Corpus linguistics, newspaper archives and historical research methods

Chinmay Tumbe
Economics Area, Indian Institute of Management, Ahmedabad, India

Received 31 January 2018
Revised 10 July 2018
4 November 2018
12 February 2019
Accepted 27 February 2019

Abstract

Purpose – The purpose of this paper is to demonstrate the utility of corpus linguistics and digitised newspaper archives in management and organisational history.

Design/methodology/approach – The paper draws its inferences from Google NGram Viewer and five digitised historical newspaper databases – The Times of India, The Financial Times, The Economist, The New York Times and The Wall Street Journal – that contain prints from the nineteenth century.

Findings – The paper argues that corpus linguistics or the quantitative and qualitative analysis of large-scale real-world machine-readable text can be an important method of historical research in management studies, especially for discourse analysis. It shows how this method can be fruitfully used for research in management and organisational history, using term count and cluster analysis. In particular, historical databases of digitised newspapers serve as important corpora to understand the evolution of specific words and concepts. Corpus linguistics using newspaper archives can potentially serve as a method for periodisation and triangulation in corporate, analytically structured and serial histories and also foster cross-country comparisons in the evolution of management concepts.

Research limitations/implications – The paper also shows the limitation of the research method and potential robustness checks while using the method.

Practical implications – Findings of this paper can stimulate new ways of conducting research in management history.

Originality/value – The paper for the first time introduces corpus linguistics as a research method in management history.

Keywords Management history, Corpus linguistics, Text analysis, Research methodology, Newspaper

Paper type Research paper

Introduction

With the advent of mass digitisation of historical prints over the past two decades, researchers now have access to a unique but under-appreciated source of data – Words. Billions of words have entered online repositories as hundreds of millions of pages of old books and newspapers have been digitised by firms and organisations. Words in these real-world texts form a "corpus" and corpus linguistics refers to a branch of linguistic studies that systematically analyses them (Biber *et al.*, 1998; Oakes, 1998; McEnery and Wilson, 2001). At its simplest, it can be defined as a "methodology that uses computer support – in particular, software called 'concordance programs'- to analyse authentic, and usually very large, volumes of textual data" (Mautner, 2009, p. 122). Alternatively, it may be defined as "the study of language data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts" (McEnery and Hardie, 2012). Both these definitions highlight the existence of a corpus of text that is sufficiently large in scale and machine-readable in nature. Scale is important to differentiate it from standard

Journal of Management History
Vol. 25 No. 4, 2019
pp. 533-549

1751-1348
DOI 10.1108/JMH-01-2018-0009

qualitative research software used which is used to analyse transcribed interviews. Further, concordance software in corpus linguistics can aid both quantitative and qualitative research.

Corpus linguistics grew substantially in the 1990s with a spate of publications by linguists and is now an important branch of linguistics with dedicated journals, associations, research groups and conferences. It is slowly being applied to legal studies (Cotterill, 2001; Vogel *et al.*, 2018) and political studies (Baker and McEnery, 2005; Fairclough, 2000). One study in management research compared letters to shareholders in two different years using corpus linguistics techniques to draw novel inferences (Pollach, 2012). The rising popularity of corpus linguistics can be attributed to the growing attraction of "data science" and "Big Data" research methods across disciplines. These methods essentially seek to unravel trends and patterns using large bodies of data and are used in a variety of for-profit organisations. They have their own set of limitations when the methods are used without adequate theory or applied with little understanding of the context in which the data was produced. Despite this, they are slowly being used for academic research, as exemplified in a special issue in 2018 in *Organisational Research Methods* devoted to "Big Data and Modern Data Analytics". When the basic form of data is textual in nature, text mining techniques analyse the properties of the text such as structure and grammar through machine learning, statistics, computational linguistics, natural language processing and corpus linguistics (Kobayashi *et al.*, 2018). The evolution of corpus linguistics as a method has to be seen against this backdrop of a revolution in computing power, rapidly proliferating text mining techniques and the availability of readily accessible corpora of billions of words.

Corpus linguistics is now weaving its way into historical research. A recent study, based on collaborative work between a historian and a linguist, unearthed new evidence on prostitution in seventeenth century England (McEnery and Baker, 2016). Their work approached the matter separately from the historian's and linguists perspective and then in combination, to generate new insights and hypotheses. While their corpora include books and texts, linguists have also begun exploring the potential of unravelling patterns in linguistic evolution using digitised newspaper databases (Westin and Geisler, 2002; Fries and Lehmann, 2006; Bamford *et al.*, 2013; Buntinx *et al.*, 2017).

Corpus linguistics is also being used to contribute to discourse analysis (Baker, 2006; Mautner, 2007; Orpin, 2005). Discourse analysis refers to "the analysis of language as it is used to enact activities, perspectives, and identities" (Gee, 1999, pp. 4-5). It is interested in naturally occurring language, studies the context in which the language is used and can also include non-verbal forms of interactions such as images and gestures (Wodak and Meyer, 2009). The word "discourse" itself is now part of mainstream media reflected in the usage of phrases such as the discourses on racism, caste, gender, populism, globalisation and other phenomena which have particular vocabularies attached to them. Discourse analysis can identify loaded words and unravel different perspectives on commonly used words. From a historian's perspective, it can also identify *when* certain words belong to or drop out of specific discourses. Corpus linguistics enables discourse analysis to work with a much larger quantum of data, reduces researcher bias and provides qualitative and quantitative insights on texts (Mautner, 2009). Apart from frequencies and statistical significance tests, it can qualitatively examine semantic patterns and collocational environments.

This paper argues that there is considerable scope to apply corpus linguistics to research in management and organisational history through discourse analysis, especially when the corpus consists of digitised historical newspaper databases. Firms that provide these databases now provide a number of research tools that enable historical analysis without

any additional software requirements. They can provide insights even when the entire corpus is not available to researchers in its rawest form.

The rest of the paper is arranged as follows: The next section describes the standard methods of research in corpus linguistics and the significance of digitised newspaper archives as major corpora. This is followed by an exposition of two of the most commonly used methods in corpus linguistics – term count and cluster analysis, using management concepts and newspaper archives. The limitations of the method and potential robustness checks of the method are then highlighted. This is followed by a discussion on how corpus linguistics can add to the theory and practice of management and organisational history and a final concluding section.

Corpus linguistics research methods

Types of corpora and the significance of newspaper archives

A corpus is defined as "a body of text which is carefully sampled to be maximally representative of a language or language variety" (McEnery and Wilson, 2001, p. 2). It should also be machine-readable. Corpora are available at different scales, broadly categorised as "do-it-yourself corpora and "reference corpora" (Mautner, 2009). In the first instance, researchers themselves assemble corpora from various sources. Once a corpus has been built, freely available concordance programs or software for corpus linguistics such as Antconc and Wordsmith Tools are relatively easy to use for analysis with minimal technical requirements. This is useful to address small-scale research problems.

Reference corpora refer to readily available corpora assembled by research teams over several years. Linguists have compiled extensive "corpora" or body of texts for the analysis of specific languages (Kennedy, 1998). They include databases such as the British National Corpus that contains millions of words. In 2009-2010, Google launched its freely accessible NGram Viewer enabling keyword search through millions of books it had digitised, written between 1500 and 2008. The NGram Viewer provides a simple graph as the output on the popularity of the word over time, if and only if the word appeared in at least 40 books. The simplicity of the search led to the growth of a field called "culturomics" that compared the popularity of different words and attracted research attention in reputed journals such as *Science* (Michel *et al.*, 2011). However, researchers have also outlined the pitfalls of using the Google books database, as results can be influenced by prolific authors and further, the nature of the corpus changes substantially over time with more publication of scientific literature in recent decades that uses different words (Pechenick *et al.*, 2015). For instance, the popularity of the word "autumn" as per Google Ngram declines over the twentieth century, but it is almost certain that this has more to do with the rising stock of scientific books in the corpus which refer less to "autumn" than any meaningful change in the day-to-day popularity of the word.

In contrast, newspaper archives can be superior to historical book databases, as they are more standardised and less affected by prolific writers. They arguably capture the *zeitgeist* or the spirit of the times better than hundreds of books written on different themes. So far, newspapers have been an important source of data in management and organisational history research to locate events, advertisements and triangulate narratives with other data sources. The advent of digitised newspaper databases provides a new dimension to existing archive-room research practices as software-based quantitative and qualitative research methods can now be used to discern patterns in linguistic data. Traditional methods would entail collecting or downloading historical news items and coding them into fields relevant for the research project. However, certain historical newspaper databases now provide

research tools that enable term count and cluster analysis that enable research, without having to manually collect thousands of articles.

There are over a hundred digitised historical newspaper databases around the world today, mostly related with America and Europe, but not all of them are currently amenable to linguistic research. Most databases enable basic and advanced search features but do not necessarily provide research tools. The Google News Archive Project, launched in 2006, was a promising database, but it collapsed mid-way and its future remains uncertain. Wordbanks Online consists of more than 500 million words of mostly American and British text, including a sub-corpus of 60 million words from the articles of *The Times*, a daily British newspaper (Mautner, 2009). ProQuest and Gale Cengage Learning are currently two data providers that have developed research tools and contain a corpus of over 35 and 15 million digitised pages, respectively, covering over 50 English language newspapers in the USA and UK, and also a few in other countries such as China and India.

The large sample size of these newspaper archive databases is ideal for linguistic research. For instance, the ProQuest *Times of India* historical database from 1838 to 2007 covers over seven million news items, of which nearly half are categorised as "articles" and a fourth have been classified as display and classified advertisements. Other items in the database include editorials, stock-quotes, letters to the editor, obituaries, weather notices and so forth. The Gale Cengage *Financial Times* historical database from 1888 to 2010 contains over four million articles. Even the non-daily *Economist* historical database from 1843 to 2014 contains nearly half a million articles. Both these newsprints are based in Britain. In America, the ProQuest historical newspaper collection on the *New York Times* (1851-2015) contains 11 million articles and the *Wall Street Journal* (1889-2001) contains 3 million articles.

Concordance software and analysis

At the heart of corpus linguistics lies the concordance software that can analyse large volumes of textual data. With digitised newspaper archives, it is always possible to download a set of articles and load them onto such a software for further analysis. Concordance software provide quantitative evidence in the form of absolute and relative word frequencies. On the co-occurrence of words, they also provide statistical significance measures such as *t*-scores and Mutual Information (MI) scores. *t*-scores signify the certainty of collocation of two words and MI scores show "whether there is a higher-than-random probability of the two items occurring together" (Mautner, 2009, p. 125).

To illustrate this, we use an example provided by Mautner (2009) for an analysis of the word "unemployed" that occurred in *The Times* database of Wordbanks Online 567 times. Table I shows the *t*-scores and MI scores of the collocates. The top ranked *t*-scores are

Table I.
Illustrative example of concordance software analysis of the word "unemployed"

Collocate	*t*-score	Collocate	MI score
1. An	6.65	1. Steelmen	13.46
2. Are	6.23	2. Househusband	11.78
3. People	5.78	3. Unemployable	11.23
4. Who	5.72	4. Housewives	8.73
5. And	4.84	5. 4m	8.07
6. Term	4.21	6. Youths	7.90

Source: Mautner (2009, Table V.1), based on The Times corpus of Wordbanks Online

grammatical items that are usually not of interest. It measures the certainty of association between two words. The MI scores, however, reveal important information on words that appear more often together than separate. That is, "an" may appear with "unemployed" yielding a high *t*-score but a low MI score because "an" also appears along with other many other words in a sentence. In the particular context of Britain, words like steelman, househusband, housewives and even a particular statistic of "4m" appear tightly associated with "unemployed". Here, corpus linguistics provides a useful way to do discourse analysis, by reducing the bias of selectively using words that researchers deem to be collocated with "unemployed".

Concordance software can also display output amenable for qualitative research. It can show a list of sentences in which the word of interest is collocated with another specified word, thereby yielding inference on whether the word implies a positive or negative connotation or some other attribute, a process known as establishing semantic prosody. Similarly, it can show the flow of words used in conjunction with the key word of interest, a process known as establishing semantic preference. Concordance software can, therefore, identify semantic prosody and preference in qualitative research and *t*-scores, MI scores and absolute and relative frequencies in quantitative research (Mautner, 2009). In the research tools provided by digitised newspaper archives, frequency analysis is known as term-count analysis and analysis of collocation is referred to as cluster analysis. The next section describes the tools of these historical databases and how they can be leveraged in research.

Applications of corpus linguistics using newspaper archives

In the case of certain digitised newspaper archives, it is possible to analyse the textual data with available research tools without loading them on to any particular concordance software. Two methods stand out: term-count analysis and cluster analysis.

Term-count analysis

One of the simplest research methods of corpus linguistics is frequency count of the word or phrase that is being searched. When the database has a time dimension, these frequency counts can attain a distribution over time and enable an analysis in absolute terms. The time frequency could be as short as on a daily basis for daily newspapers and can be aggregated at longer durations such as months, years and decades. Absolute term counts do not reveal a clear trend because the underlying corpus may change its size and nature over time. This is why normalised term counts or relative frequency counts are more insightful, as they account, to a certain extent, for the changes in the underlying corpus.

Before analysing historical newspaper archives, it is worth commenting on the more easily accessible Google NGram Viewer. The Google NGram Viewer output presents a graph with years on the x-axis and the relative frequency count or the number of times the keyword appeared in a year as a percentage of the total corpus. Figure 1 shows a sample output from the NGram Viewer for some of the key actors in management history – "worker", "owner", "manager", "entrepreneur" and "shareholder". It shows that in the nineteenth century, the word "owner" was much more popular than the other words and that its popularity declined over the twentieth century with a brief uptick towards the end of the century. The evolution of the word "manager" shows a different trend, that of a gradual increase between 1860 and 1920, stagnation for four decades and then growth over the past five decades such that it has a higher count that "owner" today. "Worker" follows the same broad trend of the "manager" till 1930, then increases in popularity such that is the most popular of the five keywords between 1950 and 1990. The popularity then declines in the subsequent two decades. Compared to "owner", "manager" and "worker", the two words –

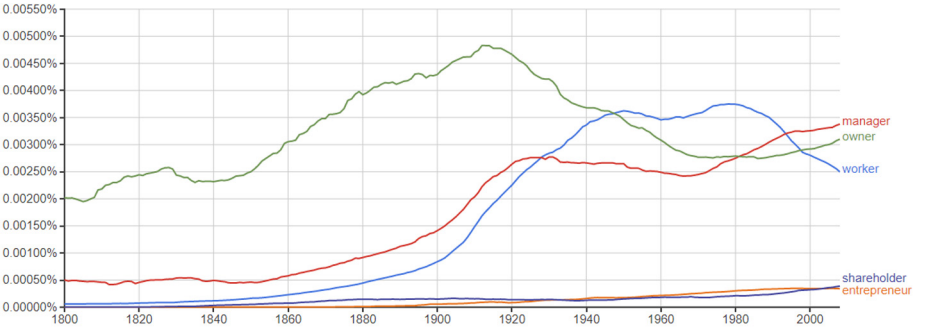


Figure 1.
Google NGram keyword search on key actors in management history

Note: Y-axis shows relative frequency count of keyword in corpus
Source: Google NGram Viewer, Scaling Factor-10, Corpus English (2009)

shareholder and entrepreneur – show less popularity but a rising graph over time. While broad ideological shifts towards labour and capital such as the rise of the welfare state in the 1930s and the move towards liberalised market-based economies in the late 1980s can explain the evolution of the term "worker", the evolution of the "manager" is counter-intuitive and invites further research. That is, if one accepts the Google database to accurately reflect popularity.

We can now perform a similar analysis using newspaper archives. The ProQuest and Gale historical newspaper databases provide a histogram of search hits by time. The histogram itself reveals the number of news items that contain the search keyword and not the number of times the keyword has appeared in totality. That is, if a keyword appears five times in an article, then that article is counted as one hit for the keyword. News items can be filtered by document type such that researchers can focus on "articles" or "advertisements" or the relevant type of document in the research project. Researchers can manually code the frequency counts by looking up the displayed histogram or in the case of Gale Cengage, download the data in spread-sheet format.

Absolute term counts would be misleading as the size of the underlying corpus has changed over time. For instance, Figure 2 shows the rising number of articles published in the *Times of India* and *Financial Times* in most decades over the past century and the brief dip in the 1940s on account of World War II. As a result, a simple keyword search would almost always show a rising graph with a dip in the 1940s. As a measure of normalisation or relative frequency, one can therefore compute the *percentage* of articles or specific news items containing the keyword in a given time period. This statistic when presented over time serves as a powerful way to depict the evolution of its usage, similar to the output shown by the Google NGram Viewer.

Figure 3 mimics the keywords displayed in Figure 1 using a different corpus – the articles published in the Financial Times – and shows some similar as well as conflicting results. The word "Worker" in particular appears to exhibit a similar inverted-U shape pattern in the end of the twentieth century in both databases though the inflection points are slightly different. The word "entrepreneur" has relatively lower usage and exhibits a pickup at the end of the twentieth century in both databases. The word "shareholder" has a much higher usage in the *Financial Times* as may be expected and rises rapidly from the 1970s in both databases. The word "owner" has a much lower usage in the *Financial Times* corpus as compared to the Google Books corpus and shows different evolution trends. The word

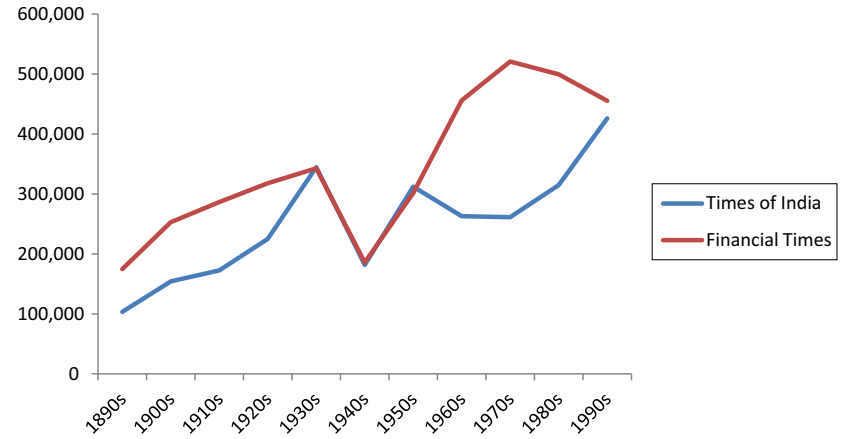


Figure 2.
Total articles published in each decade, 1890-2000

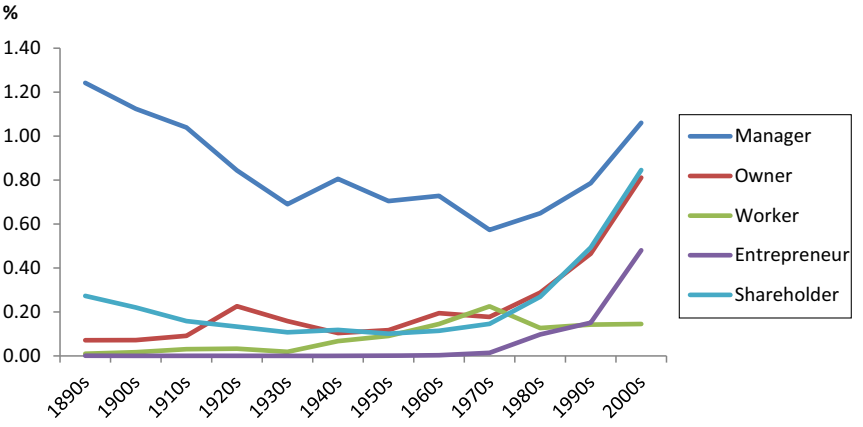


Figure 3.
Percentage of articles in the financial times historical databases containing keywords

Source: Gale Financial Times Historical Database (1888-2008)

"manager" was used extensively in the *Financial Times* before the 1930s as designations of people mentioned in firm's annual statements, classified as articles in the database. Once this practice stopped, we observe a similar trend in both the databases – stagnation till the 1970s and then a take-off. It is instructive that in both databases, the relative frequency count of the "manager" today is highest among the five words.

Figures 4 to 6 show the term-count trend analysis for the word "management" using the *Financial Times*, *Economist*, NY *Times, Wall Street Journal* and Google Books corpora. The figures show broadly similar narratives of "management" picking up from the middle of the twentieth century. The inflection point in the Google database lay in the 1930s, while in the newspaper archives, it lay in the 1960s. Though subsequent decades in the UK saw more usage of the word in the *Financial Times* than in the *Economist*, partly because of a dedicated column on management, the upward trend is unmistakable. The absolute

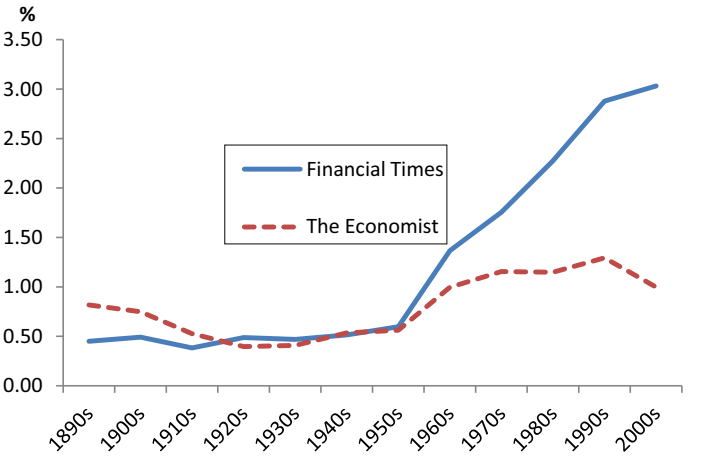


Figure 4.
Percentage of articles in the financial times and economist historical databases containing "Management"

Source: Gale Financial Times and Economist Historical Database

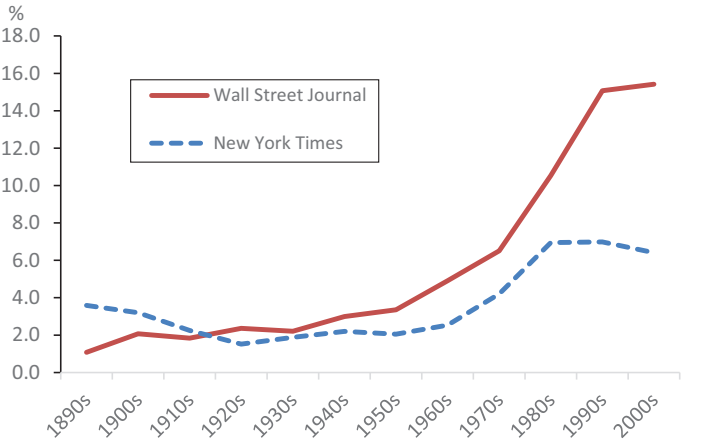
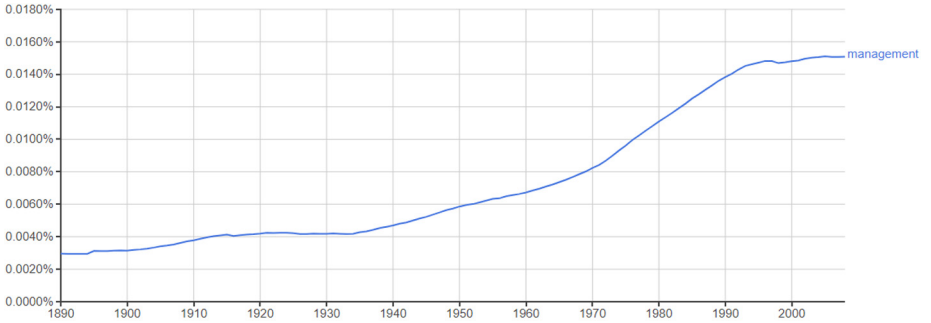


Figure 5.
Percentage of articles in the *Wall Street Journal* and New York Times containing "Management"

Source: ProQuest Historical Newspaper Database of *Wall Street Journal* and New York Times

magnitude has also consistently been much higher in the USA than in the UK with "management" occurring in around 15 per cent of articles in the *Wall Street Journal* compared to only 3 per cent in the *Financial Times* by the beginning of the twenty-first century.

This exercise of comparing management keywords in different corpora is useful to discern trends that may be considered as robust, versus trends that may be superfluous or deserve further scrutiny. Indeed, broadly similar trends in all five databases in the keywords mentioned above after the 1950s appear to be quite striking. Such an exercise is therefore



Note: Y-axis shows relative frequency count of keyword in corpus
Source: Google NGram Viewer, Scaling Factor-10, Corpus English (2009)

Figure 6.
Google NGram keyword search on "Management"

useful to locate broad patterns and relative popularity of words and can serve as an innovative way to understand the evolution of management concepts and its differences across countries. It can ultimately serve the purpose of discourse analysis in management and organisational history.

One application of term-count analysis in management history research has been done in the context of marketing history in India (Tumbe and Ralli, 2018). As Table II shows, that study traced the usage of the following keywords – selling, marketing, advertising and branding – in India using the *Times of India* historical database. It found that "marketing" slowly grew in importance over the twentieth century and finally displaced "selling" as the dominant word used in the literature. It also mapped the evolution of the terms with the economic ideology of the state. Further, the trend analysis discovered a sharp rise in "marketing" in the 1930s that was found to be associated with the establishment of rural marketing boards and marketing officers in the agricultural sector. This opened a new line of research on marketing history and countered the existing literature and intuition which claimed that "marketing" arrived in India in the 1960s through Western-oriented management school curricula. Similarly, by comparing trends on marketing keywords in India and Britain, inferences were gathered on the pace of adoption of modern marketing methods. Term-count trend analysis with multiple databases, therefore, has the potential to foster cross-country research in management and organisational history.

Another similar application of term-count analysis was done on the evolution of retailing in India (Tumbe and Krishnakumar, 2018). The words searched were retail, bazaar, shop, shopkeeper, shopping, consumer cooperative, fair price shop, supermarket and shopping mall. The term-count analysis clearly showed the relationship between state patronage towards particular forms of retail in certain eras – that being the identified discourse – and the relative presence or absence of those words in those eras. It showed the relative decline of the "bazaar" form of retail in India since the late nineteenth century and the constancy of the word "shopkeeper" in the general English-language discourse of consumption in India over more than a century.

Cluster analysis

Equally illuminating is the research tool on cluster analysis provided by the data provider Gale. The tool uses an algorithm to process the subjects, titles and roughly the first hundred top results linked with the keyword. It then provides a visual output that shows words

Decade	% of Articles in the Times of India with the word...				Total Articles ("000)
	Selling	Advertising	Marketing	Branding	
1860s	1.0	0.2	0.0	0.0	15
1870s	1.0	0.3	0.0	0.0	81
1880s	1.4	0.2	0.0	0.0	123
1890s	1.3	0.2	0.0	0.0	103
1900s	1.0	0.2	0.0	0.0	154
1910s	1.2	0.3	0.1	0.0	173
1920s	1.6	0.4	0.3	0.0	225
1930s	3.2	0.2	0.7	0.0	345
1940s	3.1	0.1	0.6	0.0	182
1950s	3.0	0.2	0.8	0.0	312
1960s	2.6	0.3	0.8	0.0	263
1970s	2.1	0.4	1.5	0.0	261
1980s	2.1	0.7	1.8	0.0	315
1990s	2.8	1.3	2.4	0.1	426
2000s	2.3	1.3	2.5	0.2	444

Source: Tumbe and Ralli (2018, Table I)

Table II.
Trends in marketing-related keywords in the Times of India, 1860-2007

closely associated with the keyword that is being searched. It is essentially an exercise in collocation, described earlier.

Figures 7 to 9 show the words closely associated with "management" in the articles of *The Economist* for the periods 1850-1900, 1900-1950 and 1950-2000, respectively. It shows that "bank" was the main word used in articles related with management in the late nineteenth century and its relative decline over the next century. Similarly, sectors like agriculture, mining and railways were no longer associated with "management" in relative terms by the end of the twentieth century. "Meeting" and "Labour" appeared in these charts only in the early twentieth century, while unions appeared in both halves of the twentieth century. "Managers" appeared in the early twentieth century and rose to more prominence by the late twentieth century, while "Board" appears to be a significant term in all three sub-periods. More generally, locations and sectors gave way to words like "business", "managers", "government", "industry" and "market" as key markers of management. Management "Guru" and "Gurus" appeared in the late twentieth century as did negative words such as "bad" and "wrong."

Cluster analysis is useful not only in confirming priors but also revealing new trends. For instance, the Chandlerian paradigm in business history credits the railways as the first major sector where management and managerial skills came forth, but as Figures 7 and 8 suggest, banks were equally important, if not more important, as sites of "management" in Britain.

Limitations

The validity of corpus linguistics is highly dependent on the size and nature of the corpus that is analysed. Newspaper archives are useful because of their large sample and standardisation, but they have their own limitations.

Size and nature of corpus

Larger corpora are preferred to smaller corpora but both can be analysed to inspect robustness and validity. For corpora that are hand collected from newspaper archives,

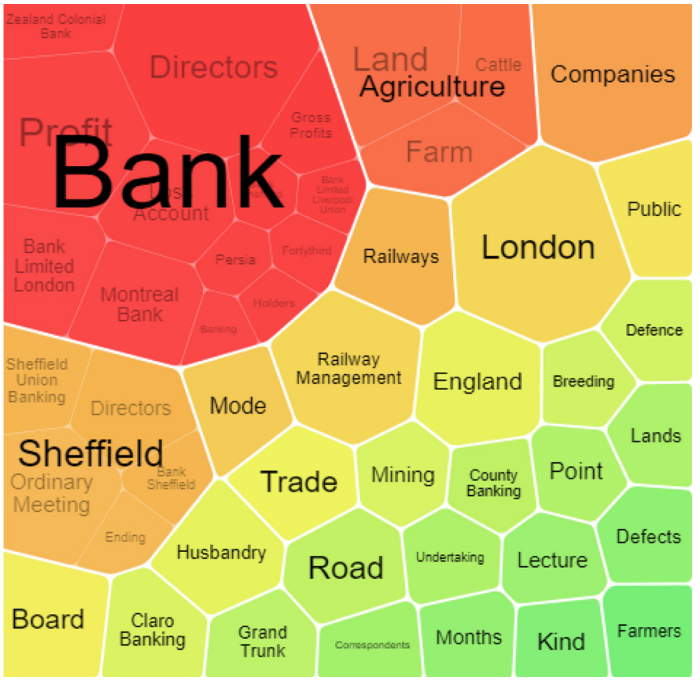


Figure 7.
Cluster analysis of "management" in *The Economist* articles, 1850-1900

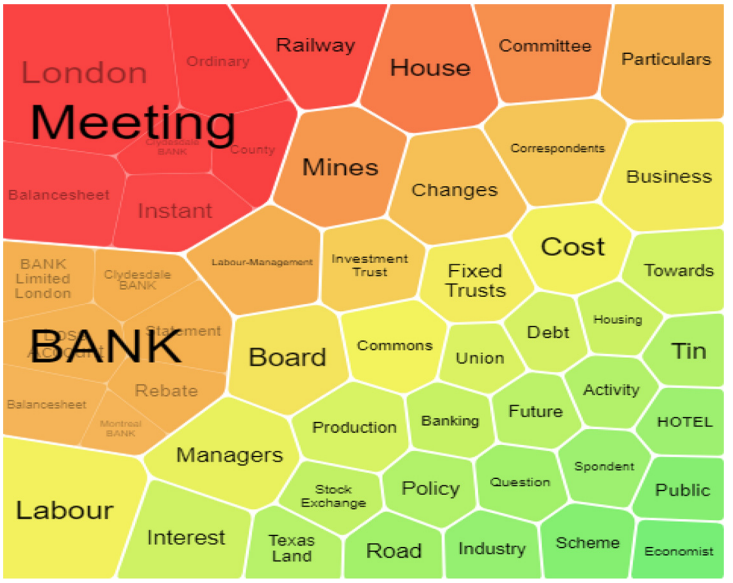


Figure 8.
Cluster analysis of "management" in *The Economist* articles, 1900-1950

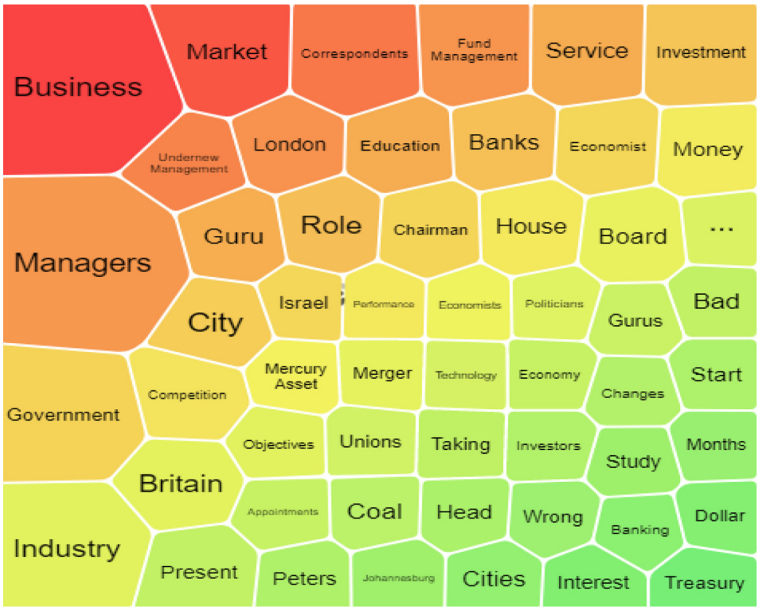


Figure 9.
Cluster analysis of "management" in *The Economist* articles, 1950-2000

researchers need to be aware of the universe from which the texts are being sampled and the representativeness of the underlying sample in terms of language use, dialect, time period and genre. For management and organisational history, the *Economist* and *Financial Times* serve as two excellent databases to inspect validity as they both tend to use similar English and cover topics of interest in the fields of business and management, largely pertaining to Britain. However, the validity of analyses using the *Economist* would be generally lower than that in the *Financial Times* because of its substantially smaller corpus. In choosing the newspaper, representation of issues is also a matter of the researcher's concern. One study on the relationship between political activism and stock prices in the US based on newspaper archives (without corpus linguistics methods), for instance, argued that the New York Times was more representative of the issue than the Wall Street Journal (King and Soule, 2007).

Choice of words

Researchers have to choose one or many of the forms of the keyword – noun, adjective, adverb, etc. In general, if the word form is less popular than other forms, then it tends to hold more validity in corpus linguistics. Words can have multiple meanings and interpretation of trends is also clearer when the word in question has a very restricted meaning. A simple search of "right" and "left" to understand shifts in ideological stances would lead to far too many errors in interpretation as both words can be used in different ways. A measure of robustness would be to inspect variation in trends between search hits on singular and plural forms of the same noun or searching with the first letter capitalised and again in a non-capitalised form. This "plurality" and "capital-letter" test can be used to pare down keywords of interest in a research project.

OCR errors

Digitisation of old prints is usually done through "Optical Character Recognition" (OCR) and there are chances of errors in recognition, especially in old prints. Certain characters may be swapped leading to new meanings, for instance "fin" could be read as "sin" as the "s" and "f" in many older font type faces are similar. As Figure 9 shows, the software erroneously recognises "[. . .]" as a word. A simple robustness check would be to download a small sample of articles containing the keyword in each sub-period and inspect the usage of the keyword to ascertain accuracy.

Corpus linguistics and the theory and practice of management and organisational history

Historical research is based on "distinct epistemic assumptions" that need to be carefully understood by organisational scholars interested in learning from the past (Wadhwani and Bucheli, 2014, p. 3). Over the past decade, several studies have noted the need for more transparency and discussion of historical research methods in management and organisational studies. It has been argued that three basic elements – source criticism, triangulation and hermeneutics – should inform the description of sources used in historical management research (Kipping *et al.*, 2014). Further, four distinct research strategies of organisational history have been pointed out – corporate history, analytically structured history, serial history and ethnographic history – closely tied with dualisms in explanation, evidence and temporality (Rowlinson *et al.*, 2014). Additionally, *periodisation* has been identified as a core function of historical approaches to organisational research (Fear, 2014). As a method of research, corpus linguistics based on newspaper archives is well suited to address several of these issues.

Corpus linguistics can serve as a powerful tool for triangulation of sources in three of the four research strategies identified in organisational history – corporate history, analytically structured history and serial history – with lesser scope for usage in ethnographic history. Corporate history refers to objectivist, holistic narrative accounts of corporate entities and is often built using historical documents pertaining to the firm or a set of firms based on business archives (Rowlinson, 2004). Given a large corpus of text based on documents internal to the firm such as bulletins, memos, letters and reports, corpus linguistics is not entirely different from standard methods used in qualitative research that codify and analyse text. However, in conjunction with the usage of the firm name in newspaper archives, researchers now have a potential tool for triangulation of sources and evidence. For instance, the first mention of the firm name in a major newspaper and the evolution of its relative frequency over time can be an important indicator of the firm's growth or engagement with public relations. If these trends coincide with what researchers have found in documents within the firm, then it serves as an important measure of triangulation. And if it does not, then it merits a closer introspection of which narrative to use, as the historian has to decide "which accounts he or she will use, and why" (Howell and Prevenier, 2001, p. 69).

Analytically structured history refers to "narrating theoretically conceptualised structures and events" (Rowlinson *et al.*, 2014, p. 250) and corpus linguistics, as shown above in the two examples of research on marketing history in India, is useful to understand the broad evolution of key words and the making or unmaking of a particular discourse. One of those studies used corpus linguistics and newspaper archives in conjunction with two corporate archive-based narrative accounts to chart out a robust periodisation of four eras of marketing in twentieth century India (Tumbe and Ralli, 2018). It is in the periodisation of analytically structured history, where the power of newspaper-based corpus linguistics can be well appreciated. Finally, in serial history where facts are repeatable and analysed as

such (as in Anteby and Molnar, 2012), the quantitative dimensions of corpus linguistics can be used across newspaper databases to discern robust trends versus frivolous ones.

Corpus linguistics is also important for "hermeneutics" or relating sources with their original contexts. Discourse analysis for instance has been used in management history to understand how the meaning of words have changed over time (Khaire and Wadhwani, 2010). Through a newspaper based cluster analysis, a set of words or phrases can be mapped out, for the key firm or concept of interest at distinct points of time, to understand how the nature and meaning of the word has changed over time. As the corpus of newspaper texts is much larger than most textual sources available to researchers using corporate of market level data, the method is potentially more robust or, in any case, serves as a means of triangulation of data hand-collected by researchers.

Digital archives of newspapers used for corpus linguistics research should also be subject to "source criticism". Within the newspaper, researchers should observe if the keywords are appearing in particular sections of the newspaper such as editorials or advertisements across different time periods. If, for instance, a word appears more in advertisements in the early twentieth century and almost exclusively in editorials in the late twentieth century, then researchers would have to carefully understand the hermeneutics behind the observed pattern. The newspaper's history should also be carefully studied – Who produced it? What were its stated objectives? Which events did they tend to cover more? The difference between a general newspaper and a financial newspaper, as observed earlier with the New York Times and Wall Street Journal is important to understand the differences observed in corpus linguistics analysis.

Conclusion

Corpus linguistics has been noted to be useful for two types of management and organisation studies (Pollach, 2012). First, lexical patterns can be located, quantified and compared across different and big samples to find similarities, and second, it has been recommended for narrative studies on organisations. However, in conjunction with newspaper archives that provide the added dimensionality of time measured in decades or centuries, it also raises the possibility of creative adaptation in the field of management and organisational history.

Writing good management history requires careful attention to a variety of data sources and multiple and competing narratives and interpretations. It has been likened to the act of a "craft" by a potter at the wheel (Grattan, 2008). This paper introduces a relatively new methodology in the toolkit of the potter or the management historian – corpus linguistics – which offers a promising "Big Data" way of locating trends in management concepts and interpreting management semantics over time. Like any "Big Data" method, it has its share of limitations, guided by the context in which the data was produced and the epistemic assumptions directing interpretations. Cognizant of these limitations, it does have several uses. It can generate a simple time-trend chart during exploratory research and also find important dates or periods where words and phrases gain or lose fashion. Cluster analysis enables a mapping of important words associated with the keyword of analysis. Comparative analysis across newspaper databases can generate new insights on evolution and adoption of management practices in cross-country settings. It can thus be a powerful aid in discourse analysis by not only identifying discourses but by also locating changes in discourses over time. It also provides a means of triangulation and periodisation in corporate, analytically structured and serial histories, though newspapers should be subject to source criticism, especially with respect to hermeneutics.

While much of the existing corpora are in the English language, similar efforts can be pursued by researchers to push for newspaper digitisation and access to research. For example, Buntinx *et al.* (2017) have conducted linguistic analysis of two French newspapers that contain four million articles and two billion words over two hundred years. As this method grows in popularity, it is likely that the number of research tools will expand as researchers ask for more customised solutions. It would also entail more inter-disciplinary connections with the field of linguistics. Management scholars, historians and linguists could unite – they have nothing to lose but their words.

References

Anteby, M. and Molnar, V. (2012), "Collective memory meets organizational identity: remembering to forget in a firm's rhetorical history", *Academy of Management Journal*, Vol. 55 No. 3, pp. 515-540.

Baker, P. (2006), *Using Corpora in Discourse Analysis*, Continuum, London.

Baker, P. and McEnery, T. (2005), "A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts", *Journal of Language and Politics*, Vol. 4 No. 2, pp. 197-226.

Bamford, J., Cavalieri, S. and Diani, G. (2013), "Variation and change in spoken and written discourse: perspectives from corpus linguistics", *Dialogue Studies*, Vol. 21.

Biber, D., Conrad, S. and Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge.

Buntinx, V., Bornet, C. and Kaplan, F. (2017), "Studying linguistic changes over 200 years of newspapers through resilient word analysis", *Frontiers in Digital Humanities*, Vol. 4 No. 2, pp. 1-10.

Cotterill, J. (2001), "Domestic discord, rocky relationships: semantic prosodies in representations of marital violence in the O. J. Simpson trial", *Discourse and Society*, Vol. 12 No. 3, pp. 291-312.

Fairclough, N. (2000), *New Labour, New Language?*, Routledge, London.

Fear, J. (2014), "Mining the past: historicizing organizational learning and change", in Bucheli, M. and Wadhwani, R.D. (Eds), *Organizations in Time: History, Theory, Methods*, Oxford University Press, Oxford, pp. 169-191.

Fries, U. and Lehmann, H.M. (2006), "The style of 18th century English newspapers: lexical diversity", in Brownlees, N. (Ed.), *News Discourse in Early Modern Britain*, Peter Lang, pp. 91-104.

Gee, J.P. (1999), *An Introduction to Discourse Analysis; Theory and Method*, Routledge, London.

Grattan, R. (2008), "Crafting management history", *Journal of Management History*, Vol. 14 No. 2, pp. 174-183.

Howell, M. and Prevenier, W. (2001), *From Reliable Sources: An Introduction of Historical Methods*, Cornell University Press, Ithaca.

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*, Longman, London and New York, NY.

Khaire, M. and Wadhwani, R.D. (2010), "Changing landscapes: the construction of meaning and value in a new market category- modern Indian art", *Academy of Management Journal*, Vol. 53 No. 6, pp. 1281-1304.

King, B.G. and Soule, S.A. (2007), "Social movements as extra-institutional entrepreneurs: the effects of protests on stock price returns", *Administrative Science Quarterly*, Vol. 52 No. 3, pp. 413-442.

Kipping, M., Wadhwani, R.D. and Bucheli, M. (2014), "Analyzing and interpreting historical sources: a basic methodology", in Bucheli, M. and Wadhwani, R.D. (Eds), *Organizations in Time: History, Theory, Methods*, Oxford University Press, Oxford, pp. 305-329.

Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihok, G. and Hartog, D.N.D. (2018), "Text mining in organizational research", *Organizational Research Methods*, Vol. 21 No. 3, pp. 733-765.

McEnery, A. and Baker, H. (2016), *Corpus Linguistics and 17th Century Prostitution: Computational Linguistics and History*, Bloomsbury.

McEnery, T. and Hardie, A. (2012), *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press, Cambridge.

McEnery, T. and Wilson, A. (2001), *Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Mautner, G. (2007), "Mining large corpora for social information: the case of elderly", *Language in Society*, Vol. 36 No. 1, pp. 51-72.

Mautner, G. (2009), "Checks and balances: how corpus linguistics can contribute to CDA", in Wodak, R. and Meyer, M. (Eds), *Methods of Critical Discourse Analysis*, Sage Publications, London, pp. 122-143.

Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A. and Aiden, E.L. (2011), "Quantitative analysis of culture using millions of digitized books", *Science*, Vol. 331 No. 6014, pp. 176-182.

Oakes, M.P. (1998), *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Orpin, D. (2005), "Corpus linguistics and critical discourse analysis: examining the ideology of sleaze", *International Journal of Corpus Linguistics*, Vol. 10 No. 1, pp. 37-61.

Pechenick, E.A., ; Danforth, C.M. and Dodds, P.S. (2015), "Characterizing the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution", *PLoS ONE*, Vol. 10 No. 10, pp. 1-24.

Pollach, I. (2012), "Taming textual data: the contribution of corpus linguistics to computer-aided text analysis", *Organizational Research Methods*, Vol. 15 No. 2, pp. 263-287.

Rowlinson, M. (2004), "Historical analysis of company documents", in Cassell, C. and Symon, G. (Eds), *Essential Guide to Qualitative Methods in Organizational Research*, Sage, London, pp. 301-311.

Rowlinson, M., Hassard, J. and Decker, S. (2014), "Research strategies for organizational history: a dialogue between historical theory and organization theory", *Academy of Management Review*, Vol. 39 No. 3, pp. 250-274.

Tumbe, C. and Krishnakumar, S. (2018), "From bazaar to big bazaar: environmental influences and service innovation in the evolution of retailing in India, c. 1850-2015", *Journal of Historical Research in Marketing*, Vol. 10 No. 3, pp. 312-330.

Tumbe, C. and Ralli, I. (2018), "The four eras of 'marketing' in twentieth century India", *Journal of Historical Research in Marketing*, Vol. 10 No. 3, pp. 294-311.

Vogel, F., Hamann, H. and Gauer, I. (2018), "Computer-assisted legal linguistics: corpus analysis as a new tool for legal studies", *Law and Social Inquiry*, Vol. 43 No. 4, pp. 1340-1363.

Wadhwani, R.D. and Bucheli, M. (2014), "The future of the past in management and organization studies", in Bucheli, M. and Wadhwani, R.D. (Eds), *Organizations in Time: History, Theory, Methods*, Oxford University Press, Oxford, pp. 3-32.

Westin, I. and Geisler, C. (2002), "A multi-dimensional study of diachronic variation in British newspaper editorials", *International Computer Archive of Modern and Medieval English*, Vol. 26, pp. 133-152.

Wodak, R. and Meyer, M. (2009), "Critical discourse analysis: history, agenda, theory and methodology", in Wodak, R. and Meyer, M. (Eds), *Methods of Critical Discourse Analysis*, Sage Publications, London, pp. 1-33.

About the author

Chinmay Tumbe is an Assistant Professor at the Indian Institute of Management Ahmedabad (IIM-A). His research interests lie in business and economic history and migration and urban studies. Chinmay Tumbe can be contacted at: chinmayt@iima.ac.in