), since the lexical changes are gradual and relatively stable, the fluctuations in word frequency are relevant, and their study will improve our comprehension of the social changes and their consequences (Zięba 2018). However, this method comes with severe limitations and can be useful only under the condition of cautious handling and testing. Otherwise, there is a high potential to gather garbled or false results due to badly formed questions being asked of data or the nature of the text(s) under study (Berry 2012). It is important to stress that no result from text mining should be taken at face value by historians. Results must be checked and confirmed, and this often involves manually delving into the text under study (see section Limitations and Methodology).

Literature Review

Since its introduction in 2010, GNV has been widely described and applied both in the social and natural sciences (Zięba 2018). Berry (2012) describes it as an example of "the way in which code and software become the conditions of possibility for human knowledge". Rutten et al. treat it as a tool to overcome a "chronological distance, or time lag, between books and their subject matter in studies of memory" (Rutten et al. 2013). Michalski et al. (2012) suggest the GNV could be used "as a fast prototyping method for examining time-based properties over a rich sample of literary prose".

Linguists used it to investigate biomedical domain literature in respect of terminology changes. In social studies, it was used to prove that moral ideals and virtues decreased significantly in the American public conversation, to analyse the concepts of happiness across time and cultures, to trace the roots of industrial

[7]Google Apis: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html. [Accessed 23 December 2021]

ecology education to the 1960s and 1970s, to study the relations of science and capitalism, to trace the history of marketing and to introduce the concept of information overload (Zięba 2018).

As already mentioned, Michel et al. (2011a) showed that the corpus enables investigators to study cultural trends quantitatively. The authors inquire into collective memory, compare the rise and fall of fame of the most well-known people, and uncover censorship in Nazi Germany. Michel et al. showed that this approach could provide insights into fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. The authors examined timelines for four diseases: influenza, cholera, HIV and poliomyelitis. In the case of flu, peaks in cultural interest showed excellent correspondence with known historical epidemics (the Russian flu of 1890, the Spanish flu of 1918 and the Asian flu of 1957) (Michel et al. 2011b).

Newberry et al. (2017) use Google Ngram Viewer to analyse changes in the English language from the 12th to the 21st century. Greenfield (2013) tested with GNV her theory on the influence of individualism on the individual's values, behaviour, and psychology.

Acerbi et al. (2013) explored the presentation of emotions in books through the twentieth century using GNV. The authors conclude that stressful and violent historical events leave traces in the expression of emotions in books, so it is possible to detect "happy" and "sad" periods of history, depending on the representation and use of words for certain emotions through books (Acerbi et al. 2013). Overall, GNV has allowed scholars to shed further light on various topics such as gender differences (Twenge et al. 2012), emotions (Mohammad 2012), personality (Roivainen 2015), cognition (Virues-Ortega and Pear 2015), psychotherapy (Rossi at al. 2013), moral values (Mooijman et al. 2018), education (Roivainen 2014), nature (Kesebir and Kesebir 2017) and the development of individualism and collectivism (Grossmann and Varnum 2015).

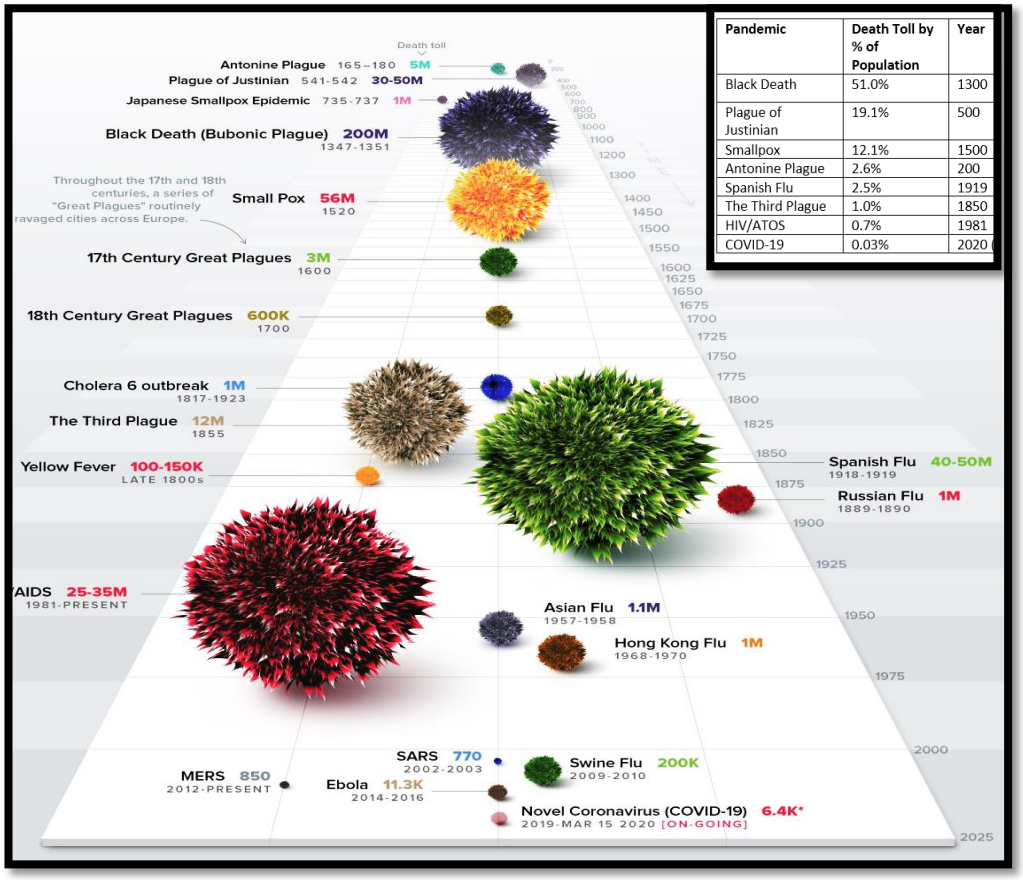
Epidemics through History

Epidemics and pandemics have always been a part of human life. Since the existence of man, there have been infectious diseases. According to Harari (2014), infectious diseases start when a person begins living sedentary; stops collecting and hunting. The First Agrarian Revolution cost man various diseases and contagions. The man no longer moved; he began to breed, keep animals and live in one place, which became an excellent prerequisite for developing diseases (Harari 2014). The spread of trade and the interaction of a growing number of people has led to epidemics, and in those times, it was not even known what humanity was facing. As humanity became more civilised, with the emergence of larger cities and population growth, exotic trade routes, and increased contact with different people, animals and ecosystems, the emergence of pandemics became greater.[8]

[8]Lider.hr: https://lider.media/poslovna-scena/svijet/infografika-sve-pandemije-kroz-povijest-130435. [Accessed 23 October 2021]

The infographic below (Figure 1) outlines some of the deadliest pandemics in human history, from the Antonine Plague that struck the Roman Empire from 165 to 180 to today's current events and coronaviruses.

Figure 1. *Timeline of Historical Pandemics*



Pandemic	Death Toll by % of Population	Year
Black Death	51.0%	1300
Plague of Justinian	19.1%	500
Smallpox	12.1%	1500
Antonine Plague	2.6%	200
Spanish Flu	2.5%	1919
The Third Plague	1.0%	1850
HIV/ATOS	0.7%	1981
COVID-19	0.03%	2020

Source: Visual capitalist, CDC, WHO, BBC, Encyclopedia Britannica (https://lider.media/poslovna-scena/svijet/infografika-sve-pandemije-kroz-povijest-130435), edited by author.

By the end of the 16th century, influenza was likely beginning to become understood as a specific, recognisable disease with epidemic and endemic forms. Since pandemic 1781–1782, starting in China, influenza became associated with sudden outbreaks of febrile illness (Potter 2001). Around the world, during the pandemics of 1889 (Russian flu) and 1918/1919 (Spanish flu), between 50 and 100 million people are estimated to have died (Spinney 2018; see: Kucharski 2020). A direct comparison between the pre-pandemic and the coronavirus cannot be made. The world at the time did not know what made people die, and viruses as the cause of the disease were discovered only in 1933. But these pandemics still have something in common: they have thrown humanity into a deep crisis. That is why we wonder if the experiences from historical records about pandemics can help us prepare for the actual pandemic and the time after the pandemic.

The problem also arises in differentiation between Flu and COVID-19. Flu and COVID-19 are contagious respiratory illnesses, but different viruses cause

them. COVID-19 is caused by infection with a coronavirus first identified in 2019, and flu is caused by infection with influenza viruses (CDC n.d.). Similarities are that both COVID-19 and flu can have varying signs and symptoms, ranging from no symptoms (asymptomatic) to severe symptoms. Common symptoms that COVID-19 and flu share include: fever or feeling feverish/having chills; cough; shortness of breath or difficulty breathing; fatigue (tiredness); sore throat; runny or stuffy nose; muscle pain or body aches; headache; vomiting and diarrhoea (CDC n.d.).

Russian Flu - An Earlier Coronavirus Pandemic?

According to Van Ranst, a flu-like illness that caused loss of taste and smell in the late 19th century was probably caused by a coronavirus that still causes the "common cold" in people today (King 2021). Van Ranst states that the COVID-19 virus will follow a similar pattern and become a continuously circulating, or "endemic" virus, joining four other human coronaviruses that infect people with common cold symptoms. "The virus OC43 is still around. It is now responsible for common colds (…). And probably in some elderly people, it can lead to severe illnesses (…). COVID-19 is now the most intensely studied virus ever. These other viruses received far less attention" (King 2021).

Vijgen et al. (2005) showed that at the same time, historical records showed a highly infectious respiratory disease with a high mortality rate affecting cattle herds around the world (see: Crookshank 1897). Today, the same similar disease is known as contagious bovine pleuropneumonia (Vijgen et al. 2005). In the XIX century, the clinical symptoms of CBPP would have been difficult to distinguish from those of BCoV pneumonia. Most industrialised countries mounted massive culling operations between 1870 and 1890 and were able to eradicate the disease by the beginning of the XX century (Storz et al. 1996). According to Vijgen et al. (2005), during the slaughtering of CBPP-affected herds, there was ample opportunity for the culling personnel to come into contact with bovine respiratory secretions. Around the period in which the BCoV interspecies transmission would probably have taken place, a human epidemic ascribed to influenza was spreading worldwide.

The 1889-1890 pandemic probably originated in Central Asia and was characterised by malaise, fever, and pronounced central nervous system symptoms (Vijgen et al. 2005). Indisputable evidence that an influenza virus was the causative agent of this epidemic was never obtained due to the lack of tissue samples from that period (Vijgen et al. 2005). However, post epidemic analysis in 1957 of the influenza antibody pattern in sera of 50 to 100 years old indicated that H2N2 influenza antibodies might have originated from the 1889-1890 pandemic (Mulder and Masurel 1958). According to Vijgen et al. (2005), dating the most recent common ancestor of BCoV and HCoV-OC43 to around 1890 is one argument. Another argument is that central nervous system symptoms were more pronounced during the 1889-1890 epidemic than in other influenza outbreaks (Anonymous 1958). It has been shown that HCoV-OC43 can be neuroinvasive (Arbour et al. 2000).

The work of Brüssow and Brüssow (2021) reported that medical reports from Britain and Germany on patients suffering from the Russian flu share several characteristics with COVID-19. Most notable are multisystem affections comprising respiratory, gastrointestinal and neurological symptoms, including loss of taste and smell perception. In COVID-19 and unlike in influenza, mortality was seen in elderly subjects, while children were only weakly affected (Brüssow and Brüssow 2021).

The Russian flu pandemic claimed the lives of an estimated 1 million humans from a world population of 1.5 billion people and represented thus one of the great epidemics of the 19th century (Valleron et al. 2010). The pandemic spread was extremely rapid, with a starting point at St Petersburg in December 1889 (Valleron et al. 2010id). The UK and Scottish cities were hit only six weeks later. The mean basic reproduction rate was 2.15, and the highest reproduction rates were observed at Stuttgart, St Petersburg, and Amsterdam (Valleron et al. 2010).

The Russian flu was described as influenza because viruses were still unknown at the time. Since the oldest influenza viruses were isolated and kept as laboratory stocks only since the 1930s, direct evidence for linking influenza viruses with the Russian flu is lacking (Brüssow and Brüssow 2021). In contrast, direct virological proof for the attribution of the Spanish flu from 1918 to 1919 to an influenza virus has been achieved by finding pathological samples and corpses of pandemic victims buried in permafrost soils, followed by reviving this pandemic influenza virus in the laboratory (Brüssow and Brüssow 2021).

To address the question of whether the clinical symptoms reported for the Russian flu patients better fit "an influenza virus infection or a trans-species infection h a bovine coronavirus or another infectious agent," Brüssow and Brüssow (2021) used two comprehensive contemporary reports on the Russian flu pandemic from Britain and Germany. According to Parsons (1890), Brüssow and Brüssow (2021) concluded that many observations described in the Parsons report resemble more characteristics of COVID-19 than those of influenza. Notable are light affection in adolescents and age as a risk factor for mortality: "Influenza was a disease especially fatal to elderly persons" (Parsons 1890). "Pulmonary inflammation was the most frequent cause of death and affected the very old and the previously diseased" (Parsons 1890).

Kousoulis and Tsoucalas (2021) also concluded that some characteristics of the 1889 pandemic resemble more coronavirus affection than classical influenza. Further insight is provided by an Encyclopaedia Britannica entry on "Influenza" published in 1911 (Encyclopædia Britannica 1911). According to Encyclopaedia Britannica from 1911, "influenza melancholia is twice as frequent as all other forms of insanity put together. Other common after-effects are weakness or "loss of the special senses, particularly taste and smell" (The German "Verein für Innere Medizin") Report issued in 1892 at Berlin[9] also lists loss of smell and taste.

According to Van Ranst, "incidences like COVID-19 happened all the time, but we did not notice them" - medicine detects viruses more frequently (King 2021). "If some of these outbreaks, like SARS in 2003, happened one hundred

[9]Leyden and Guttmann, 1892: https://collections.nlm.nih.gov/catalog/ nlm:nlmuid-64820270R-bk.

years ago, then it would not have been noticed, and it would be a local outbreak" (King 2021). In the context of the current pandemic, it is surprising that the COVID-19 virus was sequenced so quickly, especially when considering that one of the most common cold viruses, OC43, had not even been sequenced until 2003 by Mark Van Ranst et al. (King 2021).

It is, of course, difficult to formulate a hypothesis for a microbiological aetiology of a pandemic that occurred 133 years ago, at an epoch when viruses were still unknown. But differentiating an influenza virus infection from a COVID-19 patient purely on the clinical ground is a problematic task for a physician today (Brüssow and Brüssow 2021) because the symptoms overlap. As we have already stated, the most important observations of the loss of smell and taste (anosmia and ageusia) were made during the Russian flu pandemic and with COVID-19. Since anosmia and ageusia are now used as relatively reliable clinical diagnostic markers for COVID-19 (Bénézit et al. 2020), one is tempted to attribute this specific symptom seen in the Russian flu pandemic patients more to a coronavirus than to an influenza virus infection.

According to a thesis from Van Ranst (King 2021) and a reformulated hypothesis by Telenti et al. (2021), the world faced 1890 a coronavirus pandemic. Due to the mentioned limitations, it is impossible to have medical evidence. Therefore, we have looked for evidence in history using the method of Digital Humanities and GNV below.

Methods

In our work, we have used the new updated English corpora (2019) to exploit the advantages of improved OCR and better underlying library and publisher metadata. We chose to work on an English (both British and American) corpus, as it is the most extensive database available so far. We have also used both the German (2019) and Russian corpora (2012) for comparison and verification of results.

Our approach is based on analysing the so-called pandemic-related words during history. The objective was to calculate the ratio of increasing to decreasing trends in the changes in frequencies of the words representing the Russian flu symptoms and compare similarities between the development of the Russian flu and COVID-19. If the desired term or set of words is entered in the search engine, for example, the word "epidemic", the graphic display on the Y-axis shows what percentage of words in the selected corpus over the years (X-axis) make up the word.

It is important to emphasise that the smoothing factor "3" we use in the paper shows the average for each year, considering the three previous and three upcoming years. The validity of the data obtained is guaranteed by normalising the data with the number of published books each year (Michel et al. 2011a). As previously mentioned, GNV provides five operators that the researcher can use to combine n-grams: "+, -, /, *, :" With the "wild card", a searcher can ask for information that is not pre-defined by other search keywords. That can lead to an

exploration of hidden patterns (Ophir 2016). The wild card can be applied to the next adjacent word and different patterns. When the researcher puts a "*" in place of a word, the Ngram Viewer display the top ten substitutions. For instance, to find the most popular words following "University of", the researcher should search for "University of *".[10] For our study, this operator is helpful because it shows that the term "loss of smell" is most often mentioned in combination with the term "loss of taste". In addition, we see that both terms are used frequently during the Russian flu.

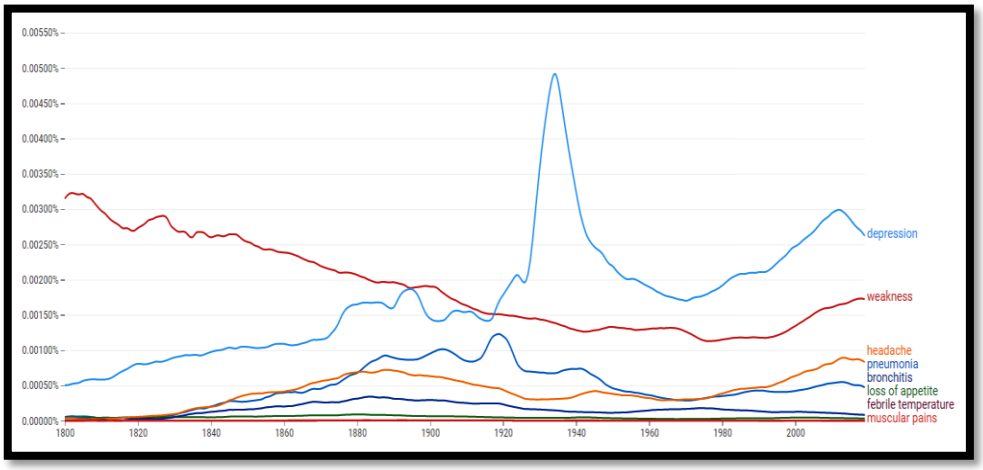
Results and Discussion

In the section Literature review, we have listed some of the sources we discovered using GNV to confirm the thesis that the Russian flu was a coronavirus infection, i.e., that COVID-19 is not a unique phenomenon. In the following, we show how to use NGram concerning pandemics throughout history and lessons for today. The first example (Figure 2) relates to the above symptoms that GNV correctly records, which is the first evidence of the reliability of this approach.

Figure 2 shows the increase in the mention of symptoms "pneumonia; weakness; loss of appetite; headache; bronchitis; febrile temperature; depression and muscular pain" in the English book corpora at the time of the outbreak of the Russian flu (1889-1891). We chose the years 1880, 1890, 1900 and 1910 to show the frequencies of mentioning symptoms in the period before the outbreak of the Russian flu and in the period after. Figure 2 indicates that NGram is a reliable tool for monitoring social trends in the past.

[10]Google NGram View: https://books.google.com/ngrams.

Figure 2. *Frequencies for the Symptoms "Pneumonia; Weakness; Loss of Appetite; Headache; Bronchitis; Febrile Temperature; Depression and Muscular Pain" Mentioned in Newspapers, Magazines and Books from 1800 to 2019 in the English Corpus*



1880	
pneumonia	0.0006702667%
weakness	0.0020688213%
loss of appetite	0.0000884414%
headache	0.0006833671%
bronchitis	0.0003105028%
febrile temperature	0.0000030374%
depression	0.0016435527%
muscular pains	**0.0000065477%**

1890	
pneumonia	0.0008937157%
weakness	0.0019650761%
loss of appetite	0.0000766940%
headache	0.0007065139%
bronchitis	0.0003090883%
febrile temperature	0.0000032282%
depression	0.0015848841%
muscular pains	**0.0000069686%**

1900	
pneumonia	0.0009507938%
weakness	0.0018963951%
loss of appetite	0.0000644783%
headache	0.0006304106%
bronchitis	0.0002929753%
febrile temperature	0.0000024288%
depression	0.0014456849%
muscular pains	**0.0000071599%**

1910	
pneumonia	0.0008439744%
weakness	0.0016346441%
loss of appetite	0.0000587604%
headache	0.0005355317%
bronchitis	0.0002515750%
febrile temperature	0.0000021021%
depression	0.0015407777%
muscular pains	**0.0000065491%**

Source: author's creation based on Google Ngram (http://books.google.com/ngrams).

Figure 3. *Frequencies for the Words "Anosmia" and "Ageusia" from 1800 to 2019 in the English Corpus*

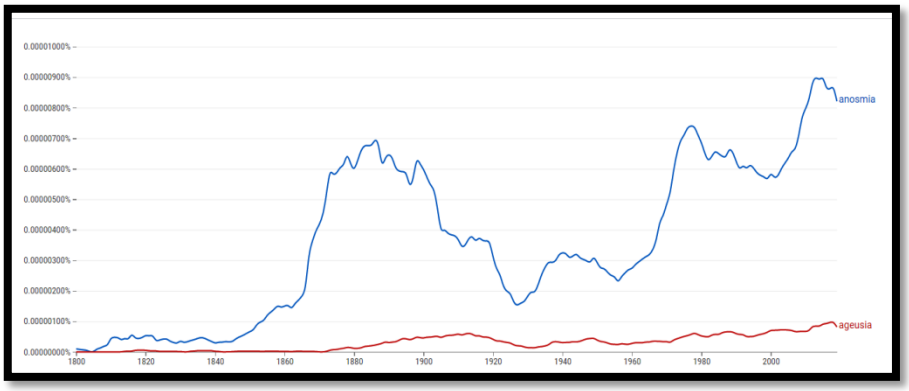


Figure 3 shows the rapid increase in the mention of the term "anosmia" (loss of smell) and "ageusia" (loss of taste) in English book corpora at the time of the outbreak of the Russian flu and immediately after it (1889-1891).

Figure 4. *Frequencies for the Words "Loss of Smell" and "Loss of Taste" from 1700 to 2019 in the English Corpus*

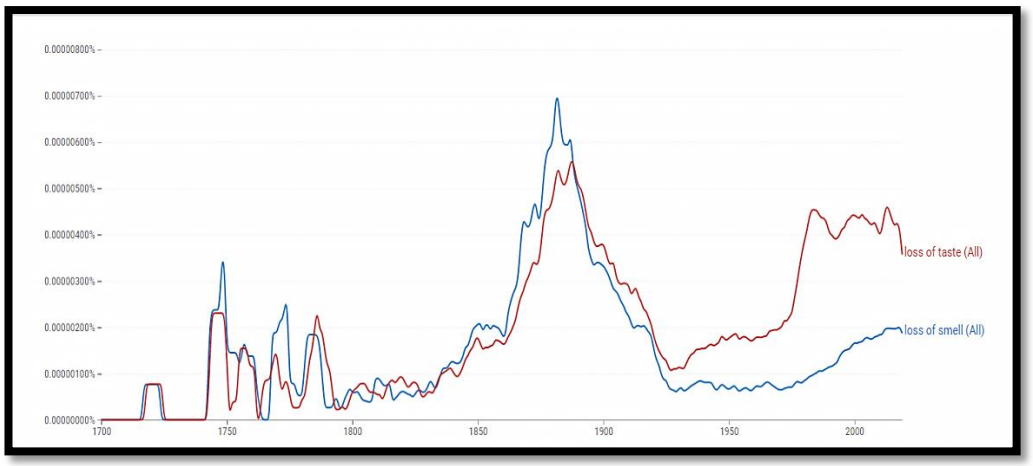
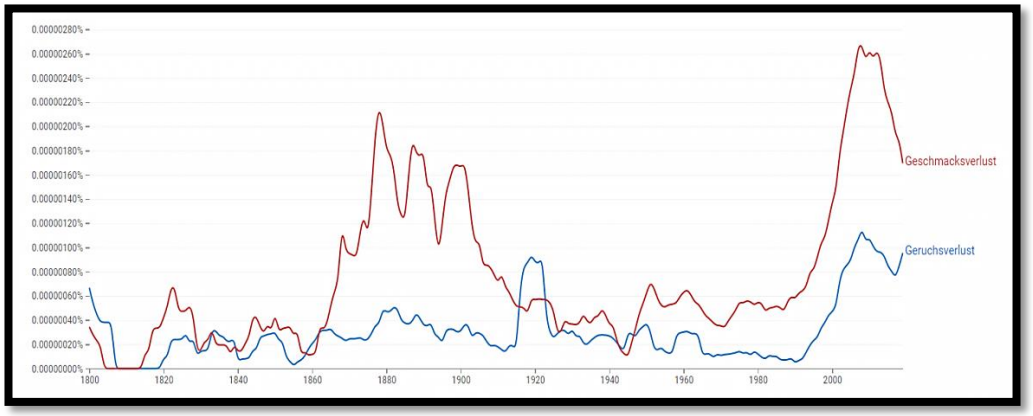


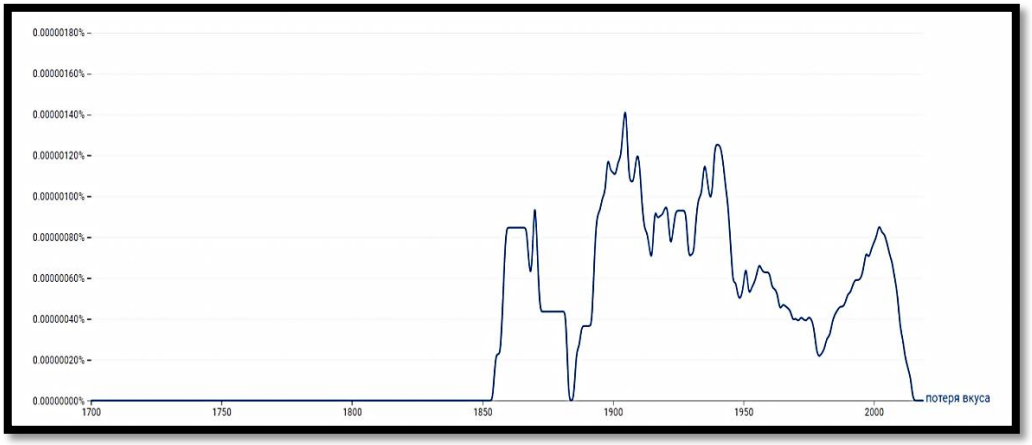
Figure 4 shows the increase in the mention of the term "loss of smell" and "loss of taste" in English book corpora (including newspapers and magazines) at the time of the outbreak of the Russian flu and immediately after it (1889-1891). We can see the same development in German and Russian book corpora (Figures 5 and 6).

Figure 5. *Frequencies for the Words "Geruchsverlust" (Loss of Smell) and "Geschmackverlust" (Loss of Taste) from 1700 to 2019 in the German Corpus*



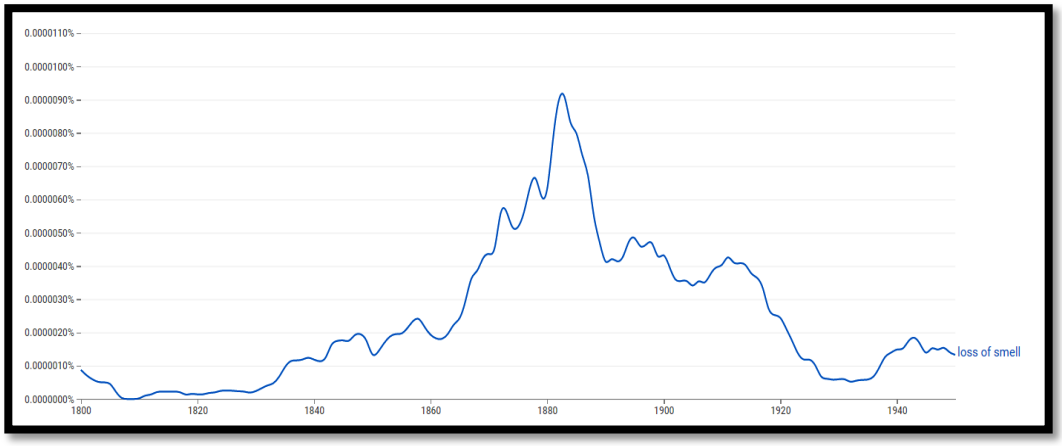
In contrast to the German and English one, the Russian corpus (Figure 6) indicates censorship because the terms quickly disappear from the public space after their sudden appearance, i.e., it is no longer mentioned in newspapers or books. The possibility of censorship is also mentioned in the work of Brüssow (2021).

Figure 6. *Frequencies for the Words "Потеря Вкуса (Loss of Taste) "from 1700 to 2019 in the Russian Corpus*



The English One Million option allows searches that limit books to 6,000 in any given year. Google has made attempts to select books randomly, but at the same time to maintain the subject distributions for each year (University of London n.d.). Figure 7 also shows, in this case, an apparent increase in the use of the term "loss of smell" in books during the Russian flu.

Figure 7. *Frequencies for the Term "Loss of Smell" from 1800 to 2019 in the Corpus English One Million*



Further similarities between Russian flu and COVID-19 are that COVID-19 has, as mentioned, its main fatality in the elderly; this was also noted for the Russian flu pandemic (Rozen 2020). While the peak mortality in the Russian flu pandemic was among the elderly, substantial mortality was also seen in adults with comorbidity, but children suffered only mild symptoms similar to the current COVID-19 pandemic (Rozen 2020).

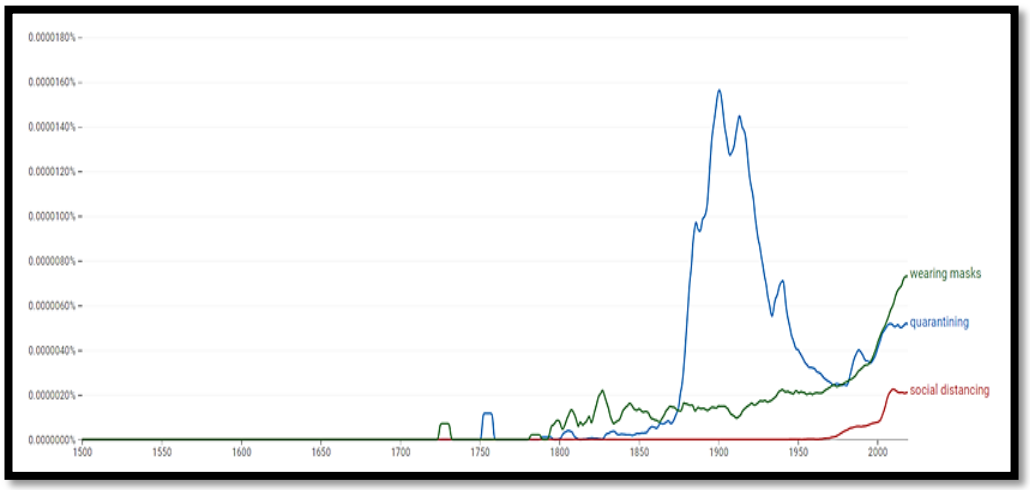
In our study, by applying NGram, we also evaluated historical reports from newspapers and scientific and medical journals. GNV recorded more than 600 news articles about the Russian flu from 42 newspapers (Paris - Le Temps, Le

Matin, Berlin - Vossische Zeitung, London - The Times, and many Austrian newspapers and medical journals such as The Lancet). The high attack rate of Russian flu can be read in the newspapers that reported the closure of schools, universities and factories because a large part of the staff fell ill. Reports quoted by the newspapers noted that mortality rates had increased by 30% compared to the same period of the pre-pandemic year.

The past pandemic has elements relevant to the COVID-19 pandemic, showing the measures that we undertake today and the same as they did in 1918 – social distancing, wearing masks, quarantining, and travel restrictions (King 2021). But just as individuals forget about the past, so do societies (Halbwachs 1992). Studying past pandemics shows that the pandemic stops on its own. According to mentioned historical records, a pandemic's 'natural' length is two to five years (Spinney 2018). In the absence of treatments and a vaccine, both the Russian and the Spanish flu ran and stopped after two to three years. The wearing of masks was during the Spanish flu understood to be of significant importance in preventing infection (Martin et al. 2007). However, "herd immunity" was not necessary to stop the pandemic (Brian 2021).

Despite the similarities, several differences distinguish the COVID-19 situation from the Russian flu. In contrast to its widespread use during the Spanish flu pandemic of 1918, face masks were not used during the Russian flu pandemic (Spinney 2018).

Figure 8. *Frequencies for the Terms "Quarantining", "Social Distancing", and "Wearing Masks" from 1500 to 2019 in the English Corpus*



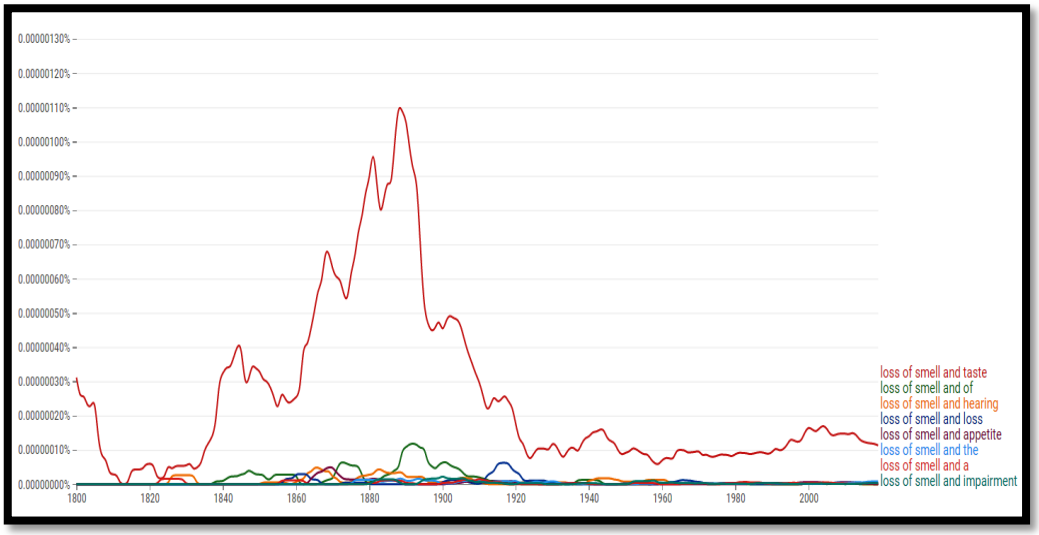
NGram (Figure 8) shows us evidence that during the Russian flu wearing masks was less used than during the period of Spanish flu. This is another proof that NGram correctly records social trends. The term "social distancing" is a newer word coin, so it is not surprising that it was not mentioned in the 19th century, while in the case of the term "quarantining", we see that this term was intensive mentioned in the middle of the XVIII century (bubonic plague between 1738 and 1740) and that it is intensively mentioned during both the Russian and the Spanish flu. Croatia first introduced quarantine, i.e., Dubrovnik, in the middle

of the 14th century.[11] However, since the printing press was invented in the middle of the 15th century, such a record cannot be registered by the NGram (this should be borne in mind in the case of many other discoveries and historical events).

Public health measures during the 1889 pandemic consisted mainly of school closures and hygiene advice (handwashing) that GNV also records. Intensive care medicine was 1889 practically non-existent, and the best medical advice of the time was early bedrest and antipyretics (Brüssow 2021).

Figure 9 below shows the benefit of the operator "*" application that enables function: most often mentioned followed words. We can see that the most frequently followed words for the phrase "loss of smell" is "loss of taste", which indicates similarities between the Russian flu and COVID-19.

Figure 9. *Frequencies for the Words "Loss of Smell and *" from 1800 to 2019 in the English Corpus Showing Most Often Mentioned Followed Words Using the Operator "*"*



Note: Below the graph, GNV shows year ranges for query terms, and by clicking on those, the query is directly submitted to Google Books. It is important to note here that one can choose between newspapers, magazines and books.

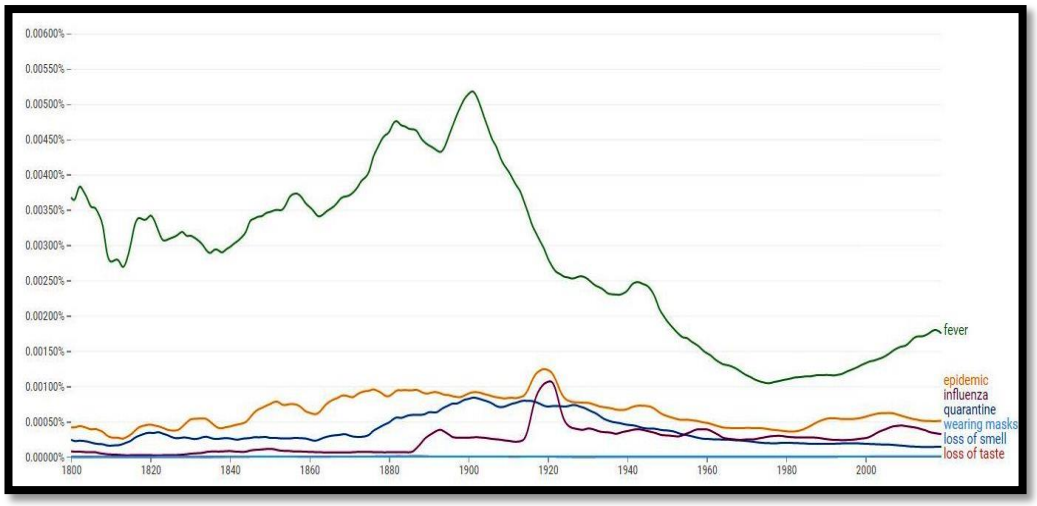
Table 1 in the following presents numeric frequencies for the pandemic-related words "fever", "epidemic", "influenza", "quarantine", "wearing masks", "loss of smell", "loss of taste" from 1800 to 2019 in the English corpus (in %) and the Figure 10 shows GNV display for the frequencies.

[11] *Opća i nacionalna enciklopedija*, Zagreb 2006.

Table 1. *Numeric Frequencies for the Words "Fever", "Epidemic", "Influenza", "Quarantine", "Wearing Masks", "Loss of Smell", "Loss of Taste" from 1800 to 2019 in the English Corpus (in %)*

Russian flu	**1880**	**1889**	**1890**	**1891**	**1900**
loss of smell	0.0000049357	0.0000043904	0.0000042023	0.0000040161	0.0000028211
loss of taste	0.0000040433	0.0000047123	0.0000044648	0.0000043490	0.0000033861
fever	0.0045785303	0.0044563019	0.0044123274	0.0043885018	0.0051403633
epidemic	0.0008526997	0.0008944025	0.0008965982	0.0009177670	0.0008983375
quarantine	0.0004794893	0.0005985671	0.0006383930	0.0006295467	0.0008166632
influenza	0.0000663529	0.0002639855	0.0002983151	0.0003359815	0.0002702021
wearing masks	0.0000015962	0.0000013616	0.0000013950	0.0000013922	0.0000015227

Figure 10. *GNV display - Frequencies for the Words "Fever", "Epidemic", "Influenza", "Quarantine", "Wearing Masks", "Loss of Smell", "Loss of Taste" from 1800 to 2019 in the English Corpus*



We can see that the frequency of the words "loss of smell" and "loss of taste" rapidly increased during the Russian flu and that the mention of this symptom fell sharply after the pandemic stopped.

Figure 11 shows that in the case of symptom "loss of taste," the frequency rose from 0.0000040433 % in 1880 to 0.0000047123 % in 1889 and the mention of this symptom fell sharply after the pandemic stopped in 1900 (0.0000033861%). In the case of symptom "loss of smell," the frequency decreased from 0.0000043904% in 1889 to 0.0000028211% in 1900.

Figure 11. *Frequencies for the Words "Fever", "Epidemic", "Influenza", "Quarantine", "Wearing Masks", "Loss of Smell", "Loss of Taste" in 1880, 1889, 1890, 1891 and 1900 in the English Corpus*

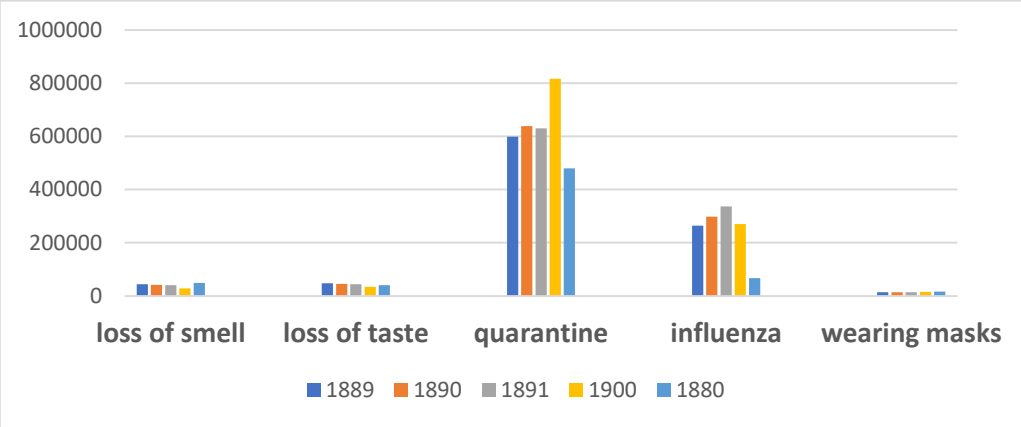
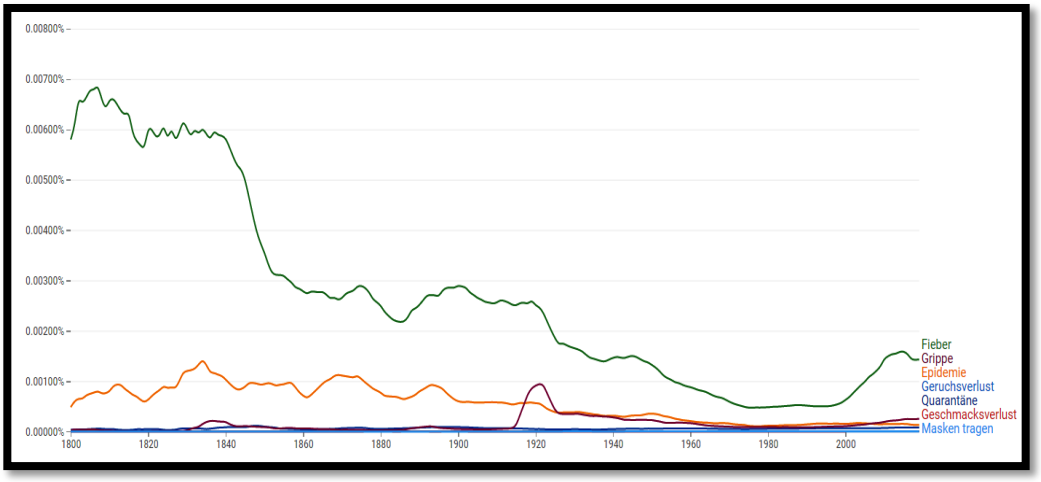


Table 2 presents numeric frequencies for the pandemic-related words from 1800 to 2019 in the German corpus (in %) and the Figure 12 shows GNV display for the frequencies.

Table 2. *Frequencies for the Words "Fieber (Fever)", "Epidemie (Epidemic)", "Grippe (Influenza)", "Quarantäne (Quarantine)", "Masken tragen" (Wearing Masks), "Geruchsverlust (Loss of Smell)", "Geschmacksverlust (Loss of Taste)" from 1800 to 2019 in the German Corpus (in %)*

Russische Grippe	**1889**	**1890**	**1891**	**1900**	**1880**
Geruchsverlust	0.0000004145	0.0000003559	0.0000003501	0.0000003071	0.0000003501
Geschmacksverlust	0.0000017517	0.0000018015	0.0000014463	0.0000016600	0.0000014463
Fieber	0.0024394190	0.0025105012	0.0026102443	0.0029010263	0.0026102443
Epidemie	0.0007245716	0.0008011005	0.0008347561	0.0005970067	0.0008347561
Quarantäne	0.0000705233	0.0000763311	0.0000766396	0.0000863325	0.0000766396
Grippe	0.0000711845	0.0000807742	0.0000877451	0.0000481621	0.0000877451
Masken tragen	0.0000006314	0.0000004534	0.0000003115	0.0000012416	0.0000003115

Figure 12. *Frequencies for the Words "Fieber (Fever)", "Epidemie (Epidemic)", "Grippe (Influenza)", "Quarantäne (Quarantine)", "Masken tragen" (Wearing Masks), "Geruchsverlust (Loss of Smell)", "Geschmacksverlust (Loss of Taste)" from 1800 to 2019 in the German Corpus*



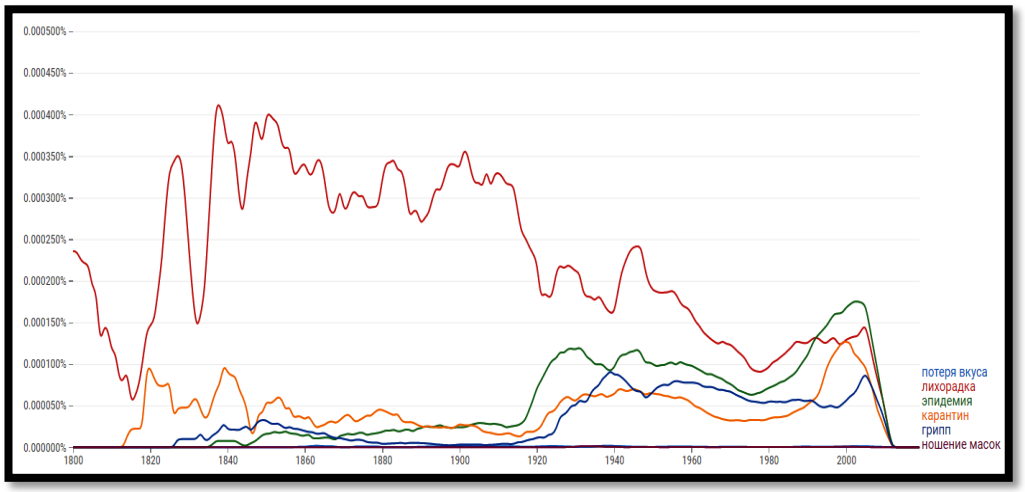
In the German corpus the frequency for "Geschmacksverlust" (loss of taste) rose from 0.0000014463% in 1880 to 0.0000018015% in 1889 and decreased rapidly after the pandemic (1900 = 0.0000016600%). The most rapid change in the German corpus between the years 1890 and 1900 can be noted at the term "Epidemie" (epidemic) (1890 = 0.0008011005%; 1900 = 0.0005970067%).

Table 3 presents numeric frequencies for the pandemic-related words from 1800 to 2012 in the Russian corpus (in %) and the Figure 13 shows GNV display for the frequencies.

Table 3. *Frequencies for the Words "лихорадка" (Fiver), "эпидемия" (Epidemic), "грипп" (Influenza), "карантин" (Quarantine), "ношение масок" (Wearing Masks), "Потеря обоняния" (Loss of Smell), "потеря вкуса" (Loss of Taste) from 1800 to 2012 in the Russian Corpus (in %)*

Russian flu	1889	1890	1891	1900	1880
Потеря обоняния	0.0000005041	0.0000005041	0.0000005041	0.0000001579	0.000000000
потеря вкуса	0.0000004682	0.0000004682	0.0000006787	0.0000011834	0.000000000
лихорадка	0.003102872	0.003102872	0.0002993711	0.0003607911	0.0003471586
эпидемия	0.0000247684	0.0000247684	0.0000278790	0.0000251270	0.0000191910
карантин	0.0000277171	0.0000277171	0.0000254140	0.0000297613	0.0000473150
грипп	0.0000065101	0.0000065101	0.0000057692	0.0000044993	0.000052766
ношение масок	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

Figure 13. *Frequencies for the words "лихорадка" (fiver), "эпидемия" (Epidemic), "грипп" (Influenza), "карантин" (Quarantine), "ношение масок" (Wearing Masks), "Потеря обоняния" (Loss of Smell), "потеря вкуса" (Loss of Taste) from 1800 to 2012 in the Russian Corpus*



In the Russian corpus the frequency for "loss of taste" rose from 0% in 1880 to 0.0000004682% in 1889 and decreased rapidly after the pandemic (1900 = 0.0000011834%). The frequency for "loss of smell" rose from 0.000000000% in

1880 to 0.0000005041% in 1889 and decreased rapidly after the pandemic (1900 = 0.0000001579%).

Table 4. *Comparison of the Symptom "Loss of Taste" in the English, German and Russian Book Corpus (in %)*

loss of taste	1880 (decrease)	1889 (increase)	1890 (increase)	1891 (increase)	1900 (decrease)
English corpus	0.0000040433 %	0.0000047123	0.0000044648	0.0000043490	0.0000033861
German corpus	0.0000014463	0.0000017517	0.0000018015	0.0000014463	0.0000016600
Russian corpus	0.000000000	0.0000004682	0.0000004682	0.0000006787	0.0000011834

The comparison presented in Table 4. clearly shows that all the three corpora (English, German and Russian corpus) we used for the analysis show that the symptoms of "loss of taste" before and after the outbreak of the Russian flu pandemic were mentioned in the literature, newspapers and magazines to a much lesser extent then it was during the pandemic. The same is noticeable in almost all other symptoms and social trends.

The frequency of the words "fever", "epidemic", "influenza", "quarantine", "wearing masks", "loss of smell", "loss of taste" increased rapidly during the Russian flu from 1899 to 1891, which is especially noticeable in the German and Russian book corpus. In the case of symptom "loss of taste" in the English corpus, the frequency rose from 0.0000040433% in 1880 to 0.0000047123% in 1889. One cannot but notice that the mention of this symptom fell sharply after the pandemic stopped in 1900 (0.0000033861%).

In the Russian corpus, the frequency rose from 0.000000000% in 1880 to 0.0000004682% in 1889 and decreased rapidly after the pandemic (1900 = 0.0000011834%). In the German corpus the frequency rose from 0.0000014463% in 1880 to 0.0000018015 % in 1889 and decreased rapidly after the pandemic (1900 = 0.0000016600%). These results prove our thesis that GNV is a reliable tool for monitoring social trends during pandemics and a very useful window into history.

Of the other social trends we have analysed using GNV, we would highlight the terms: "economic crisis", "unemployment", and "hunger". None of these terms shows a significant frequency deviation compared to the period immediately before and after the epidemic. From the historical point of view (GNV as the window of history), we conclude that a significant crisis does not need to occur after the COVID-19 pandemic. Judging by the collective memory of humanity and the insights we have gained using GNV, the virus will undoubtedly weaken over time. The results of GNV show that the pandemic in this decade will turn into an endemic or common cold and will stay with us like other types of flu.

Limitations

The possibilities and limitations of using the GNV for research have been controversially discussed (Younes and Reips 2019). Although many Google Ngram studies indicate scientific recognition, several papers address methodological issues (Gooding 2012). The data set from GNV has been criticised for its reliance on inaccurate OCR, an overabundance of scientific literature, and large numbers of incorrectly dated and categorised texts (Pechenick 2015). Because of these errors, and because it is uncontrolled for bias, according to Zhang (2017), it is risky to use this corpus to study language or test theories. Since the data set does not include metadata, it may not reflect the general linguistic or cultural change and can only hint at its effect (Younes and Reips 2019).

The main points of criticism relate to insufficient OCR, particularly concerning semantic scanning errors (which can affect words such as fail and sail due to similarities in the letters "f" and "s") (Pechenick et al. 2015) and messy metadata that may lead to the display of word frequencies in the wrong or unrelated time intervals (Gooding 2012). The last criticism is that the percentage considers published manuscripts regardless of their importance (Kratzer 2019).

Hilpert and Gries (2009) warn that a statistical measure that would help determine if the observed frequencies differ from the mean more than expected should be incorporated in more complex studies. Mayer-Schönberger and Cukier express concern about machines replacing human activities and decision-making (see: Younes and Reips 2019). Boyd and Crawford also raise critical questions about big data: "Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? (…) The era of Big Data has only just begun, but it is already important that we start questioning the assumptions, values, and biases of this new wave of research" (Boyd and Crawford 2012).

Several authors have problematised the GNV corpus and raised doubts about its representation of natural language and its development over time (Pechenick 2015). Chumtong and Kaldewey (2017) highlight that what makes the GNV a valuable research tool is not primarily its accuracy but rather its potential for "quick-and-dirty heuristic analysis". Davis (2014) recognised the dataset as remarkable but perceived the interface too simplistic. He claimed it did not allow for collocations in searches, searching by wildcards and meaningful use of parts of speech.

It also appears that GNV does not consider the different contexts in which the analysed words are set in, and contexts carry the meaning the cause of which we are unable to determine. The fact that the frequency of a word rises does not necessarily mean that the concept is valued more but that it is discussed extensively (Zięba 2018). The GNV enables viewing the excerpts from which the analysed words come; however, as collecting such data has not been automated yet, and would have to be done manually for all words in millions of contexts, it seems implausible to incorporate such information into the study, even if for reasons of time and space (Zięba 2018). Needless to say, either an individual or a larger team cannot study any of the corps manually.

According to Zięba (2018), the usage of GNV should be limited to uncomplicated studies related to word frequency. It cannot be treated as the only tool in researching complex socio-cultural transformations. However, with careful analysis of the results, the GNV does potentially improve our understanding of cultural and linguistic trends over time. With Google making its datasets available, more complex text mining tools can study the ever-growing corpus (University of London n.d.). Compared to the 2009 versions, the 2012 and 2019 versions have more books, improved OCR, improved library and publisher metadata.[12] According to Zięba (2018), even if we consider the imperfections of OCR, GNV still seems to put socio-cultural research in a context whose significance is hard to question, especially if carried out cautiously and conscientiously.

Lakoff agrees that even though the presence of most words and the changes in their frequency does not tell much about the values ascribed to certain phenomena, it may be a sign of recognition of a problem (Lakoff 2013). Younes and Reips (2019) propose how to address these concerns by introducing several methodological procedures such as cross-validations via the examination of different language corpora, the use of word inflexions and synonyms, as well as the use of a newly-developed standardisation procedure that all aim at increasing the reliability of GNV studies.

According to Solovyev et al. (2020), there are several ways to make the GBN corpus results more reliable. On the one hand, it is impossible to correct all its errors, and on the other, perfectionism should be avoided in this field since no one knows what an ideal corpus would be like (Solovyev et al. 2020). The first one is to use all possible support data extracted from the corpus and use synonyms (Younes and Reips 2019). Younes recommends studying each word and its three synonyms selected from the relevant dictionaries of synonyms (Younes and Reips 2019). Sometimes it is pertinent to perform comparative studies and see how the same or close meaning terms are used in different corpora presented in GNV (Solovyev et al. 2020). The second way to enhance the results is to pre-process the GNV raw data. Solovyev et al. show that the GNV corpus can be regarded as representative for the following reasons. It is the most extensive corpus ever existed, including texts of various types and genres written by people of different ages, sex and diverse backgrounds. Such diverse texts, their length and size, serve as a solid empirical foundation for linguistic and related studies (Solovyev et al. 2020).

Conclusions

This paper showed that the Google Ngram (GNV) can give us useful insights into the history of pandemics and that the tools of Digital Humanities can discover hidden patterns in history. With the help of GNV, we have analysed the epidemiological literature on the Russian flu pandemic development for hints on how the COVID-19 might develop in the following years. We showed indications

[12]Google NGram View: https://books.google.com/ngrams.

that the COVID-19 is not a unique phenomenon because the Russian flu might be a coronavirus infection. This thesis still cannot be confirmed, requiring further historical and medical research.

According to our study, the GNV clearly shows the influence that social changes have on word frequency. The most important observation of similarities between the Russian flu pandemic and COVID-19 is the loss of smell and taste (anosmia and ageusia). The frequency of the words "fever", "epidemic", "influenza", "quarantine", "wearing masks", "loss of smell", "loss of taste" increased rapidly during the Russian flu from 1899 to 1891, which is especially noticeable in the German and Russian book corpus. The mention of symptoms and the pandemic-related words fell sharply after the pandemic stopped.

Other social trends we have analysed using GNV "economic crisis", "unemployment", and "hunger" do not show a significant deviation in frequencies compared to the period immediately before and after the epidemic. We conclude that a historical perspective shows that a substantial crisis does not need to occur after the COVID-19 pandemic. Judging by the collective memory of humanity and the insights we have gained using GNV, the virus will undoubtedly weaken over time. The results of GNV show that the pandemic in this decade will turn into an endemic or common cold and will stay with us like other types of flu.

These results prove our thesis that GNV is a reliable tool for monitoring social trends during pandemics and a very useful window into history. This study has also shown how to overcome the binderies between the social sciences and the humanities. The results of this study open a discussion on the usefulness of the GNV insights possibilities into past socio-cultural development, i.e., epidemics and pandemics that can serve as lessons for today. We have shown hidden patterns of conceptual trends in history and their relationships with current development in the case of the pandemic COVID-19. Despite the numerous indications we have demonstrated, we are aware that the hypothesis still cannot be confirmed and that it is necessary to require further historical and medical research. The main challenge was to correctly interpret patterns discovered by digital analysis and discern correlations, causes and relations between historical events and current development.

The benefit of this method could help complement historical medical records, which are often woefully incomplete. However, this method has serious limitations and can be useful only under cautious handling and testing. Despite its limitations, the GNV research based on an over 500 billion word corpus is prone to produce valuable results when approached with great care and consideration according to the restrictions brought by this method and will certainly find application in many research areas in humanities and social sciences in future.

References

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The expression of emotions in 20th century books. *PLoS ONE* 8(3): e59030.

Anonymous (1958) Influenza 1889 and 1957. *Lanceti*: 833–835 (cited in Vijgen et al. 2005).

Arbour N, Day R, Newcombe J, Talbot PJ (2000) Neuroinvasion by human respiratory coronaviruses. *Journal of Virology* 74(19): 8913–8921.

Bénézit F, Le Turnier P, Declerck C, Paillé C, Revest M, Dubée V, et al. (2020) Utility of hyposmia and hypogeusia for the diagnosis of COVID-19. *The Lancet. Infectious Diseases* 20(9): 1014–1015.

Berry DM (2012) The social epistemologies of software. *A Journal of Knowledge, Culture and Policy* 26(3–4): 379–398.

Boyd D, Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.

Brian G (2021) *COVID-19 update: knowledge is power, but compassion is lacking*. Retrieved from: https://www.myeloma.org/blog/covid-19-update-knowledge-power-compassion-lacking. [Accessed 20 January 2022]

Brüssow H (2021) What we can learn from the dynamics of the 1889 'Russian flu' pandemic for the future trajectory of COVID-19. *Microbial Biotechnologie* 14(6): 2244–2253.

Brüssow H, Brüssow L (2021) Clinical evidence that the pandemic from 1889 to 1891 commonly called the Russian flu might have been an earlier coronavirus pandemic. *Microbial Biotechnologie* 14(5): 1860–1870.

Burdick A, Drucker J, Lunenfeld P, Presner T, Jeffrey S (2012) *Digital_Humanities*. The MIT Press.

CDC (n.d.) *Similarities and differences between flu and COVID-19*. Retrieved from: https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm. [Accessed 23 September 2021]

Chumtong J, Kaldewey D (2017) *Beyond the Google Ngram Viewer: bibliographic databases and journal archives as tools for the quantitative analysis of scientific and meta-scientific concepts*. FIW Working Paper 08. Bonn.

Crookshank EM (1897) Infectious pleuro-pneumonia. In EM Crookshank (ed.), *A Textbook of Bacteriology Including the Etiology and Prevention of Infective Diseases*, 239–248. Philadelphia: W. B. Saunders.

Davis M (2014) Making Google books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics* 19(3): 401–16.

Parsons (1890) *Report on the influenza epidemic of 1889-90 - Great Britain*. Local Government Board, Henry Franklin Parsons – (Google Books).

Gooding P (2012) Mass digitisation and the garbage dump: the conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing* 28(3): 425–431.

Greenfield PM (2013) The changing psychology of culture from 1800 through 2000. *Psychological Science* 24(9).

Grossmann I, Varnum ME (2015) Social structure, infectious diseases, disasters, secularism, and cultural change in America. *Psychological Science* 26(3): 311–324.

Halbwachs M (1992) *On collective memory*. Translated by LA Coser. Chicago: University of Chicago.

Harari YN (2014) *(Sapiens) A brief history of humankind*. London.

Hilpert M, Gries S (2009) Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24(4): 385–401.

Jurić T (2021a) Medical brain drain from South-eastern Europe: using digital demography to forecast health worker emigration. *Journal of Medical Internet Research* (Nov).

Jurić T (2021b) Google trends as a method to predict new COVID-19 cases and socio-psychological consequences of the pandemic. *Athens Journal of Mediterranean Studies* 8(1): 67–92.

Karch M (2021) *How to use the Ngram viewer tool in Google books*. Retrieved from: https://www.lifewire.com/google-books-ngram-viewer-1616701. [Accessed 23 December 2021]

Kardaš L (2020) *Uporaba Google Ngrama u društvenim znanostima*. (Using Google Ngram in the social sciences). Master Thesis. Zagreb: Hrvatsko katoličko sveučilište.

Kesebir S, Kesebir P (2017) A growing disconnection from nature is evident in cultural products. *Perspectives on Psychological Science* 12(2): 258–269.

King A (2021) *Why history suggests COVID-19 is here to stay*. Retrieved from: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/qa-why-history-suggests-covid-19-here-stay. [Accessed 20 December 2021]

Kousoulis AA, Tsoucalas G (2017) Infection, contagion and causality in Colonial Britain: the 1889-90 influenza pandemic. *Le Infezioni in Medicina* 25(3): 285–291.

Kratzer G (2019) *Google Ngram*. Retrieved from: https://gilleskratzer.netlify.app/post/ngram/. [Accessed 23 December 2021]

Kucharski A (2020) *The rules of contagion: why things spread – And why they stop*. 1st Edition. Basic Books.

Lakoff R (2013) *What words don't tell us*. Retrieved from: http://blogs.berkeley.edu/author/rlakoff/. [Accessed 23 December 2021]

Latour B (2014) Rematerializing humanities thanks to digital traces. In *Digital Humanities 2014 - Opening Night Sciences Paris*. Retrieved from: https://www.youtube.com/watch?v=4L2zRoKS0IA&ab_channel=UNILUniversit%C3%A9deLausanne. [Accessed 23 December 2021]

Lieberson S, Horwich J (2008) Implication analysis: a pragmatic proposal for linking theory and data in the social sciences. *Sociological Methodology* 38(1): 1–50.

Lin Y, Michel J-B, Lieberman Aiden E, Orwant J, Brockman W, Petrov S (2012) Syntactic annotations for the Google books Ngram corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 169–174. Jeju, Republic of Korea, 8-14 July 2012. Association for Computational Linguistics.

Martin MCJ, Bootsma N (2007) The effect of public health measures on the 1918 influenza pandemic in U.S. cities 2007. *Proceedings of the National Academy of Sciences* 104(18):7588–7593.

Michalski B, Krishnamoorthy M, Lau TY (2012) *Temporal analysis of literary and programming prose*. Retrieved from: https://bit.ly/37K63Wj. [Accessed 23 December 2021]

Michel JB, Shen YK, Presser Aiden A, Veres A, Gray MK, Brockman W, et al. (2011a) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182.

Michel JB, Shen YK, Presser Aiden A, Veres A, Gray MK, Brockman W, et al. (2011b). *Supporting online material for quantitative analysis of culture using millions of digitized books*. Available at: www.sciencemag.org/cgi/content/full/science.1199644/DC1.

Mohammad SM (2012) From once upon a time to happily ever after: tracking emotions in mail and books. *Decision Support Systems* 53(4): 730–741.

Mooijman M, Meindl P, Oyserman D, Monterosso J, Dehghani M, Doris JM, et al. (2018) Resisting temptation for the good of the group: binding moral values and the moralisation of self-control. *Journal of Personality and Social Psychology* 115(3): 585–599.

Mulder J, Masurel N (1958) Pre-epidemic antibody against 1957 strain of Asiatic influenza in serum of older people living in The Netherlands. *The Lancet* 1(7025): 810–814.

Newberry MG, Ahern CA, Clark R, Plotkin JB (2017) Detecting evolutionary forces in language change. *Nature* 551(Nov): 223–226.

Ophir S (2016) Big data for the humanities using Google Ngrams: discovering hidden patterns of conceptual trends. *First Monday* 21(7): 7–4.

Oza P (2020) Digital humanities. In GP Japee, P Oza (eds.), *Multidimensionality of the Concept & Function of Digital Publisher*. Apple Books.

Pechenick EA, Danforth CM, Dodds PS, Barrat A (2015) Characterising the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10(10): e0137041.

Potter CW (2001) A history of influenza. *Journal of Applied Microbiology* 91(4): 572–579.

Roivainen E (2014) Changes in word usage frequency may hamper intergenerational comparisons of vocabulary skills: An Ngram analysis of wordsum, WAIS, and WISC test items. *Journal of Psychoeducational Assessment* 32(1): 83–87.

Roivainen E (2015) Personality adjectives in twitter tweets and in the Google books corpus. An analysis of the facet structure of the openness factor of personality. *Current Psychology* 34(4): 621–625.

Rojas Castro A (2017) *Big data in the digital humanities. New conversations in the global academic context*. AC/E Digital Culture 2017 Annual Report, 62–71.

Rossi E, Mortimer J, Rossi K (2013) Therapeutic hypnosis, psychotherapy, and the digital humanities: the narratives and culturomics of hypnosis, 1800–2008. *American Journal of Clinical Hypnosis* 55(4): 343–359.

Rozen TD (2020) Daily persistent headache after a viral illness during a worldwide pandemic may not be a new occurrence: lessons from the 1890 Russian/Asiatic flu. *Cephalalgia* 40(13): 1406–1409.

Rutten BPF, Hammels C, Geschwind N, Menne-Lothmann C, Pishva E, Schruers K, et al. (2013) Resilience in mental health: linking psychological and neurobiological perspectives. *Acta Psychiatrica Scandinavica* 128(1): 3–20.

Sisley R (1891) The epidemic of 1889-1890. Bokhara. St. Petersburgh. Berlin. In R Sisley (ed.), *Epidemic Influenza: Notes on its Origin and Method of Spread*, 47–53. London, United Kingdom: Longmans, Green, and Co.

Solovyev VD, Bochkarev VV, Akhtyamova SS (2020) Google Books Ngram: problems of representativeness and data reliability. In A Elizarov, B Novikov, S Stupnikov (eds.), *Data Analytics and Management in Data Intensive Domains*. *Communications in Computer and Information Science*, volume 1223. Cham: Springer.

Spinney L (2017) *The Spanish flu of 1918 and how it changed the world*. 1st Edition. Public Affairs.

Storz J, Stine L, Liem A, Anderso GA (1996) Coronavirus isolation from nasal swab samples of cattle with signs of respiratory tract disease after shipping. *Journal of the American Veterinary Medical Association* 208(9): 1452–1456.

Telenti A, Arvin A, Corey L, Corti D, Diamond MS, García-Sastre A, et al. (2021) After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* 596(Jul): 495–504.

Twenge JM, Campbell WK, Gentile B (2012b) Male and female pronoun use in US books reflects women's status, 1900–2008. *Sex Roles* 67(9–10): 488–493.

University of London (n.d.) *An introduction to text mining*. Retrieved from: https://port.sas.ac.uk/mod/book/view.php?id=554&chapterid=331. [Accessed 23 December 2021]

Valleron AJ, Cori A, Valtat S, Meurisse S, Carrat F, Boëlle PY (2010) Transmissibility and geographic spread of the 1889 influenza pandemic. *Proceedings of the National Academy of Sciences of the United States of America* 107(19): 8778–8781.

Vijgen L, Keyaerts E, Moes E, Thoelen I, Wollants E, Lemey P, et al. (2005) Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *Journal of Virology* 79(3): 1595–1604.

Virues-Ortega J, Pear JJ (2015) A history of "behavior" and "mind": use of behavioral and cognitive terms in the 20th century. *The Psychological Record* 65(1): 23–30.

Ward JS, Barker A (2013) *Undefined by data: a survey of big data definitions*. arXiv.

Younes N, Reips U-D (2019) Guideline for improving the reliability of Google Ngram studies: evidence from religious terms. *PLoS ONE* 14(3): e0213554.

Zhang S (2017) *The pitfalls of using Google Ngram to study language*. WIRED.

Zięba A (2018) Google Books Ngram viewer in socio-cultural research. *Research in Language* 16(3): 357–376.