“Raw Data” Is an Oxymoron

Edited by Lisa Gitelman

The MIT Press
Cambridge, Massachusetts
London, England

been nothing odd about using the term "data" to refer to facts discerned through experimentation, but here Colson uses "data" in the usual competing sense of principles or axioms given on the basis of which methods may be devised and facts discovered.

This is what one learns from reading. But what about the data on "data"? Might a quantitative approach be possible too? Might it be possible to study the corpus of printed English books in order to discover when "data" became a common term in English, how it was naturalized from Latin, and when it achieved its various meanings? Fortunately, today we are swimming in data for lexicographic research provided by both specialized and general databases along a spectrum from stand-alone electronic books to massive archiving and scanning endeavors such as Project Gutenberg and Google Books. Some of these resources are set up in ways that generally mimic print formats. They may offer various search features, hyperlinks, reformatting options, accessibility on multiple platforms, and so forth, but, in essence, their purpose is to deliver a readable product similar to that provided by pulp and ink. Others—still relatively few—foreground the aggregate and statistical features of the textual corpora that they access, and in a few cases

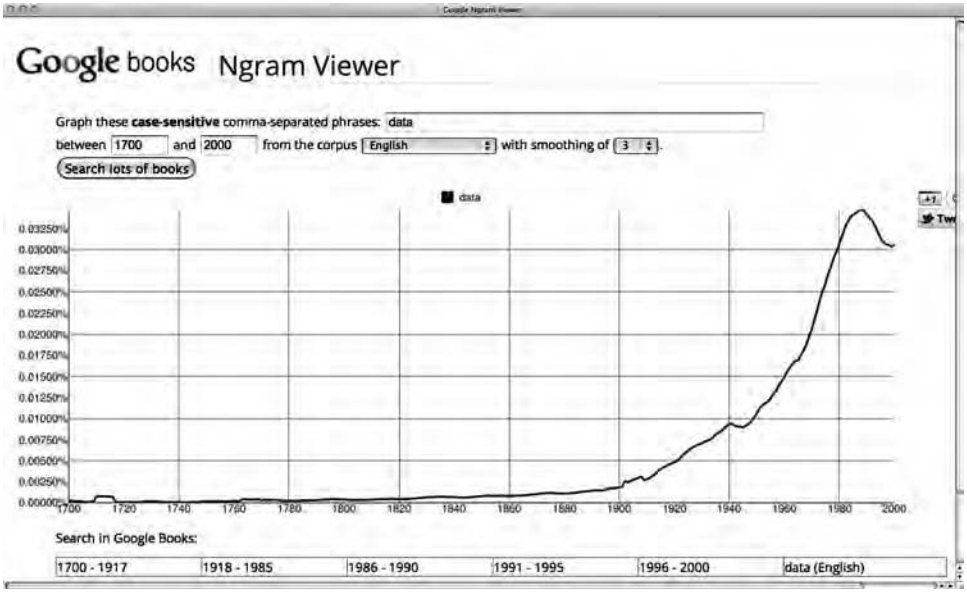

Figure 1.2 Image: Relative frequency of "data" in Google Books, by year, 1700–2000, generated by Google Ngram Viewer.

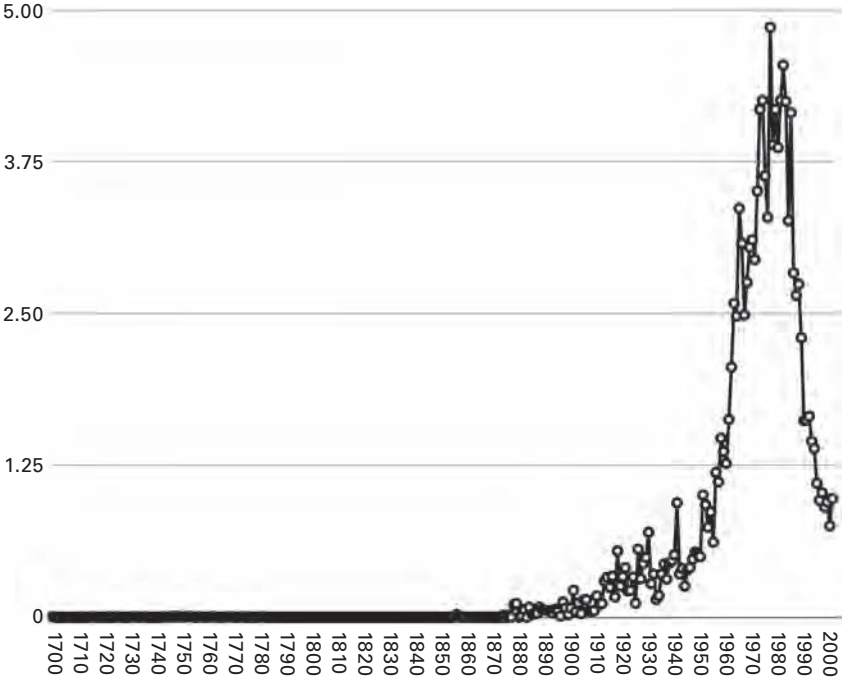

Figure 1.3 Relative frequency of "data" in Google Books corpus, 1700–2000, generated manually. *Note:* Data generated by repeated date-limited Google searches.

they do so even to the exclusion of the possibility of conventional reading, from beginning to end.

Much has been written about Google Books, but a large part of this scholarly literature has focused on the ways in which Google interacts with and places stress upon authors, publishers, libraries, and competing databases—stress that largely has to do with the fate of books in the electronic age.[15] Since the beginning of 2011, however, new attention has been focused on the research potential of Google Books as a linguistic corpus rather than as an electronic library. To facilitate research, Google has been making its book corpus accessible in two new ways: the raw data, abstracted from individual works, can be downloaded for analysis according to the interests of individual researchers, or it can be searched through a simple online interface called the Google Books Ngram Viewer. An "ngram" is a phrase consisting of a defined number of words (n): the Ngram Viewer allows corpus searches on these phrases and returns statistical

results. While the Ngram Viewer is limited in the kinds of searches it can perform, its basic trick is already impressive: presented with one or more search phrases of up to five words and a historical timeframe, the Ngram Viewer can instantly produce a graph of relative usage frequency over time.

A team of Harvard researchers led by the physicist Erez Lieberman Aiden and the biologist Jean-Baptiste Michel designed the Ngram Viewer. They introduced it with a clever publicity strategy: they aimed both low and high, promoting the Ngram Viewer as both an amusing geegaw and a tool for serious scholarly research. In their January 2011 *Science* article, "Quantitative Analysis of Culture Using Millions of Digitized Books," Michel and Aiden present the Ngram Viewer as a tool for what they call *culturomics*, quantitative cultural analysis modeled on *genomics* and the other *-omic* fields booming in the natural sciences.[16]

Michel and Aiden's publicity strategy proved successful, stirring up notice in key media venues such as the *New York Times* and in the blogosphere, where the ease of use and linking prompted a lot of kitchen culturomics. Briefly, it seemed that everyone was ngramming.[17]

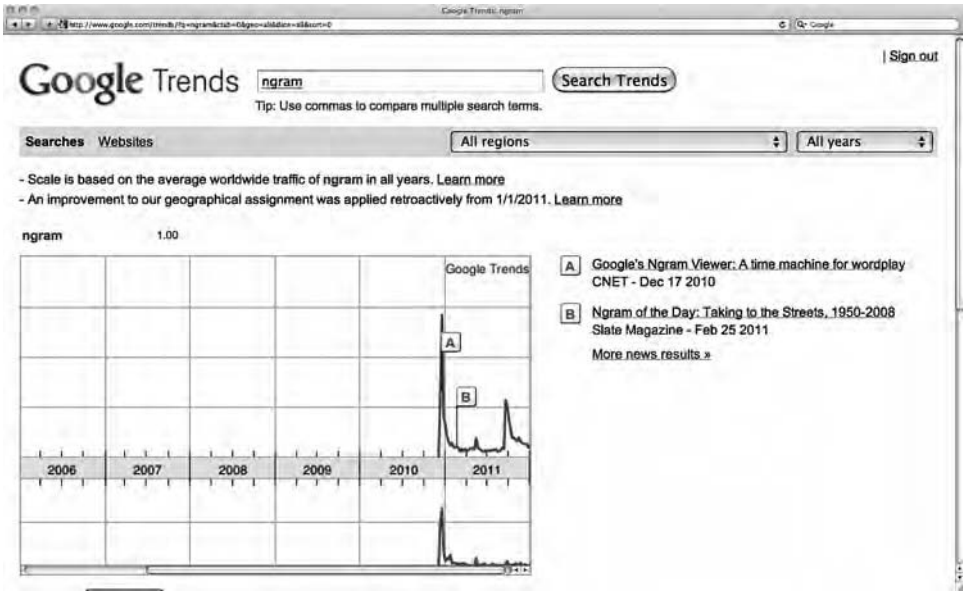

Figure 1.4 Search volume for "ngram," May 2010–December 2011, generated by Google Trends.

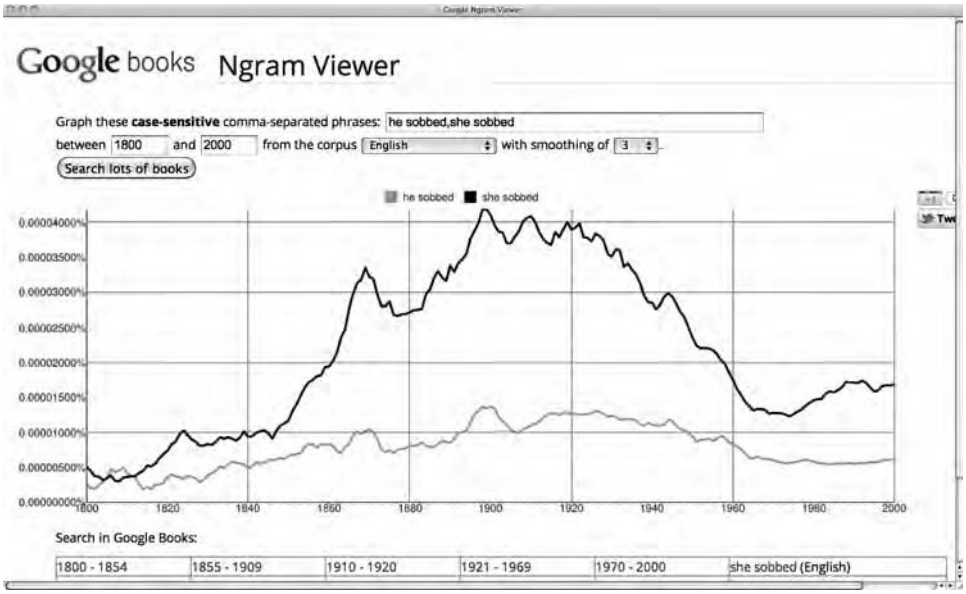

Figure 1.5 Relative frequency of "he sobbed" vs. "she sobbed" in Google Books, 1800–2000, as conceived by jezebel.com, generated by Google Ngram Viewer.

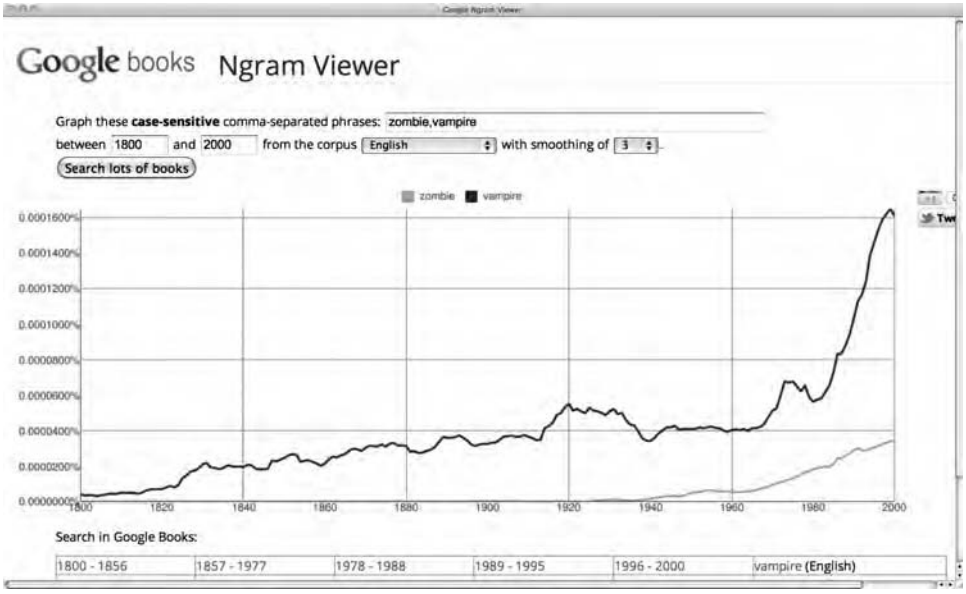

Figure 1.6 Relative frequency of "zombie" vs. "vampire" in Google Books, 1800–2000, as conceived by theatlantic.com, generated by Google Ngram Viewer.

The Harvard team got the ball rolling with some provocative diagrams of their own, plotting the changing importance in the linguistic corpus of a variety of people, events, and things. "'Galileo,' 'Darwin,' and 'Einstein' may be well-known scientists," write Michel and Aiden, "but 'Freud' is more deeply ingrained in our collective subconscious." "In the battle of the sexes, 'women' are gaining ground on the 'men.'"[18] Even *years* themselves could be tracked through the corpus, and these produced interesting regularities.

> Just as individuals forget the past, so do societies. To quantify this effect, we reasoned that the frequency of 1-grams such as "1951" could be used to measure interest in the events of the corresponding year, and we created plots for each year between 1875 and 1975. The plots had a characteristic shape. For example, "1951" was rarely discussed until the years immediately preceding 1951. Its frequency soared in 1951, remained high for 3 years, and then underwent a rapid decay, dropping by half over the next 15 years. Finally, the plots enter a regime marked by slower forgetting: Collective memory has both a short-term and a long-term component. But there have been changes. The amplitude of the plots is rising every year: Precise dates are increasingly common. There is also a greater focus on the present. For instance, "1880" declined to half its peak value in 1912, a lag of 32 years. In contrast, "1973" declined to half its peak by 1983, a lag of only 10 years. We are forgetting our past faster with each passing year.[19]

Precisely what one makes of these word-frequency trends is, of course, open to question. "Women" are not women, nor are "men" men, and there are good bureaucratic reasons unrelated to "collective memory" why 1951 would appear in documents from 1950, but the researchers argue that within the terms of the linguistic corpus the data speaks for itself.

The value of these diagrams immediately became a subject of scholarly debate. Some humanities scholars were highly skeptical; others, such as Anthony Grafton and Geoffrey Nunberg received them more favorably. Grafton invited Michel and Aiden to address the American Historical Association in two special sessions in 2011 and 2012, the second of which was substantially devoted to rebutting misconceptions including the notion that culturomics sets out to replace historians with computer programmers.[20]

More significant than the Ngram Viewer was Google's decision to make its raw data—if the term can be applied at all—available for download so that scholars could run the numbers themselves without going through the ngram interface.[21] This resource

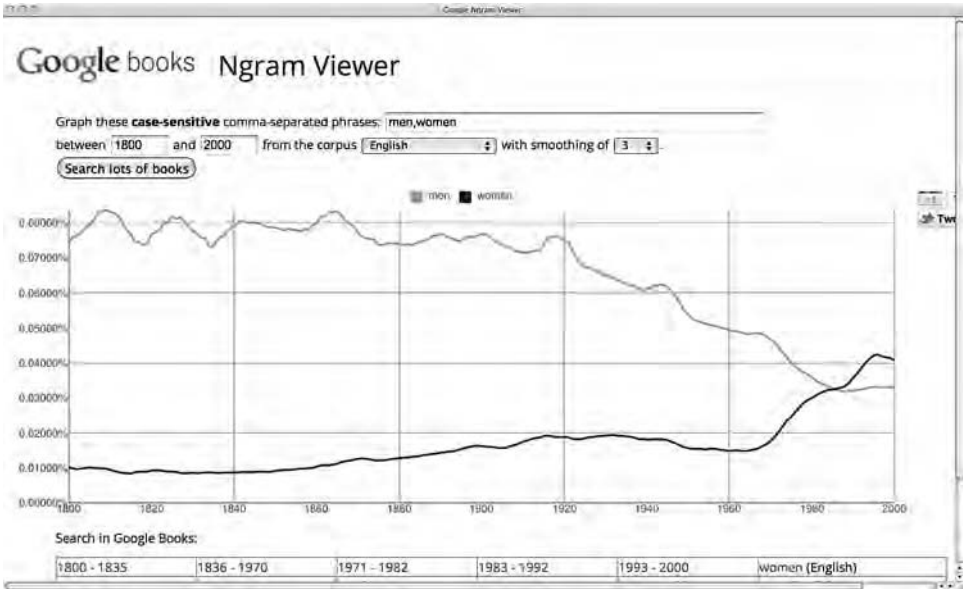

Figure 1.7 Relative frequency of "men" vs. "women" in Google Books, 1900–2000, as conceived by Michel and Aiden, generated by Google Ngram Viewer.

is likely to produce significant new research; at the same time, it should also elicit new critique.

At the time that I began research for this study, the Google Ngram Viewer was not yet available, and although it was possible to produce similar results by hand, at that time, Google Books offered neither the most obvious nor the most promising corpus with which to conduct a study such as this. As figure 1.3 demonstrates, repeating a search for the term "data" year by year and dividing the results by the results of searches for a control word in each of the same years in order to offset the effect of changing corpus size produces a curve consistent with that produced by the Ngram Viewer. This gives some indication of the promise of the corpus but only creases its surface.

In any event, I did not begin with Google Books, but rather with the subscription database Eighteenth-Century Collections Online (ECCO) from the educational publisher, Gale. ECCO is in many ways a primitive tool, and it suffers from several of the key faults for which Google Books has been criticized including inconsistent scanning quality. But ECCO has some notable advantages too. The corpus of ECCO I, based

on the English Short Title Catalogue, is large, comprising more than 136,000 unique titles, 155,000 volumes, and 26 million pages of text, backed up by an accessible analog microfilm collection from which it was generated and by well-catalogued books. A later supplement, ECCO II, raises the totals to 182,000, 205,000, and 32 million, respectively. Additionally, ECCO is well defined and much more stable than Google Books, which is changing all the time. ECCO's sources are well chosen, well known, and accessible. Its out-of-the-box search functions are more flexible. And at this point in time, the metadata is much better.

In fact, there is so much that is good about ECCO that a decade ago one might have thought ECCO would have had the kind of revolutionary effect on scholarship that Google and the culturomics advocates claim Google Books will have today. ECCO has opened new research avenues, but it hasn't made that kind of impact. In 2002, ECCO's publisher promoted it as a "research revolution." A breathless review called it a "resource that scholars will die for."[22] My graduate school friends called it "the dissertation machine."

The first thing that limited ECCO's effect, of course, is that it was not made openly available like Google Books. Additionally, though ECCO is a full-text database, it does not allow users to cut and paste text. And while users can search for words under the page images, they cannot reveal what the computer sees; they cannot see the characters that the computer recognizes in the page image. Ironically, over time ECCO's publisher has loosened its rules on downloading page images. So, for database subscribers, it has become easy and quick to download page images of full books from ECCO. Yet regular users cannot even download a single page of text as interpreted by ECCO's optical character recognition (OCR) software, which suggests that over time Gale determined there is no percentage in books, not even in digitized images of books, unless the books are already packaged as data.[23]

The future is in data.

Using ECCO, I began trying to understand the sense of "data" in Priestley. Happily, my first searches turned out to be promising. On the one hand, the ECCO results are consistent with those of Google. Speaking from a strictly quantitative point of view, the big "data" takeoff is unquestionably a post-Enlightenment phenomenon. On the other hand, ECCO shows clear trends in usage in the eighteenth century that laid the foundations for all later developments, which are difficult to perceive in Google's projections. The eighteenth century produced important new ways of thinking data, and Priestley was situated, felicitously, just exactly where those new ways of thinking happened.[24]

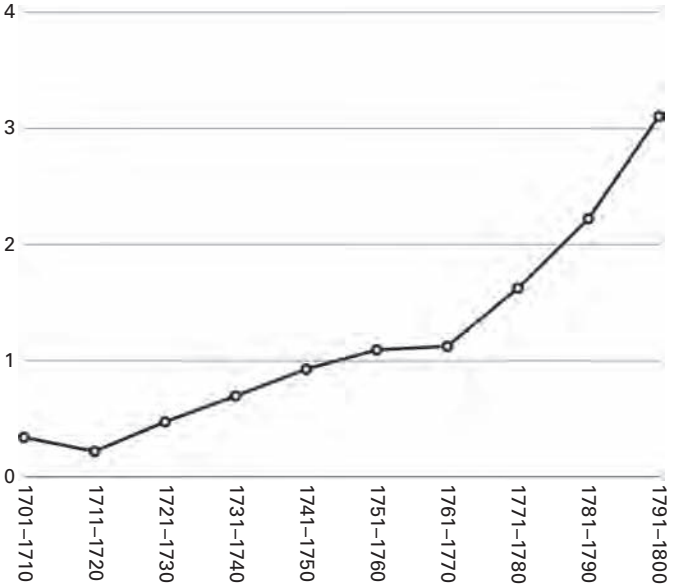

Figure 1.8 Percentage of works including the English common noun "data" in the corpus of ECCO I, 1701–1800.

The ECCO numbers are interesting, and they are also surprising in their clarity given the Google Books results, which suggest that the strong trends in the history of the term "data" begin in the nineteenth century and only accelerate definitively in the twentieth. First, from a statistical point of view, "data" was neither a rare nor an especially common term in eighteenth-century English. For comparison, a simple full-text ECCO search for the word "truth" produces hits in about 112,000 books or about 82 percent of the 136,000 total included in ECCO I. "Evidence" shows up in 66,000 books or 49 percent of total. "Fact" appears in about 35,000 or 28 percent. Even if we were to take the most generous count for "data," uncorrected for Latin usages, scanning errors, and so forth, we would find no more than 10,545 works in which "data" appears, or about 8 percent of total. And a stricter analysis of those occurrences produces a significantly smaller number, closer to 2 percent. In the eighteenth century, "data" was still a term of art.[25]

The further one goes into the data on "data," the more complicated it becomes. In my larger project, I aim to examine every usage of the term "data" in the ECCO corpus,

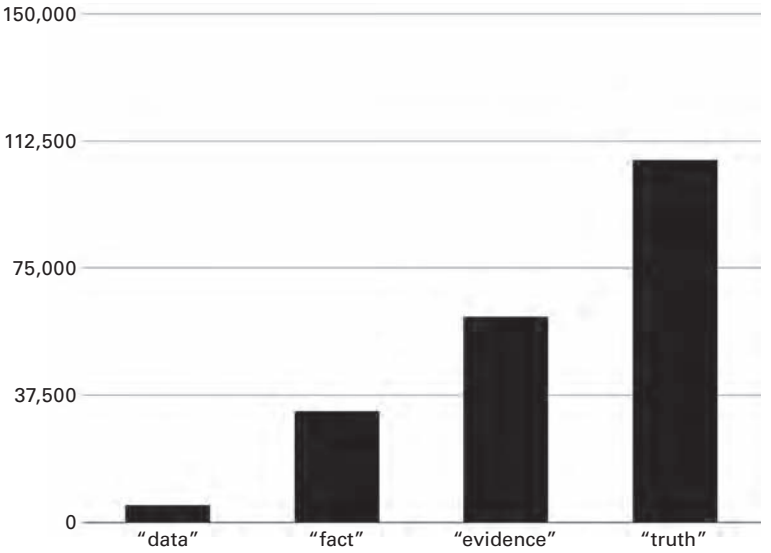

Figure 1.9 Works in the ECCO I corpus containing "data," "fact," "truth," and "evidence," 1701–1800.

not only to count for frequencies but also to examine each usage in context and to code each for semantic characteristics. The first and most pervasive problem that has turned up in this work is that a majority of usages of "data," even in the English language books in the database, turn out to be Latin. Often the Latin word *data* appears in quotations, footnotes, or conventional phrases such as *data desuper* (given from above) included in longer English texts. Other hits refer to the title of Euclid's book *Data*. Still others turn out to be scanning errors. In one instance, the search engine pulled up a reference to a certain King Data, a giant who fattened his twenty-five children by feeding them on puddings stuffed with enchanted herbs.[26] As a consequence it has been useful to examine hits individually, to sort the good from the bad and to code them, to engage in the constructive process of data making so well described in recent ethnographies of scientific practice. My own data may once have been raw, but by the time I began any serious interpretation, I had cooked it quite well.

Getting an accurate count for "data" has been a challenge. The process of scrutinizing each hit and eliminating those that were not English-language common nouns shrank the pool of viable instances. In fact, it certainly shrank the total number too far. Many works identified by ECCO as containing the word "data" in fact contain more instances

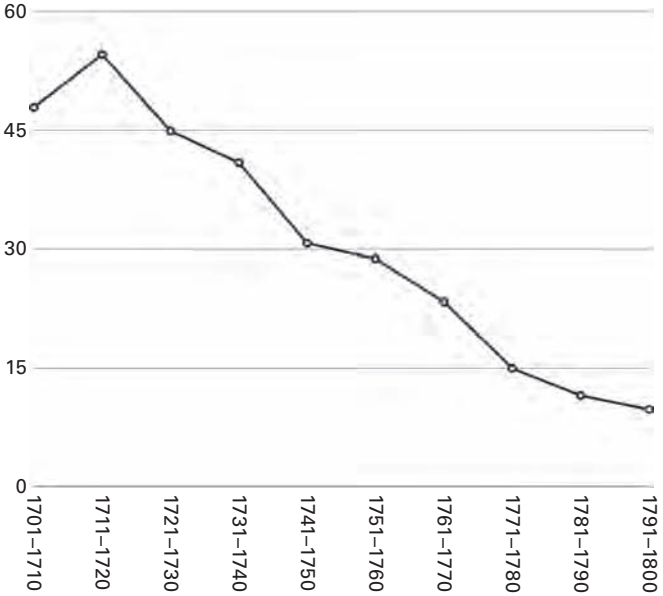

Figure 1.10 "Data" in Latin as Percentage of Total "Data" Hits in ECCO I, 1701–1800.

than ECCO shows; that is, even in works where the OCR algorithms correctly identified "data" once, they often missed it other times. And it is safe to say that there are at least as many instances in which data escaped the ECCO text search as instances in which ECCO thought it saw "data" but was mistaken. Estimating the numbers is challenging: on the one hand, there are more ways for an OCR program to overlook an instance of the word than to produce a false hit; on the other hand, since the term "data" frequently appears in a given work more than once (roughly 38 percent of the time according to my results), a significant number of OCR misses will be compensated for by correct recognitions of occurrences elsewhere in the same work.

Because the number of meaningful search hits for "data" turned out to be only about 2,300, it was possible to read them all well, to code them according to several protocols, and to produce very rich records for each instance. It was also possible to read extensively in the source works to gain a nuanced understanding of context. This has allowed me to pose a fairly wide variety of questions about the term and about key trends in its usage. And while this research is not yet complete, there are already a number of preliminary results, of which I highlight four, as follows.