Using *n*-Grams to Identify Time Periods of Cultural Influence

GREGORY P. KNIGHT and NASSEH TABRIZI, East Carolina University

An author's literary style is influenced by the cultural time period in which the author lives. The author's ideas, and the words chosen to express them, can help identify the cultural time period that most influenced the author.

Ideas are expressed in language through sequences of words called *n*-grams. Over the past several years, Google has been engaged in digitizing millions of books. As part of this endeavor, Google has created a database of *n*-grams extracted from these digitized books and has made the database available to researchers online. This is the first time ever that such an extensive repository of cultural data has been made available.

This study develops and tests an original method for utilizing Google's database to identify the cultural time period that most influenced the author of a published work. Several undisputed literary works are examined, from which sets of *n*-grams are extracted and compared against the Google database. The frequency and distribution of *n*-gram matches allow us to determine the cultural time period that most influenced the author. The method is also tested against several literary works having uncertain or disputed authorship and period of composition.

The results suggest that the method developed provides a reasonable approximation of the time period of greatest cultural influence for each book. Unexpectedly, the results tend to support conclusions reached by another researcher with regard to prior literary influences on the *Ern Malley Poems*. In addition, they lend support to early 19th-century origins for authorship of *Book of Mormon*

1. INTRODUCTION

Authors are influenced by the culture of the time period in which they live. Indeed, it is probably impossible to measure all the ways environment affects an author's output. The issues of society, the intricacies of personal relationships, the common affairs of everyday life—all are colored by the world in which the author grows and matures. To read an author's writings is to see and experience that world through the lens of the author's creativity, for even the most uniquely gifted artists cannot free

themselves entirely from the influence of their own surroundings. The environment not only provides the raw material from which the author draws but also supplies the wheel on which the author is formed.

While an author may be concerned with conveying things like ideals, morals, history, drama, and a host of other assorted and varied topics, they nevertheless all share one common fundamental element: All are expressed with *words*. Like paint to a painter, or stone to a sculptor, words are the medium through which authors practice their art. The time period and its attendant culture have a profound impact on the way authors choose individual words and group them together to effectively express their thoughts [McKee 1997].

Groups of words are sometimes referred to as *n-grams* and have been shown to be useful in identifying authorship and style [Grieve 2007; Shalhoub et al. 2010; Stamatatos 2009]. In this context, an *n*-gram is a phrase of *n* words used together as a group. For example, the phrase "in the morning" is a *3*-gram, and "speak up" is a *2*-gram (sometimes also called a *bigram* or *digram*). An author's output can be understood and analyzed as a collection of words grouped into *1*-grams, *2*-grams, *3*-grams, and so on.

Words, and the unique ways they are used in combination to capture ideas, bear the stamp of the time period and culture in which they thrive. As we continuously change the way we view and understand the world around us, we just as continuously change the way we describe it. As new ideas replace old ones, new words and phrases are born. The appearance of new inventions and discoveries is almost always accompanied by new words and phrases to describe and explain them, and as we discard old ways of thinking and doing, we likewise discard the outdated words and phrases that came with them.

Language characteristics change over time and are as dynamic as the society that uses them. One can be seen as the reflection of the other. The priorities of a society influence the things people talk and write about, and the things people talk and write about influence the priorities of society. The way people spoke and wrote English in the 18th century noticeably differs from the way they did so in the 19th and 20th centuries. Words that were common then are obscure now. Some words that are common now did not even exist then. We put our words together in new ways and stop using the old ways. The priorities of our society change, causing us to talk and write about different things in different ways.

As the popularity of words, phrases, and other elements of written communication ebb and flow over time, they leave their mark on those who preserve words in ink. Their written words can be likened to the fingerprints of a time period and the culture left behind. Those language elements that are the most influential on a society and culture will naturally be reflected in the written word of those who are part of that society and culture. Or, to look at it the other way around—an examination of the words and phrases used by an author should be indicative of the language elements that were most influential on the society and culture in which the author lived. Using *n*-grams as indicators of cultural patterns of expression can provide a way to identify a time period that influenced a particular literary work. Since these patterns of expression change over time, it should be possible to approximate the time period of greatest cultural influence on the author—if one has access to a rich collection of literary works covering a broad spectrum of time.

With the availability of the Google *n*-gram database, we have, for the first time, access to just such information.

1.1 The Google n-Gram Database

In December 2004, Google announced the "Google Print" Library Project. A few years earlier, Google had begun exploring the idea of digitally scanning every book in the world. Libraries were visited, and scanning techniques were tested and refined. By the time of their announcement, Google had formed partnerships with Harvard, Oxford, Stanford, the University of Michigan, and the New York

Public Library. Their goal was to digitize the books from these major libraries (whose holdings were estimated to number over 15 million volumes) and make their texts globally available on the worldwide web. In 2005; the project name was changed to its current designation: "Google Books." Currently, the project is scanning the collections of over 40 large libraries, along with many other books being made available directly from their publishers [Google 2015].

As part of this project, Google wanted to make it possible for users to perform textual searches against the contents of their online library. In order to accomplish this, Google employed Optical Character Recognition (OCR) to transform millions of page images into textual data. In 2007, two researchers from Harvard University—Jean-Baptiste Michel and Erez Lieberman Aiden—approached the Google Books team with the idea of making this new textual database available for research. Due to copyright restrictions, the full text of many books could not be released. However, it was possible for Google to chop the text into *n*-grams, gather statistics on the occurrences of these *n*-grams into a massive database, and make this database available for research and analysis.

From its digitized collection of over 15 million books, Google selected 5,195,769 books for inclusion in this new *n*-gram database. This subset of books was chosen based on both the quality of the digitized text produced by the OCR transformation process, as well as the reliability of each book's metadata (author, date of publication, etc.) Representing approximately 4% of all the books ever printed, the database contains volumes published from 1520 to 2008 in seven languages (Chinese, English, French, German, Hebrew, Russian, and Spanish)—a collection of over 500 *billion* words, grouped together into *n*-grams (specifically *1-* through *5*-grams), along with usage counts by year. These yearly usage counts include the total number of volumes in which each *n*-gram was found, the total number of pages within those volumes, and the total number of occurrences in those volumes overall. Only *n*-grams that occur at least 40 times within a volume were included in the Google database [Hayes 2007].

1.2   "Culturomics"

One of the most exciting consequences of the creation of this database has been the birth of a new field of study: *culturomics*. This term was first introduced in the journal *Science* in January 2011, in a article authored by Michel, Aiden, and several other researchers associated with the Google project. In their article, they define culturomics as "the application of high-throughput data collection and analysis to the study of human culture." Their research shows how large quantities of linguistic data can be utilized to aid in the study of human culture, thereby allowing researchers to "investigate cultural trends quantitatively." According to the authors, this new area of study "can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. 'Culturomics' extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities" [Michel et al. 2011; Bohanon 2011; Cohen 2010; Hand 2011; Aiden et al. 2007].

1.3   The Timeline Model

Of the many phases in the lifecycle of a published literary work, three are of particular relevance to this study: the period of cultural influence, the period of composition, and the date of publishing. Together, these comprise the Timeline Model (see Figure 1).

The timeline model relates these phases to one another with respect to time. It is somewhat simpler to explain these phases in reverse order. The *date of publishing* is simply the date that a literary work was prepared, placed into a fixed form, and generally made available to others. A well-established publish date is a prerequisite for all of the documents included in this study. Determination of the date of publishing is a straightforward task, since this information is normally included physically as part
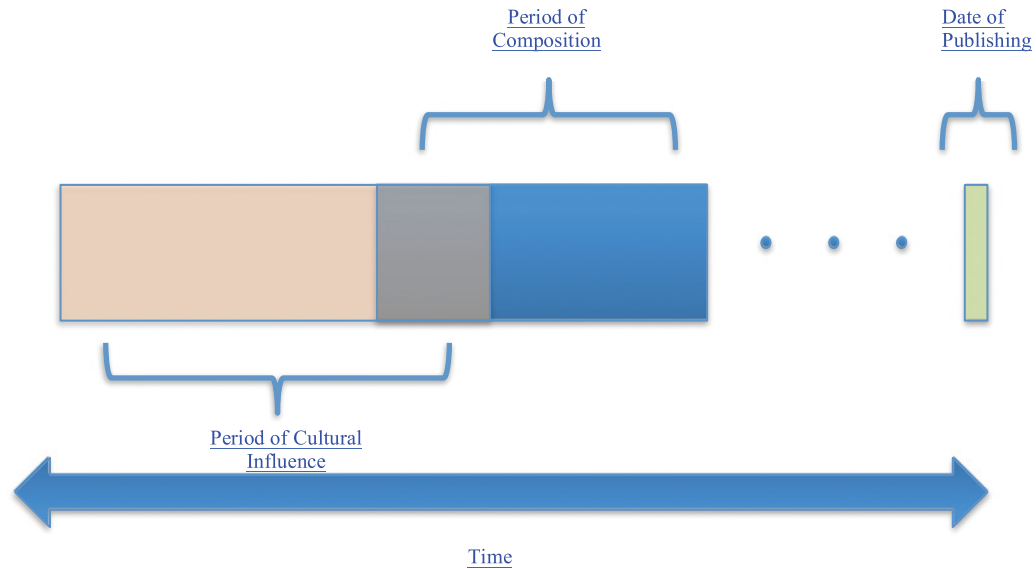
Fig. 1. The Timeline Model, containing periods of cultural influence, composition, and publishing.

of the work being published. It is generally found along with the copyright information that prefaces the text of the document.

The *period of composition* represents the period of time during which the author was actively engaged in developing and writing the work in question. Like Rome, most literary works of substance were not built in a day. Authors typically compose over periods of weeks, months, or even years. As mentioned earlier, one can generally pinpoint a specific date when a work was published, but one can almost never apply such precision to identifying a "date" of composition. Instead, it is more accurate to describe the process of composition as occurring over a period of time and always occurring before the date of publishing. Typically, the period of composition ends immediately prior to the date of publishing, but this does not necessarily have to be the case.

Finally, the first of these—the *period of cultural influence*—is the period of time during which the author was influenced (either consciously or otherwise) by the surrounding culture to the extent that the influence was made evident during the period of composition. While this period clearly precedes the period of composition, it is not at all evident when the period begins or when it ends. In fact, the two periods may overlap (as suggested in the Timeline Model). This time period is arguably the most difficult to delineate on the timeline; in fact, developing a method for discovering this period of cultural influence is the concentration of this study.

It should be noted that, for a document of a deceptive nature, the author might attempt to make the work appear as though the period of composition took place much earlier (or later) than it actually did. This is the case for documents that are forged under the name of an author from an earlier time period. From the timeline, however, it is evident that moving the period of composition artificially to a different point in time requires that the period of cultural influence be moved in tandem. The challenge for the forger, then, is to successfully imitate the cultural influence of the false time period, while at the same time quelling the influences of the true one. It is a hypothesis of this study that *n*-grams may capture the essence of these true cultural influences in such a penetrating and pervasive way to frustrate the forger's efforts to successfully expunge them.

1.4   A Method for Determining Periods of Cultural Influence

The focus of this study is the development and evaluation of an original method for determining periods of cultural influence on literary works. The main idea behind the method is based on finding *n*-gram matches in historical data. The Google *n*-gram database represents a massive record of *n*-gram usage over a long period of time—specifically, from 1520 to 2008. For each year, the database records the *n*-grams that appeared in books that were published in that year, along with how many times they occurred in those books. These data can be used to analyze any specific book by first identifying *n*-grams in the document and then locating those *n*-grams in the database. Counting the number of occurrences and the years in which they occurred helps to identify the cultural time period that influenced the author.

2.   DOCUMENTS EXAMINED IN THIS STUDY

In order to develop and test the method described in this study, it was necessary to select several books for examination. The books were divided into two groups: established documents and questionable documents. The books in the "established" group all had well-established periods of composition, while those in the "questionable" group had periods of composition that were disputed, uncertain, or deceptive to some degree. This section contains a brief background of each book and author.

2.1   Established Documents

COMMON SENSE (THOMAS PAINE, 1776)
Written by Thomas Paine, *Common Sense* was first published in Philadelphia in January 1776. Printed as a 48-page pamphlet, *Common Sense* presented a reasoned case for American independence from England. Paine's pamphlet was immensely successful and quickly became the most widely read piece of literature yet published in America—selling approximately 120,000 copies in the first year alone. Due to the treasonous nature of his subject matter, Paine published the work anonymously [Paine 1995; McCullough 2001; McCullough 2005; Hagger 2007; Chernow 2004].

FRANKENSTEIN (MARY SHELLEY, 1818)
Mary Wollstonecraft Shelley was only 18 when she wrote her first novel, *Frankenstein; or, The Modern Prometheus*. During a trip to Switzerland in 1816; Mary and her husband, poet Percy Bysshe Shelley, spent the summer with Lord Byron and other writers. In 1831, Mary recalled that "it proved a wet, ungenial summer, and incessant rain often confined us for days to the house." The group passed time by the fireplace sharing stories of the supernatural, and Lord Byron challenged each one to come up with a ghost story of their own. After retiring to bed several nights later, Mary was struck with the idea for her novel, which she published in 1818 [Shelley 1891].

THE METAMORPHOSIS (FRANZ KAFKA, 1915)
Franz Kafka was born in Prague on July 3, 1883, and died relatively young at age 40. Though he worked as an insurance agent in a state-run institution, Kafka wrote short stories in his spare time, the first being published in 1912 (*Meditation*). His most well-known short story, *The Metamorphosis*—a story about a man who awakens to find himself transformed into an insect—has remained popular ever since it was first published in 1915 and is still a standard work of study in colleges and universities around the world [Kafka 2009].

PRIDE AND PREJUDICE (JANE AUSTEN, 1813)
Jane Austen began this work in October 1797 under the working title *First Impressions* and completed her first draft nine months later. Though this was her first major novel to be written, it was actually the second to be published (*Sense and Sensibility* was first in 1811). From 1811 to 1812, Austen made

revisions to her work, shortening it, and changing its title to *Pride and Prejudice* [Faye 2002]. Her novel was immediately successful and has been a favorite ever since. Jane Austen was born in Steventon, Hampshire, England on December 17, 1775, and died at Winchester on July 18, 1817 [Austen 1813; Austen 2010; Nokes 1997].

2.2 Questionable Documents

BOOK OF MORMON (JOSEPH SMITH, JR., 1830)

The *Book of Mormon* presents itself as a history of a group of ancient Israelites who journeyed from Palestine by boat and settled the American Continent, covering the time period from approximately 600 B.C.E. to 425 C.E. This "history" was supposedly engraved on thin plates of gold in a language referred to as "reformed Egyptian" [Brodie 1945]. Joseph Smith, Jr. claimed that he found the plates buried in a hill not far from his home, having been led to the spot by an angel. Though Smith had to wait several years after his initial visit to the hill, the angel finally allowed him to take possession of the plates on September 22, 1827. Through "the gift and power of God," Smith was able to translate their contents into English and publish the text in 1830 as *Book of Mormon* [Smith 1980]. According to Smith, once he had completed the miraculous translation process, the angel returned and retrieved the plates from him. The text of *Book of Mormon* has undergone thousands of revisions since its first publication in 1830. In order to nullify their effect, the original 1830 edition of *Book of Mormon* was selected for this study. Joseph Smith was born in Sharon, Vermont, on December 23, 1805, and was murdered by a mob at Carthage, Illinois, on June 27, 1844 [Hill 1977].

CHRONICLES OF ERI (ROGER O'CONNOR, 1822)

Published in 1822, Roger O'Connor claimed that his *Chronicles of Eri* was "a true and faithful history of my country [Ireland], from the earliest times. . .. a literal translation into the English tongue, (from the Phœnican [sic] dialect of the Scythian language,) of the ancient manuscripts which have, fortunately for the world, been preserved through so many ages, chances and vicissitudes." These "ancient manuscripts" were, according to O'Connor, "faithful transcripts from the most ancient records; it not being within the range of possibility, either from their style, language, or contents, that they could have been forged." According to O'Connor, his sources included historical data up to the year 1169 C.E. O'Connor [1822].

The book was reviewed in 1941 by archaeologist R. A. Stewart Macalister, who called it "an amalgam of bombastic paraphrases of Irish annalistic matter, irreverent parodies of Biblical excerpts, 'etymologies' (which have to be seen to be believed), and wildly irresponsible inventions resembling those in the closely analogous *Book of Mormon*. . . how anyone could be left to himself, as the saying goes, so far as to take it seriously, and to waste any time over it, is a mystery inscrutable" [Macalister 1941]. The *Dictionary of National Biography* affirms, "The book is mainly, if not entirely, the fruit of O'Connor's imagination." O'Connor was born in Connorville in 1762 and died at Kilcrea (both in County Cork) on January 27, 1834 [Fitzpatrick and Lee 1895].

ERN MALLEY POEMS (JAMES MCAULEY AND HAROLD STEWART, 1945)

In 1944, two friends—James McAuley and Harold Stewart—decided to perpetrate a hoax aimed at *Angry Penguins*, an Australian magazine that published modernist poetry. The two men claimed that in a single day they invented a fictional poet named Ern Malley and wrote a collection of nonsensical poems that they attributed to him. The collection was submitted to *Angry Penguins*, where it was actually accepted as legitimate poetry and published. According to McAuley and Stewart, they did this to express their concerns over "the gradual decay of meaning and craftsmanship in poetry." The hoax was exposed shortly after the poems appeared in print in 1945, and the affair contributed to the

failure of the magazine. Curiously, the popularity of these poems continues to endure [Wilde et al. 1994; Rickard 1997; Mead 2008].

In his book, *The Sons of Clovis: Ern Malley, Adoré Floupette and a Secret History of Australian Poetry*, David Brooks proposes that these inventions of McAuley and Stewart are in fact based on a French satire by Henri Beauclair and Gabriel Vicaire entitled *Les Déliquescences d'Adoré Floupette*, published in 1885. As we shall see in Section 4, our method suggests period of cultural influence that tends to support Brooks' conclusion [Brooks 2011].

VORTIGERN AND ROWENA (WILLIAM HENRY IRELAND, 1796)

As a young man living in London in 1796, William Henry Ireland claimed to have discovered a lost play by Shakespeare entitled *Vortigern and Rowena*. In reality, Ireland had forged the play, along with several other miscellaneous Shakespearean documents he claimed to have found, such as contracts, receipts, letters, and licenses. Though some challenged the authenticity of the play, others were convinced, and the play was produced at the Drury Lane on April 2, 1796. It was an immediate failure, eliciting "ridicule and laughter from the audience with its crude action and inept dialogue." Kahan points out that this play "was a failure, in part, not because it was bad, but because it was so much of the eighteenth century that it could hardly be of any other." Ireland was born in London on August 2, 1775, and died in London on April 17, 1835 [Ireland 1799; Ireland and White 1874; Kahan 1998; Kahan 2001; Mair 1938; Grebanier 1965; Schoenbaum 1991; Campbell and Quinn 1966].

3. METHOD

This study develops an original method for using extensive historical *n*-gram usage data to identify the cultural time period that most influenced the author of a given literary work. Until the availability of the Google *n*-gram database in 2011, the ability to use literary works to identify and track cultural influences over broad time periods was not feasible. To do so would have required the researcher to carefully read and assimilate literally millions of books—a task that is not humanly possible (for example, reading 10 books per day for 80 years covers only about 292,000 books). By extracting *n*-grams from literary works and examining their distribution throughout the Google database, a reasonable and objective estimation of these time periods can be determined. After brief discussions on the problem of forward contamination, and the reasons for selecting specific *n*-gram classes for evaluation, this section proceeds to describe the method in detail.

3.1 Preventing Forward Contamination

While counting the number of *n*-gram occurrences and the years in which they occurred helps to identify the cultural time period that influenced the author, a problem arises if we attempt to examine data recorded after the book was published. If the book under examination happens to be one of the books in the database, then we will be counting *n*-gram occurrences that include instances from the very book that is under examination. In a sense, the presence of the book in the database "contaminates" our sample. The total number of matches will be artificially inflated because *n*-grams from the book itself are included in the database.

The problem can be even worse. Suppose that the book being examined has proven to be a very popular book. If so, then it has (by definition) had an effect on culture. Other authors may have been influenced by it, even to the point of quoting from it. The more popular the book was, the more it was referred to and quoted, and the more extensive the contamination it caused.

The crucial element to the solution of this problem is the book's publish year, as it divides "clean" data from potentially "contaminated" data. Any historical data recorded on or after the publish year of the book are subject to this potential contamination, while the data recorded before the publish date are

free from this concern. For lack of a better term, we will refer to this phenomenon as *forward contamination*. Accordingly, in order to ensure that the data used in the study are free from forward contamination, only data recorded before the publish year are considered for each book included in the study. For example, if a book was published 1813, only data collected from 1700 to 1812 will be considered.

3.2 Selection of N-gram Classes

In order to provide a degree of confirmation for our results, multiple classes of n-grams have been included. While the Google database is composed of collections of *1*-, *2*-, *3*-, *4*-, and *5*-grams, this study will examine only *3*-, *4*-, and *5*-grams. This restriction accomplishes several important goals.

First, by excluding the *1*- and *2*-grams, the amount of data to be downloaded and analyzed is kept to a manageable level. As previously mentioned, the Google database is massive, containing over 5 million books. Even with some classes of n-grams excluded from consideration, the study still took months to complete. Including them would have made the study unfeasible given the time period allotted.

Second, by choosing to exclude these sets of n-grams, the sizes of the locally generated "n-gram matches" databases were reduced, along with the time required to process against them. This is because, as the order of n-grams increase, the probability of finding matching occurrences of such n-grams decreases. For example, there is a much higher probability of finding the *2*-gram "in the" across multiple volumes, than there is of finding the *5*-gram "Michael ate his gingerbread cookie" in as many volumes. The number of *3*-gram matches can be as much as 50 times greater than the number of *5*-gram matches for the same book and time period. To put it simply, the longer the phrase, the more unique it is likely to be; more unique means fewer matches; and fewer matches means fewer data to analyze.

That being said, it is, of course, almost always more desirable for the purposes of statistical analysis to have too much data than too little. Due consideration should be given to results obtained from the analysis of the separate classes of n-grams when the amount of data available for study varies significantly among them. Since *4*-grams regularly yield more matches than *5*-grams, and *3*-grams yield more than *4*-grams, generation and processing of *3*-gram matches normally produces the greatest quantities of data for analysis.

Third, analyzing multiple classes of n-grams allows cross checking of results and provides a degree of validation. If, after examining frequencies of *3*-grams extracted from a specific book under investigation, we arrive at a reasonable estimate for a time period of cultural influence, then it is reasonable to ask if we would have obtained similar results had we examined *4*- or *5*-grams instead. Including these multiple classes of n-grams therefore allows us to answer such questions by comparing results. Comparable values obtained from the examination of multiple n-gram classes helps to confirm the correctness of our results.

3.3 Determining the Time Period of Greatest Cultural Influence

The main idea behind the method developed in this study is based on finding n-gram matches in historical data. The Google database records the separate n-grams that appeared in books that were published during the years 1520 to 2008, along with how many times they occurred in those books per year. These data can be used to analyze any specific book by first identifying n-grams in the document and then locating occurrences of those n-grams throughout the database. Counting these occurrences and the years in which they were recorded enables us to identify the cultural time period that influenced the author.

The analytical method proposed in this study accomplishes this goal by defining and utilizing several concepts: (1) the n-gram popularity factor, (2) the aggregate yearly popularity, (3) high aggregate yearly
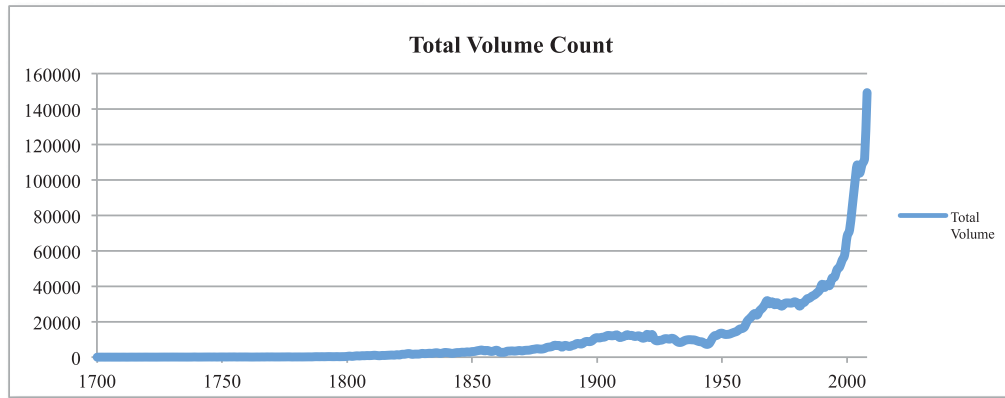
Fig. 2.   Total number of published volumes per year (as included in the Google database).

popularity, and (4) a sustained period of high aggregate yearly popularity. These concepts are discussed in detail below.

N-GRAM POPULARITY FACTOR

The fundamental component of this method is the n-*gram popularity factor*. Its purpose is to quantify how "popular"—and, therefore, how indicative of culture—a given *n*-gram was during any given year. While the class-specific Google *n*-gram databases contain "match count" data for each *n*-gram by year, these data cannot be directly used to accurately determine popularity. For example, one might suppose that if an *n*-gram appeared 100 times more often in 2004 than it did in 1778, then that *n*-gram might be considered to be 100 times more popular in 2004 than it was in 1778. But that conclusion is based on the assumption that the total number of volumes was the same for both years. If there were 1,000 times as many volumes published in 2004 as there were in 1778, then the *n*-gram would actually be (as we shall demonstrate below) about 1/10th as popular in 2004 as it was in 1778.

The key idea here is that the total number of matches must be considered in terms of the total number of volumes published, because the total number of volumes published does not remain constant from year to year and has a definite effect on the number of matches one should expect to find for any particular *n*-gram. In general, for the time period covered in this study, the total number of volumes published increases every year (see Figure 2). Also, there is a very high degree of positive correlation ($R = 0.994$) between the number of volumes published per year and the total number of matches per year (see Figure 3). This should come as no surprise. The more volumes published in any given year, the more likely a given *n*-gram will appear in those volumes; therefore, the *n*-gram is more likely to have a greater overall match count in a year in which more volumes are published.

The solution, then, to being able to quantify *n*-gram popularity (or, equivalently, the degree to which it is indicative of culture) in more absolute terms is to express popularity in terms of matches per volume, as shown in Equation (1):

$$\text{Popularity Factor}_{(\text{n-gram, Year})} = \text{Total n-gram matches}_{(\text{n-gram, Year})} / \text{Total volumes published}_{Year} \quad (1)$$

Using this definition, we can return to our example above. In the first instance, let us suppose that a specific *n*-gram appeared 7 times in 1778 and that there were a total of 108 volumes published that year. In this case, the *n*-gram popularity factor would be 7/108 or 0.0648. In the second instance, we suppose the *n*-gram appeared 700 times in 2004, when there was a total of 108,423 volumes printed. In this case, the popularity factor would be 0.00646, or nearly 1/10th the value for 1778. We conclude
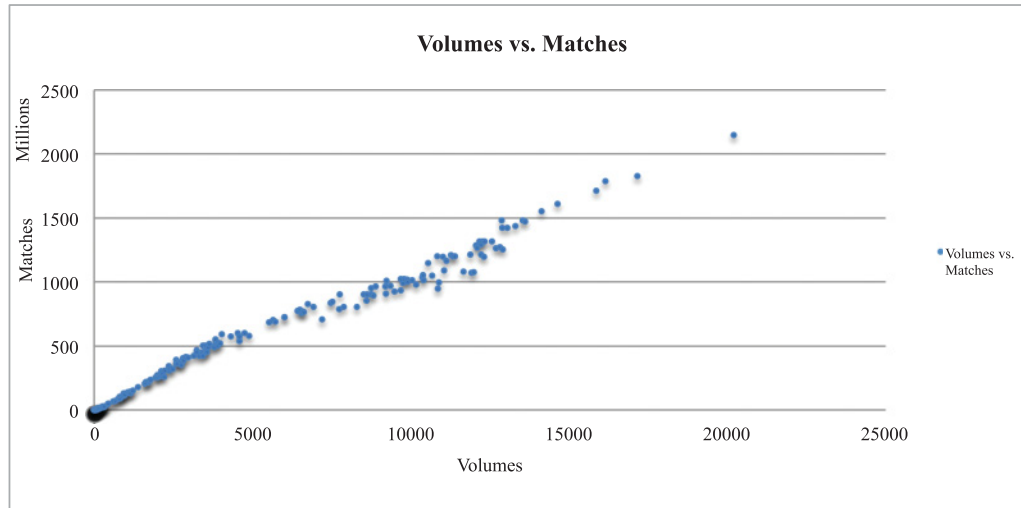
Fig. 3. Positive correlation of volumes with matches per year. R (linear correlation coefficient) = 0.994.

that this *n*-gram was only about 1/10th as popular in 2004 as it was in 1778 even though its match count was 100 times greater.

AGGREGATE YEARLY POPULARITY

The second important concept in this method is that of *aggregate yearly popularity*. During the data acquisition phase, *n*-grams that were extracted from a particular document were located in the Google database. Occurrences of each *n*-gram were identified by the year of the occurrence, along with totals for matches, pages, and volumes. For any specific document, many of its separate *n*-grams can have occurrences in the same year, so we end up with years having varied sets of *n*-gram matches for the document.

For any given document and year, the popularity factors shown in Equation (2) for each extracted *n*-gram with matches in that year can be added together, giving the aggregate *n*-gram popularity for that year:

$$\text{Aggregate Popularity}_{(\text{Document}, \text{Year})} = \sum \text{Popularity Factor}_{(\text{Document } n\text{-gram}, \text{Year})}. \tag{2}$$

Using this aggregate yearly popularity, we can extend the concept of "popularity" to include years, as well as *n*-grams. In other words, we can designate, per document, which years are more "popular"— and, hence, more indicative of cultural influence—than are others.

HIGH AGGREGATE YEARLY POPULARITY

The third concept is that of *high aggregate yearly popularity*. Now that we can calculate the aggregate popularity for any particular document and year, we can begin to determine those years that have "high" popularity. In order to do this, we determine the regression equation for the set of data points defined by the yearly aggregate popularity. We then define a year as having "high" aggregate popularity if its associated data point lies above the regression line. This can be stated more formally as in
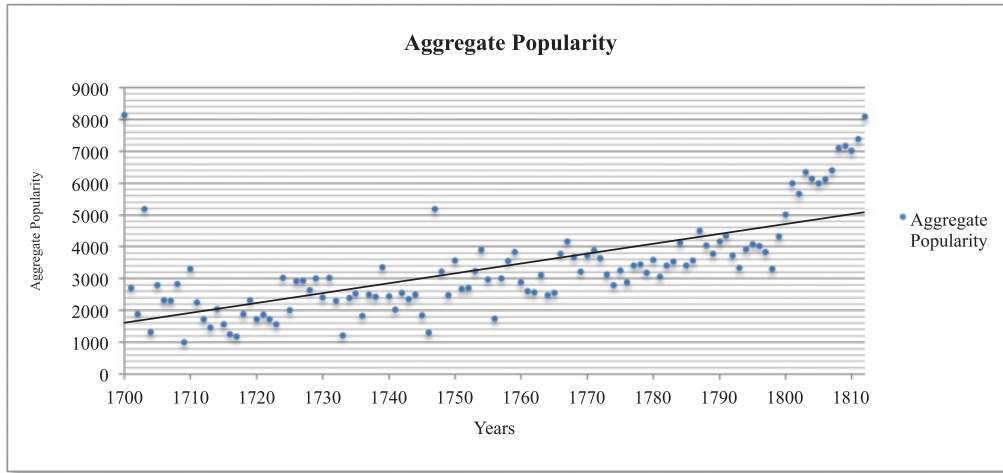
Fig. 4.   A sample scatter plot of aggregate popularity, with regression line.

Equation (3):

{Years with high aggregate popularity}$_{\text{Document}}$ = {x: Aggregate Popularity$_{\text{Document}_x}$ > $\beta_0 + \beta_1$x}, where
β_0 = *y*-intercept of the regression line
β_1 = Slope of the regression line

(3)

This can be illustrated with an example. Figure 4 shows a scatter plot of aggregate yearly popularity for a specific document and *n*-gram class. The regression line for this set of data points is depicted in the graph and has the following equation:

$$\hat{y} = -51220 + 31.078\text{x}. \tag{4}$$

Here, $\beta_0 = -51220$ and $\beta_1 = 31.078$. According to the definition above, each point that lies above the regression line indicates a year with "high" aggregate popularity. Using our scatter plot, we can see that the aggregate popularity for the year 1746 is 1304.11, while the associated value given by the regression equation is 3042.188. Since the actual value is less than the predicted value, 1746 can be said to have a "low" aggregate popularity. On the other hand, since the actual value for the year 1801 is 5992.9 (which is higher than the predicted value of 4751.478), 1801 can be said to have "high" aggregate popularity.

SUSTAINED PERIOD OF HIGH AGGREGATE YEARLY POPULARITY
The final concept for this method involves the definition of a *sustained period of high aggregate yearly popularity*. We saw from the preceding section that a regression line can be used to identify years having "high" aggregate popularity. Since one of the goals of this method is to provide a time period of greatest cultural influence, these data must begin to be viewed in terms of "time periods" or ranges of years. In addition, we must be able to define "high" aggregate popularity in terms of those time periods.

To begin, we examine yearly aggregate popularity data in time periods of 10 years, beginning with the latest year in the dataset, and proceeding backwards through time. Again, we refer to Figure 4. The latest year in this dataset is 1812. So, we can define the following time periods over this dataset:

Table I. Results—Period of Cultural Influence. Results Were Rounded to The Nearest Year

Document	Year Published	Period of Composition	Est. Range: 3-grams	Est. Range: 4-grams	Est. Range: 5-grams	Period of Influence
Pride and Prejudice	1813	1811–1812	1803–1812	1803–1812	1803–1812	1803–1812
The Metamorphosis (English)	1915	1912–1912	1805–1874	1805–1874	1825–1874	1805–1874
The Metamorphosis (German)	1915	1912–1912	1835–1864	1825–1864	1855–1864	1825–1864
Frankenstein	1818	1816–1817	1808–1817	1808–1817	1808–1817	1808–1817
Common Sense	1776	1775–1776	1766–1775	1766–1775	1766–1775	1766–1775
Chronicles of Eri	1822	1821–1822?	1802–1811	1802–1811	1802–1811	1802–1811
Ern Malley Poems	1945	1944–1944	1805–1884	1805–1894	1825–1884	1805–1894
Vortigern and Rowena	1796	1793–1795	1786–1795	1786–1795	1786–1795	1786–1795
Book of Mormon	1830	1827–1830?	1810–1829	1810–1829	1810–1829	1810–1829

*1703–1712   *1723–1732   1743–1752   1763–1772   1783–1792   *1803–1812

1713–1722   1733–1742   1753–1762   1773–1782   1793–1802

We have 11 time periods, each exactly 10 years in duration (the asterisks will be explained momentarily). Since our dataset includes exactly one data point per year, we have exactly 10 data points per time period. About half of the data points in the dataset should fall above the regression line, and half should fall below. This means that roughly, within each of the time periods defined, one would expect 5 data points to lie above the regression line and 5 to lie below it. For the purpose of this study, any time period having more than 5 data points above the regression line will be considered to have "high" aggregate popularity. In the list presented above, those time periods marked with an asterisk are those with high aggregate popularity.

A "sustained" period of high aggregate popularity, then, will be any contiguous group of these periods. This is true even if the contiguous group contains only one period. Since there may actually be several of these contiguous periods, the method specifies that we choose only the "latest" or most recent of these periods (i.e., the one "closest" in time to the publish date) as the time period of greatest cultural influence. Again referring to the list above, we can see that there are no contiguous groups containing more than one period. Of these, the "latest" is 1803–1812, and we therefore consider this to be the time period of greatest cultural influence upon the author of the document.

The period of cultural influence will be identified by both (1) a specific time period of influence and (2) a year of peak influence. In the case of most documents, this period of greatest cultural influence will likely be positioned either slightly before or coincident with the period of composition. However, in the case of documents whose period of composition is uncertain, disputed, or otherwise questioned, the two periods may differ substantially and have no overlap. Indeed, as we will see in the next section, such a disparity may be an indicator of possible dissimulation.

4. RESULTS

This section discusses the results returned from application of the method to the documents selected. As designed, the method analyzes each document against three different *n*-gram classes and returns two types of results: (1) a specific time period of influence and (2) a year of peak influence. These results are presented in Tables I and II, respectively, along with other pertinent information about each document.

Table II. Results—Peak Year of Cultural Influence. Results Were Rounded to the Nearest Year

Document	Year Published	Peak Year: 3-grams	Peak Year: 4-grams	Peak Year: 5-grams	Mean Peak Year	(Mean − Published)
Pride and Prejudice	1813	1807	1807	1807	1807	−6
The Metamorphosis (English)	1915	1841	1842	1851	1845	−70
The Metamorphosis (German)	1915	1850	1846	1860	1852	−63
Frankenstein	1818	1813	1813	1813	1813	−5
Common Sense	1776	1770	1770	1770	1770	−6
Chronicles of Eri	1822	1811	1811	1811	1811	−11
Ern Malley Poems	1945	1845	1853	1857	1852	−93
Vortigern and Rowena	1796	1791	1791	1790	1791	−5
Book of Mormon	1830	1820	1820	1820	1820	−10

It should be noted that, among the documents in the "questionable" group, the *Ern Malley Poems* and *Vortigern and Rowena* have established periods of composition, while the periods of composition for *Chronicles of Eri* and *Book of Mormon* are considered questionable. This is due mainly to the fact that the authors of the former works are confessed forgers, while the authors of the latter works never confessed to being such. Hence, if they were in fact practicing deception in the production of their books, there is no reason to assume that they were not also being deceptive with regard to the period of composition.

4.1 Interpretation of Results

This section briefly examines these results as they apply to each document in the study.

FRANKENSTEIN

Mary Shelley composed *Frankenstein* after her visit to Switzerland, during the period 1816–1817, and afterwards published it in 1818. Our method estimates 1808–1817 as the period of greatest cultural influence, with 1813 as the mean peak year.

COMMON SENSE

Thomas Paine composed *Common Sense* in late 1775, during the period 1775–1776. after which he published it in 1776. Our method estimates 1766–1775 as the period of greatest cultural influence, with 1770 as the mean peak year.

PRIDE AND PREJUDICE

Jane Austen composed *Pride and Prejudice* during the period 1811–1812, by extensively reworking her initial draft entitled *First Impressions*. She then published her novel in 1813. Our method estimates 1803–1812 as the period of greatest cultural influence, with 1807 as the mean peak year.

THE METAMORPHOSIS (ENGLISH AND GERMAN VERSIONS)

Franz Kafka wrote *The Metamorphosis* originally in German. In order to observe the performance of our method across different languages, this document was examined in both German and English. The results from both documents were comparable but are unlike any of the other documents in this group. In the case of *The Metamorphosis*, the period of cultural influence precedes the period of composition by several decades, not just several years. In fact, this was another reason for examining this document in its native language—to verify that the process of translation did not adversely affect the results of the method. Since the results are quite similar, we conclude that it is unlikely either one is in error.

The method suggests that the period of cultural influence is 1825–1864, at least 19 years before Kafka was born (1883). This result does not compare favorably with the results from the other documents in the group, where the period of cultural influence is always positioned within the author's lifetime. However, the results are not unreasonable. It is plausible that authors who are immersed within the literature of an earlier time period will evidence this influence through their works.

In order to establish a basis for this conjecture, it was deemed prudent to research Kafka's background more closely. In the biographical notes of one edition of *The Metamorphosis*, we read that "Kafka acquired some knowledge of the French language and culture; one of his favourite authors was [Gustave] Flaubert.... After elementary school, [Kafka] was admitted to the rigorous classics oriented state Gymnasium" [Kafka 2011]. Flaubert lived from 1821 to 1880 and produced most of his output during the mid-1800s [Jessup and Ives 1903]. Kafka's "rigorous classics oriented" education would suggest that, as a youth, he was immersed in the writing style of a much earlier period. These two facts would tend to suggest that Kafka's literary style might have been influenced by an earlier cultural period.

VORTIGERN AND ROWENA

William Henry Ireland composed *Vorigern and Rowena* during the period 1793–1795 and then published it in 1796. Our method estimates 1786–1795 as the period of greatest cultural influence, with 1791 as the mean peak year.

CHRONICLES OF ERI

Roger O'Connor composed *Chronicles of Eri* probably during the period 1821–1822 and then published it in 1822. Our method estimates 1802–1811 as the period of greatest cultural influence, with 1811 as the mean peak year. While there is a gap of 10 years between the end of the period of cultural influence and the beginning of the period of composition, it should be emphasized that the exact dates for the period of composition are not well established. We conclude that these results are reasonably consistent with our expectations.

BOOK OF MORMON

Book of Mormon was supposedly "translated" during the period 1827–1830, after which it was published in 1830. Though it is somewhat unclear who actually composed *Book of Mormon*, the original edition explicitly names Joseph Smith, Jr., as the "Author and Proprietor" [Smith 1830]. Our method estimates 1810–1829 as the period of greatest cultural influence, with 1820 as the mean peak year. These results are reasonable and consistent with our expectations.

The results are quite inconsistent, though, with the claims of its author, who alleged a period of composition from 600 B.C.E to 425 C.E. [Richards 1976]. While it is well beyond the scope of this study to examine documents from antiquity, it is certainly within its scope to test and compare documents with origins in 18th- and 19th-century culture. The scatter plots generated from *n*-gram aggregate popularity data effectively demonstrate the similarities among documents composed around the early part of the 19th century. Indeed, the *3*-gram scatter plots of aggregate popularity for *Pride and Prejudice*, *Frankenstein*, *Chronicles of Eri*, and *Book of Mormon* are strikingly similar—so much so, in fact, that one would be hard pressed to distinguish among any of them (see Figures 5–8). The scatter plots for *4*- and *5*-grams are no less similar. These data strongly suggest that *n*-grams from *Book of Mormon* fit perfectly within the cultural influence period of the early 19th century. It would be difficult to explain how a culture from approximately 2,000 years earlier could so perfectly imitate these early 19th-century *n*-gram frequency distributions.
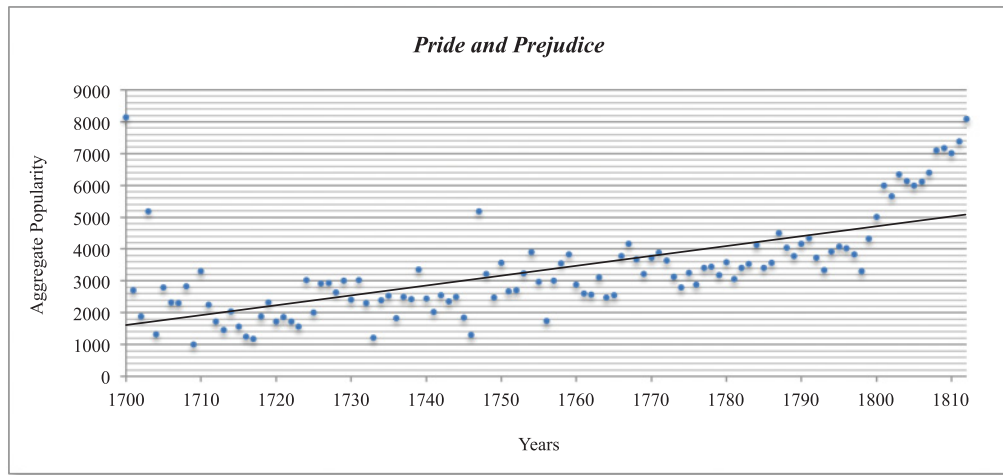
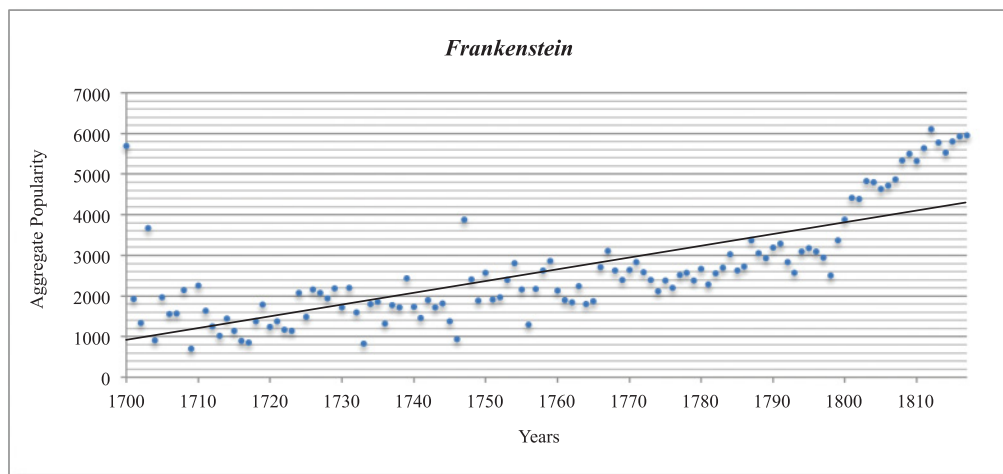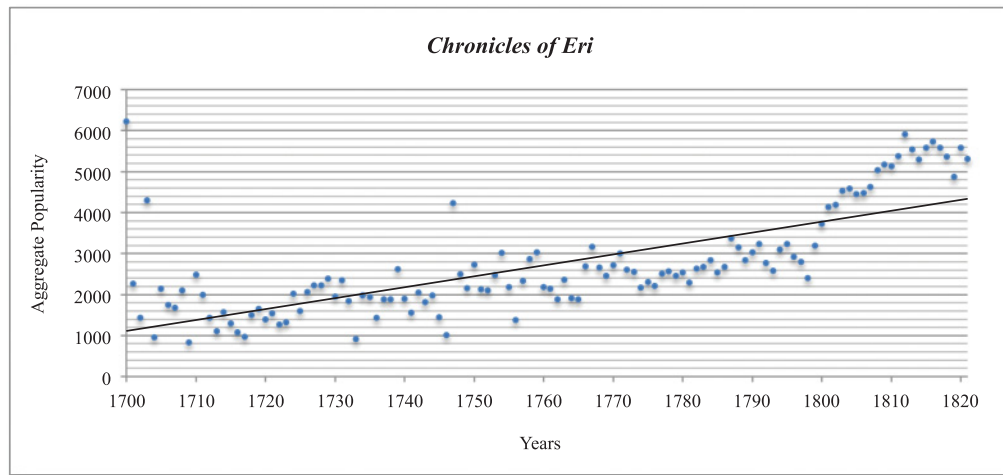Fig. 5.   Aggregate yearly popularity for *Pride and Prejudice 3*-gram data.

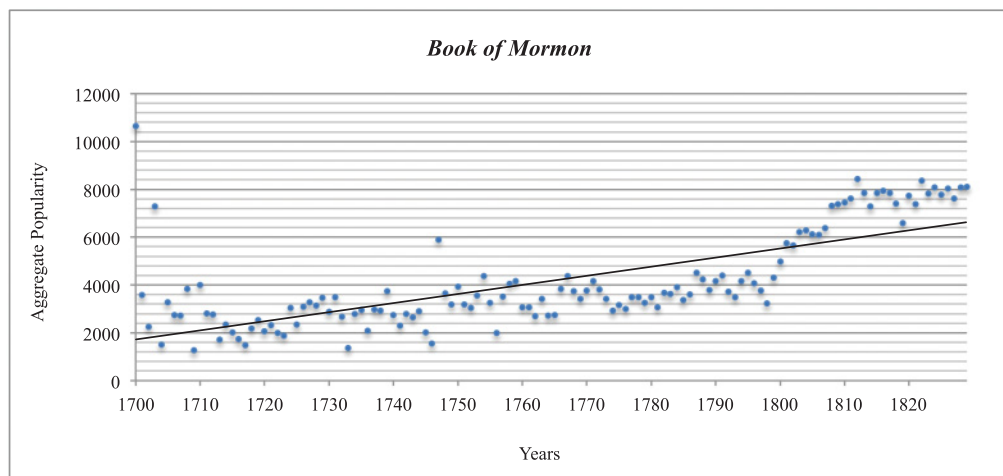Fig. 6.   Aggregate yearly popularity for *Frankenstein 3*-gram data.

ERN MALLEY POEMS

The interpretation of results from the *Ern Malley Poems* presents an even greater challenge. James McAuley and Harold Stewart invented the fictitious character Ern Malley and composed a set of poems that they attributed to him. The pair wrote this small collection of poems in 1944 and was successful in getting them published the following year. Our method estimates 1805—1894 as the period of greatest cultural influence, with 1852 as the mean peak year. These results are drastically inconsistent with our expectations (see Figure 9).

   The results of our study indicate that the authors of the *Ern Malley Poems* were most strongly influenced by a cultural time period approximately 100 years earlier. Though this result appeared shockingly erroneous at first, it actually supports similar results from another researcher concerned with the question of sources for the material found in the *Ern Malley Poems*. In his recently published

Fig. 7.   Aggregate yearly popularity for *Chronicles of Eri 3*-gram data

Fig. 8.   Aggregate yearly popularity for *Book of Mormon 3*-gram data.

book, *The Sons of Clovis: Ern Malley, Adoré Floupette and a Secret History of Australian Poetry*, David Brooks conducts a detailed examination of the background and creation of these poems [Brooks 2011]. Brooks claims to have firmly established that the poems by McAuley and Stewart "were modeled upon a French precedent, a parody of the Symbolist poets (Mallarme, Rimbaud, Verlaine and others) written by Henri Beauclair and Gabriel Vicaire and published in 1885 under the name of their own nonexistent poet, Adore Floupette." Specifically, Brooks states that "the poems are framed on Mallarmé's Afternoon of a Faun" [Anderson 2011].

Interestingly, the French author Stéphane Mallarmé wrote his *L'après-midi d'un faune* during the period 1865–1876—a period of composition that is in near perfect agreement with the results of our
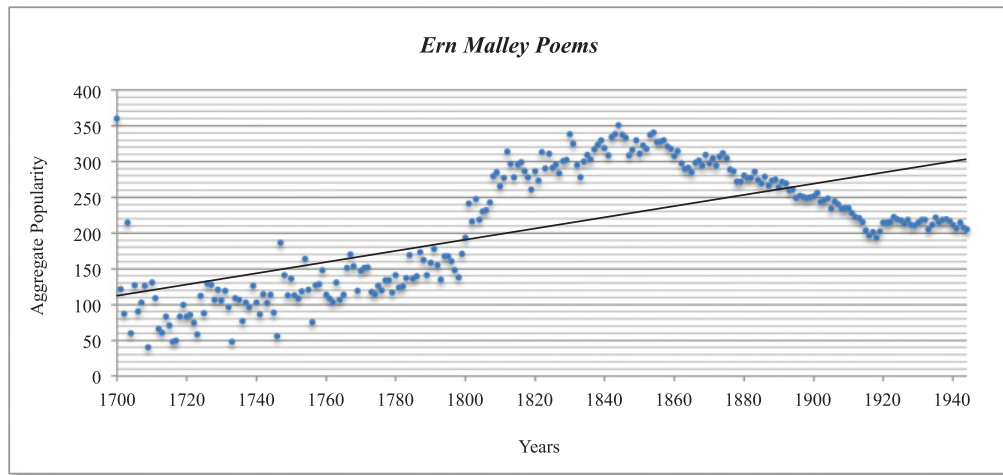
Fig. 9.   Aggregate yearly popularity for *Ern Malley Poems 3*-gram data.

method [Weinfield 1994]. Accordingly, the results of our method appear to support Brooks' conclusions on the origins of the *Ern Malley Poems*, while approaching the problem from a different perspective.

5.   CONCLUSION

This study would not have been possible without the aid of modern computer systems. The development of the method described herein relied heavily on the proper application of software engineering techniques. In particular, this study required that a very large database be processed and analyzed as efficiently as possible. Several software elements had to be designed, implemented, tested, and executed in order to process and analyze the historical *n*-gram data. Devising methods for downloading, partitioning, importing into a local database management system, and processing this massive database involved the use of advanced software engineering skills.

Based on the analysis of the individual results returned for each separate document, we conclude that the method developed does provide a reasonable estimate of the period of cultural influence. Not only does the method provide consistent results for both the "established" and "questionable" documents, it also hopefully provides new insight into the question of the early 19th-century origins of *Book of Mormon*, as well as support for modern conclusions on possible sources for the *Ern Malley Poems*. The results strongly suggest that *n*-grams can be used as viable linguistic constructs for analyzing periods of cultural influence and their effects on literary works.

REFERENCES

Erez Lieberman Aiden, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449, 7163 (Oct. 11, 2007), 713–716. DOI: 10.1038/nature06137

Don Anderson. 2011. Ern, it turns out, has a French cousin. *The Australian* (October 1, 2011).

Jane Austen. 1813. *Pride and Prejudice: A Novel*. T. Egerton, Military Library, Whitehall, London.

Jane Austen. 2010. *Pride and Prejudice: An Annotated Edition*, Patricia Meyer Spacks, (Ed.). Harvard, 3–4.

John Bohannon. 2011. Google Books, Wikipedia, and the Future of Culturomics. *Science* 331, 6014 (Jan. 14, 2011), 135. DOI: 10.1126/science.331.6014.135

Fawn M. Brodie. 1945. *No Man Knows My History: The Life of Joseph Smith*. Vintage Books, New York, NY, 50.

David Brooks. 2011. *The Sons of Clovis: Ern Malley, Adore Floupette and a Secret History of Australian Poetry*. University of Queensland Press, Brisbane, Australia.

Oscar James Campbell and Edward G. Quinn, eds. 1966. *The Reader's Encyclopedia of Shakespeare*. Thomas Y. Crowell Company, New York, NY, 386–387.

Ron Chernow. 2004. *Alexander Hamilton*. The Penguin Press, New York, NY, 70.

Patricia Cohen. 2010. In 500 Billion Words, New Window on Culture. *New York Times*, A3 (Dec. 17, 2010).

Deirdre Le Faye. 2002. *Jane Austen: The World of Her Novels*: Frances Lincoln Limited, London, 178.

W. J. Fitzpatrick and Sidney Lee, ed. 1895. *Dictionary of National Biography, Volume 41*. MacMillan and Company, New York, NY, 407–408.

Google. 2015. About Google Books. Retrieved November 7, 2015, from https://www.google.com/intl/en/googlebooks/about/index.html.

Bernard D. N. Grebanier. 1965. *The Great Shakespeare Forgery*. W. W. Norton, New York, NY.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Lit. Ling. Comput.* 22, 3.

Nicholas Hagger. 2007. *The Secret Founding of America*. Watkins Publishing, London, 149.

Eric Hand. 2011. Word Play. *Nature* 474 (Jun. 17, 2011), 436–440. DOI: 10.1038/474436a

Brian Hayes. 2007. Bit lit. *Am. Sci.* 99, 3 (May-June 2011), 190. DOI: 10.1511/2011.90.190

Donna Hill. 1977. *Joseph Smith: The First Mormon*. Signature Books, Salt Lake City, UT, 35, 410–416.

William Henry Ireland. 1799. *Vortigern, an Historical Tragedy, in Five Acts*; *Represented at the Theatre Royal, Drury Lane*. J. Barker, London.

William Henry Ireland and Richard Grant White. 1874. *The Confessions of William Henry Ireland, Containing the Particulars of his Fabrication of the Shakespeare Manuscripts; Together with Anecdotes and Opinions of Many Distinguished Persons in the Literary, Political, and Theatrical World*. James W. Bouton, New York, vii–xxxi, 135.

Alexander Jessup, ed., and George Burnham Ives, trans. 1903. *Little French Masterpieces*. The Knickerbocker Press, New York and London, ix–xxv.

Franz Kafka. 2009. *The Metamorphosis and Other Stories (Oxford World's Classics)*, Joyce Crick, trans. Oxford University Press, vii–xxxiii.

Franz Kafka. 2011. *The Metamorphosis*. CSF Publishing, New Mexico.

Jeffrey Kahan. 1998. *Reforging Shakespeare: The Story of a Theatrical Scandal*. Associated University Presses, Inc., Cranbury, NJ, 11.

Jeffrey Kahan. 2001. Shakespeare and the forging of belief. *Crit. Quart.* 32, 2 (July 2001), 1. DOI:10.1111/1467-8705.00352

R. A. S. Macalister. 1941. *Irish Hist. Stud.* 2, 7 (March 1941), 335.

John Mair. 1938. *The Fourth Forger: William Ireland and the Shakespeare Papers*. Ayer Publishing.

David McCullough. 2001. *John Adams*. Simon & Schuster, New York, NY. 96–97.

David McCullough. 2005. *1776*. Simon & Schuster, New York, NY. 250–251.

Robert McKee. 1997. *Story: Substance, Structure, Style, and the Principles of Screenwriting* (2nd. ed.). HarperCollins, New York, NY.

Philip Mead. 2008. *Networked Language: Culture & History in Australian Poetry*. Australian Scholarly Publishing, 89.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, Aiden, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 6014 (Jan. 14, 2011), 176–182. DOI: 10.1126/science.1199644

David Nokes. 1997. *Jane Austen: A Life*. University of California Press, Berkeley and Los Angeles, 2, 51.

Roger O'Connor. 1822. *Chronicles of Eri; Being the History of the Gaal Sciot Bier: or, the Irish People; Translated from the Original Manuscripts in the Phœnician Dialect of the Scythian Language*, Volume 1, 2 vols. Sir Richard Phillips and Co., London, iii-xi.

Thomas Paine. 1995. *Common Sense*. Fall River Press, New York, NY.

LeGrand Richards. 1976. *A Marvelous Work and a Wonder*. Deseret Books, Salt Lake City, UT, 72–73.

John Rickard. 1997. *Australia: A Cultural History (The Present and the Past)*. Longman Group, United Kingdom, 245.

S. Schoenbaum. 1991. *Shakespeare's Lives*. Oxford University Press.

Gregory Shalhoub, Robin Simon, Ramesh Iyer, Jayendra Tailor, and Dr. Sandra Westcott. 2010. Stylometry System—Use Cases and Feasibility Study. In *Proceedings of Student-Faculty Research Day, CSIS*. Pace University, May 7, 2010.

Mary Shelley. 1891. *Frankenstein; or, the Modern Prometheus*. George Routledge and Sons, Limited, London, v–xii.

Joseph Smith Jr. 1830. *Book of Mormon*. E. B. Grandin, Palmyra, New York, title page.

Joseph Smith. 1980. *History of the Church, Volume I*. Deseret Book Company, Salt Lake City, UT, 15–18.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *JASIST* 60, (Jan. 2009), 538–556.

Henry Weinfield, trans. 1994. *Collected Poems / Stéphane Mallarmé*. University of California Press, Berkeley and Los Angeles, 179.

William Henry Wilde, Joy W. Hooton, and B. G. Andrews, eds. 1994. *The Oxford Companion to Australian Literature*. Oxford University Press, 257.