Irish Journal of Psychological Medicine (2019), **36**, 23–27.
doi:10.1017/ipm.2017.37

SHORT REPORT

Culturomics and the history of psychiatry: testing the Google Ngram method

O. P. O'Sullivan[1,*], R. M. Duffy[2] and B. D. Kelly[2]

[1] *National Forensic Mental Health Service, Central Mental Hospital, Dundrum, Dublin, Ireland*
[2] *Department of Psychiatry, Trinity College Dublin, Trinity Centre for Health Sciences, Tallaght Hospital, Dublin, Ireland*

Objectives. Culturomics is the study of behaviour and culture through quantitative analysis of digitised text. We aimed to apply a modern technique in this field to examine trends related to the history of psychiatry. In doing so, we aimed to explore the nature of the Google Ngram methodology.

Methods. Using Google Ngram Viewer, we studied Google's corpus of over 4% of all published books and explored relevant trends in word usage.

Results. An exponential growth in the use of 'psychiatry' between 1890 and 1984 was identified. 'Sigmund Freud' was mentioned more frequently than all other prominent figures in the history of psychiatry combined. Mentions of 'suicide' increased since 1820. The impact of several DSM editions is discussed.

Conclusion. This study demonstrated the potential application of the Ngram methodology to the study of the history of psychiatry. The role of textual analysis in this field merits careful, constructive consideration and is likely to expand with technological advances.

Received 3 March 2017; Revised 29 May 2017; Accepted 30 June 2017; First published online 17 August 2017

Key words: Culturomics, history, internet, psychiatry, psychoanalysis.

Introduction

In psychiatry, more than any other area of medicine, language matters deeply. It is central to all aspects of practice, from the symptom-based definitions of mental disorders to 'talking therapies'. Culturomics is the use of software to analyse the written lexicon of a society, following trends in language over time. It provides a lens through which linguistic and cultural phenomena are observed (Michel *et al*. 2011). The Ngram technique has been used recently to illustrate the evolution of scientific writing. In an analysis of PubMed abstracts between 1974 and 2014 (Vinkers *et al*. 2015) a drastic and disproportionate increase of positive words – such as, 'robust', 'unprecedented' – relative to negative was observed over four decades. It was demonstrated quantitatively that scientific abstracts are now written using more definitive terms.

We applied a culturomic method to the history of psychiatry, observing the popularity of words and phrases in written text over time. This research was made possible by the 'Google corpus' of published works at http://books.google.com/Ngrams. Google took books from libraries and publishers around the world, scanned each page, and identified each word. Metadata for each book note when and where it was published, and whether it is fact or fiction. The initial Google corpus created in 2009 included over five million books. The English corpus included over 360 billion words and represented 4% of books ever printed (Michel *et al*. 2011). This was further expanded in 2012.

The books in the Google corpus are not a random sample, but were selected on the basis of the quality of the metadata and digitised text. Central to consistent and reliable digitisation is high-quality optical character recognition (OCR), that is how the pixels of a scanned book are converted into text. Naturally, the older a text the less likely this process is to be reliable which may affect the text's inclusion. The 2012 corpus has a number of advantages over its 2009 predecessor, including improved metadata, better OCR and analysis of phrases across page boundaries. The 2012 corpus includes books published between 1500 and 2008. Google's software allows the user to plot the frequency of a word or phrase as a percentage of all words published that year (i.e. data are normalised by the total number of words in the corpus that year). This resultant statistic is expressed as a percentage of all words in the corpus for that year and is plotted on the y-axis, with time in years on the x-axis, yielding a 'Google Ngram view'.

Within the Google corpus there are a number of sub-corpora including American English, British English,

* Address for correspondence: Dr O. P. O'Sullivan, National Forensic Mental Health Service, Central Mental Hospital, Dundrum, Dublin 14, Ireland.
(Email: owenosullivan@rcsi.ie)

English, English Fiction, Chinese, French, German, Hebrew, Spanish, Russian and Italian. Books are included in the American English corpus if they were published in the United States and in the British English corpus if published in the United Kingdom. Books are included in the English fiction corpus if a library or publisher identifies them as fiction.

Google Ngram view has been used to look at diverse topics including: astrology and phrenology (Genovese, 2015); the psychology of culture (Greenfield, 2013); stereotypes about age (Mason *et al*. 2015); and the 'Spanish' flu (Phillips, 2014).

We used Google Ngram Viewer to analyse trends over the last 500 years relating to psychiatry and societies' relationship with it. We looked specifically at the frequency of use of particular words and phrases related to psychiatry, famous figures in its history, in addition to psychiatry in fiction. We aimed to apply this new technique in order to start a discussion about its potential contributions to the historiography of psychiatry.

Methods

We analysed the English 2012 Google corpus, using UK and US samples. We also analysed the English fiction corpus for certain terms because previous research has highlighted that this may be the most accurate reflection of societies' usage of language, and tends to be least skewed by the inclusion of scientific texts since the turn of the 20th century (Brysbaert *et al*. 2011; Pechenick *et al*. 2015). Occasionally, we searched through additional languages, as outlined in the relevant sections. Unless otherwise stated, we used 'case-insensitive analysis', meaning that the analysis ignored whether letters were upper or lower case.

We used 'smoothing' to make graphs clearer. Unless otherwise stated, we used a smoothing factor of two, meaning that the word-count for any given year is the average of that year and the two years before and after it (similar to a moving average). We did not use smoothing when looking at the first recorded use of a word. To examine long-term trends, we used higher levels of smoothing.

Where multiple terms, spellings or, variations could be used for the same phrase (e.g. 'DSM 1', 'DSM-1', 'DSM-I', etc.), we did an initial search of all possible terms and analysed the one that was most frequently used. We drew a sample of prominent figures in the history of psychiatry from two historical texts (Shorter, 1997; Lieberman and Ogas, 2015). We used full names to minimise spurious findings due to individuals with the same names. We made exceptions for 'R. D. Laing' and 'C. G. Jung' as their names written as shown were used more commonly than their full names. The impact of DSM-5 (APA, 2013) could not be examined as the corpus included in this study only went as far as 2008. Mentions of the World Health Organisation's (1992) *International Classification of Mental and Behavioural Disorders (Volume 10)* could not be examined either, as its acronym (ICD) has too many alternative meanings (e.g. implantable cardioverter defibrillator).

Results

'Psychiatry'

We found that 'psychiatry' first appeared in the English corpus in 1689 and featured only five times before 1800. All of these pre-1800 mentions of 'psychiatry' occurred in the US corpus rather than the UK one. It was as late as 1870 before 'psychiatry' had an annual place in the US corpus and not until 1882 that 'psychiatry' first appeared in the English fiction corpus. It was a further 36 years before it was consistently present in the lexicon of English fiction, in 1918. Figure 1 shows trends relating to the terms 'insane', 'lunatic', 'asylums' and 'alienists'. Trends are broadly similar for all of these terms, as certain terms grew in popularity and then declined, to be replaced by others.

Analysis of all English writing in the corpus shows exponential growth in the use of 'psychiatry' between 1890 and 1984. It peaked in 1984 at 165.44×10^{-5}% of all words used. There was also huge growth in the use of 'psychiatry' in English fiction during the 20th century, from 0.11×10^{-5}% in 1900 to a peak of 24.76×10^{-5}% in 1975. Since then, there has been a reduction to 9.40×10^{-5}% in 2008.

Prominent figures in the history of psychiatry

Results are shown in Table 1. Increasing the smoothing factor to 50 allowed us to measure the influence of each figure over the 50 and 100-year periods leading up to 2008. In the 50 years preceding 2008, 'Sigmund Freud' accounted for 11.97×10^{-5}% of all two-word pairs used in English. In the 100 years leading up to 2008, he accounted for 7.29×10^{-5}% of all two-word pairs. C. G. Jung, by way of comparison, accounted for just 2.11×10^{-5}% of all two-word pairs in the 50 years leading up to 2008, and 1.12×10^{-5}% in the 100 years leading up to 2008. Overall, 'Sigmund Freud' was mentioned more frequently than all the other historical figures mentioned in Table 1 combined.

Diagnostic and Statistical Manual of Mental Disorders (DSM)

'DSM-I' was by far the most commonly used, reaching its peak in 1972. 'DSM-I' did not appear in the fiction corpus until 1990, some 38 years after it was published, and does not feature strongly in fiction at any point.

https://doi.org/10.1017/ipm.2017.37 Published online by Cambridge University Press

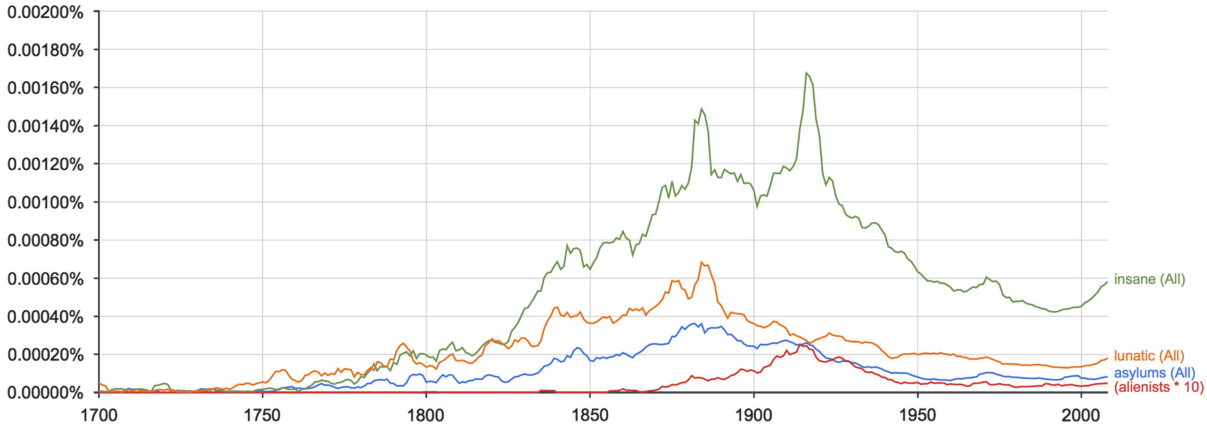

Figure 1. Percentage of words that 'insane', 'lunatic', 'asylums' and 'alienists' account for in the Google English corpus (1700–2008). *Vertical axis:* Percentage of words that 'insane', 'lunatic', 'asylums' and 'alienists' (as indicated) account for in the Google English corpus. *Horizontal axis*: Year. *Note*: The term 'alienists' was not commonly used; its trend line in this Google Ngram is multiplied by 10, which was necessary in order to make the line visible and thus demonstrate the trend, but it means that this trend line is not comparable with the others in terms of magnitude.

Table 1. *Frequency of occurrence of the names of prominent figures in the history of psychiatry in the 2012 Google corpus of published work in English*

Name	Year most written about	Percentage in year most written about $\times 10^{-5}$	Period of most rapid growth	Highest usage in fiction $\times 10^{-5}$	Year most used in fiction corpus
Sigmund Freud	1995	15.19	1938–1974	8.30	1997
C. G. Jung	1976	4.60	1941–1955	3.15	1973
R. D. Laing	1975	2.11	1969–1972	1.66	1975
Thomas Szasz	1978	0.88	1961–1974	0.27	1977
Adolf Meyer	1942	1.40	1932–1937	1.15	1907
Emil Kraepelin	2001	0.33	1940–1943	0.07	1997
Aaron Beck	2004	0.40	1987–1995	0.03	1987
Elisabeth Kübler-Ross	2008	0.11	2002–2008	0.02	2002
Philippe Pinel	1978	0.31	1925–1935	0.07	1956
Viktor Frankl	2007	0.63	1960–1973	0.43	1965
Eugen Bleuler	1972	0.30	1947–1959	0.26	1969
Scott Peck	1995	0.78	1981–1993	0.25	1999
Milton Erickson	1987	0.43	1972–1983	0.06	1988
Alois Alzheimer	2001	0.24	1981–2001	0.04	2005
Karl Jaspers	1967	2.01	1948–1966	0.49	1968
Jacques Lacan	1998	4.63	1983–1992	2.63	1996
Kurt Schneider	1933	0.14	1928–1933	n/a	n/a

The findings regarding 'DSM-I' must, however, be interpreted with caution because DSM-I was not, of course, known as 'DSM-I' at the time: it was simply 'DSM'. Unfortunately, searching for 'DSM' in the Google corpus yields *all* references to 'DSM', 'DSM-I', 'DSM-II', etc., with the result that it is not possible to use a search for 'DSM' just on its own to draw any conclusions about the first edition.

It is relatively easier to search for 'DSM-II' which, from its year of publication (1968), appears consistently in the Google corpus and was most frequently used in 1978, closely followed by 1981 and 1976. 'DSM-II' was surpassed by 'DSM-III' a full year before the latter was published in 1980, at which point 'DSM-III' was more commonly used than 'DSM-I' or 'DSM-II' ever were (Table 2). Even in 2008, use of 'DSM-III' still surpassed both 'DSM-I' and 'DSM-II' by a factor greater than 10. This reflects the particular impact of DSM-III, as described by Lieberman and Ogas (2015). In due course, the impact of DSM-III was matched by DSM-IV, which, at its peak, also accounted for just over 18×10^{-5}% of words published (see Table 2).

https://doi.org/10.1017/ipm.2017.37 Published online by Cambridge University Press

Table 2. *Occurrences of 'DSM' in the Google corpus of published works in English*

Edition	DSM-I	DSM-II	DSM-III	DSM-IV
Year of publication	1952	1968	1980	1995
Year of most frequent use	1972	1978	1991	2004
Percentage of all words in the corpus in year of most frequent use ($\times 10^{-5}$)	0.88	1.48	18.33	18.19
Years until consistently below 50% of peak	1975	1989	2000	n/a
Percentage usage in 2008 ($\times 10^{-5}$)	0.17	0.28	2.62	12.77
Year its use surpassed predecessor	n/a	1970	1979	1996

Suicide

The first time that suicide was mentioned in the English corpus was 1563. Between 1563 and 1698 there were just 11 years when it appeared in the corpus. From 1698 to 1750 it featured in small numbers but regularly, and from 1750 onwards it appeared on an annual basis. The use of the word 'suicide' has been increasing since 1820 and the rate of this increase accelerated since the 1920s. It reached a peak in 2005 when it accounted for 191.73×10^{-5}% of all words published in the corpus that year. In 2008, the final year in this study, it accounted for 184.56×10^{-5}% of words used in the English corpus. In the fiction corpus, use of the word 'suicide' has been decreasing steadily since the mid-1970s.

Discussion

The primary purpose of this paper was to apply the Google Ngram technique to the study of the history of psychiatry. We hoped that this would speak on a broader level to the potential of using 'big data' methodologies in the field of medical humanities and open a discussion about the nature and potential of future similar applications.

Lieberman and Ogas (2015) discuss how North American psychiatry was heavily influenced by Sigmund Freud, while European psychiatry followed a more biological route. Comparing the UK corpus with the US one can see this pattern very clearly. Between 1940 and 1998, 'Sigmund Freud' was substantially more frequently used in the United States compared with the United Kingdom. In 1930, at the height of the disparity, 'Sigmund Freud' was used 4.23 times more frequently in the US corpus than the UK one. By 1999, however, references to 'Sigmund Freud' in the UK corpus surpassed those in the US corpus.

The relatively greater popularity of Freud in the United States is closely linked with the history of the Jewish people in the early 1900s (Shorter, 1997) and, as already noted, the Google corpus duly demonstrates the increasing popularity of 'Sigmund Freud' in the United States in the early 1900s. In addition, however, comparison of the French, British, American and Spanish corpora with the German and Italian ones, demonstrates a marked paucity of references to 'Sigmund Freud' in the latter two countries: between 1930 and 1940, use of 'Sigmund Freud' increased in the French, Spanish, British and American corpora, but decreased in Germany and Italy.

The impact of DSM is further illustrated by the occurrences of the name of 'Robert Spitzer', a leading figure in the development of DSM (Shorter, 1997). In 1981, following publication of DSM-III in 1980, 'Robert Spitzer' accounted for 0.20990×10^{-5} of all two-word pairs in the corpus (no smoothing used). To put this in context, 1981 was the year in which the single *Endless Love* by Lionel Richie and Diana Ross was released, and in that year 'Robert Spitzer' had almost eight times as many references as 'Lionel Richie' (0.027×10^{-5}%) although not as many as 'Diana Ross' (0.39×10^{-5}%). Since then, references to 'Robert Spitzer' have been relatively constant at between 0.04 and 0.14×10^{-5}%.

It was 1587 before suicide first appeared in the fiction corpus and it only became a permanent feature from 1785. This was 11 years after the publication of Goethe's *Die Leiden des jungen Werthers (The Sorrows of Young Werther)* (1774), in which the protagonist finds himself in a hopeless love triangle ending in his suicide. This was Goethe's first major success and came to be associated with copycat suicides as fans reportedly over-identifying with the work are said to have taken their own lives by the same means giving rise to the term 'Werther effect' (i.e. copycat suicides) (Hittner, 2005). This novel may have stimulated popular interest in suicide.

This paper has several strengths. We used a new analytic technique to study a vast body of published material. While absolute figures can be difficult to interpret or contextualise, comparative statistics provide valuable information. On this basis, we contextualised mentions of 'Robert Spitzer' through a contemporary cultural comparison, producing notably surprising results reflecting the extraordinary magnitude of the debate surrounding DSM-III.

https://doi.org/10.1017/ipm.2017.37 Published online by Cambridge University Press

This paper has a number of limitations. Some of these relate to the Google Ngram methodology itself. Pechenick *et al*. (2015) offer a comprehensive review on this subject. For example, each appearance of a given word in the corpus is given equal weight, so the appearance of a word in a text that was read by 10 people is given the same weight as its appearance in a best-selling book that was read by millions. In addition, the inclusion in the corpus of scientific texts which have proliferated greatly since the 1900s means that the corpus is arguably over-influenced by this material, making it more difficult to reach conclusions about non-scientific terms. Analysis of the English fiction corpus (as outlined in parts of this paper) can help avoid some of these problems (Brysbaert *et al*. 2011; Pechenick *et al*. 2015). The arbitrary selection of texts included in the Google corpus (based on technical quality rather than popularity) is another factor, although the inclusion of over 4% of all books ever printed (Michel *et al*. 2011) still makes the Google corpus vastly greater than any other repository. We chose terms purposively, selecting terms that appeared to us to be important in the history of psychiatry. Furthermore, we compared groups of terms that were less likely to have de-contextualised uses. For example, there is a difficulty in using the Ngram method to compare relative frequencies of the words 'depression' and 'schizophrenia' as 'depression' is used in many contexts outside of mental health. Future studies may benefit from the input of a psycholinguist.

Conclusions

The analysis of the Google corpus offers particular possibilities to both clinical psychiatry and study of the discipline's history. The Ngram approach represents an interesting and provocative methodology which would benefit from further technical advances in the coming years but which also requires careful interpretive thought from an historiographical perspective if its possibilities are to be realised appropriately and in full.

Acknowledgements

None.

Financial Support

None.

Conflicts of Interest

None.

Ethical Standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committee on human experimentation with the Helsinki Declaration of 1975, as revised in 2008.

References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition) (DSM-5)*. American Psychiatric Association: Washington, DC.

Brysbaert M, Keuleers E, New B (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology* **2**, 27.

Genovese JE (2015). Interest in astrology and phrenology over two centuries: a Google Ngram study. *Psychological Reports* **117**, 940–943.

Goethe JW von (1774). *Die Leiden des jungen Werthers (The Sorrows of Young Werther)*. Weygand'sche Buchhandlung: Leipzig.

Greenfield PM (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science* **24**, 1722–1731.

Hittner JB (2005). How robust is the Werther effect? A re-examination of the suggestion-imitation model of suicide. *Mortality* **10**, 193–200.

Lieberman JA, Ogas O (2015). *Shrinks: The Untold Story of Psychiatry*. Weidenfeld & Nicolson: London.

Mason SE, Kuntz CV, McGill CM (2015). Oldsters and Ngrams: age stereotypes across time. *Psychological Reports* **116**, 324–329.

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL (2011). Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182.

Pechenick EA, Danforth CM, Dodds PS (2015). Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One* **10**, e0137041.

Phillips H (2014). The recent wave of 'Spanish' flu historiography. *Social History of Medicine* **27**, 789–808.

Shorter E (1997). *A History of Psychiatry: From the Era of the Asylum to the Age of Prozac*. John Wiley & Sons: New York.

Vinkers CH, Tijdink JK, Otte WM (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *British Medical Journal* **351**, h6467.

World Health Organisation (1992). *International Classification of Mental and Behavioural Disorders*, Vol. 10. World Health Organisation: Geneva.

https://doi.org/10.1017/ipm.2017.37 Published online by Cambridge University Press