Experiments in 17th century English: manual versus automatic conceptual history

Stephen Pumfrey, Paul Rayson and John Mariani
Lancaster University

Abstract

Previous research in conceptual history, the study of change over time of key terms and value systems, has been carried out manually using a restricted number of pre-identified texts. We propose that a method combining techniques from corpus and computational linguistics can be exploited to support conceptual history with semantic searches on a vast sample of texts. To exemplify this method, we focus on a fundamental concept in modern science, the experimental method, in order to trace when the pre-existing and primarily religious concept of *experiment* (or experience) took on its modern, scientific meaning. We contrast a manual approach using the existing Early English Books Online search interface with an automatic method using corpus linguistics software and methods to turn the transcribed portion of the same dataset into a corpus. Both approaches allow us to separate the religious and scientific senses and plot their change over time. We observe a rapid change in the meaning of *experimental* from overwhelmingly religious to largely scientific within the 1660s. However, the automatic corpus method is much more efficient and will support future scholars in carrying out iterative studies in a matter of minutes rather than through weeks of painstaking work. Such methodological innovation has the potential to support the formation of new research questions, which could not have been considered previously.

Correspondence: Dr Paul Rayson, School of Computing and Communications, Infolab21, Lancaster University, Lancaster, LA1 4WA, UK.
E-mail: p.rayson@lancaster.ac.uk

1 Introduction

In this article, we describe how the techniques from computational and corpus linguistics can be brought to bear on very large-scale historical collections. In particular, we show how they can answer big questions in conceptual history. Over the last few years, the presentation of large volumes of digital facsimiles through web interfaces has delivered data straight to the historian's desktop. A second revolution is now underway with full text versions of those historical collections being produced by keyboarding and OCR processes. The question then arises: 'what do you do with a million books?' (Crane, 2006). The first, most obvious and currently most common use, is for the historian to 'borrow' books from this digital library that s/he had already selected for study from the traditional library shelf or catalogue. However, we focus on another approach which turns the entire collection into a corpus and enables searching for patterns across multiple works using techniques from the corpus linguist's toolbox, e.g. frequency analysis, corpus annotation/tagging, concordances, and collocations. This allows the historian to explore existing research questions across more texts and

doi:10.1093/llc/fqs017 Advance Access published on 1 June 2012

in a shorter time scale. Importantly, the historian will also be able to pose new types of research questions that could not have been addressed before. As we demonstrate, the new approach facilitates corpus-driven studies where unexpected patterns in the data can drive the research in new directions. The corpus linguistic approach is distinct from information extraction and text mining methods which have been applied elsewhere to such data since the focus is on the investigation of changes in language and concepts over time and across genres in the historical corpus. In contrast, prototypical text mining methods would involve the automatic extraction of named entities (people and places), the relationships between them, semi-automatic structuring of the data, document classification and so on (Argamon and Olsen, 2009; Elson *et al.*, 2010). In the research presented here, we apply corpus linguistic techniques and provide a case study related to conceptual history and in particular the historical development of experimental science.

Linguists have been studying language change with historical corpora for some time. These linguistically sampled corpora tend to be of the order of millions of words, e.g. the Helsinki corpus contains 1,572,800 words (Kytö, 1996). However, very large-scale text collections such as Early English Books Online (EEBO) via the transcribed version from the Text Creation Partnership (EEBO-TCP) contain hundreds of millions of words e.g. 521 million in the 2006 release of 11,462 documents. Enormous data sets such as the Internet Archive and Google Books provide even greater opportunities for studying historical trends with an estimated 243 billion words in the Internet Archive (Bamman and Crane, 2011). Given the mind-boggling scale of these collections, new and updated corpus methodologies need to be developed. This is especially true because making the collections available as searchable text via the corpus methodologies will allow historians to pose new questions in conceptual history and to answer some that were previously impossible. An example of a new, previously impossible question is the very one we pose and answer here in this paper.

Conceptual history is one of several closely related sub-disciplines, with important differences from linguistic history, intellectual history, history of ideas, etymology and discourse analysis. Definitions of these sub-disciplines are disputed but discourse analysis, whether in plain, critical or discourse historical form, has a primarily synchronic focus (Wodak *et al.*, 2009).[1] Conceptual history is concerned with diachronic semantics, and particularly in change over time in the meaning of key terms and value systems such as ‘liberty’ or ‘experiment’. As such it is much broader than the concern of etymology with individual words, but narrower than linguistic history, which examines the evolution of entire languages.

The German historian Reinhart Koselleck was a pioneer of the interdisciplinary method of *Begriffegeschichte* which translates as ‘the history of concepts’. *Begriffegeschichte* focussed on a slightly later period than that of EEBO, the late seventeenth to the early nineteenth century, because its practitioners considered it to be the crucial period of transition from pre-modern to modern culture. They argued that concepts could only be understood when situated by historians in their contingent and changing cultural contexts. This rejection in *Begriffegeschichte* (or *Historische Semantik*) of both Marxist historical materialism and the history of ideas continues to unite the more general community of conceptual historians. In both cases, ideas or ideologies were typically treated as determined and unchanging. Early historians of ideas such as A. O. Lovejoy asserted the existence of unit-ideas which combined in different ways in different historical periods but which were relatively stable themselves. The history of ideas was criticized for its lack of contextual focus by members of the ‘Cambridge School’ (itself a form of conceptual history), notably the historian of political theory Quentin Skinner. Skinner combines analysis of the languages of political thought with an intellectual historian’s focus on the historically specific cultures and intellectuals who produced them.[2]

This is where conceptual historians differ from intellectual historians. Conceptual historians are less concerned with the contingent relationship between individuals and historical cultures, often specific intellectuals and their canonical texts. They analyse more general and representative discourses.

Thus Koselleck and the others who, between 1972–97, produced *Geschichtliche Grundbegriffe* and the journal *Archiv fur Begriffsgeschichte*, traced the evolution of concepts through a plethora of texts they had identified as relevant, ranging from classic texts through minor ones to popular reference works, a range which transcended the intellectual historian's notion of context but which acknowledged that concepts changed their meaning as their contexts of production, reception and use changed (Lehmann and Richter, 1996).

To return to discourse analysis, our retention of the word 'concept' is a reminder that conceptual history is different. Like discourse analysis it looks 'beyond the sentence boundary' to the word in context, even if it is more concerned to trace individual concepts which have been expressed by a number of words and phrases. But it differs fundamentally in its assumption that the analyst is able to get beyond or behind the text or discourse to underlying concepts expressed by the text.

Locating conceptual history in relation to the other sub-disciplines discussed above makes clear how computational and corpus linguistics can advance conceptual history. Koselleck and his team of researchers conducted their analysis of the evolution of concepts before the age of large electronic corpora. They had to limit their manual semantic searches to a manageable number of pre-identified texts. We can now perform automated searches and semantic analyses upon a huge sample of texts, which in the case of EEBO is approaching completeness.

In the remainder of this article, we describe our collaborative work in one area of conceptual history, i.e. the historical development of experimental science. This is outlined in Section 2 along with a brief summary of the EEBO data set. Section 3 describes the two methods, a manual method using the current EEBO web front end combined with quantitative analysis via a spreadsheet, and an automatic method using a corpus environment that has been created at Lancaster. The former method requires limited skills and tools beyond EEBO itself while the latter depends upon that advanced corpus environment. Results of the case study are interwoven in Section 3. Finally, in Section 4, we draw some conclusions from our work and the proposed methodology, before highlighting further research and extensions to the methodology.

2 Background

2.1 A brief history of EEBO and its uses

Historians who work on the early centuries of print are familiar with the problems of locating and reading rare books. The *Early English Books* (*EEB*) project reduced the problem by issuing microfilms of facsimile pages of books listed in the Short Title Catalogues. Issues began in 1938 and are once again moving towards completion, revived by the *EEBO* initiative.[3]

The scholarly experience presented by *EEB* was very similar to that afforded by a copyright library. Researchers approached the catalogue with discrete books in mind and inspected the microfiches. Other books only presented themselves if they were referred to in the books consulted, or were adjacent in the catalogue or on the microfiche. The differences between the experiences were few but not negligible. It is important to remember that change happens whenever an artefact undergoes 'remediation' from one medium to another, whether it is from a hard copy of a book to its facsimile images or from cylinder sound recordings to mp3 files. A particular loss was the impact of the physical aspects of the copy of the book: for example, how big was it, how lavishly produced, how well-thumbed, how was it bound, and how was it bound or shelved with other texts?

Proquest began work to present the PDF facsimiles through a searchable online interface in 1998, and in 1999 also became part of the TCP. This has led, in turn, to a new aim: that of converting 25,000 books into fully searchable, TEI-compliant SGML/XML texts. The first phase has been achieved and Phase II, the conversion of the remaining 44,000 distinct monographs has begun. Indeed, as we write nearly 33,000 of the 128,070 items have been converted.

As Kichuk (2007) has noted, the remediation brought about by the digitisation of EEB into EEBO has profoundly changed the ways that scholars engage with the texts.[4]

First, though of less significance in this article, the historian comes across individual books in new ways. Instead of locating a book in a full catalogue, one enters search terms concerning author, title, date, etc. Unless those terms are needlessly precise, the desired book will form part of a list, juxtaposed with other, often unfamiliar texts which may not be related in any historically significant way but only by keywords. As with any database searched by keywords (whether JSTOR, Google books or the entire web) one encounters unexpected and often intriguing works.

A second and more radical change is made possible by the expanding corpus of searchable text, which in 2011 exceeds one billion words and 30,000 items and is set to double. EEBO search screens allow one to search for the occurrence of specific words and their variants, and to identify the texts which contain them. Searched in this way, the subset of searchable texts effectively forms a single corpus, within which the distinctions between unique texts, individual authors, subject matter and dates of production are partially erased.

Inspecting such results, scholars unfamiliar with corpus linguistics experience some, but only some, of the different ways of encountering historical discourse once it has been digitized and turned into a corpus. They are confronted by many unknown books. They will be presented with a curious mix of quotations from books, collated purely according to the presence of key words. Some books will be familiar, others not; some canonical, others not; some from the fifteenth century, others from 1700; some on predictable topics, others belonging to unexpected topics or genres. In short, the texts are grouped with little reference to the metadata which humanities scholars conventionally use to make their textual selections. It should be noted, however, that in the search results EEBO forcibly reintroduces the familiar distinctions. They are presented in units of discrete texts, with metadata which allows sorting only by author, title or date.

Thus EEBO introduces some decontextualisation of early modern English writing from the conventional classificatory categories of author, subject and date, and creates some recontextualisation according to the categories of corpus linguistics. However, the constraints imposed by EEBO's search engine and methods of searching, which were designed without the requirements of linguists and conceptual historians in mind, deprive the user of much of the power of corpus-based methods. This joint article illustrates what can be done using EEBO alone, and how much more can be achieved when it is further remediated as a *bona fide* corpus.

2.2 Background—a historian of ideas discovers corpus linguistics

Pumfrey, the first author, explored for the first time the new but limited possibilities of EEBO when pursuing a very peripheral research question in 2006. In his study of dedicatory letters from scientists to patrons in Elizabethan and Jacobean England (1558–1625) he had found frequent references to two little-known classical figures, Momus and Zoilus, who represented potential critics of the scientist's work.

A valuable feature of the EEBO search page is its ability to perform Boolean searches, and to identify variant spellings and forms. From a search for 'Momu* OR Zoilu*' in the 12,000 searchable texts (at that time) he learned quickly that there were hits in 450 texts. It was time-consuming to tabulate these texts by decade, but rewarding to discover that that there were no occurrences before 1540, whereafter usage increased rapidly until the 1590s.[5] The ensuing decrease in use was more marked once Pumfrey realised the need to normalize the frequencies, to take account of the constant expansion of printing during the period. He considered his two days of manual analysis well-rewarded by the results. He had the solid conclusion that reference to Momus and Zoilus was a special feature of Elizabethan and Jacobean texts, and he could tentatively infer that writers in the period were unusually concerned with criticism of their books. These were novel results which could not have been obtained without EEBO's searchable sub-corpus. They prompted Pumfrey to make contact with Rayson and other computational and corpus linguists, and to ponder a larger project, of significance to conceptual history. That larger project informs this article, and we now describe its evolution.

A fundamental concept in modern science is that of the experimental method. Historians of early modern science know that the development of an experimental approach to knowledge of nature was a crucial part of what has been called 'the scientific revolution of the sixteenth and seventeenth centuries'—the era of EEBO. Specialists in the early history of experimental science will also be aware of a parallel linguistic development. As the performance of experiments became an increasingly important part of science, so the pre-existing word *experiment* (in English, or *experimentum* in Neo-Latin) took on its modern, scientific meaning (Schmitt, 1969).

Historians of experiment were aware that medieval and Renaissance writers, who normally wrote in Latin, had already used the word *experimentum*. However, they used it interchangeably with *experientia*. Likewise, the English word 'experiment' had the same meaning as 'experience'. Consequently English writers could report that it was their 'daily experiment' that the Sun rose in the east, or that a scientific claim accorded with their personal experiment.[6] As Peter Dear (1995) has shown, pre-modern scientists thought of 'experience' in a specific philosophical sense. They believed that nature normally behaved in law-like ways. People discovered these regularities by observing the same phenomenon over and over again. This repeated, common empirical knowledge of nature was what scientists called experience, and it was the evidential basis of their belief in natural laws. In this pre-modern model of scientific knowledge, experiments in the modern sense, rare events requiring careful design and special apparatus, did not provide evidence for reliable inference.

Existing studies understandably confined themselves almost exclusively to science texts written in Latin. Pumfrey wondered whether EEBO would permit a broader investigation of how the modern concept of experiment emerged in England and in English if he extended the manual method he had used to track Momus and Zoilus. Such an investigation was especially pertinent because it was scientists in England, following the lead of Sir Francis Bacon, who were most associated with the new experimentalism. Indeed, in 1660 Robert Boyle, a key protagonist, effectively invented and promoted the term 'experimental philosophy' (now termed 'experimental science') as a description of the new method in his seminal work *New Experiments Physico-Mechanical* (Boyle, 1660). The results were revealing, but the manual research process exposed how poorly the current EEBO website supports this kind of research, and yet how much potential there is in the EEBO corpus.

3 Two Opposing Methods and their Results

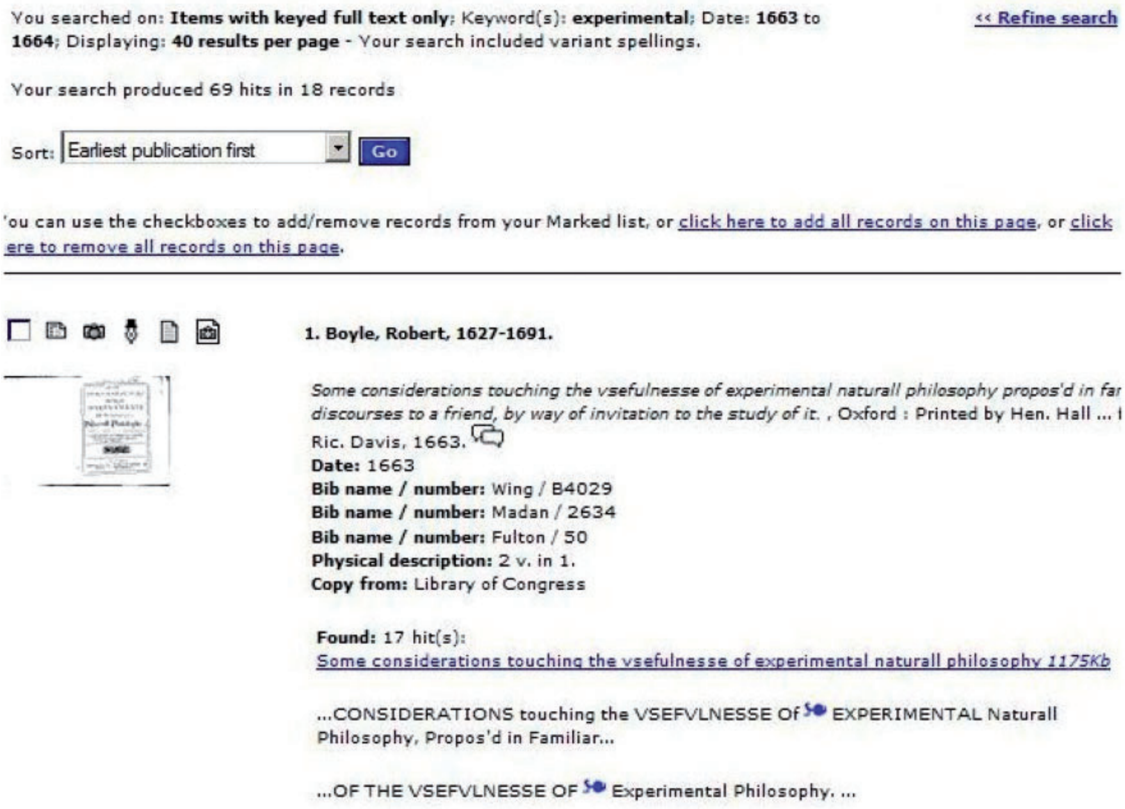
3.1 Manual method

The first problem is the slowness with which the EEBO website returns the results of a Boolean keyword search such as 'experiment*'. The results are presented in the usual form, i.e. a series of records of the relevant books, including full bibliographical metadata, plus up to five occurrences of the word, displayed in the centre of ten words of co-text. If the maximum of forty records is requested, the webpage can take more than a minute to load. Moreover, to view more than five hits in one record requires even more time as the whole text is searched and downloaded. In short, a manual method is barely feasible for any search term that appears thousands of times in hundreds of books.

As a consequence, Pumfrey had to abandon his analysis of 'experiment' in its variant forms because (in 2009) the search returned 20,909 hits in 3,241 records out of 25,271.[7]

He settled upon 'experimental' because it returned a manageable, if daunting, 2,700 hits in over 1,000 records.[8] He pasted these, with their co-text, into an Excel spreadsheet, arranged chronologically together with the metadata of date of publication, author and short title. Only the first five hits in any work are shown unless one downloads the entire text, which is slow. An additional obstacle was the presence in every co-text of special and hidden characters, such as the fish-shaped symbol in Fig. 1, which had to be deleted. The database was almost ready for analysis.

The first analysis was chronological: the hits were manually counted and classified by decade and



You searched on: **Items with keyed full text only**; Keyword(s): **experimental**; Date: **1663** to **1664**; Displaying: **40 results per page** - Your search included variant spellings.

Your search produced 69 hits in 18 records

Sort: Earliest publication first

'ou can use the checkboxes to add/remove records from your Marked list, or click here to add all records on this page, or click ere to remove all records on this page.

□ **1. Boyle, Robert, 1627-1691.**

Some considerations touching the vsefulnesse of experimental naturall philosophy propos'd in fa discourses to a friend, by way of invitation to the study of it. , Oxford : Printed by Hen. Hall ... f Ric. Davis, 1663.

Date: 1663
Bib name / number: Wing / B4029
Bib name / number: Madan / 2634
Bib name / number: Fulton / 50
Physical description: 2 v. in 1.
Copy from: Library of Congress

Found: 17 hit(s):
Some considerations touching the vsefulnesse of experimental naturall philosophy *1175Kb*

...CONSIDERATIONS touching the VSEFVLNESSE Of EXPERIMENTAL Naturall Philosophy, Propos'd in Familiar...

...OF THE VSEFVLNESSE OF Experimental Philosophy. ...

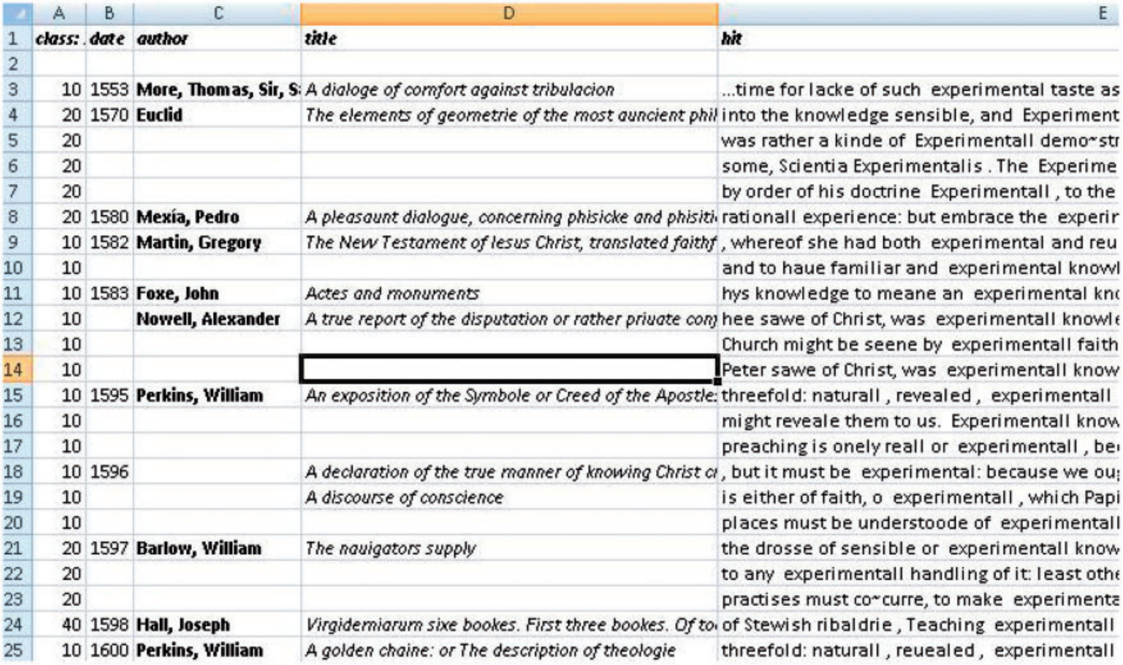
Fig. 1 Screenshot to show how EEBO returned the results of a keyword search for 'experimental' for the period 1663–64.

where appropriate, in periods of significant change, in periods of 2 years. Obviously, the raw hit count needed to be normalized. This was crudely done by painstakingly adding up the occurrence in EEBO for every time period of words whose frequency Pumfrey assumed would not change over time but which were likely to appear in texts discussing experiment. These were 'divine', 'natural' and 'et*'. [9]

The chronological analysis confirmed as expected that English writers used the word 'experimental' long before it acquired the modern scientific sense present in the example of 'experimental philosophy'. The first recorded use was in a religious context. In 1553, the first, posthumous edition of the 'martyr' Sir Thomas More's *Dialoge of comfort against tribulacion* mentioned the 'lacke of such experimental taste as God geveth here' on earth of the pleasures waiting in heaven.[10]

What was not expected was that meaning and uses of the word were overwhelmingly religious right up until the 1660s, and especially so during England's Interregnum (1649–60). They described an 'experimental Christianity'. An experimental Christian differed from the merely notional or formal Christian in that he knew Jesus experimentally, i.e. he could testify to a personal experience of the Holy Spirit dwelling in his heart. In order to plot the relationship between religious and scientific uses of the word it became necessary to classify the occurrences as religious or scientific, in effect to tag them semantically, as shown in Fig. 2. Pumfrey did this, subjectively and manually, on the basis of the ten words of co-text and, where applicable, his knowledge of the book.

The results, whilst not of first importance in this article, were remarkable, and are illustrated in Fig. 3.



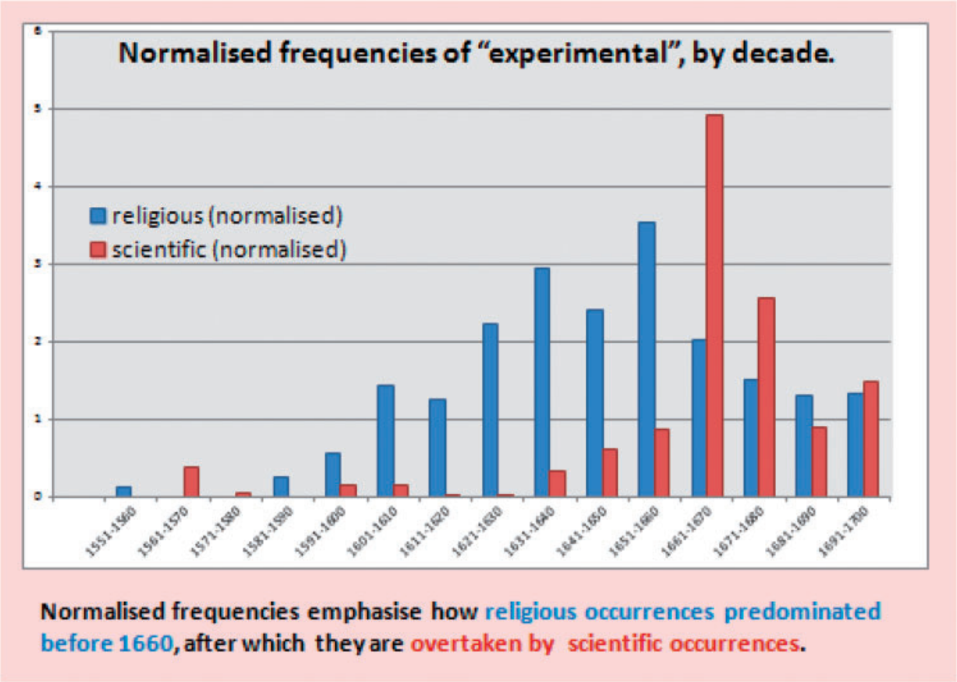
	A	B	C	D	E
1	*class:* .	*date*	*author*	*title*	*hit*
2					
3	10	1553	**More, Thomas, Sir, S**	*A dialoge of comfort against tribulacion*	...time for lacke of such experimental taste as
4	20	1570	**Euclid**	*The elements of geometrie of the most auncient phil*	into the knowledge sensible, and Experiment
5	20				was rather a kinde of Experimentall demo~str
6	20				some, Scientia Experimentalis . The Experime
7	20				by order of his doctrine Experimentall , to the
8	20	1580	**Mexía, Pedro**	*A pleasaunt dialogue, concerning phisicke and phisiti*	rationall experience: but embrace the experir
9	10	1582	**Martin, Gregory**	*The New Testament of Iesus Christ, translated faithf*	, whereof she had both experimental and reu
10	10				and to haue familiar and experimental knowl
11	10	1583	**Foxe, John**	*Actes and monuments*	hys knowledge to meane an experimental knc
12	10		**Nowell, Alexander**	*A true report of the disputation or rather priuate con*	hee sawe of Christ, was experimentall knowle
13	10				Church might be seene by experimentall faith
14	10				Peter sawe of Christ, was experimentall know
15	10	1595	**Perkins, William**	*An exposition of the Symbole or Creed of the Apostle:*	threefold: naturall , revealed , experimentall
16	10				might reveale them to us. Experimentall know
17	10				preaching is onely reall or experimentall , be
18	10	1596		*A declaration of the true manner of knowing Christ c*	, but it must be experimental: because we ou
19	10			*A discourse of conscience*	is either of faith, o experimentall , which Papi
20	10				places must be understoode of experimentall
21	20	1597	**Barlow, William**	*The nauigators supply*	the drosse of sensible or experimentall know
22	20				to any experimentall handling of it: least othe
23	20				practises must co~curre, to make experimenta
24	40	1598	**Hall, Joseph**	*Virgidemiarum sixe bookes. First three bookes. Of to*	of Stewish ribaldrie , Teaching experimentall
25	10	1600	**Perkins, William**	*A golden chaine: or The description of theologie*	threefold: naturall , reuealed , experimentall

Fig. 2 The first entries in the manually compiled database. In the first column, a classification of ten signified a religious occurrence, twenty scientific, thirty political and forty other.

The predominantly religious meaning of 'experimental' was overtaken by the new scientific meaning within a decade of the publication of Boyle's *New Experiments* in 1660. Although this manual method of analysis took more than 4 weeks, the results and the questions they raised for future research made the effort worthwhile, even though it is now clear that automated methods can achieve similar results in minutes.

3.2 Automatic method

In response to the practical problems and time limitations that the manual method imposes on the researcher, we sought to develop a more scalable automatic method. The manual method as described in the previous section exploits tools that have been designed for historians who are exploring in depth a small set of works. In contrast, corpus and computational linguistics methods often take a step away from the individual texts and provide an overview or route in via analysis of the whole collection back down to fragments (i.e. concordance lines) of the texts. Commonly used tools in corpus linguistics are WordSmith (Scott, 2008) and AntConc (Anthony, 2011) which allow the user to create a corpus from a set of files and then carry out full text searches (concordances), count words (frequency lists), see which words regularly occur together (collocations) and explore how words are spread within the corpus (distribution). Such methods can be applied to the whole corpus or a subset selected by specific metadata such as author, publication date and location. Applying corpus methods to a dataset such as EEBO requires a powerful tool that pre-indexes the collection in order to retain fast access times. In our research described here, we have employed CQPweb (Hardie, forthcoming). This is one of a family of fourth generation (web-based) concordancing tools that exploit the power of web servers to deliver the functionality of PC-based tools on a larger scale (McEnery and Hardie, 2011) to the researcher's desktop via a web browser. The corpus user then carries out analyses via the web browser



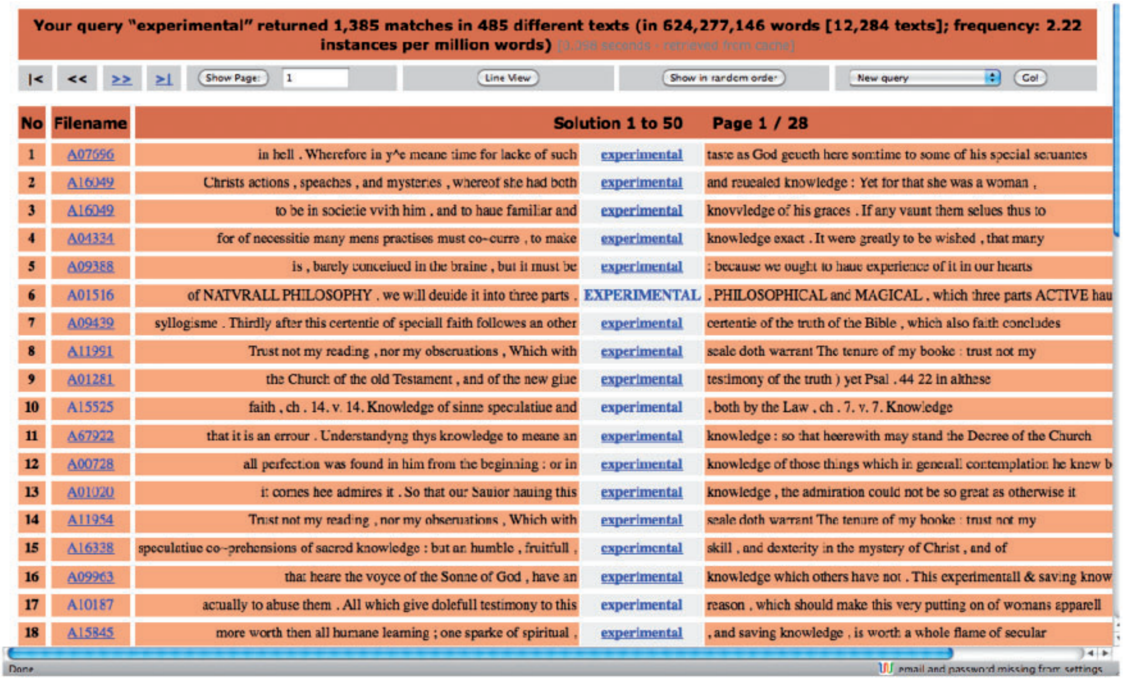
Normalised frequencies emphasise how religious occurrences predominated before 1660, after which they are overtaken by scientific occurrences.

Fig. 3 Chart of results obtained manually. The *x*-axis plots decades. The *y*-axis is linear but nominal.

interface akin to searching the web via a search engine. The CQPweb software is built on the indexing power of the Corpus Query Processor (CQP), which is part of the Open Corpus Workbench.[11]

The first task was to prepare the full text of the EEBO files for indexing in CQPweb. We had access to the EEBO-TCP data via the UK JISC agreement[12] and this was delivered on DVDs to us in SGML and XML formats. For indexing purposes, CQP requires that input text be converted into a vertical form with one word per line and any extra annotation alongside in further columns. A Perl script was developed which converted each XML file into the CQP format retaining the content of the input. For the first version of our indexing, we extracted a minimal amount of metadata in terms of the date of publication so that we could index each file by decade. A Java tool was built to assist with this part so that we could extract date information automatically wherever possible. However, several hundred files contained date fields that needed to be manually checked, for example where dates were shown in roman numerals, not fully specified or where they were extracted from other parts of the XML header. In total, we indexed 12,284 texts containing 624,277,146 words. It should be noted that even for the current state of the art in corpus linguistics, this is a 'larger than average' dataset. Only a handful of corpus collections such as those on Mark Davies' online interface[13] and Sketch Engine's Oxford English Corpus[14] are comparable in size. To some extent, we have repeated what indexing has been done for the online EEBO interface[15] but that interface, as discussed above, does not easily facilitate the application of corpus methodologies.

Once the corpus data was indexed in CQPweb, we were able to carry out searches very quickly with the web interface. Figure 4 shows a search across the full dataset for Pumfrey's chosen word 'experimental' which took <0.1 s once the above preparatory work had been undertaken. The concordance view shows a short section of the surrounding context of each occurrence as well as the filename in which the line originates. A tooltip shows the metadata



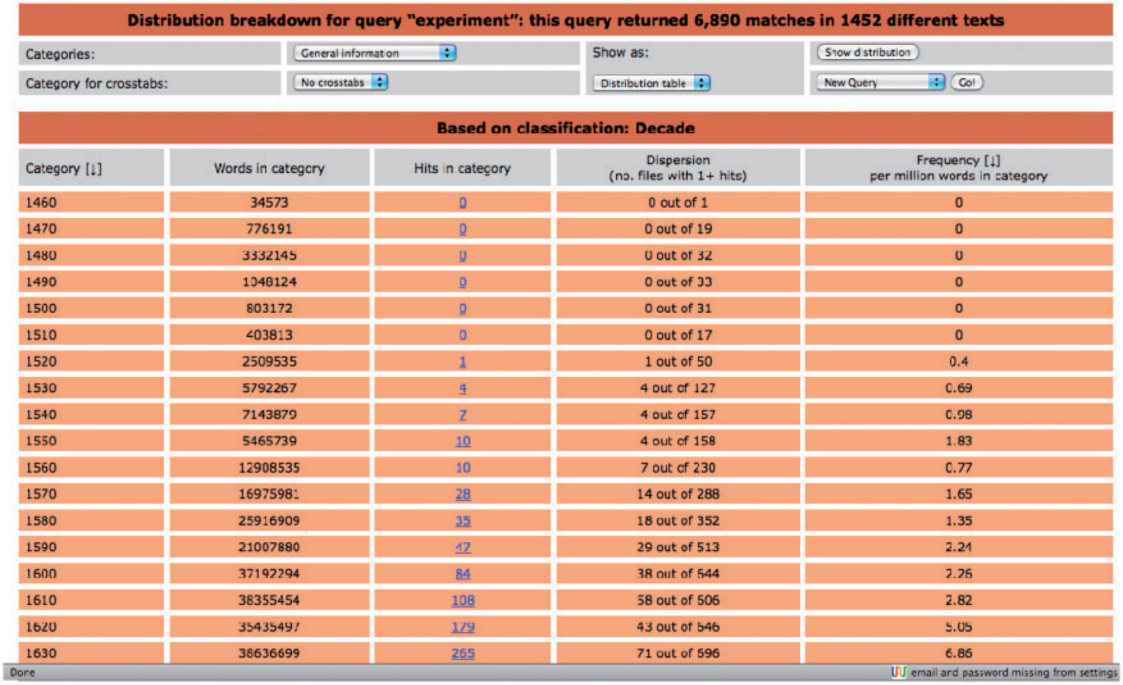
Your query "experimental" returned 1,385 matches in 485 different texts (in 624,277,146 words [12,284 texts]; frequency: 2.22 instances per million words) [0.398 seconds - retrieved from cache]

No	Filename	Solution 1 to 50		Page 1 / 28
1	A07696	in hell . Wherefore in y^e meane time for lacke of such	experimental	taste as God geueth here somtime to some of his special seruantes
2	A16049	Christs actions , speaches , and mysteries , whereof she had both	experimental	and reuealed knowledge : Yet for that she was a woman ,
3	A16049	to be in societie vvith him , and to haue familiar and	experimental	knovvledge of his graces . If any vaunt them selues thus to
4	A04334	for of necessitie many mens practises must co-curre , to make	experimental	knowledge exact . It were greatly to be wished , that many
5	A09388	is , barely conceiued in the braine , but it must be	experimental	: because we ought to haue experience of it in our hearts
6	A01516	of NATVRALL PHILOSOPHY . we will deuide it into three parts .	EXPERIMENTAL	. PHILOSOPHICAL and MAGICAL , which three parts ACTIVE hau
7	A09439	syllogisme . Thirdly after this certentie of speciall faith followes an other	experimental	certentie of the truth of the Bible , which also faith concludes
8	A11991	Trust not my reading , nor my obseruations , Which with	experimental	seale doth warrant The tenure of my booke : trust not my
9	A01281	the Church of the old Testament , and of the new giue	experimental	testimony of the truth) yet Psal . 44 22 in althese
10	A15525	faith , ch . 14. v. 14. Knowledge of sinne speculatiue and	experimental	. both by the Law , ch . 7. v. 7. Knowledge
11	A67922	that it is an errour . Understandyng thys knowledge to meane an	experimental	knowledge : so that heerewith may stand the Decree of the Church
12	A00728	all perfection was found in him from the beginning : or in	experimental	knowledge of those things which in generall contemplation he knew b
13	A01020	it comes hee admires it . So that our Sauior hauing this	experimental	knowledge , the admiration could not be so great as otherwise it
14	A11954	Trust not my reading , nor my obseruations , Which with	experimental	seale doth warrant The tenure of my booke : trust not my
15	A16338	speculatiue co-prehensions of sacred knowledge : but an humble , fruitfull ,	experimental	skill , and dexterity in the mystery of Christ , and of
16	A09963	that heare the voyce of the Sonne of God , have an	experimental	knowledge which others have not . This experimentall & saving know
17	A10187	actually to abuse them . All which give dolefull testimony to this	experimental	reason , which should make this very putting on of womans apparell
18	A15845	more worth then all humane learning ; one sparke of spiritual ,	experimental	, and saving knowledge , is worth a whole flame of secular

Fig. 4 Simple search for 'experimental'.

available for each filename, in this case the decade and the number of words in the text. A larger amount of context can be shown by clicking on the search word in the centre of the concordance. CQPweb also integrates a database and this makes frequency analysis possible. Figure 5 shows the results of the investigation which Pumfrey had abandoned as impossible, i.e. the distribution of the word 'experiment' by decade. This operation took <1 s. The figure shows the actual frequency, normalized frequency per million words and dispersion information (the number of files containing the search word). Thus, although 'experiment' is much more frequent than 'experimental', the results can be displayed almost instantly due to the indexing within the system. Figure 6 shows this information in graph form within the CQPweb tool. Powerful regular expression searches can be carried out, for example 'experiment*' would show most variants of the base form.[16]

As described in this section, the prototype EEBO interface in CQPweb makes word level searches trivially easy and very fast. However, the lack of annotation in the corpus does not currently allow the separation of religious and scientific senses of 'experiment' and related forms. In the previous section, we described how each instance was classified manually to permit further quantitative analysis. To achieve this separation of senses automatically and on a larger scale, we would need to apply a word sense disambiguation system of some sort. Although this goes beyond the scope of our initial prototype, we wished to investigate whether such a system could be made to work on the historical data contained in the EEBO corpus. Previous research has reported a dramatic reduction in the accuracy and robustness of automatic analysis tools from computational and corpus linguistics when they are applied to historical data (Archer et al, 2003; Rayson et al, 2007; Baron et al, 2009). This is due to these tools being trained on modern language data which is less 'noisy' than historical data in terms of OCR errors or spelling variation (Lopresti et al, 2009).



Distribution breakdown for query "experiment": this query returned 6,890 matches in 1452 different texts

Categories: General information

Category for crosstabs: No crosstabs

Show as: Distribution table

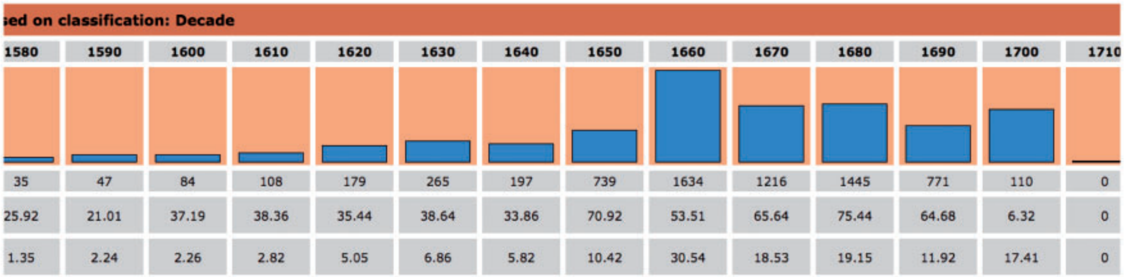
Based on classification: Decade

Category [↓]	Words in category	Hits in category	Dispersion (no. files with 1+ hits)	Frequency [↓] per million words in category
1460	34573	0	0 out of 1	0
1470	776191	0	0 out of 19	0
1480	3332145	0	0 out of 32	0
1490	1048124	0	0 out of 33	0
1500	803172	0	0 out of 31	0
1510	403813	0	0 out of 17	0
1520	2509535	1	1 out of 50	0.4
1530	5792267	4	4 out of 127	0.69
1540	7143879	7	4 out of 157	0.98
1550	5465739	10	4 out of 158	1.83
1560	12908535	10	7 out of 230	0.77
1570	16975981	28	14 out of 288	1.65
1580	25916909	35	18 out of 352	1.35
1590	21007880	47	29 out of 513	2.24
1600	37192294	84	38 out of 544	2.26
1610	38355454	108	58 out of 506	2.82
1620	35435497	179	43 out of 546	5.05
1630	38636699	265	71 out of 596	6.86

Done

email and password missing from settings

Fig. 5 Distribution table for ‘experiment’ by decade.



sed on classification: Decade

1580	1590	1600	1610	1620	1630	1640	1650	1660	1670	1680	1690	1700	1710
35	47	84	108	179	265	197	739	1634	1216	1445	771	110	0
25.92	21.01	37.19	38.36	35.44	38.64	33.86	70.92	53.51	65.64	75.44	64.68	6.32	0
1.35	2.24	2.26	2.82	5.05	6.86	5.82	10.42	30.54	18.53	19.15	11.92	17.41	0

Fig. 6 Distribution plot for ‘experiment’ by decade.

Using the gold-standard dataset that contained a manual classification for each occurrence of ‘experimental’ into a religious or scientific sense we wanted to build a trained model that could later be applied to the much larger set of occurrences for ‘experiment’ and other forms. Given that collocation is one of the key factors in determining the sense of a word in running text, we decided to investigate the concepts that were located in the close context of each occurrence. We took the concordance lines from the manual analysis and applied semantic field annotation via the USAS tagger (Rayson et al, 2004). This assigns a semantic field tag to each word or multi-word expression in the concordance lines with a high degree of accuracy in modern data. It should be noted that USAS did not make use of the manual tagging at this point to differentiate scientific and religious senses of ‘experimental’. Instead, we counted and conflated the tags on the words in the concordance lines to observe any differences between the religious and scientific senses. In the USAS tagset, the top level domain ‘S’ includes words and phrases related to religion and the top level domain ‘Y’ includes scientific terminology. Figure 7 shows the relative frequencies of the USAS top level domains for the contextual words grouped into four data points in time: 1550, 1600, 1650 and 1700. For each data point, we show two charts: manually marked religious contexts on the left and manually marked scientific contexts on the right. The progression through the four data points illustrates the reduction in religious terms (S category) over time in both religious and scientific contexts. This is most sharply observed between 1550 and 1600 in the scientific contexts. Scientific concepts (Y category) are rarer overall

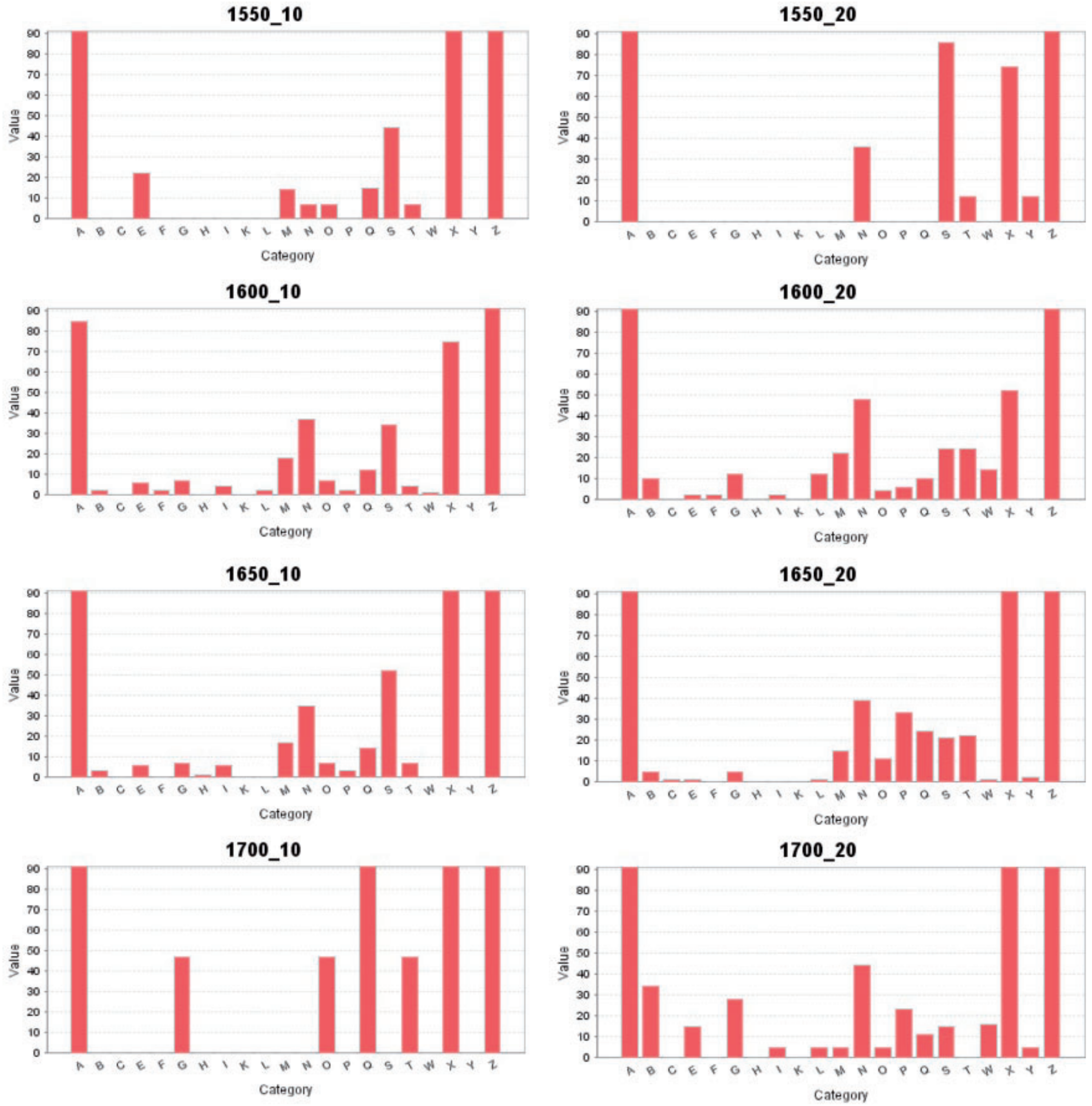


Fig. 7 Semantic contexts for religious and scientific senses of ‘experimental’.

and do not appear at all within the religious contexts, but do occur between 0% and 12% within scientific contexts.

Our initial prototypes here have shown the potential of the automatic method for exploring historical data using computational and corpus-based methods. The main advantage is that results (when properly annotated) are returned within seconds and tedious manual counting can be automated. This permits a much more iterative approach where hypotheses can be tested and investigations can proceed quickly in a way that is more driven by the data itself.

4 Conclusion

In this article, we have highlighted the need to apply computational and corpus methods to a historical corpus in order to support investigations in what can be called conceptual history. In particular, we showed how turning the EEBO dataset into a corpus and indexing it within a corpus linguistics style interface permitted the extension of a study on ‘experimental’ to the more frequent ‘experiment’. We also described our first investigations into the automatic separation of religious and scientific senses of these forms. The automatic method has been shown to replicate the manual method while being much more efficient. This speed gain will allow future scholars to carry out similar studies in an iterative fashion forming hypotheses and scanning results within minutes rather than after weeks of painstaking work. In turn, this will support the formation of new research questions that would not have been considered previously because the practical process of searching and categorizing examples is better supported.

The driver for this methodological innovation has been the original research question posed by Pumfrey. The uses of the word ‘experimental’ were overwhelmingly religious until the 1660s and then rapidly (within 10 years or so) became dominated by the new scientific meaning.

In terms of limitations, the current prototype uses the original plain versions of the EEBO-TCP files. To provide much richer search functionality, we are currently processing the full texts with a standard linguistic annotation pipeline to attach lemmas, part of speech tags and semantic analysis. In order to sustain high accuracy of these tools we plan to apply the variant detector software (VARD) (Baron and Rayson, 2009) which matches historical spelling variants to modern equivalents with high accuracy. This needs to be applied to the full running text rather than the indexed word list (as in the current EEBO web front end) so that we can fully benefit from the results with the three levels of linguistic annotation. Our prototype interface displays metadata consisting only of the file reference and the date so that we can index the files by decade. We have extracted metadata from the XML headers and need to compare this with full MARC records for EEBO to determine its consistency. The next version of our CQPweb interface will integrate the linguistic annotation, spelling variation patterns and the metadata to permit improved searching and research.

Acknowledgements

The authors would like to thank Sebastian Hoffmann, Alistair Baron, John Hutchinson, and Andrew Hardie for assistance with EEBO corpus conversion, indexing and metadata extraction, and Neil Foxlee for his comments on part of the paper. The authors would also like to thank Proquest and TCP colleagues at the University of Michigan for providing copies of the full text of the keyboarded EEBO-TCP data. This work was partly supported by the European Union CLARIN project [Grant No. 212230].

References

Anthony, L. (2011). *AntConc (Version 3.2.2) Computer Software*. Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp/ (accessed 15 May 2012).

Archer, D., McEnery, T., Rayson, P., and Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D., Rayson, P., Wilson, A., and McEnery, T. (eds), *Proceedings of Corpus Linguistics 2003*. Lancaster, UK: Lancaster University, pp. 22–31.

Argamon, S. and Olsen, M. (2009). Words, patterns and documents: experiments in machine learning and text analysis. *Digital Humanities Quarterly*, **3**(2). http://www.digitalhumanities.org/dhq/vol/3/2/000041/000041.html (accessed 20 September 2011).

Bamman, D. and Crane, G. (2011). *Measuring historical word sense variation, Proceedings of JCDL'11*. Ottawa, ON, Canada, pp. 13–17 June 2011.

Baron, A. and Rayson, P. (2009). Automatic standardisation of texts containing spelling variation. How much training data do you need? In *proceedings of Corpus Linguistics 2009*, University of Liverpool, Liverpool, July 2009.

Baron, A., Rayson, P., and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, **20**(1): pp. 41–67.

Boyle, R. (1660). *New Experiments Physico-Mechanicall, Touching the Spring of the Air*. Oxford: H. Hall.

Crane, G. (2006). What do you do with a million books? *D-Lib Magazine*, **12**(3). http://www.dlib.org/dlib/march06/crane/03crane.html (accessed 23 August 2011).

Dear, P. (1995). *Discipline and Experience. The Mathematical Way in the Scientific Revolution*. Chicago, IL: Chicago University Press.

Elson, D., Dames, N., and McKeown, K. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 138–147.

Halford, J. (2011). *Semantic Shifts: The Emergence of New Philosophy and the Construction of Meaning in the Discourses of Seventeenth Century Philosophy*. Unpublished M.A. dissertation, Lancaster University.

Hardie, A. (forthcoming). CQPweb—combining power, flexibility and usability in a corpus analysis tool.

Kichuk, D. (2007). Metamorphosis: remediation in Early English Books Online (EEBO). *Literary and Linguistic Computing*, **22**(3): 291–303.

Kytö, M. (1996). *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*, 3rd edn. Helsinki: Helsinki University Printing House. http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM (accessed 24 August 2011).

Lehmann, H. and Richer, M. (1996). *The Meaning of Historical Terms and Concepts*. Studies on Begriffsgeschichte. Washington, DC: German Historical Institute. Occasional Paper No. 15. http://www.ghi-dc.org/publications/ghipubs/op/op15.pdf/ (accessed 15 November 2011).

Lopresti, D., Roy, S., Subramaniam, L. V., and Schulz, K. (2009). Introduction to the special issue on noisy text analytics. *International Journal on Document Analysis and Recognition*, **12**(3): 139–140.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics*. Cambridge: Cambridge University Press.

Pumfrey, S. (2006). Managing Momus: following the fortuna and frequency of a trope in Early English Books Online. In Rayson, P. and Archer, D. (eds), *Workshop on Historical Text Mining*. http://ucrel.lancs.ac.uk/events/htm06/ (accessed 13 September 2011).

Rayson, P., Archer, D., Piao, S., and McEnery, T. (2004). The UCREL semantic analysis system. In *Beyond Named Entity Recognition Semantic labelling for NLP Tasks in Association With 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 25 May 2004.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the bard: evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In Davies, M., Rayson, P., Hunston, S., and Danielsson, P. (eds), *Proceedings of Corpus Linguistics 2007*. Lancaster, UK: UCREL, Lancaster University.

Schmitt, C. B. (1969). Experience and experiment: a comparison of Zabarella's view with Galileo's in De Motu. *Studies in the Renaissance*, **16**: 80–138.

Scott, M. (2008). *WordSmith Tools Version 5*. Liverpool: Lexical Analysis Software.

Wodak, R. and Meyer, M. (2009). *Methods of Critical Discourse Analysis*. London: Sage Publications.

Notes

[1] Concerning Discourse Historical Analysis (DHA), Wodak and Meyer acknowledge that a discourse must be situated in its historical context. However, unlike conceptual history, DHA does not address changes in historical contexts and associated discourses over a long period.

[2] Interestingly, he and other members of the 'Cambridge School' have often worked on political writers of the sixteenth and seventeenth centuries such as Machiavelli, Hobbes, and Locke. We are grateful to Neil Foxlee for his comments on this section.

[3] See ‘What is Early English Books Online’. http://eebo.chadwyck.com/about/about.htm (accessed 13 September 2011).

[4] We are indebted to Jacob Halford for this reference. See Halford (2011).

[5] For original results and presentation, see Pumfrey (2006). Repeating the search in September 2011, using CQPweb on 25,000 texts, confirmed the result—in under 10 s.

[6] See, for example Nathaniel Carpenter, *Geographie delineated forth in two books* (2nd edn; Oxford, 1635), p. 149. He discusses those who hold that there is a lot of subterranean water. ‘This they prove from the daily experiment of such as diggs diverse wells and deep trenches in the Earth; Who many times under the Earth; find not only many rivers and ponds, but many times happen upon so great abundance of Water, that they can neither find the bottom or bounds thereof’. See also Samuel Hammond, *The Quakers house built upon the sand* (London, 1658), p.23. He discusses whether the Bible’s advice to try all things was a license to get drunk or to have fulfilment of ‘any other sinfull lusts in our personal experiment’.

[7] In September 2011, this had risen to 23,789 hits in 3,658 records from 32,863.

[8] In September 2011, this had risen to 2,946 hits in 1,172 records. He searched for variant spellings but not variant forms. The inclusion of variant forms in September 2011 produces 4,669 hits in 1,772 records.

[9] The stem ‘et*’ was included in order to ensure that Latin texts were included in the sample.

[10] In September 2011, this remains the first occurrence.

[11] http://cwb.sourceforge.net/.

[12] http://www.jisc-collections.ac.uk/Catalogue/Overview/index/1006.

[13] http://corpus.byu.edu/.

[14] http://oxforddictionaries.com/page/oec.

[15] http://eebo.chadwyck.com/home.

[16] Note that this regular expression would not locate variants such as ‘experiment’ or, if it exists, ‘experiment’. Thus, more complex expressions like ‘experiment*’ would need to be used on a word based search.