Research in Language, 2018, vol. 16:3 DOI: 10.2478/rela-2018-0015

# GOOGLE BOOKS NGRAM VIEWER IN SOCIO-CULTURAL RESEARCH

***ANNA ZIĘBA***
Adam Mickiewicz University, Poland
azieba@amu.edu.pl

**Abstract**
The objective of this paper is to verify if Google Books Ngram Viewer, a new tool working on a database of 361 billion words in English, and enabling quick recovery of data on word frequency in a diachronic perspective, is indeed valuable to socio-cultural research as suggested by its creators (Michel et al. 2010), i.e. the Cultural Observatory, Harvard University, Encyclopaedia Britannica, the American Heritage Dictionary, and Google. In the paper we introduce a study performed by Greenfield (2013), who applies the program to her *Ecological Analysis*, and contrast the findings with a study based on similar premises, in which we follow the trends in changes in word frequency throughout the 19th and 20th centuries to observe if these changes correspond to one of the major socio-cultural transformations that took place in the studied period, i.e. mediatization. The results of this study open a discussion on the usefulness of the program in socio-cultural research.

**Keywords**: Google Books Ngram Viewer, word frequency, socio-cultural transformations, mediatization, news values

## 1. Introduction

It is tempting to believe that the arrival of a new tool giving access to a massive database, a corpus of 5,195,769 books scanned and digitized with the use of optical character recognition (OCR) will open new possibilities in many fields of science. As performing a study on such vast material has not been achievable before, a research on the corpus provided by Google Books Ngram Viewer seems a state-of-the-art endeavour that should provide reliable data.

It could also be valuable to the socio-cultural research that is based on linguistic material as such research is usually very time-consuming. Therefore, one of the merits of this tool is that it allows the researcher to spend more time on the analysis of data than on their collection.

Moreover, it might appear that since the lexical changes are gradual and relatively stable, the fluctuations in word frequency, upon which Google Books Ngram Viewer provides extensive data, are relevant and their study will improve our comprehension of the social changes and their consequences.

The paper first presents the tool and gives examples of its possible application to research in various fields as suggested by its creators. Then, a recent study in

human ecology proposed by Greenfield (2013) is introduced, a study which inspires us to perform a similar one on the relationship between one of the biggest socio-cultural transformations in the period under study, i.e. mediatization, and changes in frequency of words relevant to the subject. Next, we present the methodology of the study and the obtained data. In the later parts of the paper the results of our study are discussed and conclusions on the usefulness of Google Books Ngram Viewer in socio-cultural research are drawn.

## 2. Theory and background

### 2.1 Google Books Ngram Viewer

Linguists have hitherto worked with word frequency dictionaries or lists such as 450 million word Corpus of Contemporary American English (Davies 2010), 5 million word The American Heritage Word Frequency Book (Carroll, Davies & Richman 1971), or databases such as 1.7 billion word Dante (Atkins 2010) and WordNet (Fellbaum 2005) with little over 155,000 words. These tools seem modest in comparison with Google Books Ngram Viewer, a new tool introduced in 2010 by the Cultural Observatory, Harvard University, Encyclopaedia Britannica, the American Heritage Dictionary, and Google, as its creators constructed a corpus of 5,195,769 digitized books (4 percent of all books that have ever been published) from over 40 university libraries and individual publishers.

The texts were scanned and digitized with the use of optical character recognition (OCR). Having taken into account the quality of the texts' OCR and metadata, the team selected a group of over 5 million books for analysis to develop the corpus including 361 billion words in English, 45 billion in French, 45 billion in Spanish, 37 billion in German, 35 billion in Russian, 13 billion in Chinese, and 2 billion in Hebrew. The study was limited to the analysis of frequency of a given 1-gram, which might be understood as a single lexical unit, or an n-gram (a series of lexical units) over time, but occurring at least 40 times in the corpus. Michel et al. (2010) define the 1-gram as "a string of characters uninterrupted by a space" and an n-gram as "a sequence of 1-grams, such as the phrases 'stock market' (a 2-gram) and 'the United States of America' (a 5-gram)". The frequency was "computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year" (Michel et al. 2010).

In the article published in *Science* Michel et al. (2010) maintain that the corpus enables investigators to study cultural trends quantitatively, and that it has opened up a new field of research, namely *culturomics*, a field drawing a connection between changes in word frequency and linguistic and cultural shifts. The researchers give examples of such undertakings. They observe changes in the English lexicon studying the overall number of words to discover that the size of this language increased by over 70% in the past 50 years. They also follow the

changes in frequency of the 2077 headwords that entered the American Heritage Dictionary of the English Language in 2000 and notice that part of the words, still found in the dictionaries, were no longer used. These investigations lead them to the following conclusion: "Our results suggest that culturomic tools will aid lexicographers in at least two ways: (i) finding low-frequency words that they do not list, and (ii) providing accurate estimates of current frequency trends to reduce the lag between changes in the lexicon and changes in the dictionary" (Michel et al. 2010).

The team investigates grammatical changes and finds that frequency is an important factor in the shifts between regular and irregular forms of past verbs. The researchers also make an inquiry into collective memory, developing plots concerning the interest in various events between 1875 and 1975, in order to compare the rise and fall of fame of the most well-known people and uncover censorship in Nazi Germany.

At a later stage the researchers add a system that enables identification of parts of speech, searching for inflections or for multiple capitalization styles simultaneously and a feature called 'wildcards': for retrieving the ten most popular collocations.

Since its introduction in 2010 Google Books Ngram Viewer has been widely described and employed both in social and natural sciences. Berry (2012) describes it as an example of "the way in which code and software become the conditions of possibility for human knowledge, crucially becoming computational epistemes" (Berry 2012: 1), Rutten et al. treats it as a tool to overcome a "chronological distance, or time lag, between books and their subject matter in studies of memory" (Rutten 2013: 40) and Michalski et al. (2012) suggests the Ngram Viewer could be used "as a fast prototyping method for examining time-based properties over a rich sample of literary prose" (Michalski 2012: 1).

Google Books Ngram Viewer has been applied in various studies. Linguists used it to investigate biomedical domain literature in respect of terminology changes (Grigonyte et al. 2012), to follow word usage and cultural transformations in contemporary West Bengal (Phani 2012) and to illustrate diachronic variation of preferred adjective ordering (Hill 2012). It was also employed in social studies: Kesebir and Kesebir (2012) used it to prove that moral ideals and virtues decreased significantly in the American public conversation, Oishi et al. (2013) to analyze the concepts of happiness across time and cultures, Cockerill (2013) to trace the roots of industrial ecology education to the 1960s and 1970s, Lucier (2012) to study the relations of science and capitalism, Kumar and Sahu (2010) to trace the history of marketing, and Johnson (2011) to introduce the concept of information overload, not to mention Greenfield (2013) who applied it to a research into human ecology. It has also been used by Alcock (2012) to assess trends in the use of evolutionary concepts in non-technical literature and Crasto (2011) to study the use of the term 'bioinformatics' in literature.

Google Books Ngram Viewer also received criticism, which came mostly from Mark Davis (2014), who recognised the dataset as remarkable but perceived the interface as too simplistic. He claimed it did not allow for the use of collocates in searches, searching by wildcards and a meaningful use of parts of speech. As the datasets had been made available online by their collectors, Davis incorporated them into his work and proposed an alternative architecture and interface that enabled more complex searches (e.g. with variables for a given part of speech or providing data on complicated grammatical constructions). However, his criticism towards Google Books Ngram Viewer does not seem fully grounded as GBNV does in fact allow for searches based on speech tags or wildcard searches (though these features were not possible at the original stage).

There also appeared questions concerning the accuracy of data acquired through the use of Google Books Ngram Viewer (mostly on blogs and forums): the long 's' mistaken by OCR for 'f' (which fell out of the English typeface in early 19th century), as well as semantic and spelling changes. However, these can influence a study of word frequency in the 19th and 20th century only marginally. Therefore, even if we take into consideration the imperfections of OCR, Google Books Ngram Viewer still seems to put socio-cultural research in a context whose significance is hard to question, especially if carried out cautiously and conscientiously.

### 2.2. Greenfield's Ecological Analysis

Our study, whose objective was to inspect whether the tool is indeed suitable for investigating socio-cultural changes, was inspired by the work of Greenfield published in Psychological Science in 2013. The researcher uses Google Books Ngram Viewer to study human ecology and finds confirmation for her hypothesis concerning a shift in this ecology from rural to urban. She also maintains that cultural features indexed by word frequencies reflect what is preferred by a population.

She generates the hypotheses on a theory of social change from *gemeinschaft* into *gesellschaft* environments. She focuses on individualistic values and behaviours such as: personal choice, materialism, significance of personal property, independence and assertiveness, becoming dominant in the modern world. Greenfield assumes in her study that the gesellschaft-adapted cultural traits, indicated by relevant words in the American English corpus should grow in number, and that the gemeinschaft-adapted features studied within the same corpus should decline.

In her study Greenfield uses high-frequency words, as advised by Michel et al. (2010), with a narrow range of semantic interpretations, relevant to the theory and their synonyms. She studies the changes in frequency of the following word pairs:

- ‘oblige’ (characteristic of the gemeinschaft environment) and ‘choose’ (characteristic of the gesellschaft environment), and their noun synonyms: ‘duty’ and ‘decision’,
- ‘give’ and ‘get’ (representative of gemeinschaft and gesellschaft environments respectively) and their noun synonyms ‘benevolence’ and ‘acquisition’,
- ‘act’ (exhibiting gemeinschaft comprehension of the social world in terms of action or behaviour) and ‘feel’ (representing inner psychological processes typical of the gesellschaft domain), and their noun synonyms ‘deed’ and ‘emotion’,
- and additional concepts to illustrate the historical pattern of shifts in values: ‘obedience,’ ‘authority,’ ‘belong’ and ‘pray’ (and their synonyms: ‘conformity,’ ‘power,’ ‘join’ and ‘worship’) as depicting gemeinschaft values, and ‘child,’ ‘unique,’ ‘individual,’ and ‘self’ (and their synonyms: ‘baby,’ ‘special,’ ‘personal’ and ‘ego’) as exemplary of the gesellschaft scene.

The results of the analysis confirmed Greenfield’s stance. The relative frequency of all words characteristic of the gemeinschaft environment decreased and the words characteristic of the gesellschaft environment increased. Putting the results in the context of other studies relevant to the field and replicating the analysis for each word in the corpus of British books validated the assumptions of the researcher. Therefore, Greenfield maintains that the transformation of the American culture from rural to urban is reflected in the American cultural products, i.e. books.

## 3. The analysis

The research conducted by Greenfield encouraged us to perform a similar study with the use of Google Books Ngram Viewer. Our analysis covered one of the major socio-cultural changes, which occurred in the last two centuries, namely mediatization (Hjarvard 2008, Lilleker 2008). With the development of media, a change in communication has taken place. As a consequence, entire societies are strongly influenced or even formed by mass media (Mazzoleni and Schulz 1999, Hjarvard 2013). Following Greenfield’s example it can be assumed that this in turn should lead to changes in the frequency of words relevant to the socio-cultural phenomenon in question.

### 3.1. Methodology

The semantic key upon which we performed the analysis was based on a set of features deciding on the newsworthiness of information, i.e. *news values*. The set, originally defined by Gatlung and Ruge in 1965 and by now well established in

media studies, comprised 18 qualities: *negativity*, *recency*, *proximity*, *consonance*, *unambiguity*, *unexpectedness*, *superlativeness*, *relevance*, *personalization*, *eliteness*, *quality of attribution*, *facticity*, *continuity*, *competition*, *co-option*, *composition*, *predictability*, and *prefabrication*. We assumed that if Greenfield is unmistaken, the analysis of the frequency of words relevant to the news values should show an increase in the studied period. Therefore, we selected ten features to be included in the analysis and prepared 5-word semantic keys presented in Table 1. We chose to work on an English (both British and American) corpus, as it is the largest database available so far.

**Table 1**. The selected news values and their representative lexical items

| no. | News value | 5-word semantic key |
|---|---|---|
| 1. | negativity | awful, bad, dreadful, poor, unacceptable |
| 2. | recency | currently, lately, presently, recently, today |
| 3. | proximity | close, dear, familiar, nearby, neighbouring |
| 4. | consonance | average, common, normal, standard, usual |
| 5. | unambiguity | apparent, clear, distinct, evident, obvious |
| 6. | unexpectedness | abrupt, rapid, sudden, surprising, unexpected |
| 7. | superlativeness | best, first, last, least, most |
| 8. | relevance | essential, great, important, significant, substantial |
| 9. | continuity | constantly, continuously, regularly, steadily, still |
| 10. | predictability | anticipated, expected, predictable, probable, supposed |

Including all 18 values would not have been possible, as some of the features cannot be represented accurately. In the case of *personalization*, *facticity*, *co-option*, *composition*, and *prefabrication* the feasibility of the analysis was found to be very limited. It would be problematic to find words relating to a personal portrayal of information, intertextual references (in case of *co-option*) and *eliteness* as the objects of these strategies would be different in each text, and *facticity* would be manifested by an infinite number of names, dates and statistics. Likewise, it was difficult to identify words characteristic for a prefabricated message. Additionally, we assumed that *competition* (in the media occurring between agencies, editorial teams or journalists) and *composition* (maintaining a balance of different types of coverage) are irrelevant to our study since they affect language rather at the textual than lexical level.

Each semantic key includes 5 lexical items, listed in alphabetical order in Table 1. The items were selected as the most relevant to the given value, i.e. occurring in contexts representing the value. The keys include both adjectives and adverbs, as part of the news values are described predominantly by adjectives (*negativity*, *proximity*, *consonance*, *unambiguity*, *unexpectedness*, *superlativeness*, *relevance*, *predictability*) and part by adverbs (*recency*, *continuity*). The main criterion for selection was high-frequency (mean frequencies for all relevant items were measured).

As Greenfield points out, since Google Books Ngram Viewer works on a corpus of 361 billion words in English, the absolute percentage of any single word is naturally small. However, the focus of the study is on the change in frequency, not its height. For example, the word 'great' in Figure 8 starts in the year 1800 with a frequency of about 130 occurrences per 100,000 words but decreases to about 30 occurrences per 100,000 words by the year 2000. It seems that this change is meaningful. Moreover, we took interest only in general trends in changes in the frequency of selected items, as focusing on each rise and fall in frequency of subsequent items in a 200-year period would not increase the value of the study, on the contrary it could blur the results.

We also considered using the advanced interface proposed by Davis (2014), but since the object of the study was the standard version and since in the study we included only basic searches available in both interfaces, such endeavour seemed pointless.

### 3.2. Findings of the study

The scores for each news value, presented separately for the purpose of clarity, are illustrated in the figures below. Comparing the results we included values accurate to four decimal places. The objective was to calculate the ratio of increasing to decreasing trends in the changes in frequencies of the selected words representing each of the 10 news values. In cases when the relative change between the values for 1800 and 2000 was less then (+/-) 30% the change was deemed insignificant.

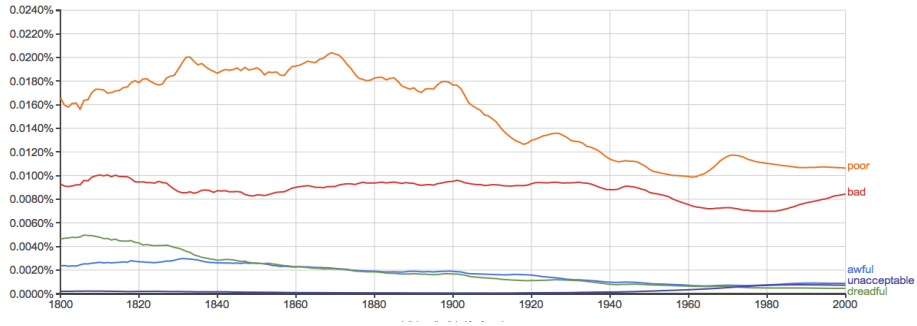


**Figure 1**. The frequency of the five words representative of *negativity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The first news value studied was *negativity*, which refers to higher rating of bad news than good news by the media. It was represented by the following words: *awful*, *bad*, *dreadful*, *poor* and *unacceptable*. Out of these five items just one (*unacceptable*) confirms the assumption that the values typical for mass media do

influence the changes in frequency of words representing these values, as the word's frequency rises from 0.0002% to 0.0007%. The frequency of the other four items either decreases (*poor*, *awful*, *dreadful*) or does not change significantly (*bad*), i.e. the relative change is little over 10%.

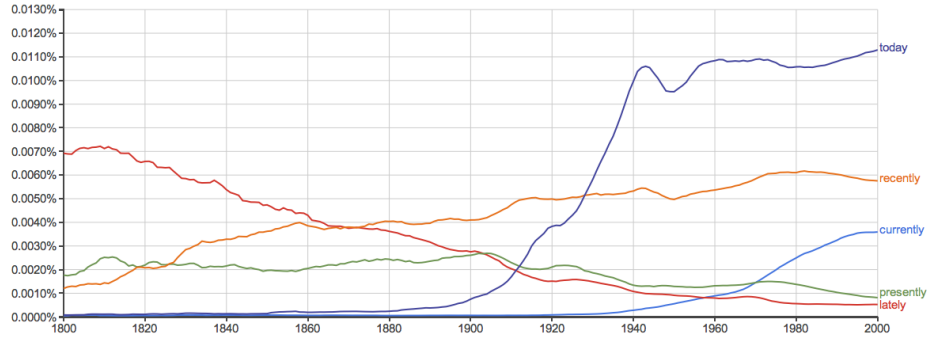


**Figure 2**. The frequency of the five words representative of *recency* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The scores for the items representing the second value, i.e. *recency*, treating of the media's preference for breaking news, are more diverse. The frequency of three words increases quite rapidly since 1800: in case of *today* it rises over 130 times, in case of *recently* – almost five times, and in case of *currently* – over 70 times. On the other hand, the frequency of the other two items falls from 0.0069% to 0.0005% in case of *lately* and from 0.0018% to 0.0008% in case of *presently*.

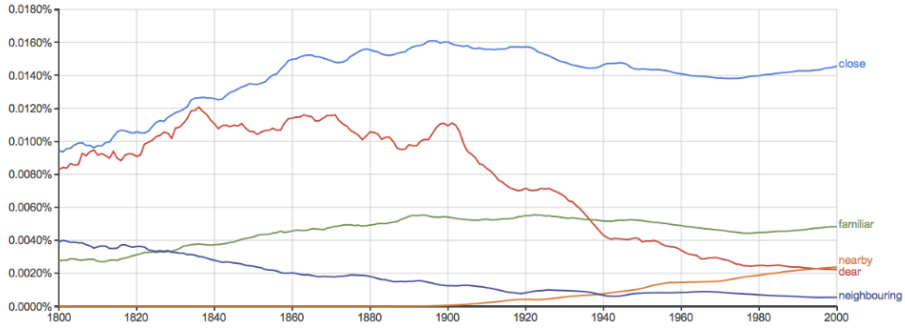


**Figure 3**. The frequency of the five words representative of *proximity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Likewise, the results for *proximity*, which relates to the closeness (either geographical or in terms of values) of the occurrence to the readers, are diverse.

Three values rise and two fall. The most rapid change concerns the word *nearby* whose frequency increases over 200 times. The relative change in case of *close* is 35% and in case of *familiar* – 71%. The frequency of *dear* decreases from 0.0083% in 1800 to 0.0022% in 2000, exhibiting relative change of -73% and the frequency of *neighbouring* changes by -87%.

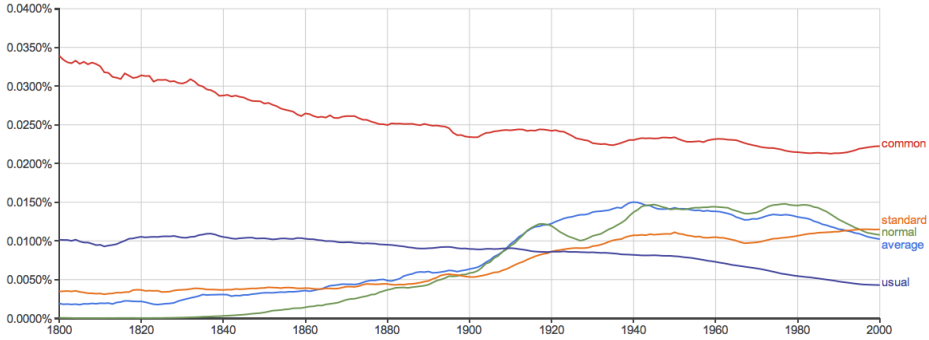


**Figure 4**. The frequency of the five words representative of *consonance* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The words relevant to *consonance*, a value referring to high newsworthiness of occurrences following regular, familiar patterns, provide a similar model: the frequency of three items rises, and the frequency of two words decreases (by over one third in case of *common* and by half in case of *usual*). The most rapid change can be noted with *normal*, whose frequency increases over 100 times. The relative changes of frequency of *standard* and *average* are also substantial: 229% and 437% respectively.

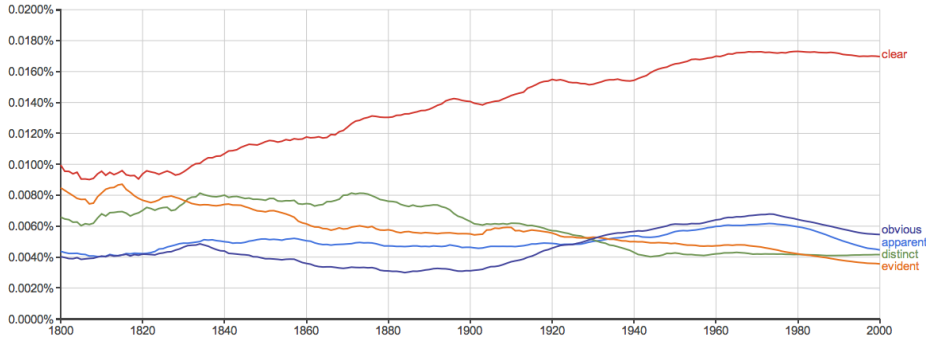


**Figure 5**. The frequency of the five words representative of *unambiguity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The next news value studied was *unambiguity*, which refers to the media's preference of clarity of the information and interpretation (preferably limited to one) of events. The biggest change was noted for the word *clear*. Its frequency rises as many as 17 times in the studied period. The frequency of *obvious* increases from 0.0040% to 0.0054% (relative change 35%) and the frequency of both

*distinct* and *evident* decreases. The relative change in the first case is -38% and in the second -59%. The frequency of *apparent* does not change significantly.

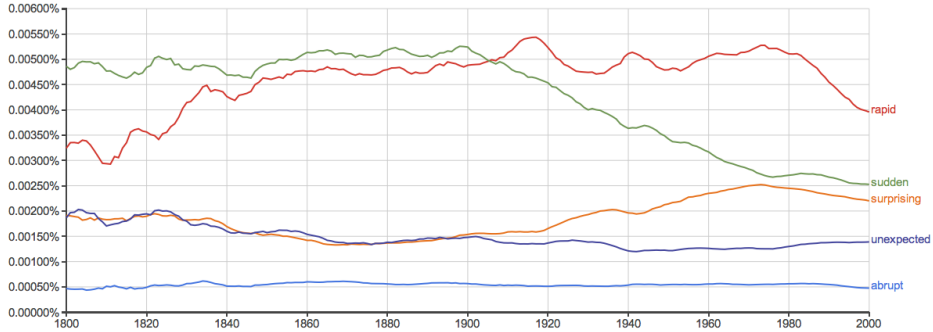


**Figure 6**. The frequency of the five words representative of *unexpectedness* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

As far as *unexpectedness* is concerned, none of the items exhibit a dramatic change in frequency. As shown in Figure 6 *abrupt*, *unexpected*, *surprising* and *rapid* note similar values in 1800 and in 2000, differing only slightly (the relative changes are: 2% for *abrupt*, -26% for *unexpected*, 16% for *surprising* and 22% for *rapid*). The frequency of *sudden* falls (-49%). The feature clearly stands in opposition to one of the previously mentioned news values, i.e. *consonance*, as it refers to extraordinary events.

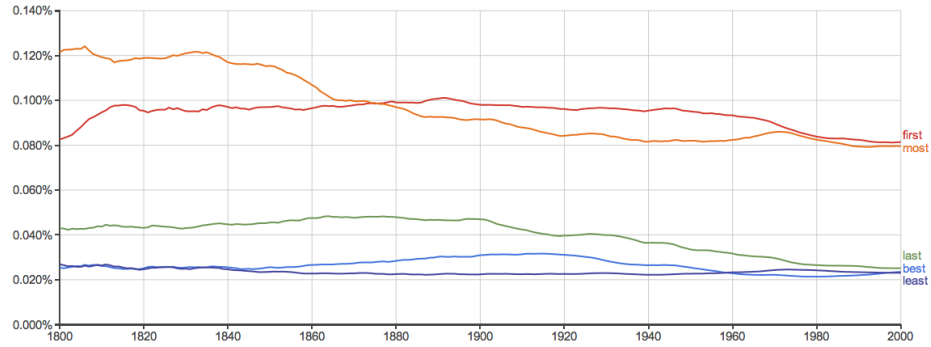


**Figure 7**. The frequency of the five words representative of *superlativeness* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Surprisingly, the frequency in the observed lexical items representative of *superlativeness*, which refers to high newsworthiness of the most spectacular occurrences, does not rise at all. It falls substantially in case of *last* (-41%) and

*most* (-34%). It also decreases in the other cases: the relative change in the frequency of *first* in the studied period is -1%, of *best* it is -7% and of *least* it is -14%.

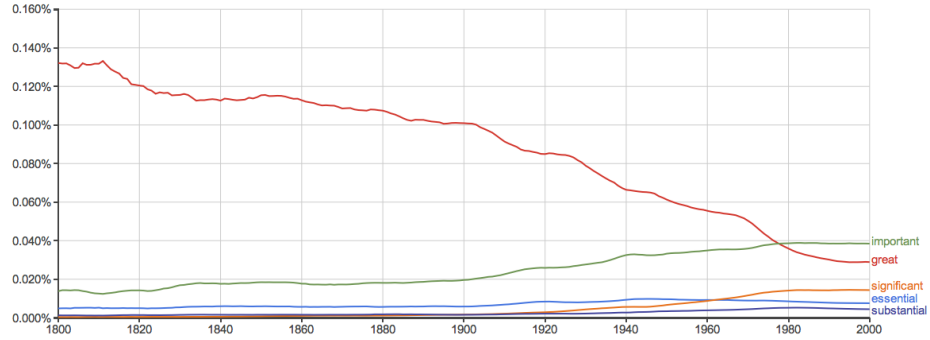


**Figure 8**. The frequency of the five words representative of *relevance* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Apart from *great*, whose frequency falls between 1800 and 2000 (relative change -78%), all words representative of *relevance*, a value referring to high newsworthiness of occurrences important to the readers, rise. The relative change in the frequency of *essential* is 53%, and the frequency of *important* and *substantial* increases over three times and of *significant* as many as over 35 times.

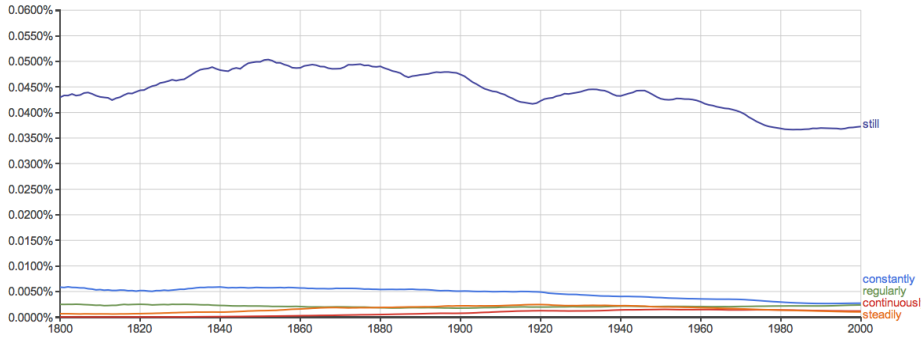


**Figure 9**. The frequency of the five words representative of *continuity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The next news value studied was *continuity*, a value underlining the relevance of information referring to previous news. The frequency of the most common word representative of this value, i.e. *still*, does not change much in the studied period. The relative change in frequency of the word is -13%. The frequency of *regularly*,

whose relative change is -4% does not fall significantly either. However, the frequency of *constantly* decreases by half and the frequency of *continuously* rises over 60 times. The frequency of *steadily* also increases (relative change 67%).

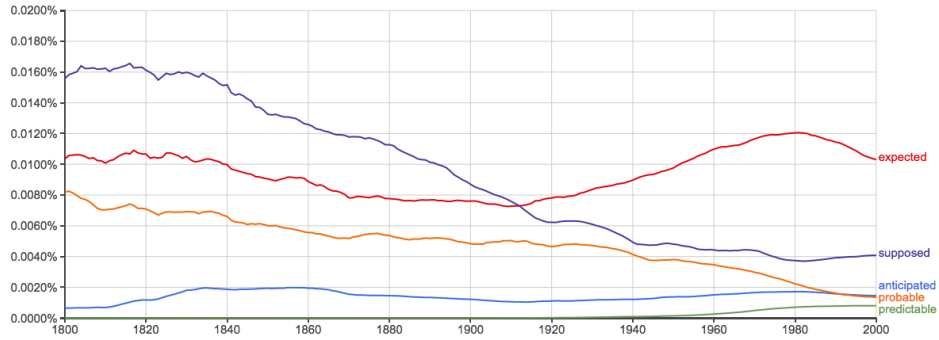


**Figure 10**. The frequency of the five words representative of *predictability* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Among the words illustrated in Figure 10 and representative of *predictability*, a value referring to high newsworthiness of occurrences that are easy to foretell, the most dramatic change in frequency was observed in *predictable*. Its frequency rises over 880 times. Another word whose frequency increases is *anticipated* (over two times). The frequency of *expected* in 2000 is very similar to that in 1800 (-0.5%) and the frequency of *probable* and *supposed* falls from 0.0082% to 0.0013% and from 0.0156% to 0.0040% respectively.

To confirm that the frequency of selected words corresponds the socio-cultural change in question we should expect an increase in all or at least most of the items. Yet, as illustrated in Table 2 an increase in the frequency of the lexical items is noted only in 20 instances (which make up 40% of all lexical items). Moreover, the frequency of the selected words falls in 18 cases (36% of all items) and does not change significantly in 12 (24%). Even if we take into account the fact that the most rapid changes concern only increases (*predictable* which rose over 880 times, *nearby* over 200 times, *today* 130 times, and *normal* over 100 times), the results for the 30 items which do not present any major increase still undermine the discussed hypothesis.

**Table 2**. The trends in frequency change of the selected lexical items between 1800 and 2000

| news value | number of words whose frequency increased | number of words whose frequency decreased | number of words whose frequency did not change significantly (less than 30%) |
|---|---|---|---|
| negativity | 1 | 3 | 1 |
| recency | 3 | 2 | 0 |
| proximity | 3 | 2 | 0 |
| consonance | 3 | 2 | 0 |
| unambiguity | 2 | 2 | 1 |
| unexpectedness | 0 | 1 | 4 |
| superlativeness | 0 | 2 | 3 |
| relevance | 4 | 1 | 0 |
| continuity | 2 | 1 | 2 |
| predictability | 2 | 2 | 1 |
| **total** | **20** | **18** | **12** |

## 4. Discussion

It seems that the Google Books Ngram Viewer though providing an extensive database and enabling a fast collection of data does not give clear evidence of the influence that social changes have on word frequency. The results of the study may be surprising, as it has been argued for long that the relationship between values fostered in a society and its language is close. It seems reasonable to assume that if culture and language are linked, one should have an impact on the other.

Undoubtedly, Google Books Ngram Viewer will find application in linguistic studies. Easy access to digitalised texts offers incomparable opportunities in lexicography, chronologization of units of language and datation of textual objects (area thoroughly studied by Wierzchoń (2008)). Its latest feature, i.e. 'wildcards', used for retrieving popular collocations can be beneficial both to foreign language teaching (e.g. the writing component), and translation, as finding the right collocations improves the naturalness of texts.

Additionally, the tool provides information on the popularity of topics of discussion in the digitalised material, yet it does not explain why the values increase or decrease. An example could be the word *family*, whose frequency rose from 0.02% to 0.03% in the studied period, even though it seems that the declining marriage rates, lower number of children being born and a growing number of divorces could suggest a devaluation of this institution. In this case one might assume that such a fundamental change is worth many a discussion and hence the increase in the frequency of the word. Certainly the data could not lead to a conclusion that there were more families in 2000 than in 1800. Moreover, the topic could also be discussed with no mention of *family*, as other words concerning this

idea could be used (*relatives*, *children*, *husband*, *wife*, etc.), which further diminishes the value of the data.

Therefore, despite the fact that the tool may be helpful in developing certain theories concerning socio-cultural phenomena, we claim that the data obtained with Google Books Ngram Viewer is not reliable enough to confirm these theories.

First, the material selected by Michel et al. includes only 4 percent of all books ever published. Even though 5 million books is a considerable number, it is only a fraction of all printed texts and hence inferences should not be drawn on this basis, as they could lead to false statements.

It also appears that Google Books Ngram Viewer does not take into account the different contexts in which the analysed words are set in, even though such contexts seem essential in any study concerning semantics. Contexts carry meaning. The fact that the frequency of a word rises does not necessarily mean that the concept is valued more, but, as mentioned above, that it is discussed extensively.

Omitting context means ignoring various lexical senses of words. Greenfield states that the decrease in the frequency of the word 'give' is symptomatic of a social change from gemeinschaft into gesellschaft values. However, if we type into the program the phrases: *give back*, *give away*, *give priority*, *give a hand* or *give birth* the trend in frequency is actually rising. Greenfield claims that selecting words with narrow range of semantic interpretations prevents incorporating into the study words in contexts irrelevant to the analysis. However, it seems that all instances should be taken into account to ensure the reliability of the study, especially as the proposed selection would be random unless done manually, which appears implausible. Moreover, it appears that the narrower the range of semantic interpretations the lower the frequency of the word, which in turn affects the analysis, as changes in low-frequent words could seem less meaningful. For the same reason Greenfield chooses high frequency words for her analysis, and for the same reason it seems difficult to explain the observed trends in word frequencies that occurred in our study of the lexical items representative of news values. The fact that the most rapid changes concerned only increases seems meaningful, but without the contexts and vast etymological knowledge we are unable to determine the cause. It is possible that the rising influence of media on societies, culture and language plays a role, but it might as well be caused by other factors: political, economic, linguistic or psychological.

Admittedly, Google Books Ngram Viewer 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 50 words in millions of contexts, it seems implausible to incorporate such information into the study, even if for reasons of time and space.

As Lakoff (2013) states in her criticism towards the approach examined in this paper, we are rarely able to say whether the changes in frequencies carry meaning

or are just accidental. She claims that even though there are words whose appearance, or increase in frequency, can be easily explained by the socio-cultural phenomena, such words are scarce and usually limited to technological innovation or political transformations.

Nevertheless, Lakoff agrees that even though the presence of most words and the changes in their frequency do not tell much about the values ascribed to certain phenomena it may be a sign of recognition of a problem. A case in point might be the appearance, and/or increasing frequency of words such as *homophobia*, *racism* or *sexism*. As Lakoff aptly notes, the fact that these terms were not used (or used incidentally) in the 19th and early 20th century does not mean that the phenomena did not exist, only that in the 20th century they were noticed and became worthy of naming and changing. And indeed, the original studies presented in the article in Science concern such terms. The occurrence in the corpus and the significance of the fluctuations in frequencies of words as e.g. *netiquette* or *World War I*, as well as names of well-know people, seem self-explanatory and therefore may entice researchers to apply data obtained via Google Books Ngram Viewer to more complex studies.

However, one should be circumspect in such undertakings especially as Hilpert and Gries (2009) warn that since the trends in frequencies are rarely unidirectional or strong enough to be intuitively clear, a statistical measure that would help determine if the observed frequencies differ from the mean more than it could be expected, should be incorporated in more complex studies.

It could be concluded in Lakoff's words that the relationship between language and the reality it refers to is complicated and difficult to embrace by merely following changes in word frequencies. The study of single lexical items (or even phrases or sentences) answers the questions posed by researchers interested in the relationship between word frequency and specific socio-cultural phenomena only partially. Not only does it fail to address the context, but also omits the meaning conveyed at the text level. To judge whether certain phenomena are represented in a language or to follow trends one should perform a thorough analysis incorporating whole texts, not just single words, into the study. Employing the examples provided above, i.e. *homophobia*, *racism* or *sexism*, we should see that a text with homophobic, racist or sexist contents would rarely include the words naming the phenomena. Therefore, if the researcher studying these phenomena took under consideration only the frequency of these items, the results would be far from reliable, as the meaning of texts may be implicit. It may lie in metaphors, intertextual allusions or even images.

Therefore, the conclusions drawn from the study by Greenfield even though reasonable, seem far-fetched as the research is based on few words indexing the contrasting values and, more importantly, it does not take into account the contexts in which the words occurred, nor does it include any sort of thematic analysis based on texts as units thereof.

## 5. Conclusion

It is a fact, that a research based on a 361 billion word corpus is prone to produce valuable results. It should help answer the questions posed by scholars both in humanities and social sciences and show the co-dependencies between certain phenomena and language at the lexical level or reveal the patterns of grammatical changes in irregular forms of verbs (as suggested by Michel et al. (2010)).

Google Books Ngram Viewer enables the researcher to put word frequency in a historical context as it shows the changes in the frequency of any selected word or group of words in time. Furthermore, its digital form is yet another reason to acknowledge it as a valuable tool as it allows the researcher to spend more time on the analysis of data than on their collection.

Nonetheless, the usage of Google Books Ngram Viewer should be limited to uncomplicated studies related to word frequency. It cannot be treated as the only tool in a research into complex socio-cultural transformations, as it does not provide extensive information on the contexts in which the words occurred and may lead to superficial, inaccurate, or less precise descriptions of studied phenomena if not confronted with a comprehensive textual analysis.

Our goal was to verify if Google Books Ngram Viewer is indeed a valuable tool in socio-cultural research as suggested both by its creators and Greenfield. Thus, we decided to perform a study similar to Greenfield's *Ecological Analysis* and follow the trends in changes in word frequency throughout the 19$^{th}$ and 20$^{th}$ centuries to observe if these changes correspond to one of the major socio-cultural transformations that took place in the studied period, i.e. mediatization. The data obtained in the course of the study suggest that the changes in word frequency do not directly depend on the rise of the role of the news values in modern societies. Additionally, a close examination of the methodology suggested by Greenfield demonstrated its shortcomings: most importantly disregarding the importance of contexts in which the words occurred and their different semantic interpretations. It seems that in a research of the relationship between socio-cultural transformations and word frequencies, a multifaceted study based on etymological considerations and incorporating thematic analysis should be performed. So far a tool enabling such an undertaking on a scale of millions or even billions of texts remains unknown.

## References

Alcock, Joe. 2012. Emergence of Evolutionary Medicine: Publication Trends from 1991-2010. *Evolutionary Medicine, 1*. doi:10.4303/jem/235572

Atkins, Sue. 2010. The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In: Gilles Maurice de Schryver (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*, 267-97. Kampala: Menha Publishers.

Atkinson, Maxine P. and Stephen P. Blackwelder. 1993. Fathering in the 20th Century. *Journal of Marriage and the Family*,*55*(4), 975–986.

Bell, Allan. 1991. *The Language of News Media*. Oxford: Blackwell Publishers Ltd.

Berelson, Bernard. 1971 [1952]. *Content Analysis in Communication*. New York: Hafner Publishing Company.

Berry, David M. 2012. The Social Epistemologies of Software. Social Epistemology: A Journal of Knowledge. *Culture and Policy, 26*(3-4), 379–398. doi:10.1080/02691728.2012.727191

Cabrera, Natasha, Tamis-LeMonda, Catherine S., Bradley, Robert H., Hofferth, Sandra, & Michael E. Lamb. 2000. Fatherhood in the twenty-first century. Child development, 71, 127–136. doi: 10.1111/1467-8624.00126

Carroll, John B., Davies, Peter and Barry Richman. 1971. *The American Heritage Word Frequency Book*. Boston: Houghton Mifflin.

Castells, Manuel. 1996. *The Rise of the Network Society, The Information Age: Economy, Society and Culture*. Malden, Oxford: Blackwell.

Chow, Esther Ngan-ling. 2003. Gender Matters Studying Globalization and Social Change in the 21st Century. *International Sociology*, *18*(3), 443–460.

Cockerill, Kristan. 2013. A Failure Reveals Success. *Journal of Industrial Ecology*, *17*, 633–641. doi: 10.1111/jiec.12049

Cowan, Ruth Schwarz. 1976. The "Industrial Revolution" in the Home: Household Technology And Social Change in the 20th Century. *Technology and Culture*, *17*(1), 1–23.

Crasto, Chiquito J. 2011. Bioinformatics for Biological Researchers – Using Online Modalities. In: Eta Berner (ed.), *Informatics Education in Healthcare*, 147–165. Birmingham: Springer.

Davies, Mark. 2005. The Advantage of Using Relational Databases for Large Corpora: Speed, Advanced Queries, and Unlimited Annotation. *International Journal of Corpus*, *10*(3), 307–334. doi:10.1075/ijcl.10.3.02dav

Davies, Mark. 2010. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–465. doi:10.1093/llc/fqq018

Davis, Mark. 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.

Edmunds, June and Bryan S. Turner. 2005. Global Generations: Social Change in the Twentieth Century. *The British Journal of Sociology*, *56*, 559–577. doi: 10.1111/j.1468-4446.2005.00083

Fellbaum, Christiane. 2005. WordNet and Wordnets. In: Keith Brown (ed.), *Encyclopedia of Language and Linguistics, Second Edition*, 665–670. Oxford: Elsevier.

Fuchs, Christian. 2008. *Internet and Society: Social Theory in the Information Age*. London: Routledge.

Greenfield, Patricia M. 2013. The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science*, *24*(9), 1722-1731. doi:10.1177/0956797613479387

Grigonyte, Gintare, Rinaldi, Fabio and Martin Volk. 2012. Change of Biomedical Domain Terminology Over Time. In: Arvi Tavast, Kadri Muischnek and Mare Koit (eds.), Human Language Technologies – The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012 (Vol. 247). IOS Press.

Hill, Felix. 2012. *Beauty Before Age?: Applying Subjectivity to Automatic English Adjective Ordering*. Proceedings of the NAACL HLT '12 2012 Student Research Workshop, 11–16. Stroudsburg, PA: Association for Computational Linguistics.

Hilpert, Martin and Stefan Gries. 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. doi: 10.1093/llc/fqn012

Hjarvard, Stig. 2008. The Mediatization of Society. A Theory of the Media as Agents of Social and Cultural Change. *Nordicom Review*, *29*(2), 105–134.

Hjarvard, Stig. 2013. *The Mediatization of Culture and Society*. Oxon: Routledge.

Hsieh, Hsiu-Fang and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Johnson, Clay A. 2011. *The Information Diet: A Case for Conscious Consumption*. Beijing, Cambridge, Tokyo: O'Reilly.

Kesebir, Pelin and Selin Kesebir. 2012. The Cultural Salience of Moral Character and Virtue Declined in Twentieth Century America. *Journal of Positive Psychology*, *7*(6), 471–480.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. London: Sage.

Kumar, Nitu and Manish Sahu. 2010. The Evolution of Marketing History: a Peek Through Google Ngram Viewer. *Asian Journal Of Management Research*, *1*, 415–426.

Lakoff, Robin. 2013. *What Words Don't Tell Us*. Retrieved May 20, 2014 from http://blogs.berkeley.edu/author/rlakoff/

LaRossa, Ralph, Gordon, Betty A., Wilson, Ronald J., Bairan, Annette and Charles Jaret. 1991. The Fluctuating Image of the 20th Century American Father. *Journal of Marriage and Family*, *53*(4), 987–997.

Lilleker, Darren. 2008. *Key Concepts in Political Communications*. London: SAGE

Lucier, Paul. 2012. The Origins of Pure and Applied Science in Gilded Age America. *ISIS*, *103*(3), 527–536.

Mazzoleni, Gianpietro and Winfried Schulz. 1999. "Mediatization" of Politics: A Challenge for Democracy? *Political Communication*, *16*(3), 247–261.

Michalski, Brian, Krishnamoorthy, Mukkai and Tsz-Yam Lau. 2012. *Temporal Analysis of Literary and Programming Prose*. Retrieved September 23, 2014 from Cornell University Library http://arxiv.org/pdf/1202.2131.pdf

Michel, Jean-Baptiste, Shen, Yuan Kui, Aiden, Aviva P., Veres, Adrian, Gray, Matthew K., The Google Books Team, Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A. Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182.

Mowery, David C. and Nathan Rosenberg. 1998. *Paths of Innovation: Technological Change in 20th-Century America*. Cambridge: Cambridge University Press.

Murray, Denise E. 2000. Protean Communication: The Language of Computer-Mediated Communication. *TESOL Quarterly*, *34*, 397–421. doi: 10.2307/3587737

Oishi, Shigehiro, Graham, Jesse, Kesebir, Selin and Iolanda C. Galinha. 2013. Concepts of happiness across time and cultures. *Personality and Social Psychology Bulletin*, *39*(5), 559–577.

Ong, Walter J. 2002. *Orality and Literacy: The Technologizing of the Word*. London, New York: Routledge.

Phani, Shanta, Lahiri, Shibamouli and Arindam Biswas. 2012. Culturomics on a Bengali Newspaper Corpus. *International Conference on Asian Language Processing*, 237–240. doi: 10.1109/IALP.2012.68

Roseneil, Sasha and Shelley Budgeon. Cultures of Intimacy and Care beyond 'the Family': Personal Life and Social Change in the Early 21st Century. *Current Sociology*, *52*(2), 135–159.

Rutten, Ellen, Fedor, Julie and Vera Zvereva. 2013. *Memory, Conflict and Social Media*. Abingdon: Routledge.

Schoen, Robert and Vladimir Canudas-Romo. 2006. Timing Effects on Divorce: 20th Century Experience in the United States. *Journal of Marriage and Family*, *68*, 749–758. doi: 10.1111/j.1741-3737.2006.00287

Stemler, Steve. 2001. An Overview of Content Analysis. *Practical Assessment, Research & Evaluation*, *7*(17). 137–146.

Thurlow, Crispin, Lengel, Laura and Alice Tomic. 2004. *Computer Mediated Communication*. London, New Delhi, London: Sage.

Ullmann, Stephen. 1962. *Semantics: an Introduction to the Science of Meaning*. Blackwell: Oxford.

Volti, Rudi. 1988. *Society and Technological Change*. New York: St. Martin 's Press.

Weber, Robert P. (ed.). 1990. *Basic Content Analysis*. London, New Delhi, London: Sage.

Wellman, Barry, Quan-Haase, Anabel, Boase, Jeffrey, Chen, Wenhong, Hampton, Keith, Díaz, Isabel and Kakuko Miyata. 2003. The Social Affordances of the Internet for Networked Individualism. *Journal of Computer-Mediated Communication*, *8*. doi: 10.1111/j.1083-6101.2003.tb00216

Wierzchoń, Piotr. 2008. *Fotodokumentacja, chronologizacja, emendacja: teoria i praktyka weryfikacji materiału leksykalnego w badaniach lingwistycznych*. [*Photo-documentation, chronologization, emendation: theory and practice of lexical material verification in linguistic studies*] Poznań: Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza.

Wood, Andrew F. and Matthew J. Smith. 2005. *Online Communication: Linking Technology, Identity, and Culture (Second Ed.)*. Mahwah, NJ: Lawrence Erlbaum & Associates.