MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-437

Analysis of Historical Medical Phenomena Using Large N-Gram Corpora

Zdenko Kasáč[a, b], Stefan Schulz[a]

[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria,*
[b] *Faculty of Medicine, Masaryk University, Brno, Czech Republic*

Abstract

Historically, numerous indirect references to real world phenomena have been conserved in literature. High-quality libraries of digitized books and their derivatives (like the Google NGram Viewer) have proliferated. These tools simplify the visualization of trends in phrase usage within the collective memory of language groups. A straightforward interpretation of these frequency changes is, however, too simplistic to draw conclusions about the underlying reality because it is affected by several sources of bias. Although these resources have been studied in social sciences and psychology, there is still lack of user-friendly, yet rigorous methods for analysis of phenomena relevant for medicine. We present a methodological framework to study relationships of observable phenomena quantitatively over periods, which span over centuries. We discuss its suitability for knowledge extraction from current and future large-scale, book-derived, n-gram collections.

Keywords:

Historiography; Semantics; Publications

Introduction

It is a common in science for the phenomena of interest to be observed by proxy. Appropriate tools and methods are needed whenever direct measurement or inspection are unavailable or too complicated. A well-known example in the field of climatology is the analysis of ice core records for the estimation of atmospheric carbon dioxide [1]. Measuring the proportion of air gases trapped in polar ice allowed approximating parameters of atmospheric conditions along the last 800,000 years. Such information has redefined our understanding of the Earth's climate. In this case, nature did the scientist's job of archiving samples of air for hundreds of thousands of years, until our species developed and created a method to analyse the precious archived sample bank.

We seek to develop an analogous method for studying relevant *phenomena* for medicine and health. The described method would be also easy to exploit in other fields, e.g. in social sciences. A *phenomenon* might be anything (e.g. event, thing, situation, idea or sensation) that might be observed. On one hand, using the current epidemiological tools to research past phenomena would introduce recall bias and hide many environmental exposures because one simply forgets what happened many years ago. In other words, asking individuals about their expositions much earlier in their lives just would not work. For similar reasons, asking the respondents about the risk factor exposure of their great-grandparents would not work at all. It is therefore hard to analyse long-running health-related phenomena of the past directly. On the other hand, the extraction of retrospective data from medical records is overly limited, because the current form of medical documentation is recent, having started approaching its current form only in the run of the last century [2]. In addition, medical language has constantly evolved, as did the concepts in physiology and pathology. Many aspects of health today considered important, and thus relevant for research, were granted low priority in earlier decades. Diagnostic criteria frequently change, and diagnostic tools evolve rapidly, which further complicates the evaluation of long-term epidemiological data. Finally, the availability of medical care constitutes an additional bias whenever studying medical records.

Therefore, to approximate and plot the temporal change in presence of a phenomenon, a source of coherent data is required. An ideal source must retain six properties of the ice core, *viz.* i) Populated with *contemporary* data; ii) Conserved *unchanged* until analysed; iii) *Sufficiently large* for an evaluation iv) *Attributable* to a certain population; v) The circumstances of data entry are *homogeneous;* vi) The process of data entry is *self-organized*, not institutional (to avoid group bias).

It seems that such a dataset exists. It is scattered across the world in the form of books in libraries. These books were written by authors who unwittingly – associatively (in fiction) or deliberately (non-fiction) reported on what they themselves saw/experienced (direct) or heard/read about (indirect observation) prior to writing. Once the books are digitised, their content can be reprocessed into databases of yearly n-gram frequencies [3]. By "*n-gram",* we mean a space-delimited sequence of 1-grams. In this text, we use *1-gram* interchangeably with *token* to denote a string of characters not separated by a space – mostly *a word* like "syphilis" or "antibiotic". Regarding syphilis, Figure 1 depicts the frequency of its mention in literature, growing in the era of industrialization and peaking during wars, accompanied by its then most popular mercurial therapy. Both the disease and the useless therapy were slowly extinguished after the discovery of penicillin.

Figure 1 – The three plots, each sum of n-gram frequencies, from top: syphilis; traditional therapy; and effective therapy.

In the following, we propose and discuss a method to study real-world phenomena by systematic analysis of their mentions in literature over time. After cautious exclusion of potential confounders, we expect this mirroring in literature can be used for a quantitative visualization of the change in the real-world presence of certain phenomena – e.g. illnesses, observable factors of environment, etc. – as shown in the following.

Methods

Source

We used the freely available English (E), German (G) and Russian (R) Google Books n-gram data (V.20120701) as an input for used examples, documented in each caption as *Src:E/G/R*. Just the English version of the library used to create the Google n-grams contained 361 billion words derived from a nearly 5.2 million books large subset of Google Books library – approximately 4% of all books ever published [4]. The creators of Google NGram published descriptive statistics of 1-grams, 2-grams, 3-grams, 4-grams and 5-grams (text fragments, which contained 0-4 spaces) in these books. This resulting n-gram database consists of yearly counts of every n-gram and counts of books containing it. Our method requires precise metadata of the resource, in particular, precise publication date. Some kinds of publications notoriously contain older text, thus their publication date is imprecise. An example of these are periodicals and books that aggregate more works – such as anthologies or collected works by a given author. To improve the accuracy of the publication dates the corpus the creators had removed these. After the filtering, a sample of book metadata had been verified by a human annotator (n=1000, sampling five books for every year between 1801 and 2000) and the rate of books with a publishing date outside a five-year range of tolerance was found to amount to 5.8% [4]. Though not required for the proposed method but justified by the propaedeutic character of this paper, we selected only few marker n-grams for the examples. In most cases, we averaged data over three years for our figures. If the span of moving average differed, we specified it in the caption (as MAvg=*n*), where *n* is the number of years. As explained later in this section the absolute ordinate value (n-gram frequency) is not relevant and is therefore usually relative, but always of linear scale starting from zero.

There is a broad topic spectrum of books in libraries and they do not come from a single group of authors. We appreciate this as such "institutional" origin could amplify group think bias [5]. There is yet another reason why we prefer books to news articles or abstracts of research papers: Due to Heaps' Law [6], the longer the text (book vs. article), the higher the number of distinct n-grams in it. Similarly, also the number of topic-unrelated reports will follow the text length sub-linearly.

Figure 2 – "Radiogram" and "roentgenogram". The ordinate is NOT scaled.

Arguably, trends in word usage are not merely due to phrases and words going in and out of fashion as seen in Figure 2. It is also, because their referents (the real-world *phenomena*) change in frequency and salience. Huge amounts of data have been generated by analyses of millions of digitized books [4]. The data has been applied in other fields, such as in language research [4, 7-8] , social sciences such as measuring social functions and even predicting wars [9-10] and in psychology for measuring emotions, individualism or misery [11-13]. Yet it has remained virtually untouched by medical research.

One might wonder why this opportunity has been missed – it could have at least served as a cost-effective tool for the generation of population level hypotheses or for further observation of already known causal relationships.

There are objective reasons for this mistrust. Apart from its novelty, it may be explained by the uncertainty about whether the detected trends and correlations in word frequencies are really caused by the (i) changes in the domain of reference (real-world phenomena), or by (ii) linguistic and other confounders (Figure 2). We will discuss both in the following.

The division we just introduced is slightly different but not entirely unlike the division provided by Michel et al. [4] into "two central factors": (i) cultural and (ii) linguistic change, such as the changing likelihood the author describes the referent by "Xray" vs. "radiogram" [8]. We had to define these groups differently, as many cultural factors – albeit important for others – confound or hide what we seek.

Let us shortly discuss the major linguistic origins of signal cancellation, noise or confounding. Aside from new words coming (*word birth*) to compete and eventually replace the current ones (which might face *word extinction*) just like seen in Figure 2, they might *acquire new spellings or meanings*, *lose* old referents and *undergo shift* in their meanings (see Figure 4). Interestingly, not only vocabulary gradually changes, so do parameters of the language dynamics. To provide an example, in Figure 3 we show a marked shortening of our *collective working memory* over the past two centuries. This supports the finding reached by other means by Petersen et al. [8]. Although we do not expect this to interfere with the method, it might still be important to keep this plot in mind when interpreting any trends in n-grams generated from that period.

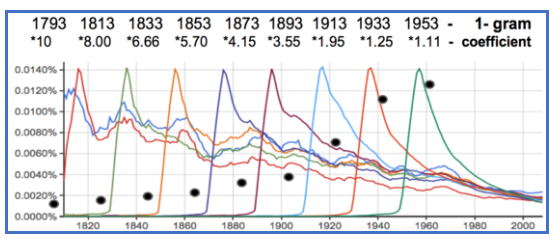

Figure 3 - Frequencies of 9 strings marking years starting with "1793" by increments of 20. A linear coefficient was used, so that the peak of each string would reach f=0.014%. The black dots that follow each peak show inverse values of the coefficient used (roughly approximating the original amplitudes). The time for the recall to fall to 25% of the peak frequency notably shortens over the last 200 years. Apart from the "shortening of recall" a common "oscillation in recall" of these markers with a period of 20-25 years is visible. (Src:E).

In order to study the underlying real world phenomena, we should suppress the confounding linguistic effects. We can do this by appropriate selection of only high-precision *marker n-grams*, as described in the following section. Afterwards, we can approximate the environmental factor – the phenomenon.

A *phenomenon* in a domain is mirrored in literature by a *marker set* = a set of high-precision marker n-grams. E.g., phenomena of means of transportation in the real world are mirrored in books by a *marker set* $S_{\text{Means of Transportation}} = S_{\text{MoT}}$.

Marker set is not to be confused with the term *topic* used in n-gram analysis by others [14-15]. While a *topic* is defined by a group of n-grams that denote various phenomena and which typically occur together in a discourse on a theme, a *marker set* includes only the tokens that have a high precision (not just recall or sensitivity) in denoting the *phenomenon* of interest.

In our example, we assign a few case-insensitive strings to the example marker set for **"Means of transportation" (S_{MoT})**:

$$S_{Means\ of\ transportation} = S_{MoT} \ni \begin{Bmatrix} sS_{Human\text{-}powered\ MoT}, \\ sS_{Animal\text{-}powered\ MoT}, \\ sS_{\text{Motor-powered } MoT}, \\ \ldots etc. \end{Bmatrix}$$

We used "…etc." as a wildcard for further possible *marker subsets* (sS_{xyz}) or n-grams, such as "travel" or "transport", which have good precision to S_{MoT} but do not fit in any of the *subsets*. In contrast, a token like "balloon" would not properly fit into S_{MoT}. It is imprecise for **MoT** because it could just as well denote a hot air balloon as an inflatable toy balloon.

N-grams belonging to a *marker set* can be assigned to categories called subsets (sS). We demonstrated this by dividing *the set* S_{MoT} into the following subsets and populating them:

$$sS_{Human\text{-}powered\ MoT} = sS_{H\text{-}MoT} \ni \{"bicycle", "ride\ a\ bicycle", "skateboard", "roller\ skates", "kayak", "by\ foot", "to\ walk" \ldots etc.\}$$

$$sS_{Animal\text{-}powered\ MoT} = sS_{A\text{-}MoT} \ni \{"dogcart", "horsecar", "riding\ horse", "dog\text{-}drawn", "brancard", "howdah" \ldots etc.\}$$

$$sS_{\text{Motor-powered } MoT} = sS_{M\text{-}MoT} \ni \{"passenger\ car", "electric\ vehicle", "aircraft", "airplane" \ldots \text{etc.}\}$$

Some of the n-grams truly belong to a marker set only during a certain era; e.g. more and more bicycles are now equipped with an electric motor. Therefore, if we are to define a subset: $sC_{Exclusively\ Human\text{-}powered}$, we would have to replace "bicycle" with a higher-precision synonym, or alternatively, limit the analysis to the time when "bicycle" meant solely an exclusively human-powered vehicle. For this reason, it is important to state the mutual exclusivity of the *marker subsets*.

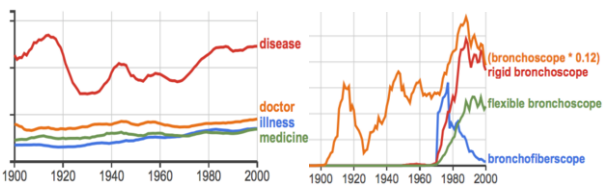

Figure 4 – a) Left: "disease" has low and "illness" has high precision for human health problems. b) Right: The umbrella term "bronchoscope" strives not only during wars, but also with the arrival of a "flexible bronchoscope". Mentions of "rigid bronchoscope" were absent (1898-1958/66). (Src:E)

A typical example of an n-gram unfit for a *marker set* due to a non-intuitively low *marker precision* is "rigid bronchoscope", shown in Figure 4b. Since the introduction of bronchoscopy until 1960s, all bronchoscopes used to be rigid. In spite of this, or better to say: exactly for this reason, the term "rigid bronchoscope" did not need to exist. With introduction of fibre optics, flexible bronchoscopes became standard. To denote these flexible ones specifically, authors could now use terms "bronchofiberoscope" and "flexible bronchoscope", but often, they would use the umbrella term. Due to this bandwidth steal, the "flexible bronchoscope" seems to be less frequent. Similarly, the name "First World War" (instead of "Great War") came to use only after advent of the Second World War – when a distinction was required. Such terms might decrease recall.

Now it is clear that for an n-gram to be added to a *set*, it must have a high *marker precision* – not to be confused with *meaning precision*, the mere absence of polysemy. *Marker precision* is a property of an n-gram that defines how well it reflects the phenomenon of interest across the whole period of analysis, by its frequency change in a given corpus, without detecting unrelated phenomena. If only highly precise *marker* n-grams are present in the *set*, the *set* will mirror its phenomenon in a reliable way.

Non-precision might be also constant in time. This applies also for "disease" (see Figure 4a) which is astonishingly unspecific to human health, being also used for plants and animals. In contrast, the less popular "illness" shows both meaning and marker sensitivity to human medicine. We show another example in Figure 5, which highlights how cultural and political differences between the studied language groups have to be accounted for, like state censorship in the Soviet Union.

Figure 5 – The effects of politics on the literature [until 1917], first mentions during the February Revolution (blue, "свободный рынок", Src:R) and soaring interest in the following 8 years, immediately interrupted in the year of Lenin's death (1924). In red is the English "free market" (Src:E) which continued rising steadily until the mid '90s. (MAvg=1)

In the following paragraphs, we explain the steps by which noise and confounding effects can be minimised. As a result, the *marker sets* and *subsets* should be "clean" enough to relate quantitatively to the changes in the underlying phenomena.

Data pre-Processing

Creation of the "candidate" n-gram sets and subsets:

After selection of the phenomena for the observational analysis, we can define the future *marker sets* and *subsets* for these; e.g. as shown above: [C_{MoT},$sC_{H\text{-}MoT}$, $sC_{H\text{-}MoT}$ and $sC_{H\text{-}MoT}$].

Mechanistic token selection [inclusion criteria]

Here we describe how to populate the main candidate sets of n-grams by *potential marker n-grams*. Then we propose how these could be cleaned from "noisy" n-grams by applying exclusion criteria on each of the "candidate" n-grams.

Firstly, we select the *potential marker* n-grams by:

1. extraction from defined and cited literature,
2. search of a comprehensive dictionary or thesaurus,
3. variation of the grammatical or lexical form,
4. a proposal from any the co-authors, validly argued by the rules 1-3. and/or validated through full-text analysis.

Even though *meaning precision* mostly implies *marker precision*, this is not always the case. As an alternative to the "mechanistic" selection of words described above, in some contexts it would be also viable to use token selection by questionnaire as done by others [13]. This approach is more useful for hard-to-define parameters or phenomena, such as emotions, attitudes or concerns because we cannot easily extract these from a dictionary. Even here, it must be reflected that like any other dimension of language these parameters are subject to change. If a word, such as "terrific" is charged with a positive emotion in 2016, this does not necessarily imply it was positive in 1916.

Definitions

In the next step, the timescale of analysis is defined and exclusion criteria are defined, which must filter out:

1. N-grams discovered that frequently denote unrelated phenomena (*homographs*) or have *low precision.*

2. N-grams with a high noise to signal ratio, as with rare n-gram tokens or in older books due to OCR errors.

Where needed, the manual verification of precision for an n-gram might be performed by human annotator on a sample of full-text books – similar as performed by Michel et al. [13].

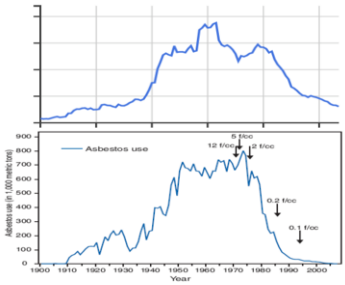

*Figure 6 – **(1900-2007)**: Upper plot shows a single bigram shape C_c={"asbestos cement"} (Src:E). Bottom plot shows asbestos use and legally permissible exposure limits as arrows in the USA [16].*

Analysis

A measurement of n-gram frequency in a single point in time does not predict the contemporary expression of the related phenomenon in the real world by itself. It is also influenced by the salience of the phenomenon and the popularity of a given expression at that time. This introduces noise and makes a trend of the single marker n-gram hard to interpret. The shifts in popularity and the resulting noise can be suppressed by using the *marker sets,* thanks to the increased sample size.

As only the dynamics of the frequency of an n-gram tells the interesting story, the absolute amplitude of its trend is irrelevant for our analysis. We can linearly scale it without interference. In other words, the information is in the shape of curve, such as the positions of the local maxima, minima and trends. To extract knowledge, we propose the following approach.

For the first of the *marker sets* to be quantified (the *reference set*, usually the simplest one) and its subsets; if any, the plots of the member n-grams are added by simple summation.

Linear Fitting

Firstly, we plot the frequencies for each n-gram that belongs to the other *sets* (referred to as *independent sets*). We adjust their amplitudes by mutually independent coefficients, defined by a fitting algorithm. This algorithm uses the linear (!) coefficients with the goal to find the best alignment of the *independent set* to the *reference set* – measured by G (goodness-of-fit, see the section: Measurement).

If studying causalities with delayed effects (Figure 7), a proper adjustment of a whole *(sub-) sets* on the x-scale may be required as well, before proceeding with the following section.

Measurement

Now as we have fitted the curves of the *independent sets* S_i (or sets) to the *reference set* (S_r), we can measure two parameters. The first parameter is the **goodness-of-fit** *(0 < G < 1)*. For a curve of each *set S* (or *subset sS*), the area under its curve A_c (or A_{sC}) is calculated. We define the goodness-of-fit as:

$$G = \left(A_{S_r} \cap A_{S_i} \right) \; / \; \left(A_{S_r} \cup A_{S_i} \right)$$

An ideal correlation of two *marker sets* would have a goodness-of-fit *G=1*. Furthermore, a value of G for a set compared with a constant function (e.g. *x=1*) presents a calibration value *G=t* for the given set. *G=t* means "exactly no measurable association". G lower than the calibration value is a negative correlation, higher G is a positive correlation.

The second parameter, called ***share***, can be derived if *G is close to* 1. We define it as the fraction of the height of the *reference set* made up by a fitted *independent subset* in a time point. This correlates with the partial association of the respective sub-phenomenon (a category) in the respective time. Thus, we can use it to reflect a partial influence in case of a multifactorial causal relationship.

Results

In Figure Six, we compare asbestos use according to CDC with the frequency of mentions of asbestos cement in literature. The similarity of shape (growth in 1910, 1920s, and 1940) is still clear. Unlike the CDC data above, the mentions of "asbestos cement" are not limited to industrial and construction use of asbestos. The n-gram data can also report on demolitions, encounters in environment, daily use, etc. These would well explain the asynchrony in curve descent as even nowadays it is easy to come across aged asbestos containing materials. On the other hand, our single token did not account for other uses than reinforcement of concrete (such as braking systems, stove lining, etc.). Analysis using more n-grams would address this.

Figure 7 –Two simple marker sets in 2012 German n-gram data, using unweighted summation. S_a = {"Eternit", "Asbestzement"} and S_m={"Mesotheliom", "Pleuramesotheliom"}.

Here comes up another aspect, *viz*. a possible amplification of "asbestos cement" by medical and legal literature over the last decades. The recognized causality causes a backwards flowing influence, where these terms do increasingly co-occur. Unlike the underlying aetiological connection, which is delayed, this amplification is immediate: when an author states that a mesothelioma was caused by exposure to asbestos cement. Two possible solutions are at hand: i) to use a resource that allows for excluding books containing both n-grams, thus suppressing the effect, or ii) to limit the inference to the "naïve" era, when the causality was not widely known. Not doing so would decrease sensitivity of the method and the goodness-of-fit.

To extend the example of asbestos, we used two simple sets S_a and S_m (see Figure 7) on a different population (German language group). Here, the patterns expected for this causality are obvious. There are three major peaks in frequency of the *marker set* S_a (1913, 1937, and 1996). For mesothelioma, there are two peaks (1962, 1978) followed by a rising trend filling the last 2 decades of this plot. German epidemiological data confirm this late growth [17]. More interestingly, the two peaks for the *marker set* S_m follow first two peaks of *marker set* S_a with a lag of 49 and 41 years, respectively. This is in accord with the generally accepted latency of incidence for this malignancy between 40 and 50 years [17-18]. We expect the next peak of S_m between 2036 and 2046.

Discussion

The precision of the method critically depends on the elimination of unrelated influences. Therefore, very popular but lexically ambiguous (homonymous) tokens need to be sacrificed (Figure 4a). Even when excluded, these n-grams might soar in

popularity at expense of used markers, which could lower the sensitivity. Thus, in the language where this happens, correlations of certain phenomena might stay hidden (false-negativity). There is certainly much space for improvement of the existing n-gram corpora, as pointed out by Pechenick et al. [19]. One of the good critiques raised, is that the dataset does not measure *true* popularity of ideas, but the frequency with which they are generated [19]. In our case, what is bug for culturomics becomes a feature: popular beliefs and fashions would only confound what we study, *viz.* the incidental observations by authors of books, whether related or unrelated to the topic or author popularity[1].

A second critique raised against the Google n-gram corpora, less relevant in our context, points out the unequal distribution of popular vs. scientific content underlying the English corpus [19]. For our goal, the genre is of little importance if the size of a sampling unit (length of book vs. article) is retained and attention to corpus homogeneity is paid when highly technical terms are used as marker n-grams for observations. An unequal genre/topic distribution could interfere with the sensitivity of this method. However, we doubt the validity of this criticism; namely, we doubt that there is indeed a growing fraction of scientific literature in the 2012 English n-gram resource. A simple search in English Fiction for the token "research" shows a steep and continuous rising trend for the entire 20th century. This suggests that the growth might not stem from a sampling error but rather from a similar culturomical process as shown in [10] – a social function shift in favour of research.

A more relevant critique points at the OCR errors, which are quite prevalent in the pre-1800 texts [4, 19]. This may be improved in the future by more reliable OCR techniques. In this case, which would enable digging far into the past, new evidence regarding lexical changes, social stigmatization, censorship etc. may be found.

Conclusions

We described the main issues in the extraction of information about observable real-world phenomena from changes of n-gram frequencies in large time-indexed n-gram corpora, with the focus on establishing a standardized method for exploitation of these data in, but not exclusively, the public health research. We plan to make a user-friendly tool based on this method available to other researchers and authors.

We demonstrated some of the possible applications on examples of use relevant for the public health and we have shown how already some very simple *marker sets* might well reflect the related real-world phenomenon in its parameters.

Acknowledgements

This work has been done as a part of Erasmus+ stay with a financial support of European Commission. We are thankful to Martin Komenda and Jiri Pavlacky for their valuable feedback. This work would not have been possible without support of Bronislava Kasáčová, Zdenko Medior Kasáč and Jana Petrášeková.

[1] Maybe this text does not really belong to the "culturomics" field created by the authors of the NGram Viewer, as the culture is not our focus. Thus, a better fitting name for this approach would be "exponomics" as it examines the observable exposures of a population (the "exponome" of a population - as opposed to the exposome, which measures all the exposures of an individual).

References

[1] D. Lüthi, M. Le Floch, B. Bereiter, T. Blunier, J.M. Barnola, U. Siegenthaler, D. Raynaud, J. Jouzel, H. Fischer, K. Kawamura, T.F. Stocker, High-resolution carbon dioxide concentration record 650, 000–800,000 years before present, *Nature* **453** (2008), 379-382

[2] R.F. Gillum, From papyrus to the electronic tablet: A brief history of the clinical medical record with lessons for the digital age, *Am J Med* **126** (2013), 853-857

[3] Y. Goldberg, J. Orwant, A dataset of syntactic-ngrams over time from a very large corpus of English books. *Second Joint Conference on Lexical and Computational Semantics* **1** (2013), 241-247

[4] J.B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, E. Lieberman Aiden, Quantitative analysis of culture using millions of digitized books, *Science* **331** (2011), 176-182

[5] I.L. Janis, Groupthink: psychological studies of policy decisions and fiascoes, Houghton Mifflin, Boston, 1982

[6] L. Lü, Z. Zhang, T.Zhou, Zipf's Law leads to heaps' aw: analyzing their relation in finite-size systems, *PLOS ONE* **5** (2010), e14139

[7] L. Borin, D. Dubhashi, M. Forsberg, R. Johansson, D. Kikkinakis, P. Nugues, Mining semantics for culturomics: towards a knowledge-based approach, *ACM* (2013), 3-10

[8] A.M. Petersen, J. Tenebaum, S. Havlin, H.E. Stanley, Statistical laws governing fluctuations in word use from word birth to word death, *Sci Rep* **2** (2012), 313

[9] C. Besse, A. Bakhtiari, L. Lamontagne, Forecasting Conflicts using N-gram Models, (2012), 124-127

[10] S. Roth, Fashionable functions: A Google Ngram view of trends in functional differentiation (1800-2000), *Int J Technol Hum Interaction* **10** (2014), 35-58

[11] A. Acerbi, V. Lampos, P.R. Garnett, A. Bentley, The expression of emotions in 20th century books, *PLOS ONE* **8** (2013), e59030

[12] R.A. Bentley, A. Acerbi, P. Ormerod, V. Lampos, Books average previous decade of economic misery, *PLOS ONE* **9** (2014), e83147

[13] J.M. Twenge, W.K. Campbell, B. Gentile, Increases in individualistic words and phrases in American books, 1960–2008, *PLOS ONE* **7** (2012), e40181

[14] D.M. Blei, Probabilistic topic models, *Communications of the ACM* **55** (2012), 77-84

[15] D. Hall, D. Jurafsky, C.D. Manning, Studying the history of ideas using topic models, *ACL* (2008), 363-371

[16] Centers for Disease Control and Prevention (CDC), Malignant mesothelioma mortality, *MMWR Morb Mortal Wkly Rep* **58** (2009), 393-396

[17] V. Neumann, S. Loseke, D. Nowak, F.J.F. Herth, A. Tannapfel, Malignant pleural mesothelioma, *Dtsch Arztebl Int* **110** (2013), 319-326

[18] S. Prazakova, P.S. Thomas, A. Sandrini, D.H. Yates, Asbestos and the lung in the 21st century: an update: Asbestos and the lung, *Clin Resp J* **8** (2014), 1-10

[19] E.A. Pechenick, C.M. Danforth, P.S. Dodds, Characterizing the Google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution, *PLOS ONE* **10** (2015), e0137041

Address for correspondence

Zdenko Kasáč, 375720@mail.muni.cz