N-GRAM ANALYSIS IN THE ENGINEERING DOMAIN

Martin Leary[1], Geoff Pearson[1], Colin Burvill[2], Maciej Mazur[1], Aleksandar Subic[1]
(1) RMIT University, School of Aerospace, Mechanical and Manufacturing Engineering, Australia (2) University of Melbourne, Department of Mechanical Engineering, Australia.

ABSTRACT
New technologies have enabled the digitization and linguistic analysis of a vast number of books published throughout history. This technology has enabled a step-change in the opportunities to understand the interests of the authors and by doing so provide insight into the aspirations of society throughout published human history. Such analysis provides an unprecedented opportunity, however there are numerous analysis pitfalls due to fundamental technology limitations and misunderstanding of the analysis outcomes. This work defines the technologies which have enabled this opportunity and, in doing so, identifies potential risks of erroneous outcomes. A broad scope analysis of the engineering design domain is presented for the first time.

Keywords: ngram; n-gram; optical character recognition; OCR; linguistic analysis; technology review.

1 ENABLING TECHNOLOGIES
The convergence of several enabling technologies has resulted in an unprecedented opportunity to study word usage within a vast number of publications throughout history. These technologies are:
- automated book scanning
- optical character recognition
- n-gram methods

1.1 Automated book scanning

Recent technical developments have enabled automated capture of book pages to provide high resolution digital files. These devices are applied for commercial book database acquisition, for example Google Books [1,2], as well as for philanthropic projects such as Project Gutenburg [3].

1.2 Optical character recognition

Early text digitization projects required the manual conversion of text by a human operator. Such methods are time consuming and prone to human error. However, manual digitization allows intelligent text interpretation and may be necessary for non-standard fonts, damaged texts or for hand-written manuscripts. Optical Character Recognition (OCR) is the automated acquisition of alphanumeric data from digital images. The development of OCR algorithms has been motivated by several domains, including bar-scanning technologies for inventory management, number plate and traffic sign recognition as well as text digitization.

Within the domain of OCR research, the acquisition of text data involves a number of opportunities and challenges. Unlike number plate or road sign recognition, printed text is typically available on a high contrast background and little filtering is required to differentiate the printed text from the page. Furthermore, the majority of printed works use a relatively small number of standard fonts, thereby reducing the technical burden required for character recognition. Challenges associated with text digitization include the accommodation of special characters such as mathematical symbols. The larger the subset of feasible OCR characters the more likely that a character can be misinterpreted. For example the Greek letter α may be misinterpreted as an italicized English alpha letter *a*, and vice-versa. Examples of OCR errors are presented in Section 5.2. The likelihood of such errors can be

reduced by context sensitive algorithms which restrict the range of feasible characters depending on the likelihood of their occurrence. Such a determination may be difficult without *a priori* knowledge of the published work. However, there may be opportunities for automation of this decision based on the associated Dewey decimal number. For example, the existence of the Greek letter is more likely within the Science field (Dewey decimal class 500) than in the Literature field (Dewey decimal class 800).

OCR algorithms reference a standard library of potential characters based on a database of character samples. Printing errors and physical contamination (wear, soiling) of the page may result in incorrect OCR outcomes. Manual analysis may be used to compliment OCR in the elimination of such errors. The efficiency of manual methods may be increased by flagging high risk OCR outcomes, for example where there are multiple possible characters or when an identified word is not part of a standard dictionary.

Manual correction of OCR outcomes has been effectively applied by crowd-sourcing methods. For example, the OCR outcomes of historical newspapers include a high rate of error due to poor quality printing and physical contamination. This process is currently underway within the Australian Newspapers Digitisation Program [4], where these errors are being manually corrected by groups of volunteer enthusiasts. 40 million articles from out of copyright Australian newspapers are planned to be manually corrected by July 2011.

1.3 N-gram methods

Digitized text provides an opportunity for en mass analysis of the occurrence of words, symbols, and associated combinations [5]. The latter can be achieved by directly searching for word combinations within the texts of interest, however this method is infeasible for a non-trivial dataset. An efficient method of enabling a large number word combination searches is by the use of n-grams. N-grams are a look up table of word combinations. For example, the following text:

The quick brown fox jumped over the lazy dog

Results in the following n-grams of order 2, also known as 2-grams or bi-grams:

The quick
quick brown
brown fox
fox jumped
jumped over
over the
the lazy
lazy dog

And the following n-grams of order 3, also known as 3-grams or tri-grams:

The quick brown
quick brown fox
brown fox jumped
fox jumped over
jumped over the
over the lazy
the lazy dog

The resulting n-gram look up table allows rapid assessment of word combinations within a range of texts, however the initial computational cost of defining the n-gram database is high, especially as the allowable n-gram order increases. For this reason the n-gram order is restricted for non-trivial datasets.

Individual n-gram algorithms must select how to handle specific scenarios including; the use of capital letters, punctuation marks, hyphenated words, and extraneous text (such as header and footer text). These scenarios must be understood to accommodate potential analysis error.

2 SOURCES OF ERROR

Potential errors in n-gram analysis include: erroneous OCR outcomes (Section 1), and misinterpretation of n-gram data. Misinterpretation of n-gram data can result in unintended false-positive and false-negative outcomes, including misunderstandings associated with:

- data and presentation
- corpus
- n-gram algorithm
- spelling and local idiom
- transience of meaning
- erroneous correlation

2.1 Data presentation

Various n-gram viewers access and present the available data using a wide range of methods. The n-gram viewer used in this work is the Google Book Ngram Viewer [1], which presents the occurrence of the searched n-gram as a normalised percentage of the total number of identified n-grams. Detailed technical information about the Google Book Ngram project, database and viewer are presented in [2]. As a first presentation, the unigrams *engineering*, *design* and *technology* are presented (Figure 1). To aid in identifying long term trends, a rolling average (smoothing) may be applied. Figure 2 presents the same data as Figure 1, with a default smoothing value of 3 (as applied to the following results in this work).

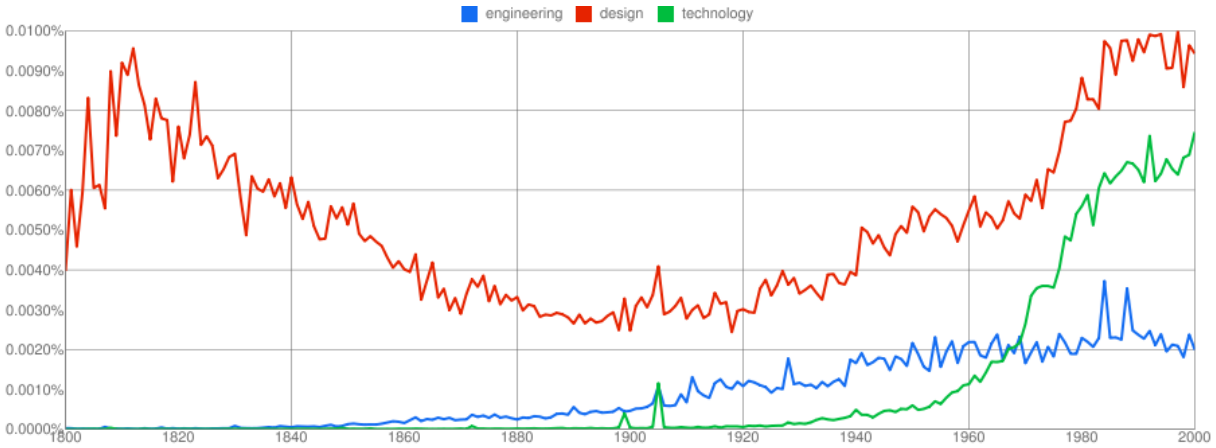

Figure 1. Google Books Ngram Viewer result for *engineering*, *design* and *technology*. Smoothing value = 0 [1].

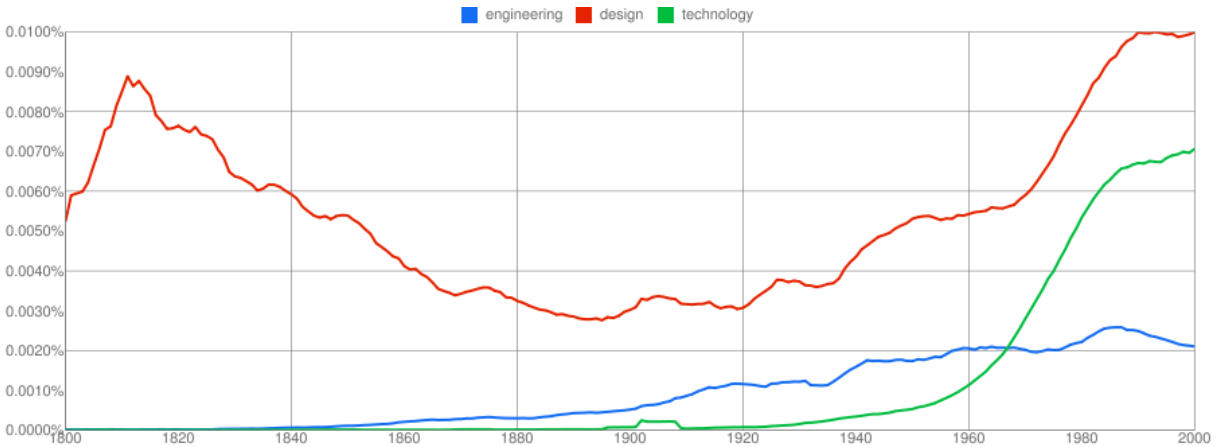

Figure 2. Google Books Ngram Viewer result for *engineering*, *design* and *technology*. Smoothing value = 3 [1].

2.2 Corpus

The available n-gram database may be stratified according to various corpora. For example, the Google Book Ngram Viewer includes corpus of American English, British English, Chinese, French, German, Hebrew, Spanish and Russian, as well as various other specialized strata [1].

The corpus applied is of critical importance to the associated outcomes. For example, Figure 3 includes the same n-grams of Figure 2, but is restricted to the German corpus. It is evident that these terms were not commonly part of the German lexicon until relatively recent times when they were integrated from the English language usage.

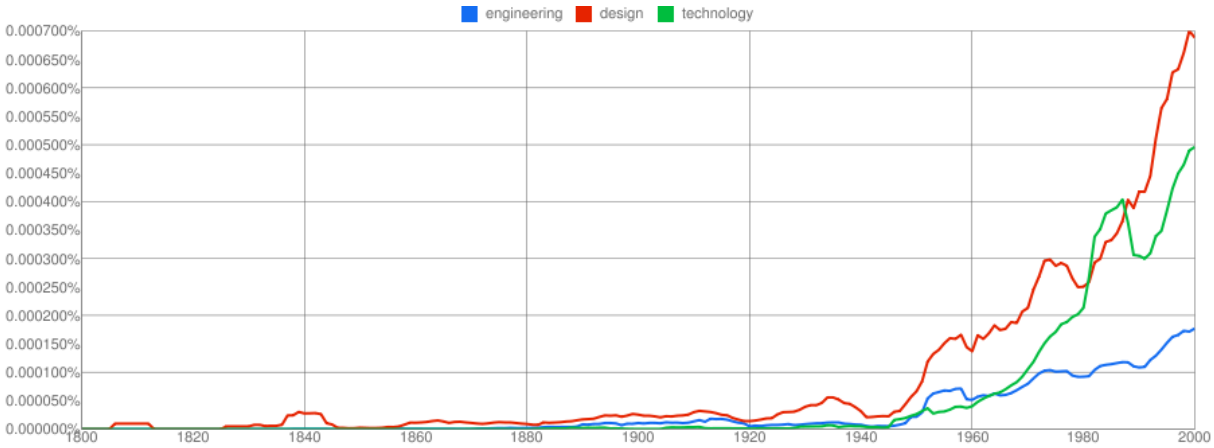

Figure 3. Google Books Ngram Viewer result for *engineering*, *design* and *technology* within the German corpus [1].

2.3 n-gram algorithm

Various n-gram algorithms may interpret specific cases differently. For example, Google Book Ngram Viewer [1], is case-sensitive, consequently the search term *Engineering design* returns a different outcome to *engineering design*, or *ENGINEERING DESIGN* (Figure 4).

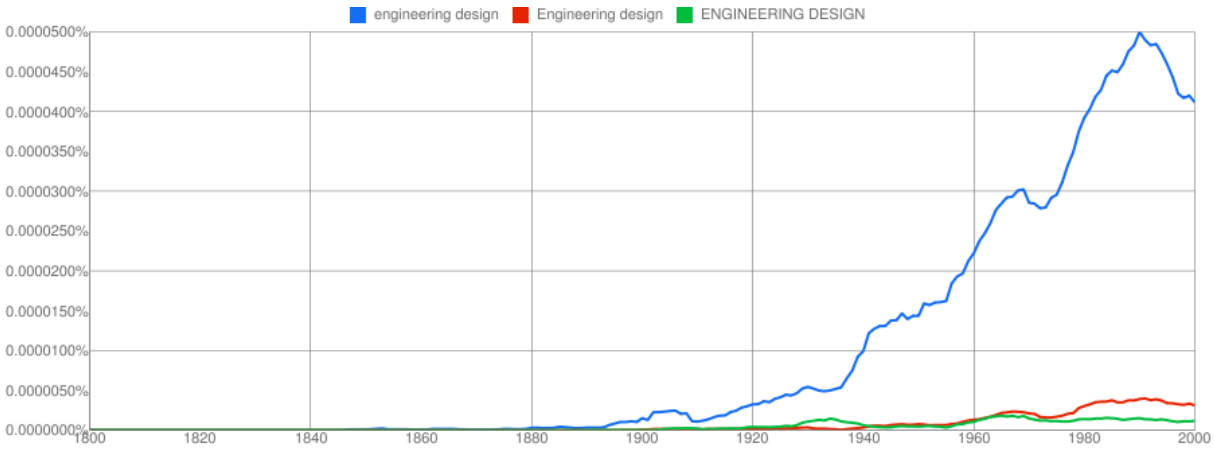

Figure 4. Google Books Ngram Viewer result for *Engineering design*, *engineering design* and *ENGINEERING DESIGN* [1].

2.4 Spelling and local idiom

Local idiom and spelling preferences can results in erroneous n-gram analysis. Within the technical domain there are many such examples, for example *aluminium* (British English) and *aluminum* (US English) (Figure 3). If an overall estimate of the global relevance of search term is required it is necessary to ensure that local spellings and idiom are accommodated.

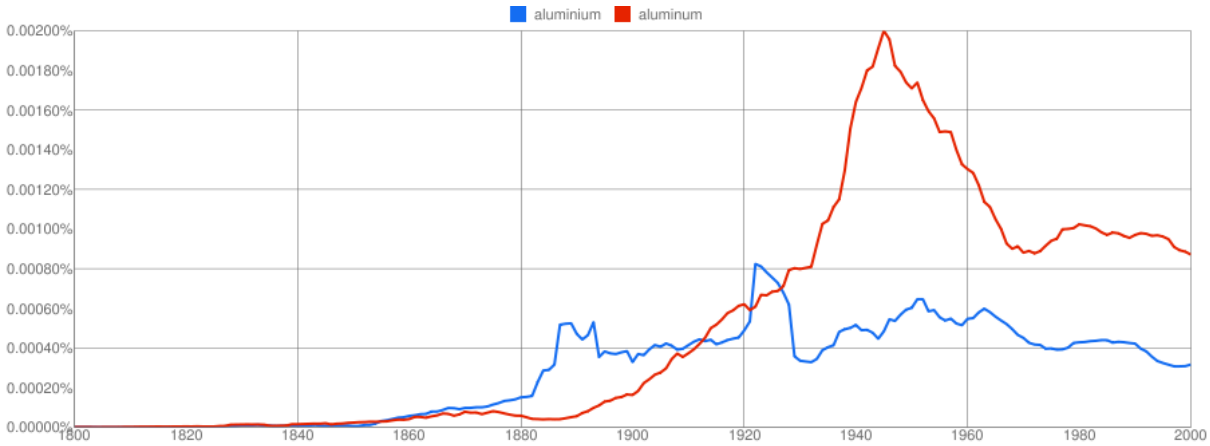

Figure 5. Google Books Ngram Viewer result for *aluminium* and *aluminum* [1].

2.5 Transience of meaning

N-gram datasets can extend over a significant time period. Consequently there exists a risk of misinterpretation due to transience of meaning. For example, Figure 6 displays results for search terms *car* and *automobile*. It is not surprising that the term *automobile* rises abruptly with the commercialisation of the horseless carriage around the early 20th century, and also that the term *car* is more commonly accepted than automobile in modern parlance. It may come as a surprise that *car* was in common usage as early as the 1800s. This potential misinterpretation can be clarified by reference to the Oxford English Dictionary.

The Oxford English Dictionary (OED) is the oldest extant dictionary to define language by examples of their historical use, rather than by a summary based on expert opinion. Thereby the OED provides an excellent opportunity to clarify search terms by confirming language use through history. For example, the OED identifies that in the early 1800s the term *car* was used to mean any form of personal transport, including for example, a litter. Over time, this term has been appropriated as a standard term for motorized personal transport.

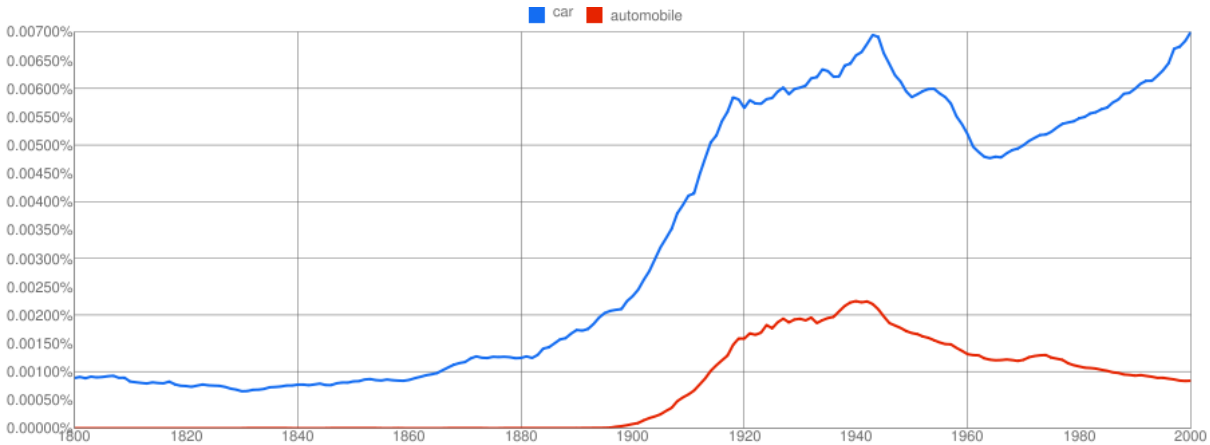

Figure 6. Google Books Ngram Viewer result for *car* and *automobile* [1].

2.5 Erroneous correlation

A significant cause of misinterpretation is the erroneous correlation of the occurrence of an n-gram with the subject of interest. This is an example of a false-positive result and occurs when the analyst is unaware (or unable to eliminate) the overlap of an unintended n-gram from another domain. An example of erroneous correlation is provided in Section 3.2.

3 N-GRAM ANALYSIS IN THE ENGINEERING DOMAIN

Keeping in mind the identified limitations, a series on n-grams of relevance to the engineering domain are outlined. This broad scope analysis provides initial insight into the historical use of the language of technology, and provides a basis for continuing work within the field.

3.1 Engineering disciplines

A broad overview of a range of engineering disciplines identifies that *civil engineering* entered the English language lexicon in the early 1800s and has subsequently grown steadily, despite periodic fluctuations. The term *mechanical engineering* became known around 1850 and followed a similar trajectory. The term *electrical engineering* was published later but increased rapidly in relevance. These dominant disciplines have reduced in n-gram frequency in recent years, with more specialized disciplines such as *biomedical engineering* and *environmental engineering* gaining importance.

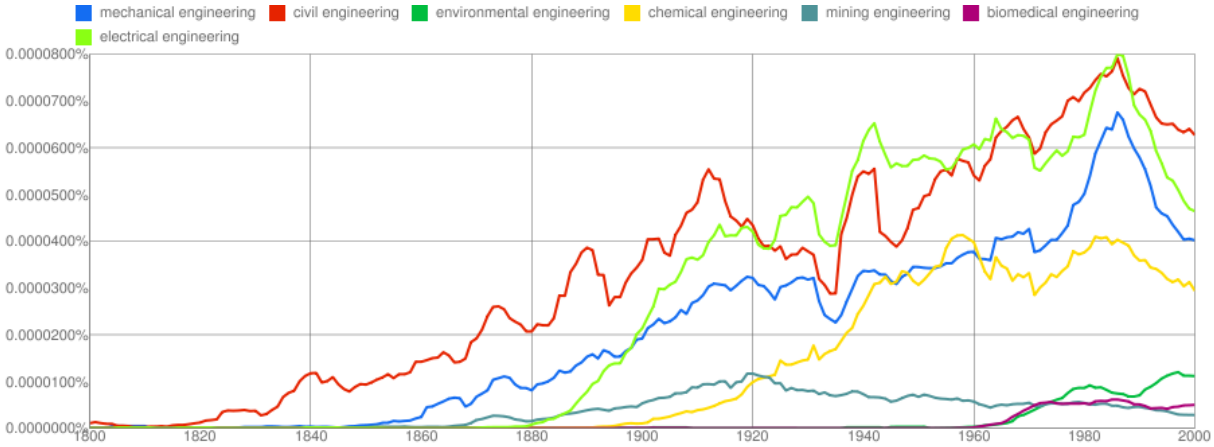

Figure 7. Google Books Ngram Viewer result for *mechanical engineering*, *civil engineering*, *environmental engineering*, *chemical engineering*, *mining engineering*, *biomedical engineering*, *electrical engineering* [1].

3.2 Environment and resources

Figure 8 provides some insight into concern for the environment and the non-renewable nature of fossil fuel resources. It is apparent also that the acronym *GHG* is becoming a common abbreviation of greenhouse gas, however, a detailed analysis would need to assess whether this may in fact be a false-positive due to another usage of the term *GHG*. The term *peak oil* requires a larger scale (Figure 9) to identify that the term was introduced in the 1940s and diminished in importance to be revised later in the 1980s.

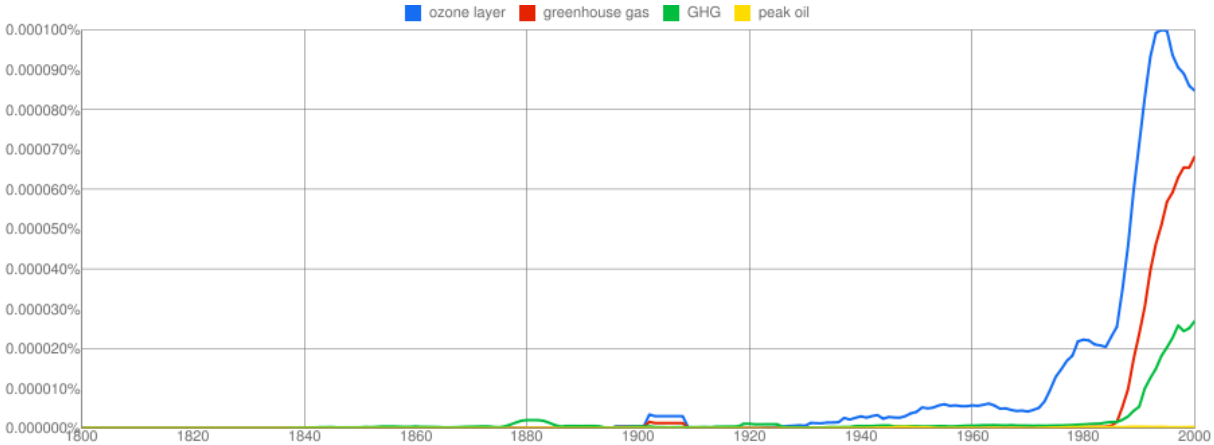

Figure 8. Google Books Ngram Viewer result for *ozone layer*, *greenhouse gas*, *GHG*, *peak oil* [1].

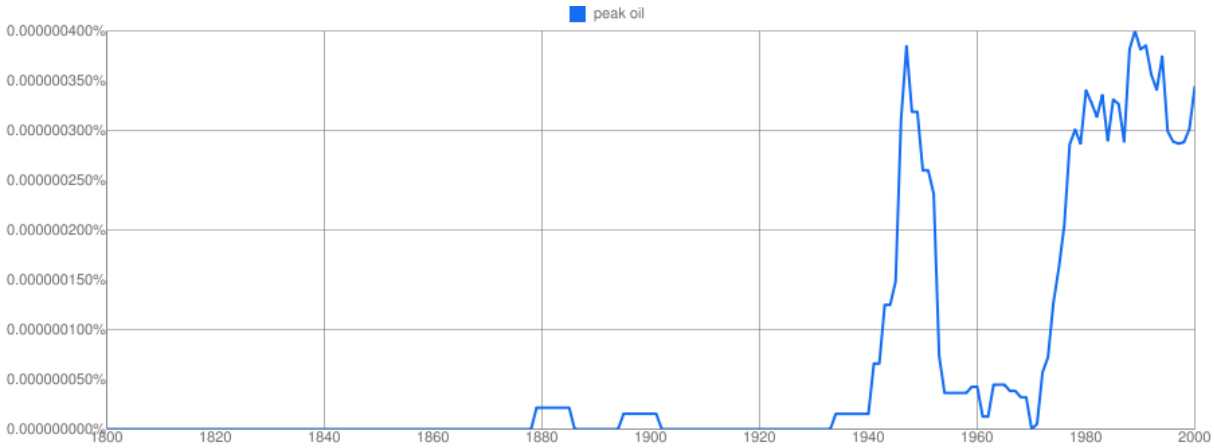

Figure 9. Google Books Ngram Viewer result for *peak oil* [1].

Figure 9 highlights a common source of misinterpretation. Namely that the bi-gram *peak oil* returns a positive result in 1880 and 1900, well before Hubbert's [4,5] seminal work on diminishing fossil fuel reserves. This misinterpretation is likely due to: erroneous OCR outcomes [1] or erroneous correlation (Section 2.5). On initial inspection, the following candidates for misinterpretation have been identified:

- an erroneous correlation with "Pikes Peak Oil, Gas and refining company" referenced by the Indiana Secretary of State (1904)
- an OCR error which appears to have digitized "pear oil" as "peak oil" [8]

3.3 Power technologies

The use of the bi-gram *steam power* has been fairly consistent since the accelerated uptake of steam power during the industrial revolution. Milestones in steam power include Watts' design refinement of the steam engine in the late 1700s and Trevithick's introduction of high-pressure steam in the early 1800s. Steam power is generally considered to be a reliable and non-contentious form of power production, and this is reflected in the consistency of usage.

This contrasts with the bi-gram *nuclear power*, which is relatively volatile in its frequency of usage. It is notable that maximum usage of the bi-gram corresponds with the peak of the Cold War in the mid-1980s, when fears of nuclear confrontation were at their highest. A smaller peak occurs around the time of the Cuban missile crisis in the early 1960s. The peak usage of the phrase seems to occur around 1986, which may be linked to the Chernobyl nuclear power plant disaster in that year.

It is interesting that the terms *wind power* and *solar power* peak at roughly the same time as *nuclear power*. However, in terms of relative interest in these bi-grams, the results of Figure 10 are misrepresentative due to use of the suffix "power". For example, the frequency of the bi-gram *solar energy* is much greater than *solar power*, but follows similar usage trends (Figure 11). However, comparison of *nuclear energy* to *nuclear power*, and *wind energy* to *wind power* does not reveal the same pattern. This highlights the idiosyncratic nature of language use, and the potential pitfalls of this method of enquiry, especially when attempting to quantify the relative importance of different fields of interest.

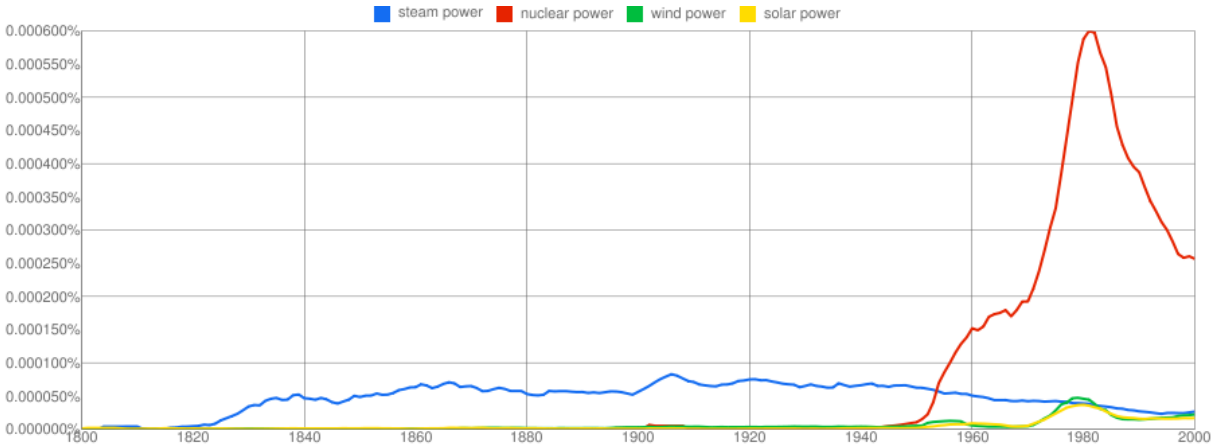

Figure 10. Google Books Ngram Viewer result for *steam power*, *nuclear power*, *wind power and solar power* [1].

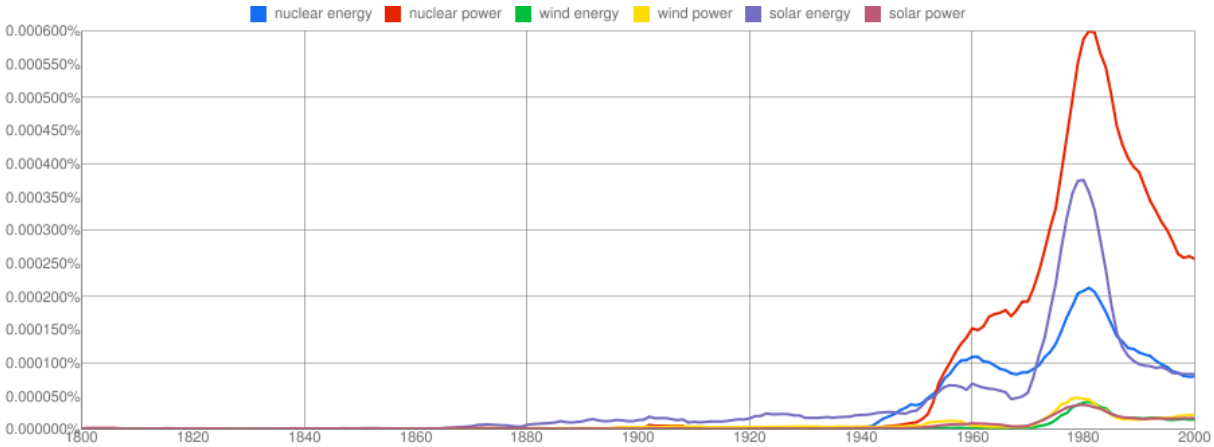

Figure 11. Google Books Ngram Viewer result for *nuclear energy*, *nuclear power*, *wind energy*, *wind power*, *solar energy*, *solar power* [1].

3.4 Material technologies

Figure 12 provides some insight into the history of material technologies. Polymer technologies have developed since the 1930s. Composite materials have developed continuously through human history, which earlier in history included natural composites such as mud and horse hair, and later included technical composites such as kevlar and carbon fiber reinforced polymers [6].

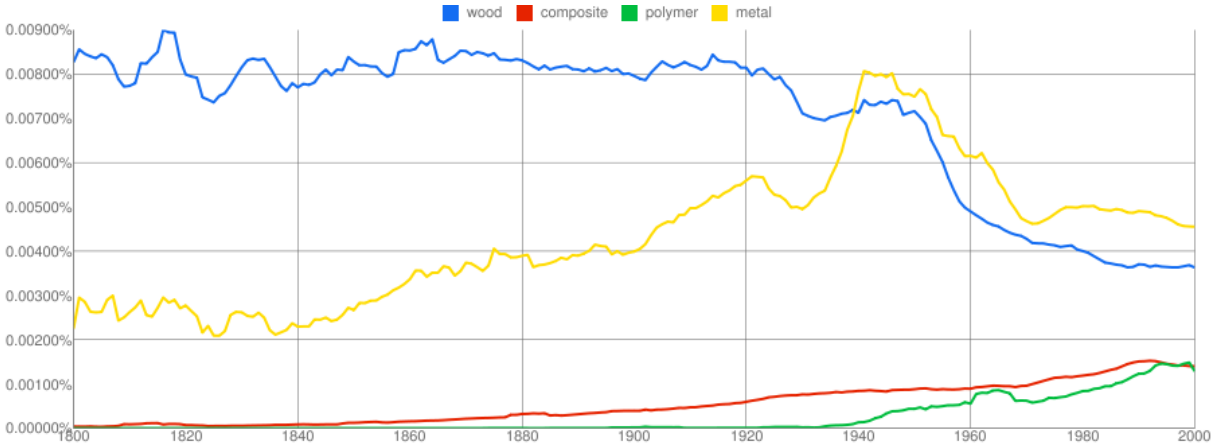

Figure 12. Google Books Ngram Viewer result for *wood*, *composites*, *polymer*, *metal* [1].

3.5 Automotive and vehicle technologies

Figure 12 shows usage trends of terms relating to power output and maximum speed. While each of these terms could also be relevant to non-automotive fields such as aerospace and rail, distinct trends are still apparent. Peak interest in such n-grams corresponds with the timing of World War II, when competitiveness and performance was of paramount importance to the design community.

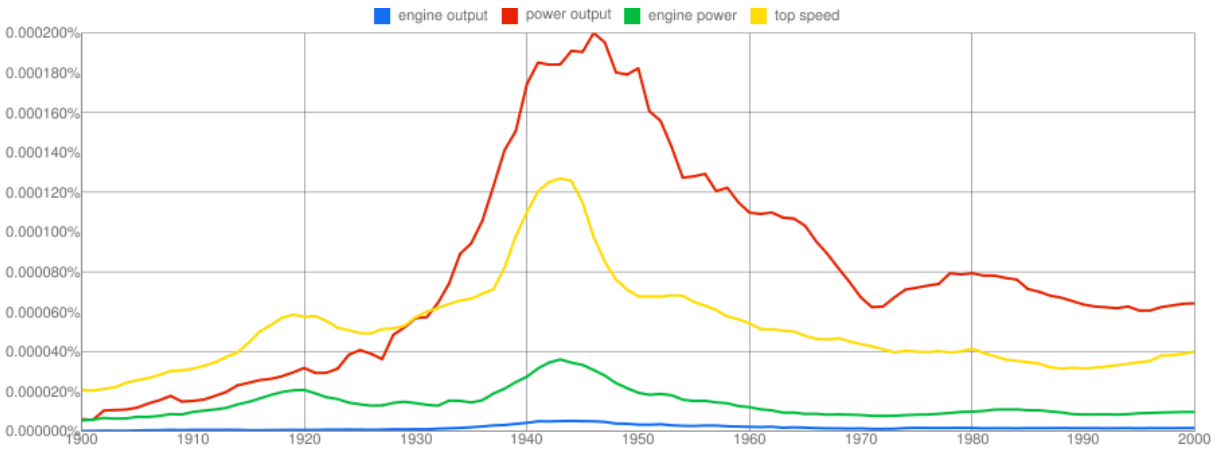

Figure 12. Google Books Ngram Viewer result for *engine output*, *power output*, *engine power*, *top speed* [1].

Figure 13 reflects the emergence of more recent criteria by which vehicles are evaluated, as society becomes more aware of the environmental and social impacts of the motor vehicle. While issues relating to power tended to decline and then plateau after the war, new concerns such as *vehicle safety*, *fuel efficiency* and *vehicle emissions* have risen in importance. A sharp increase in discussion of vehicle safety occurs from around 1965, the year that lawyer and vehicle safety activist Ralph Nader [9] released the seminal text "Unsafe at Any Speed". The fuel crisis in 1972 launched a rapid escalation in interest in *fuel efficiency*, whilst more recently *vehicle emissions* is a bi-gram of interest.

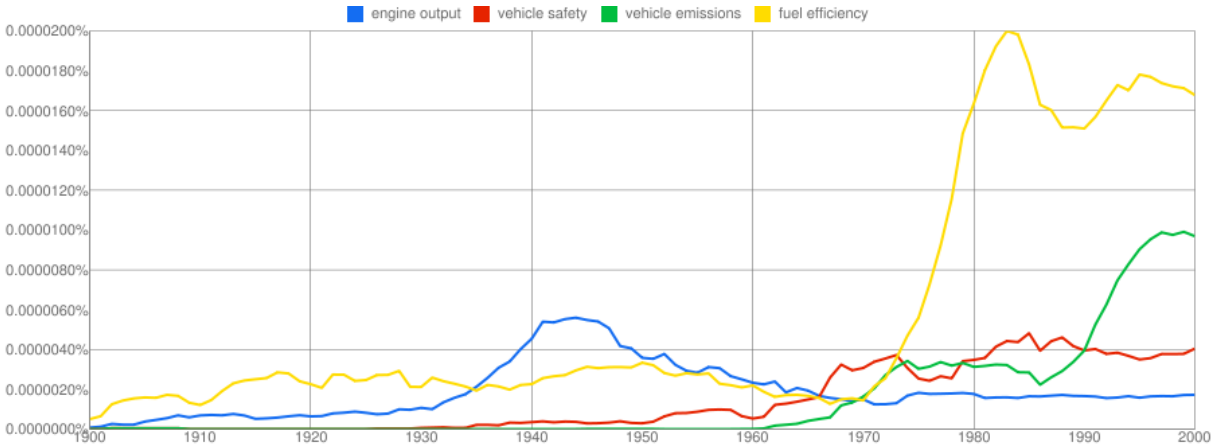

Figure 13. Google Books Ngram Viewer result for *engine output*, *vehicle safety*, *vehicle emissions*, *fuel efficiency* [1].

4 CONCLUSIONS

Novel data acquisition and analysis technologies have enabled a step-change in capability to understand the interests and aspirations of authors throughout published human history. This unprecedented opportunity is potentially stymied by errors which may occur due to erroneous OCR and misinterpretation of analysis results. The former are continuously being reduced by new technologies, the latter is the responsibility of the analyst. To minimize the risk of erroneous outcomes this work identifies a series of potential misunderstandings inherent to n-gram analysis. Subsequently this work provides the first application of the extensive Google Books Ngram Viewer to the engineering domain. A novel, broad scope analysis of the engineering design domain is presented for the first time, providing an initial insight into the historical use of language of technology, and providing a basis for continuing work within the field.

REFERENCES

[1] Google, 2010, Google Books Ngram Analysis, http://ngrams.googlelabs.com/info, accessed December 2010.

[2] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden, 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. Science Magazine (Published online: 16 December 2010).

[3] Project Gutenburg, 2010, http://www.gutenberg.org/wiki/Main_Page, accessed December 2010.

[4] Rose, H., 2009. A success story - Australian Newspapers Digitisation Program Subjects, Journal of Technical Services in Libraries, Archives and Museums. 3-Dec-2009.

[5] Microsoft, 2010, Web N-gram Services, http://web-ngram.research.microsoft.com/info/, accessed December 2010.

[6] Ashby, M.F. Materials and the Environment: Eco-informed Material Choice, Butterworth-Hieneman, 2010.

[7] Hubbert, M.,K., 1956, Nuclear Energy and the Fossil Fuels, Presented before the Spring Meeting of the Southern District, American Petroleum Institute, Plaza Hotel, San Antonio, Texas, March 7–8-9, 1956.

[8] Cooley, A. J. 1892. Cooley's cyclopædia of practical receipts and collateral information in the arts, manufactures, professions, and trades. J. & A. Churchill.

[9] Nader, R. 1965, Unsafe at Any Speed, Grossman Publishers, New York.