Using *n*-grams to identify time periods of cultural influence

by Gregory P. Knight

December 2012

Director of Thesis: Dr. Nasseh Tabrizi

Major Department: Computer Science

Abstract

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 a well-known alternate theory on the authorship of the* Book of Mormon.

Using *n*-grams to identify time periods of cultural influence

A Thesis

Presented to the Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Software Engineering

by

Gregory P. Knight

December 2012

Using *n*-grams to identify time periods of cultural influence

by

Gregory P. Knight

APPROVED BY:

DIRECTOR OF THESIS: _____

M. H. Nassehzadeh Tabrizi, PhD

COMMITTEE MEMBER: _____

Junhua Ding, PhD

COMMITTEE MEMBER: _____

Sergiy Vilkomir, PhD

CHAIR OF THE DEPARTMENT OF COMPUTER SCIENCE:

_____

Karl Abrahamson, PhD

DEAN OF THE GRADUATE SCHOOL:

_____

Paul J. Gemperline, PhD

ACKNOWLEDGEMENTS

I would like to thank two people who helped make this study a reality. First, I wish to express my most sincere appreciation to my advisor, Dr. Nasseh Tabrizi, for his endless supply of encouragement, genuine personal concern, and expert guidance. His commitment to the ideals of scholarly research has been an inspiration to me throughout my studies at ECU.

And I especially wish to express my warmest appreciation to my dear wife, Cathy, for her patience and support throughout this long and demanding project, and for her expert assistance in not only proofreading and correcting my manuscript, but also for locating key biographical information on Jane Austen and Franz Kafka that were crucial to the completion of this study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Authors, *n*-grams, and Cultural Influence

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. Environment not only provides the raw material from which the author draws; it supplies the wheel upon 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. 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 [1].

Groups of words are sometimes referred to as *n-grams*, and have been shown to be useful in identifying authorship and style [2, 3, 4]. 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 "too heavy" 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 is noticeably different than 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 fingerprints of time period and 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.

1.2 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 forty large libraries, along with many other books being made available directly from their publishers [5].

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 up 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 [6].

1.3 "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 paper authored by Michel, Aiden, and several other researchers associated with the Google project [7]. In their paper, 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" [7].

1.4 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).

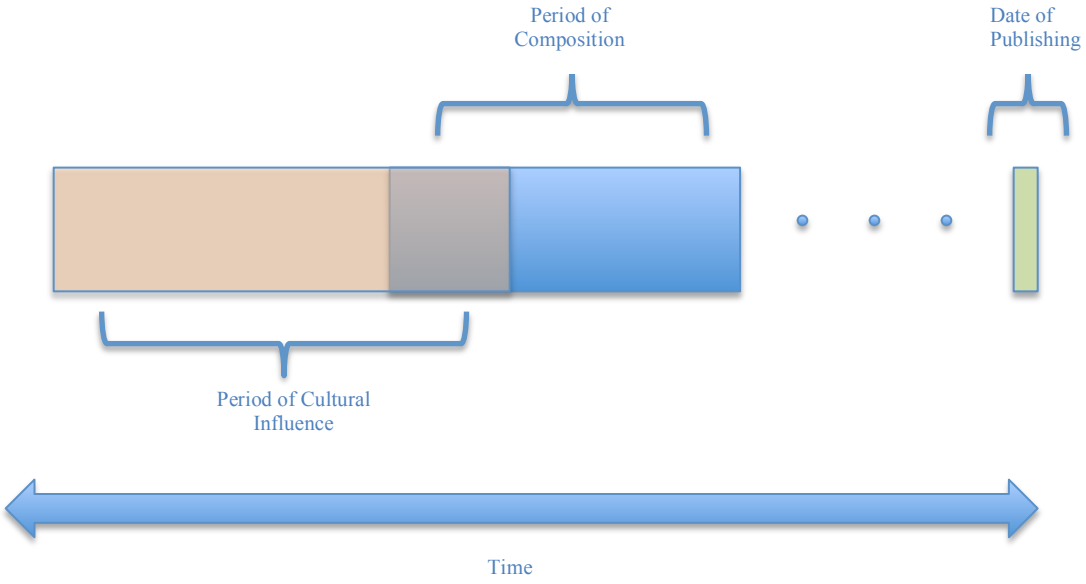


Figure 1: The timeline model, containing Periods of cultural influence, composition, and publishing.

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 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 figure). 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 so as to frustrate the forger's efforts to successfully expunge them.

1.5 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.

1.6 Structure of this Study

The remainder of this study is structured as follows. Chapter 2 introduces the documents to be examined. They are divided into two groups: an "established" group, and a "questionable" group. The documents in the former group are established literary works of undisputed authorship and period of composition. The documents in the latter group, on the other hand, are of questionable or disputed authorship and period of composition. A brief description of each document is included.

Chapter 3 covers the topic of data acquisition and preparation. It lists and describes the steps necessary to download the Google database, prepare the documents, extract *n*-grams from the documents, and locate the extracted *n*-grams in the Google database.

Chapter 4 is concerned with developing the method for determining the time period of greatest cultural influence. After defining and solving the problem of forward contamination, as well as establishing the rationale for *n*-gram class selection, this chapter develops and describes the underlying concepts of the method in detail. These include the *n*-gram popularity factor, aggregate yearly popularity, high aggregate yearly popularity, and sustained period of high aggregate yearly popularity.

Chapter 5 traces the actual application of the method and examines the results obtained. It includes samples of procedures, routines, database tables and scripts that were used to implement the method on a local system. Data collected from all phases of the process are examined and weighed, and the results are evaluated and interpreted.

Chapter 6 presents the conclusions reached from the interpretation of the results from Chapter 5.

1.7 Limitations

The study has several limitations. These are covered in detail below.

Limitations concerning the Google database

Since the Google database contains historical data exclusively from 1520 to 2008, any historical *n*-gram usage data outside these bounds are beyond the scope of this study. The study itself adds the additional restriction to books published on or after 1700. The Google database contains a diminishing number of volumes per year as one proceeds backwards in time. While the study was being conducted, and the data was being processed and analyzed, it became apparent that the older the records, the fewer the number of volumes published per year. Between 1520 and 1700, the number of volumes per year is quite small, and analytical methods produce results with too much variance to be considered reliable.

Since one of the goals of this study was to develop a method against a database rich with historical data, it was felt that the year 1700 was an acceptable lower boundary. In fact, other researchers have limited their studies of the Google database to data collected for the years 1800-2000, for similar reasons [6]. Only *3*-, *4*-, and *5*-grams were studied (the reasons for this are covered in Chapter 4). Also, only *n*-grams that occurred at least 40 times within a single volume are included in the Google database (this is a limitation imposed by Google, not by this study).

The English and German subsets of the Google database, version 20090715, were chosen for use since, at the time, it was the only version available. During the course of conducting this study, however, a newer version became generally available from Google (20120701). As the study was well under way at that point, it was deemed impractical to download and process against the newer database files. It is proposed that any continuation of this study, or another subsequent study of Google's database, utilize the most recently published version.

Language

With the exception of one document, all the books in the study are limited to English. The exception is *The Metamorphosis*, a book originally written in German. To perform a more comprehensive analysis of this book, both the original German and a popular English translation were included. Consequently, both the English and German versions of the Google *n*-gram database had to be downloaded and processed. It is important to note that the method proposed in this study is not limited to English; any of the languages provided in the Google database can be used.

Documents

Though it would have been desirable to include a great number of books in this study, time and storage constraints made this impractical. As has already been noted, the Google database is exceptionally large – containing over 500 billion words. Aside from the time required to download the database itself (which took nearly a week), the time required to process even the meager number of books included in this study was prohibitive. Just processing the books against the *3*-gram set took weeks. Accordingly, the number of books chosen for inclusion was kept to a manageable level of 9. In order to test the method against datasets of different sizes, books of differing lengths were chosen. It was desirable to ascertain whether the method performed equally well against short books versus large books.

The development of the method proposed in this study requires that certain information be known about each book being examined. Each book must be a published work, and its publish date must be well documented. For books in the "established" group, an approximate period of composition must be known. For books in the "questionable" group, the period of composition must be uncertain – either because it is unknown, disputed, or is known to have been purposely misrepresented.

Due to copyright restrictions, each book selected must either be in the public domain, or be free from restrictions that would preclude its inclusion in this type of study.

1.8 Tools Used

Given the amount of data examined by this study, the task would have proven impossible without the aid of modern computer systems. Even with their use, the downloading and data processing phases of

this study took months. To assist with the task of searching hundreds of millions of records, several tools were used (see Table 1).

Tool	Description
iMac / Mac OS X	OS version 10.6.8; 2.7 GHz Intel Core i5; 4 GB RAM 1333 MHz DDR3; Apple Corporation
PC / Windows 7	Windows 7 SP1; AMD Athlon II X2 235e Processor 2.70 GHz; 4 GB RAM; Microsoft Corporation
kfNgram	*n*-gram extraction software; KWiCFinder Code.
MySQL Workbench	Version 5.2.39 Revision 8757; Oracle Corporation
MySQL	Version 5.5.254-osx10.6
Microsoft Excel for Mac 2011	Version 14.2.3 (120616); Microsoft Corporation
Safari (browser)	Version 5.1.7 (6534.57.2); Apple Corporation

Table 1: Tools used.

CHAPTER 2: DOCUMENTS EXAMINED IN THIS STUDY

2.1 Overview

In order to develop and test the method described in this study, it was necessary to select several books for examination. The criteria for selecting these books were presented in the previous chapter. The books were divided into 2 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 chapter contains a brief background of each book and author.

2.2 Established Documents

Common Sense (Thomas Paine, 1776)

Written by Thomas Paine, *Common Sense* was first published in Philadelphia in January of 1776. Printed as a forty-eight-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 [8].

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 [9].

***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 [10].

***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 1811-1812, Austen made revisions to her work, shortening it, and changing its title to *Pride and Prejudice* [20]. Her novel was immediately successful and has been a favorite ever since. Jane Austen was born at Steventon, Hampshire, England on December 17, 1775 and died at Winchester on July 18, 1817 [11].

2.3 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" [12]. 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 the *Book of Mormon* [13]. 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 at Sharon, Vermont on December 23, 1805, and was murdered by a mob at Carthage, Illinois on June 27, 1844 [14].

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. [15].

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" [16]. The *Dictionary of National Biography* affirms, "The book is mainly, if not entirely, the fruit of O'Connor's imagination." O'Connor was born at Connorville in 1762, and died at Kilcrea (both in County Cork), on January 27, 1834 [17].

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 [30, 31, 32].

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 the Chapter 4, our method suggests period of cultural influence that tends to support Brooks' conclusion [18].

***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 [19, 33].

CHAPTER 3: DATA ACQUISITION AND PREPARATION

3.1 Overview

This study develops a method for determining cultural influence based on historical *n*-gram usage. However, the application of the method requires that preliminary steps be taken to prepare the historical data for analysis. This section describes the steps necessary in order to acquire the data and prepare it for processing. These steps are:

- Download the Google database
- Edit and prepare the documents
- Extract *n*-grams from the documents
- Find the extracted *n*-grams in the Google database

These are described in detail below.

3.2 Download the Google *n*-gram database

The first step in this method involves downloading the Google database to a local system for analysis. As mentioned earlier, Google has chosen to make their database available in several languages (see Figure 2). This study limits itself exclusively to the English and German databases (However, it should be noted that the method described here is applicable to any of the language databases available). The databases are further broken down into *n*-gram classes, each of which is presented as a collection of compressed, comma-separated-value (CSV) files. These classes, and the number of CSV files in each, are displayed in Figure 3.

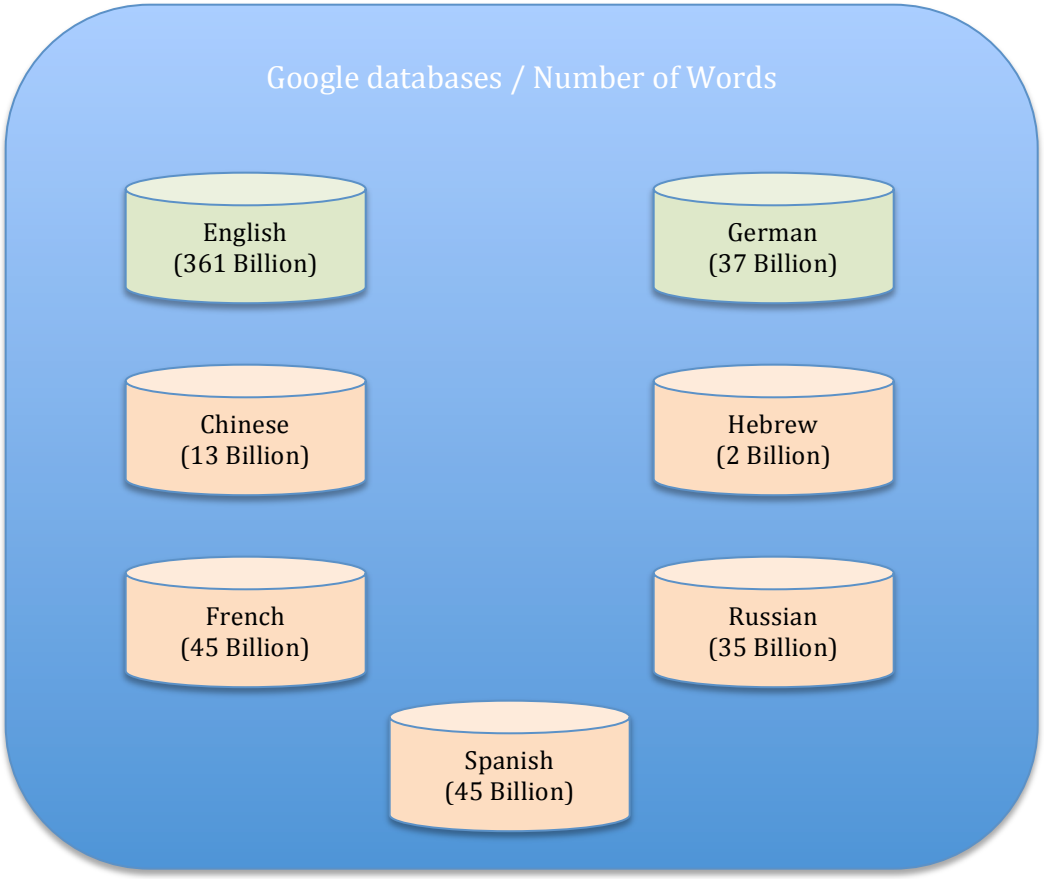


Figure 2: Google *n*-gram databases available for download (Note: English is broken down into five separate databases: English, English One Million, American English, British English and English Fiction).

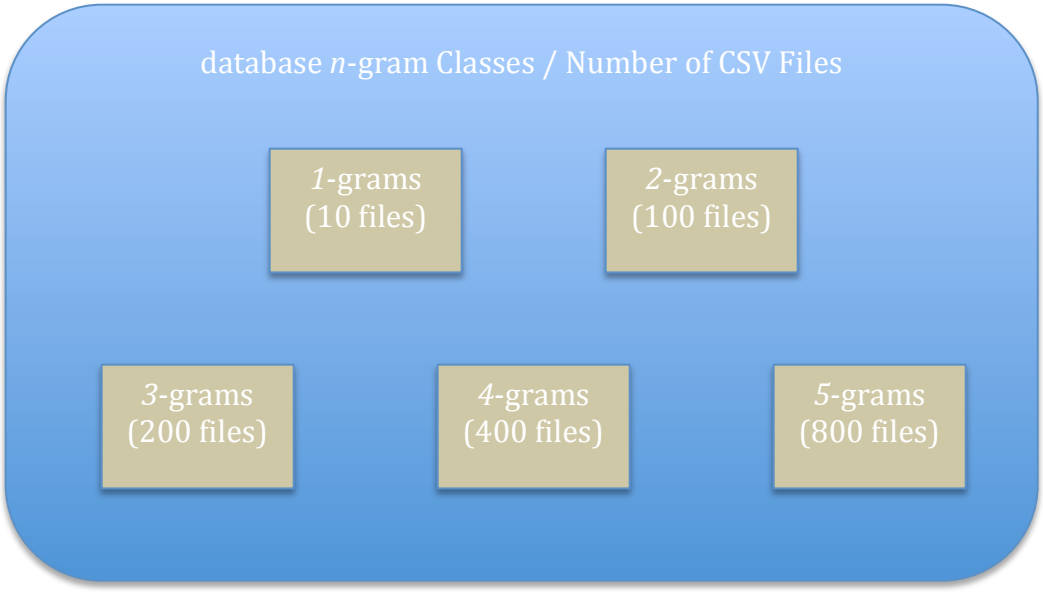


Figure 3: Google *n*-gram classes, along with the number of compressed files in each.

The Google database is comprised of *n*-gram classes *1* through *5*, but only classes *3* through *5* are utilized in this study (the rationale for which is discussed in Chapter 4). In addition, a "total counts" file is available for download in order to provide the total number of volumes included in the database by publish year. This means that a total of 1,401 files per language must be downloaded from the Google website and stored locally. Due to the size and number of files to be downloaded, it is desirable that an automated procedure be created and executed to accomplish this task.

In order to perform analysis against these files after they have been downloaded, each file must be extracted (or uncompressed) and then imported into a suitable local database management system (DBMS). This requires that table definitions first be created in the local DBMS that match the layout of the CSV files from Google – including such information as *n*-gram class, *n*-gram, year, match count, page count, and volume count. Due to the size of the databases being imported, and in order to clearly distinguish between the data for each *n*-gram class, it is recommended that three separate tables be defined per language – one for each *n*-gram class. Since each CSV file contains a distinct subset of the entire selected language database, each file may be imported incrementally into tables defined in the local DBMS (see Figures 4-5). Again, due to the number and size of the files involved, it is recommended that an automated process be created for importing the CSV files.

3.3 Edit and Prepare Documents

Before extracting *n*-grams from the documents under study, any necessary editing or preparation should be performed. This includes:

- Editing for formatting
- Editing for content

The first step (editing for formatting) involves placing the documents into a format that is accessible to the *n*-gram extraction software. For example, a document originally available in PDF format may have to be converted to MS-DOS text format. The requirements for this process, of course, depend on the limitations and capabilities of the *n*-gram extraction software selected.

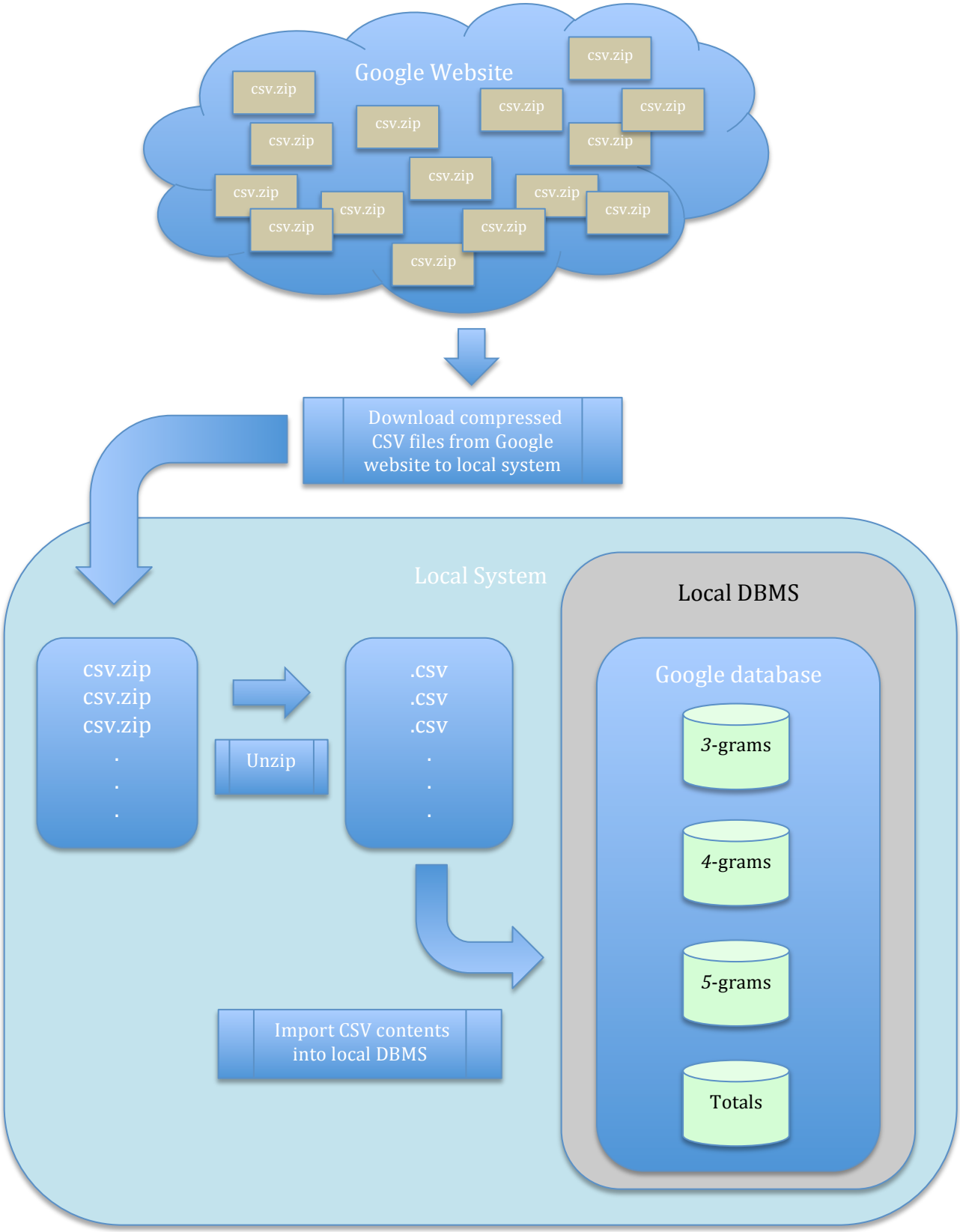


Figure 4: Downloading the Google database and importing into the local DBMS.

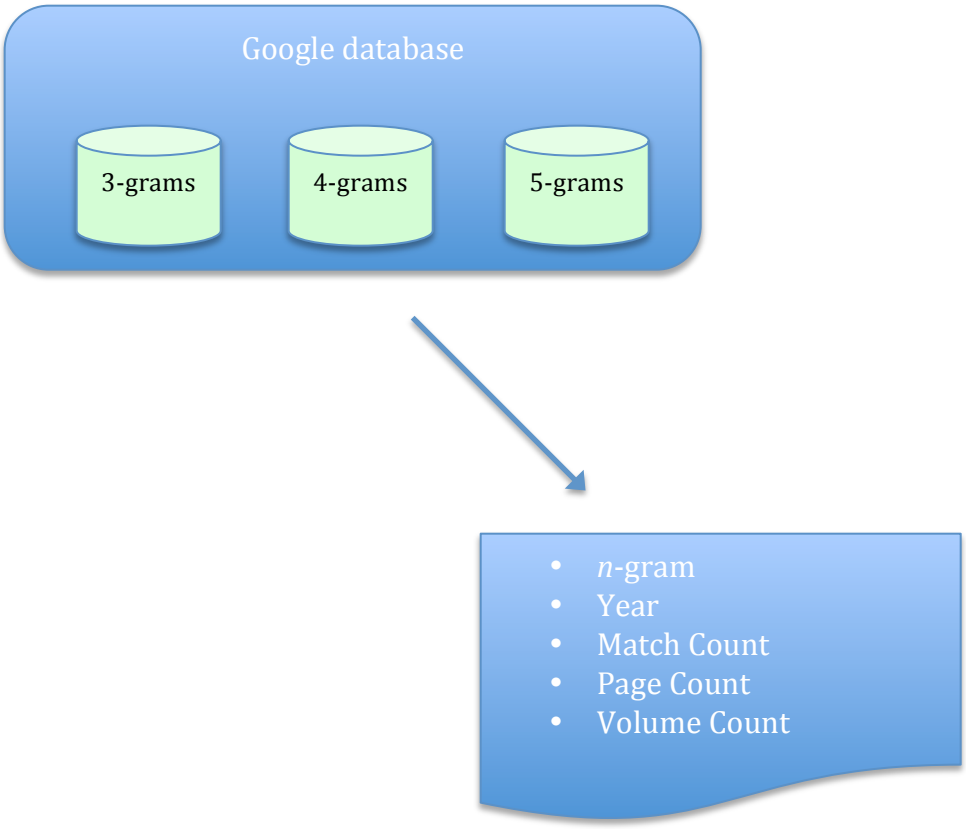


Figure 5: Table layout for storing Google *n*-gram database.

The second step (editing for content) is equally important, but may require careful analysis in order to accomplish. As a general rule, the fewer changes made to the content of a document, the better. In as much as is reasonable, the original document should be left unaltered. However, there are situations that indicate when editing of a document is in order. Once such indicator is the inclusion of non-original "wrapping" text. This is generally comprised of text that is added at the beginning and/or the end of the document, and is not part of the original text of the document. Common instances of this type of extraneous material can be found in modern editions of much older texts. These may include modern copyright notices, footnotes and even "Introductions" contributed by modern writers – any of which would likely contain words and phrases that are anachronistic to the original work, and could possibly bias our analysis in favor of a later time period. In these types of situations it is suggested that, if possible, a first edition or other original version of the document be procured. If this is not possible, attempts should be made to remove these modern additions to the text.

Another indicator is the verbatim inclusion of significant amounts of material from other authors and/or earlier time periods. This condition is much more difficult to correct, and requires a high level of familiarity with the text of the document under study. The reason for concern here is clear: since the purpose of this method is to arrive at a reasonable and unbiased estimation of the time period in which the document was authored, the presence of substantial amounts of material from earlier time periods may artificially bias our results towards the earlier time period. As a general rule, if material of this type is discovered within the document, and the amount of material is substantial, it should be removed before the document is submitted for *n*-gram extraction.

3.4 Extract *n*-grams from Documents

Once the documents have been edited for both format and content, they may be submitted for *n*-gram extraction. This process involves scanning each document individually in order to accomplish the following:

- Identification of each unique *3*-, *4*-, and *5*-gram embodied in the document's text
- Collection of these lists of *n*-grams into the local DBMS for analysis

The identification process can be performed by special software designed for that purpose. Such software applications will accept a document as input, and will produce as output a list of all unique *n*-grams of a specific class contained within the text, along with total counts of their occurrences. The list of *n*-grams produced is then imported into the local DBMS (see Figure 6). This requires that a table (or tables) be defined in the local DBMS to capture information such as document id, *n*-gram class, and *n*-gram (see Figure 7).

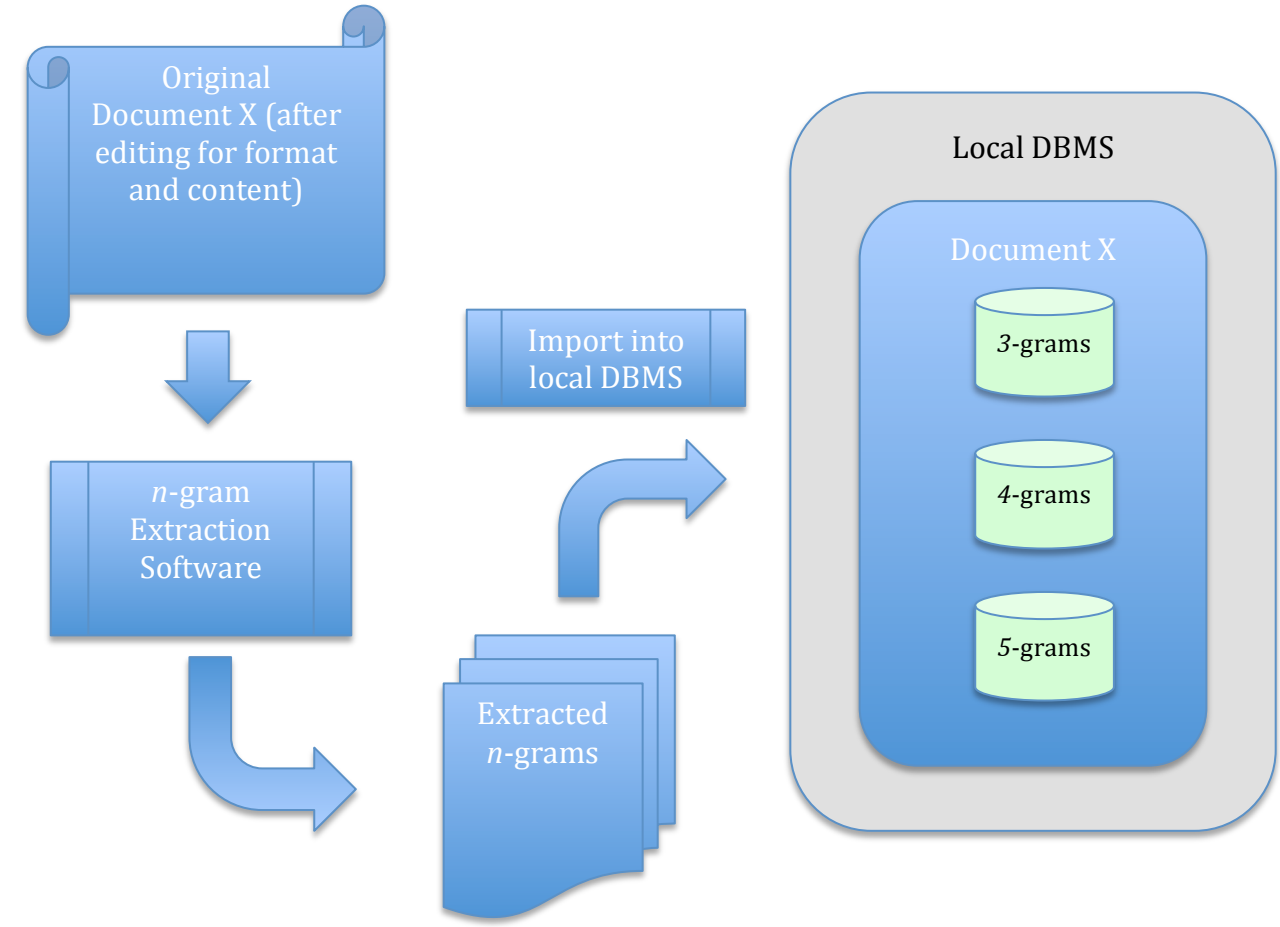


Figure 6: Extracting *n*-grams from documents and importing into the local DBMS.

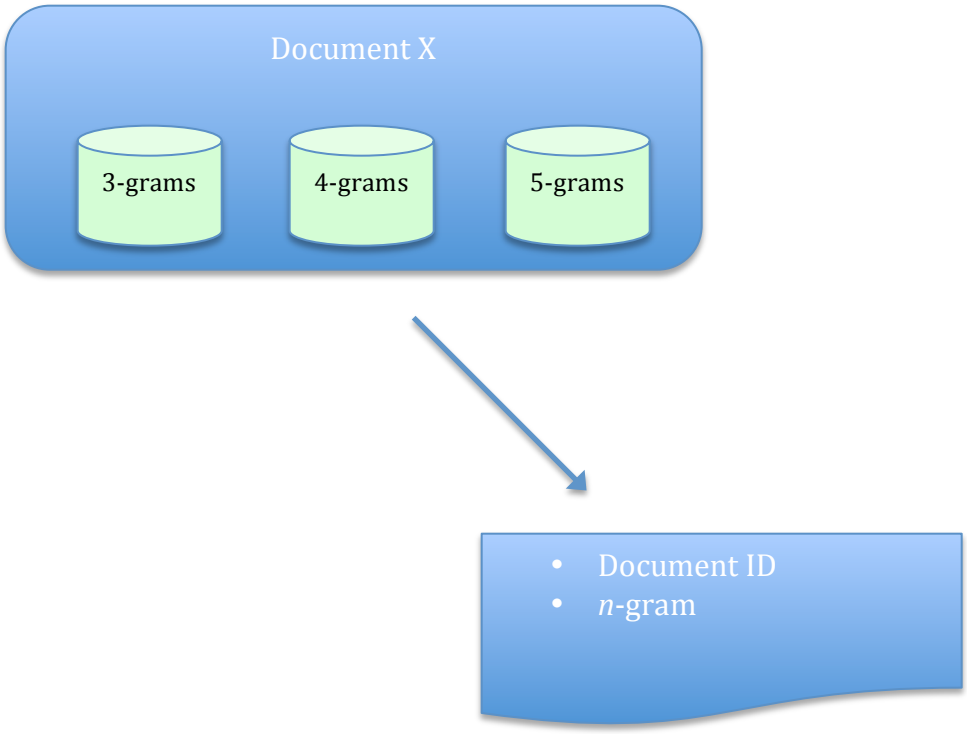


Figure 7: Table layout for collecting *n*-grams extracted from documents.

3.5 Find Extracted *n*-grams in the Google database

At this point in the process, the local DBMS contains the designated portions of the Google database along with the extracted *n*-grams from the documents to be studied (see Figure 8). Our final step in preparation for actually applying the method is to take each extracted *n*-gram and locate all of its occurrences within the corresponding Google database. Each match is noted in a set of tables created for this purpose (see Figure 9). It is recommended that a separate table be created for each *n*-gram class, and that it contain information such as document id, year, *n*-gram, match count, page count, and volume count (see Figure 10). As with others steps, an automated process should be created to detect these matches and record them in the appropriate tables.

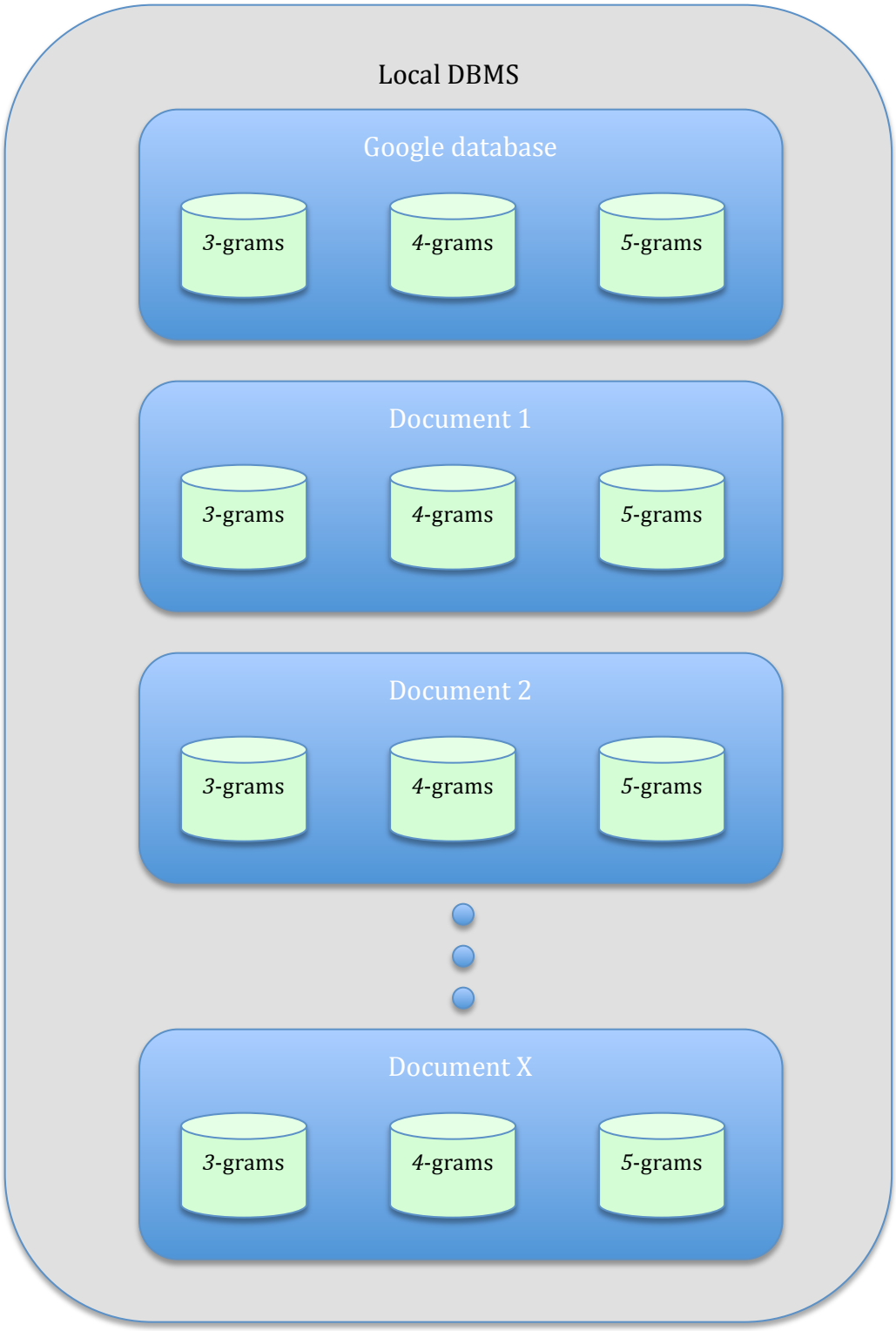


Figure 8: Contents of local DBMS after importing the Google database and *n*-grams extracted from the documents.

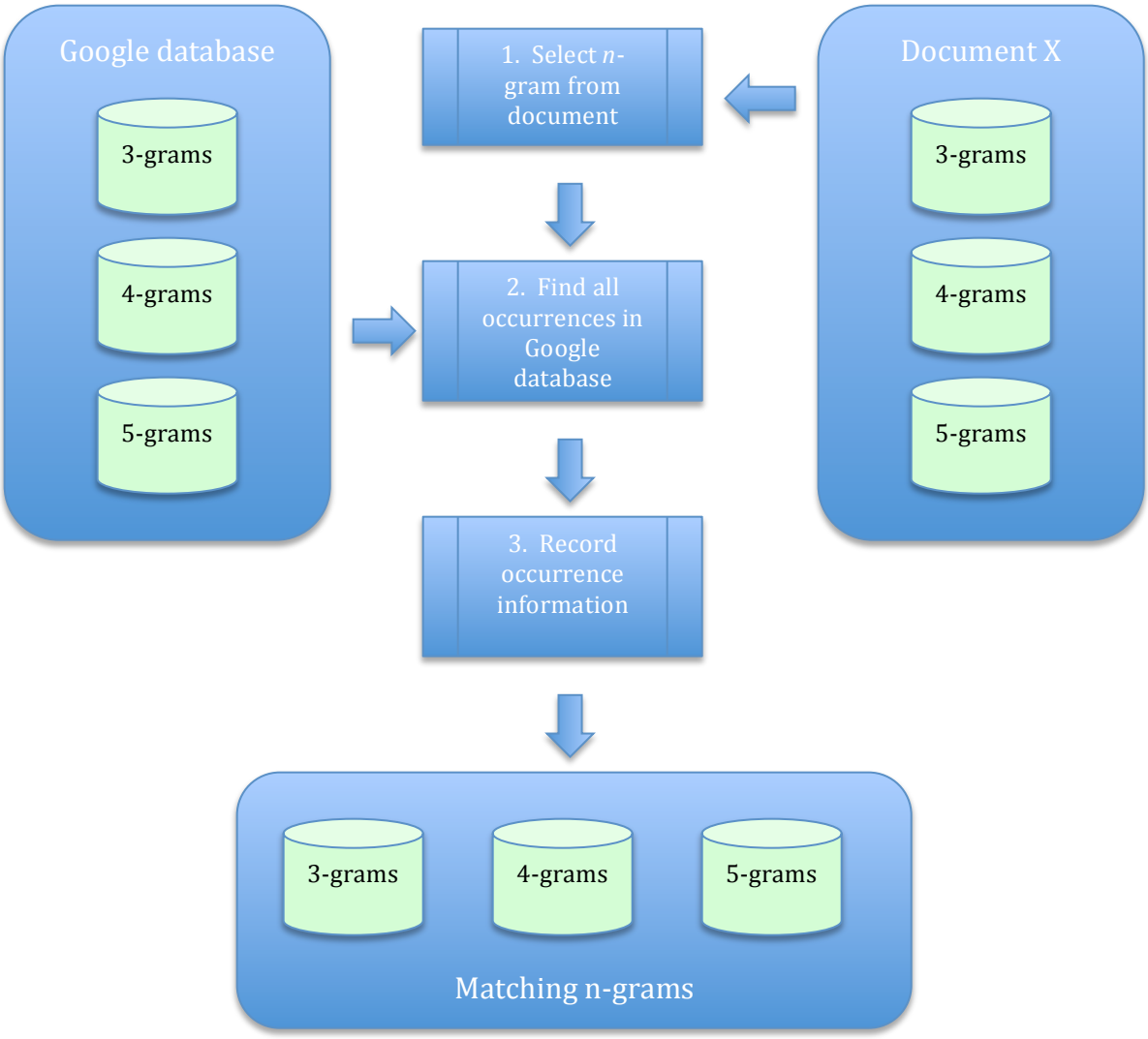


Figure 9: Finding document *n*-gram matches in the Google database and recording them in separate tables.

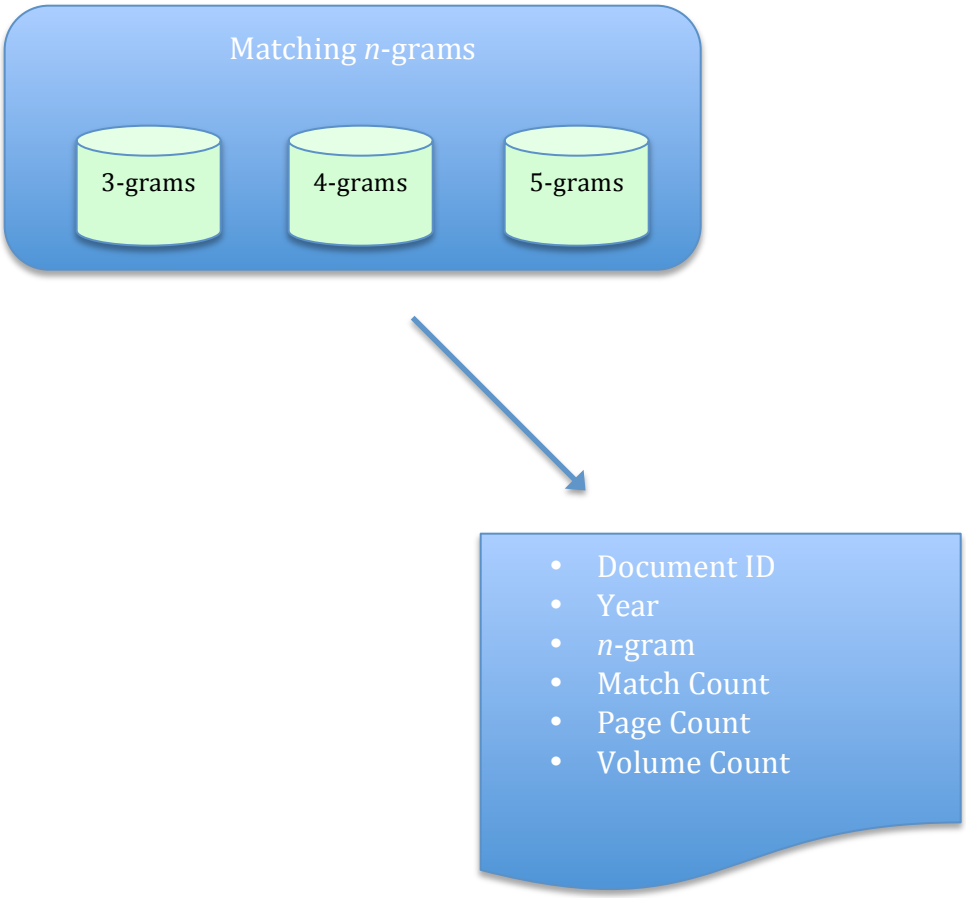


Figure 10: Table layout for collecting matching *n*-gram information.

CHAPTER 4: METHOD

4.1 Overview

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 determination 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 chapter proceeds to describe the method in detail. After the method is described, a process for its evaluation is proposed.

An initial attempt at developing a method was based on the concept of a *transient* n-*gram* (see Appendix A for details of this method). While its results were encouraging, the transient *n*-gram method suffered from weaknesses inherent in its design that restricted its applicability. It was abandoned in favor of a stronger approach based on *n*-gram popularity, which is the main focus of this study.

4.2 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 insure that the data used in the study are free from forward contamination, only data recorded before the publish year is 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.

4.3 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 comprised 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 execute. 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 (explained in detail below) 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 left 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 less matches; and less matches means less 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, 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.

Lastly, the inclusion of multiple classes of *n*-grams helps to determine if a particular class of *n*-gram is more useful than another in determining approximate date of authorship. At the outset, it is not apparent that one class of *n*-gram is inherently better suited to the task than is another; however, the fact that *3*-gram matches are more numerous could provide us with a higher degree of confidence in our results.

4.4 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 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, this 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 one-tenth 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 11). 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 12). 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 follows:

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

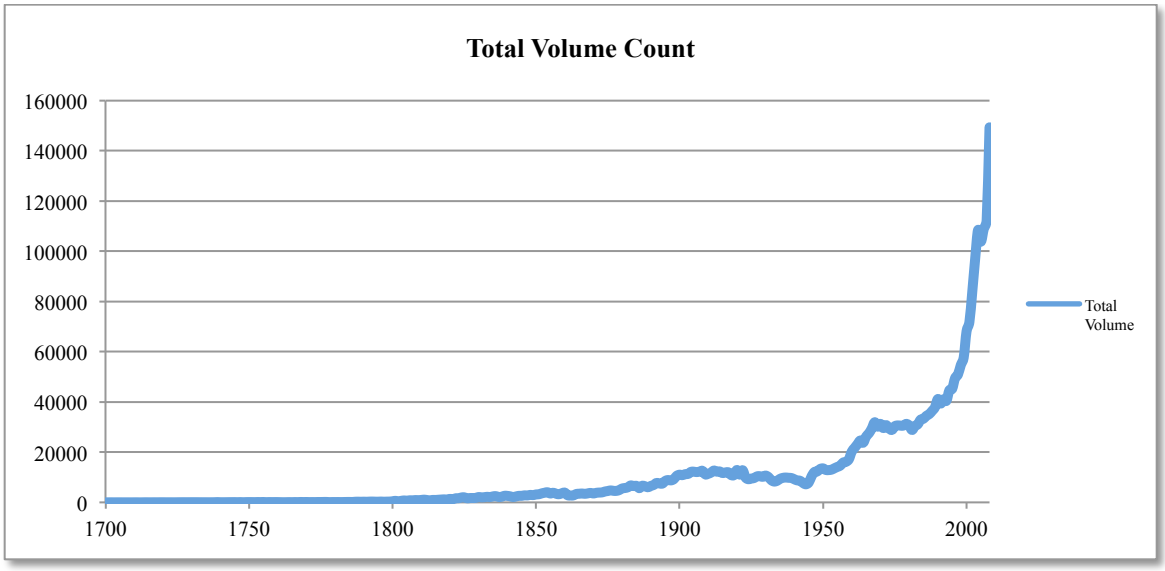


Figure 11: Total number of published volumes per year (as included in the Google database).

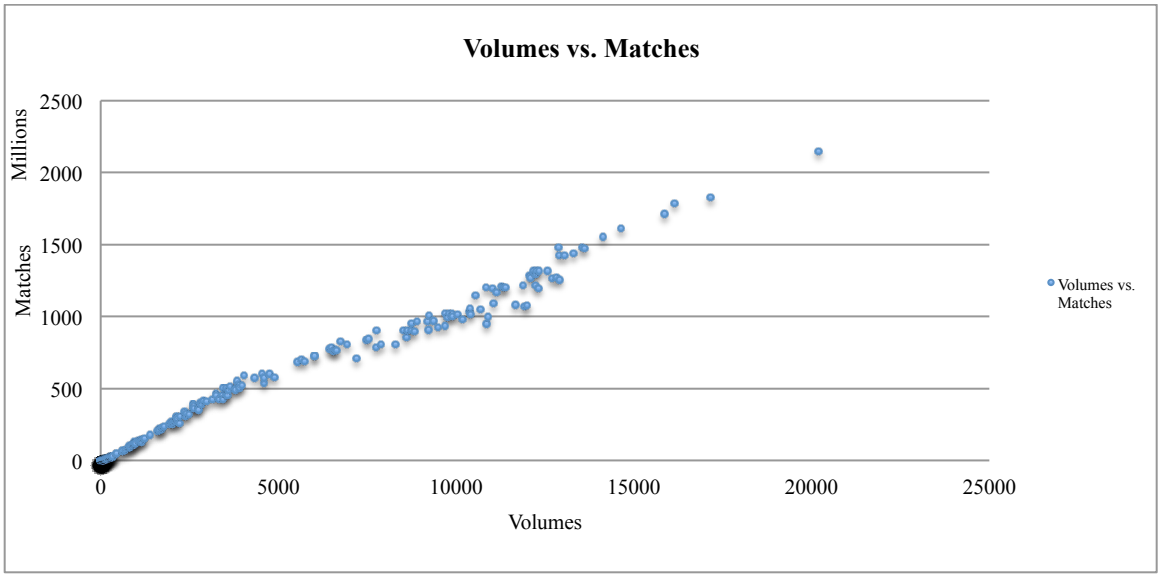


Figure 12: Positive correlation of volumes with matches per year. R (linear correlation coefficient) = 0.994.

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 one-tenth the value for 1778. We conclude that this *n*-gram was only about one-tenth 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 and 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 that we end up with years having varied sets of *n*-gram matches for the document.

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

$$\textit{Aggregate Popularity}_{(\textit{Document, Year})} = \Sigma\ \textit{Popularity Factor}_{(\textit{Document n-gram, Year})}$$

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 will determine the regression equation for the set of data points defined by the yearly aggregate popularity. We will 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 follows:

$\{Years\ with\ high\ aggregate\ popularity\}_{Document} = \{x: Aggregate\ Popularity_{(Document, x)} > \beta_0 + \beta_1 x\}$, *where*

β_0 = y-*intercept of the regression line*
β_1 = *Slope of the regression line*

This can be illustrated with an example. Figure 13 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:

$$\hat{y} = -51220 + 31.078x$$

Here, β_0 = -51220 and β_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.

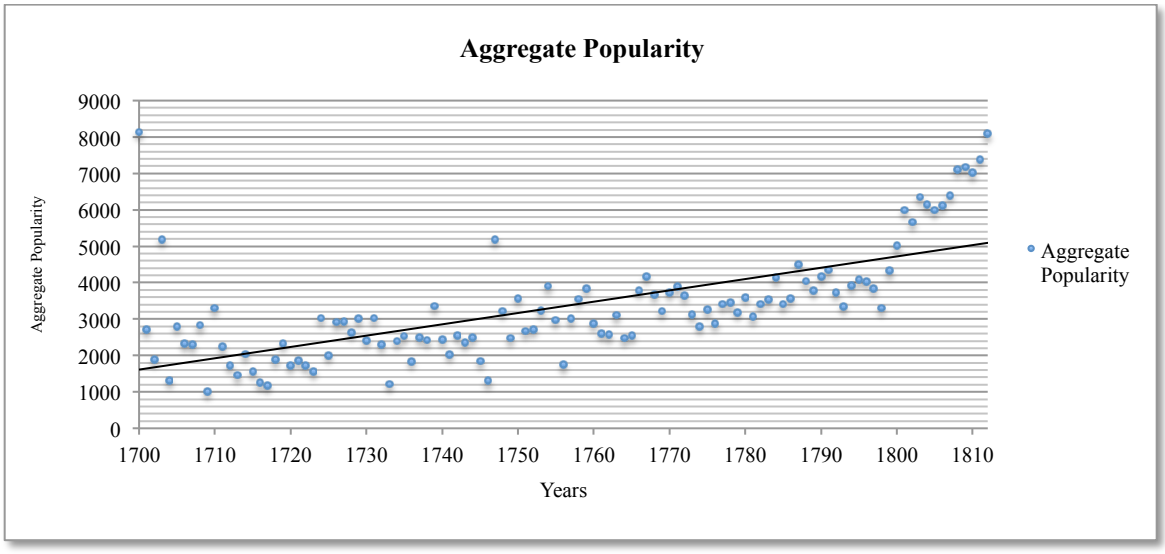


Figure 13: A sample scatter plot of aggregate popularity, with regression line.

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, this 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 will 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 can refer to Figure 14. The latest year in this dataset is 1812. So, we can define the following time periods over this dataset:

*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 a specific time period and a peak year. 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 Chapter 5, such a disparity may be an indicator of possible dissimulation.

4.5 Method Evaluation

Before proceeding with details of the method, it is appropriate and necessary to define a process for evaluating the method itself. One of the crucial elements of this study involved establishing a way to measure the performance of the devised method. It is one thing to develop a method that gives an approximate date or time period; it is another to develop a process that can objectively determine whether or not the approximate date or time period given is "close" to being correct or reasonable. Such a process should naturally include a set of criteria against which our method can be objectively evaluated. Table 2 includes a list of research hypotheses to assist us with this evaluation.

We propose a method evaluation that first applies the method to a group of documents for which the period of composition is known with a high degree of certainty (the "established" group), one document at a time, and determines how well its results fit the timeline model in consideration of hypotheses H_1 and H_2. The method is then applied in a similar manner against the documents in the "questionable" group. After the individual results from each document have been evaluated, the combined results are analyzed as a whole using H_1 through H_4, in order to see how the method performed overall. Finally, H_5 and H_6 are tested against the combined results simply to learn more about the relationship between an author's age and the length of the period of cultural influence, as well as how consistently the model performed against documents of different sizes.

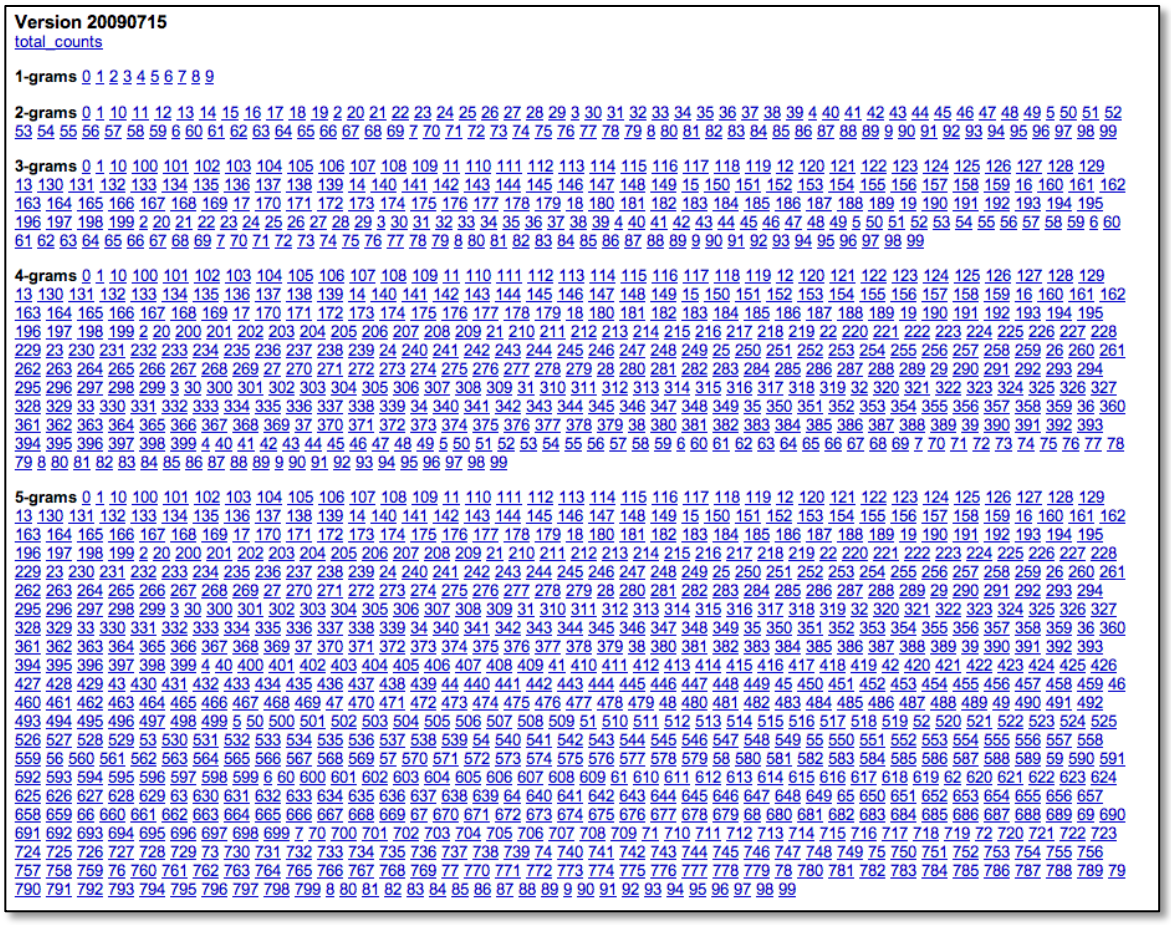
Hypothesis No.	Description
H_1	The period of cultural influence should begin before the beginning of the period of composition.
H_2	The period of cultural influence should end either before or during the period of composition.
H_3	As the birth year of the author increases, the length of the period of cultural influence should increase.
H_4	As the birth year of the author increases, the start date of the period of cultural influence should increase.
H_5	As the age of the author increases, the difference between the start of the period of cultural influence and the Publish Year should increase.
H_6	As the document file size increases, the variance between the results returned from different *n*-gram classes should decrease.

Table 2: Research hypotheses.

CHAPTER 5: METHOD APPLICATION AND RESULTS

5.1 Downloading the Google *n*-gram database

The first step in the process was to download the Google *n*-gram database to a local system. As mentioned previously, Google makes the databases available as a collection of compressed, comma-separated-value (CSV) files. These are presented as several hundred HTML links in a web page at their download site (see Figure 14).



Version 20090715
total_counts

1-grams 0 1 2 3 4 5 6 7 8 9

2-grams 0 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49 5 50 51 52
53 54 55 56 57 58 59 6 60 61 62 63 64 65 66 67 68 69 7 70 71 72 73 74 75 76 77 78 79 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92 93 94 95 96 97 98 99

3-grams 0 1 10 100 101 102 103 104 105 106 107 108 109 11 110 111 112 113 114 115 116 117 118 119 12 120 121 122 123 124 125 126 127 128 129
13 130 131 132 133 134 135 136 137 138 139 14 140 141 142 143 144 145 146 147 148 149 15 150 151 152 153 154 155 156 157 158 159 16 160 161 162
163 164 165 166 167 168 169 17 170 171 172 173 174 175 176 177 178 179 18 180 181 182 183 184 185 186 187 188 189 19 190 191 192 193 194 195
196 197 198 199 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49 5 50 51 52 53 54 55 56 57 58 59 6 60
61 62 63 64 65 66 67 68 69 7 70 71 72 73 74 75 76 77 78 79 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92 93 94 95 96 97 98 99

4-grams 0 1 10 100 101 102 103 104 105 106 107 108 109 11 110 111 112 113 114 115 116 117 118 119 12 120 121 122 123 124 125 126 127 128 129
13 130 131 132 133 134 135 136 137 138 139 14 140 141 142 143 144 145 146 147 148 149 15 150 151 152 153 154 155 156 157 158 159 16 160 161 162
163 164 165 166 167 168 169 17 170 171 172 173 174 175 176 177 178 179 18 180 181 182 183 184 185 186 187 188 189 19 190 191 192 193 194 195
196 197 198 199 2 20 200 201 202 203 204 205 206 207 208 209 21 210 211 212 213 214 215 216 217 218 219 22 220 221 222 223 224 225 226 227 228
229 23 230 231 232 233 234 235 236 237 238 239 24 240 241 242 243 244 245 246 247 248 249 25 250 251 252 253 254 255 256 257 258 259 26 260 261
262 263 264 265 266 267 268 269 27 270 271 272 273 274 275 276 277 278 279 28 280 281 282 283 284 285 286 287 288 289 29 290 291 292 293 294
295 296 297 298 299 3 30 300 301 302 303 304 305 306 307 308 309 31 310 311 312 313 314 315 316 317 318 319 32 320 321 322 323 324 325 326 327
328 329 33 330 331 332 333 334 335 336 337 338 339 34 340 341 342 343 344 345 346 347 348 349 35 350 351 352 353 354 355 356 357 358 359 36 360
361 362 363 364 365 366 367 368 369 37 370 371 372 373 374 375 376 377 378 379 38 380 381 382 383 384 385 386 387 388 389 39 390 391 392 393
394 395 396 397 398 399 4 40 41 42 43 44 45 46 47 48 49 5 50 51 52 53 54 55 56 57 58 59 6 60 61 62 63 64 65 66 67 68 69 7 70 71 72 73 74 75 76 77 78
79 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92 93 94 95 96 97 98 99

5-grams 0 1 10 100 101 102 103 104 105 106 107 108 109 11 110 111 112 113 114 115 116 117 118 119 12 120 121 122 123 124 125 126 127 128 129
13 130 131 132 133 134 135 136 137 138 139 14 140 141 142 143 144 145 146 147 148 149 15 150 151 152 153 154 155 156 157 158 159 16 160 161 162
163 164 165 166 167 168 169 17 170 171 172 173 174 175 176 177 178 179 18 180 181 182 183 184 185 186 187 188 189 19 190 191 192 193 194 195
196 197 198 199 2 20 200 201 202 203 204 205 206 207 208 209 21 210 211 212 213 214 215 216 217 218 219 22 220 221 222 223 224 225 226 227 228
229 23 230 231 232 233 234 235 236 237 238 239 24 240 241 242 243 244 245 246 247 248 249 25 250 251 252 253 254 255 256 257 258 259 26 260 261
262 263 264 265 266 267 268 269 27 270 271 272 273 274 275 276 277 278 279 28 280 281 282 283 284 285 286 287 288 289 29 290 291 292 293 294
295 296 297 298 299 3 30 300 301 302 303 304 305 306 307 308 309 31 310 311 312 313 314 315 316 317 318 319 32 320 321 322 323 324 325 326 327
328 329 33 330 331 332 333 334 335 336 337 338 339 34 340 341 342 343 344 345 346 347 348 349 35 350 351 352 353 354 355 356 357 358 359 36 360
361 362 363 364 365 366 367 368 369 37 370 371 372 373 374 375 376 377 378 379 38 380 381 382 383 384 385 386 387 388 389 39 390 391 392 393
394 395 396 397 398 399 4 40 400 401 402 403 404 405 406 407 408 409 41 410 411 412 413 414 415 416 417 418 419 42 420 421 422 423 424 425 426
427 428 429 43 430 431 432 433 434 435 436 437 438 439 44 440 441 442 443 444 445 446 447 448 449 45 450 451 452 453 454 455 456 457 458 459 46
460 461 462 463 464 465 466 467 468 469 47 470 471 472 473 474 475 476 477 478 479 48 480 481 482 483 484 485 486 487 488 489 49 490 491 492
493 494 495 496 497 498 499 5 50 500 501 502 503 504 505 506 507 508 509 51 510 511 512 513 514 515 516 517 518 519 52 520 521 522 523 524 525
526 527 528 529 53 530 531 532 533 534 535 536 537 538 539 54 540 541 542 543 544 545 546 547 548 549 55 550 551 552 553 554 555 556 557 558
559 56 560 561 562 563 564 565 566 567 568 569 57 570 571 572 573 574 575 576 577 578 579 58 580 581 582 583 584 585 586 587 588 589 59 590 591
592 593 594 595 596 597 598 599 6 60 600 601 602 603 604 605 606 607 608 609 61 610 611 612 613 614 615 616 617 618 619 62 620 621 622 623 624
625 626 627 628 629 63 630 631 632 633 634 635 636 637 638 639 64 640 641 642 643 644 645 646 647 648 649 65 650 651 652 653 654 655 656 657
658 659 66 660 661 662 663 664 665 666 667 668 669 67 670 671 672 673 674 675 676 677 678 679 68 680 681 682 683 684 685 686 687 688 689 69 690
691 692 693 694 695 696 697 698 699 7 70 700 701 702 703 704 705 706 707 708 709 71 710 711 712 713 714 715 716 717 718 719 72 720 721 722 723
724 725 726 727 728 729 73 730 731 732 733 734 735 736 737 738 739 74 740 741 742 743 744 745 746 747 748 749 75 750 751 752 753 754 755 756
757 758 759 76 760 761 762 763 764 765 766 767 768 769 77 770 771 772 773 774 775 776 777 778 779 78 780 781 782 783 784 785 786 787 788 789 79
790 791 792 793 794 795 796 797 798 799 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92 93 94 95 96 97 98 99

Figure 14: The English database version 20090715 as presented for download in the Google user interface.

Even though this study is limited to examination of only *3*-, *4*-, and *5*-grams, the size and number of files to be downloaded made it impractical to perform the process manually. Accordingly, an Automator script was constructed to perform the task unattended (see Figure 15).

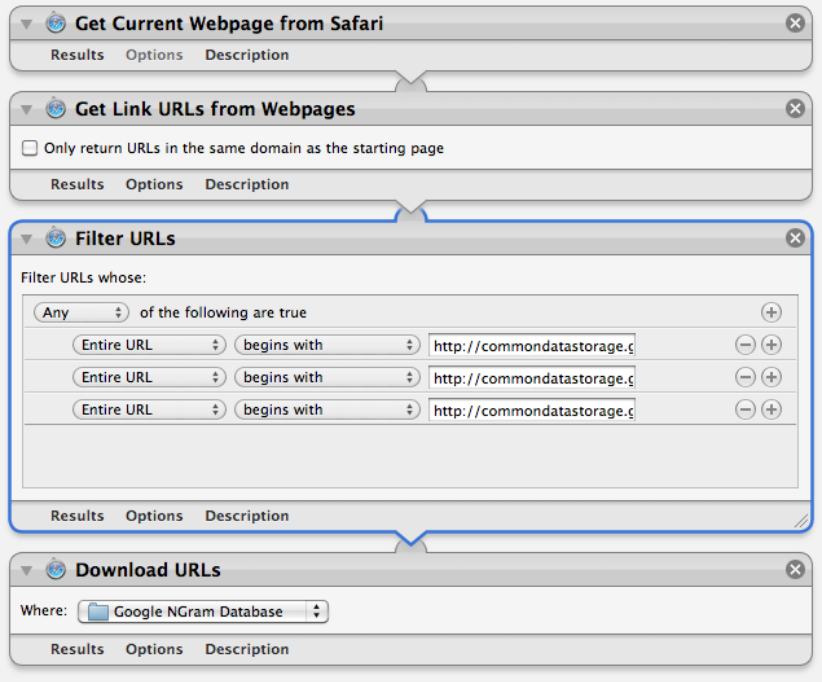


Figure 15: Automator workflow for downloading Google database files.

The Automator script (implemented as a sequence of pipes and filters) was composed of the following steps:

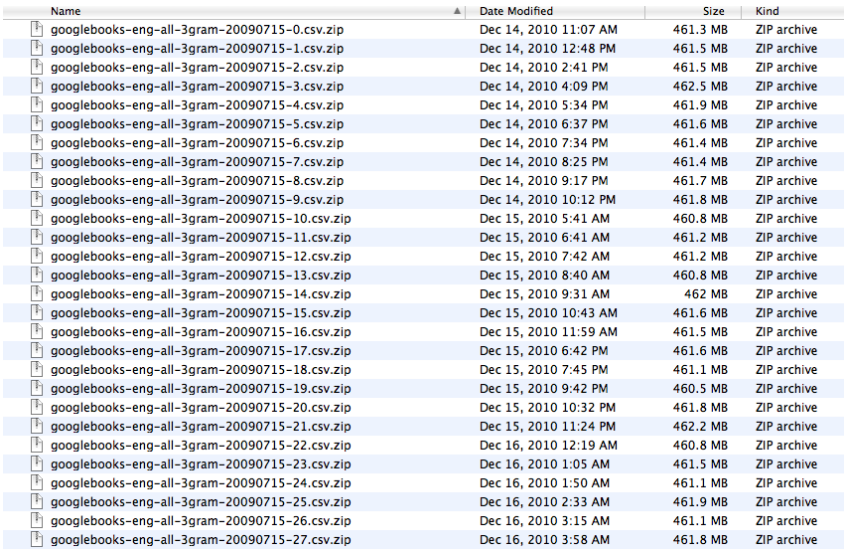
1. Get the current webpage from the Safari browser.
2. Get the link URLs from the selected webpage.
3. Filter the URLs, selecting only those that are actual download links for the *3*-, *4*-, and *5*-gram databases.
4. Download the selected URLs.

When executed, the script downloaded the compressed CSV files to the local system (see Figure 16).

At this point, the method called for decompressing the downloaded files and importing them into the local DBMS. The amount of data involved, however, forced a modification to this process. If the local system has ample disk space, the process outlined in Chapter 3 can be followed unmodified: all downloaded files can be decompressed, and then all of the resulting CSV files can be imported into the local DBMS. However, as has already been mentioned, the Google database is massive, requiring 310.91 GB in its compressed form alone. Since the compression ratio achieved is approximately 7:1, decompressing the entire database at once requires approximately 2.176 TB of available disk space. Such a

requirement is impractical for most local systems, and indeed was found to be prohibitive for the local system used in this study.

In order to address this constraint, the process was modified so as to combine the task of importing the Google database with the task of finding *n*-gram matches. This process in its modified form will be described in greater detail in section 5.4.



Name	Date Modified	Size	Kind
googlebooks-eng-all-3gram-20090715-0.csv.zip	Dec 14, 2010 11:07 AM	461.3 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-1.csv.zip	Dec 14, 2010 12:48 PM	461.5 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-2.csv.zip	Dec 14, 2010 2:41 PM	461.5 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-3.csv.zip	Dec 14, 2010 4:09 PM	462.5 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-4.csv.zip	Dec 14, 2010 5:34 PM	461.9 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-5.csv.zip	Dec 14, 2010 6:37 PM	461.6 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-6.csv.zip	Dec 14, 2010 7:34 PM	461.4 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-7.csv.zip	Dec 14, 2010 8:25 PM	461.4 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-8.csv.zip	Dec 14, 2010 9:17 PM	461.7 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-9.csv.zip	Dec 14, 2010 10:12 PM	461.8 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-10.csv.zip	Dec 15, 2010 5:41 AM	460.8 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-11.csv.zip	Dec 15, 2010 6:41 AM	461.2 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-12.csv.zip	Dec 15, 2010 7:42 AM	461.2 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-13.csv.zip	Dec 15, 2010 8:40 AM	460.8 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-14.csv.zip	Dec 15, 2010 9:31 AM	462 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-15.csv.zip	Dec 15, 2010 10:43 AM	461.6 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-16.csv.zip	Dec 15, 2010 11:59 AM	461.5 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-17.csv.zip	Dec 15, 2010 6:42 PM	461.6 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-18.csv.zip	Dec 15, 2010 7:45 PM	461.1 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-19.csv.zip	Dec 15, 2010 9:42 PM	460.5 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-20.csv.zip	Dec 15, 2010 10:32 PM	461.8 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-21.csv.zip	Dec 15, 2010 11:24 PM	462.2 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-22.csv.zip	Dec 16, 2010 12:19 AM	460.8 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-23.csv.zip	Dec 16, 2010 1:05 AM	461.5 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-24.csv.zip	Dec 16, 2010 1:50 AM	461.1 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-25.csv.zip	Dec 16, 2010 2:33 AM	461.9 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-26.csv.zip	Dec 16, 2010 3:15 AM	461.1 MB	ZIP archive
googlebooks-eng-all-3gram-20090715-27.csv.zip	Dec 16, 2010 3:58 AM	461.8 MB	ZIP archive

Figure 16: Partial list of compressed CSV files after downloading from the Google website.

In addition to the 1,400 CSV files containing detailed *3*-, *4*-, and *5*-gram counts by year, the *1*-gram "total counts" file was also downloaded and decompressed manually. A table was created that matched the layout of the CSV file, and the data was imported into the local DBMS by executing a SQL script (see Figures 17a-b).

```
CREATE TABLE `totals_1_grams' (
        `year'          int(11) DEFAULT NULL,
        `match_count'   int(11) DEFAULT NULL,
        `page_count'    int(11) DEFAULT NULL,
        `volume_count'  int(11) DEFAULT NULL
        )
```

Figure 17a: SQL used to create the **totals_1_grams** table.

```
load data local infile 'googlebooks-eng-all-totalcounts-20090715.txt'
into table test.totals_1_grams
fields terminated by '\t'
lines terminated by '\n'
(year, match_count, page_count, volume_count);
```

Figure 17b: SQL used to load *1*-grams totals into the **totals_1_grams** table.

5.2 Editing and Preparing Documents

All of the documents selected for this study were readily available in the public domain, and were easily located and downloaded from the Internet. The documents selected are listed in Table 3, along with their respective resulting file sizes in bytes. The relative size of each file is depicted in Figure 18. Since the *n*-gram extraction tool used in this study accepted files in plain text format, those files not already in text format were first converted. A few files contained extraneous metadata at the beginning and/or end of the document that was removed.

Book of Mormon was the only document that required more extensive editing before submission to the *n*-gram extraction process. This was due to the inclusion of substantial amounts of verbatim material from other authors of earlier time periods. Specifically, *Book of Mormon* quotes extensive passages from the King James Version of the Bible, totaling about 519 verses (see Table 4). These passages were removed in their entirety in order to guard against artificial bias towards an earlier time period.

Document	File Size (bytes)
Book of Mormon	1,350,985
Chronicles of Eri	733,993
Pride and Prejudice	682,851
Frankenstein	418,592
The Metamorphosis (German)	125,989
The Metamorphosis (English)	118,582
Common Sense	108,352
Vortigern and Rowena	88,154
Ern Malley Poems	15,820

Table 3: Documents selected for inclusion in the study, along with resulting file sizes.

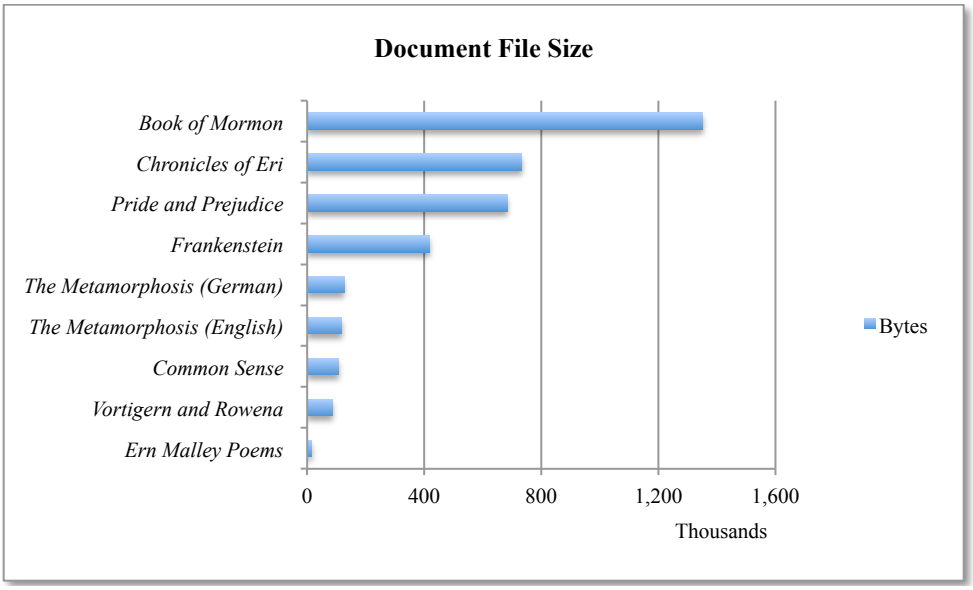


Figure 18: Relative document file sizes.

Biblical Passage	Pages in Book of Mormon 1830 Edition	Verses in current Book of Mormon Edition
Isaiah 48-49	1 Nephi 6 (52-56)	1 Nephi 20-21
Isaiah 50-52:2	2 Nephi 5 (74-78)	2 Nephi 7-8
Isaiah 2-14	2 Nephi 8-10 (86-102)	2 Nephi 12-24
Isaiah 53	Mosiah 8 (185-186)	Mosiah 14
Matthew 5:3-7:27	3 Nephi 5-6 (479-485)	3 Nephi 12-14
Isaiah 54	3 Nephi 10 (501)	3 Nephi 22
Malachi 3-4	3 Nephi 11 (503-504)	3 Nephi 24-25

Table 4: Biblical passages removed from *Book of Mormon* prior to *n*-gram extraction.

5.3 Extracting *n*-grams from the Documents

After the documents were edited for formatting and content, they were ready for the *n*-gram extraction process. This process involved supplying each document to the *n*-gram extraction software in order to create separate lists of all unique *3*-, *4*-, and *5*-grams in the document. After each list of *n*-grams was compiled, it was imported into the local DBMS.

The *n*-gram extraction software used in this study was kfNgram version 1.3.14 (July 11, 2007) by William H. Fletcher, available from KWiCFinder.com. This version of the software executed under

Microsoft Windows 7, accepted either text or HTML files for input, and produced text files for output. The software allowed user control of several processing parameters. These are listed and described in Table 5.

The resulting lists of extracted *n*-grams were then imported into the local DBMS. A temporary table was defined that matched the layout of the text files generated by the *n*-gram extraction software. Another master file was created to collect all extracted *n*-grams from all documents. Each list of extracted *n*-grams was imported into the temporary table, and then selected from there for insertion into the master table. The SQL used to define these tables and populate them is shown in Figures 19a-c. Some basic statistics on the *n*-grams extracted are shown in Tables 6a-b, and Figures 20a-b.

Parameter	Description	Value(s)
"nGrams"	The *n*-gram class to be extracted.	3, 4, 5
"Floor"	The minimum or threshold frequency an *n*-gram must have to be included. "1" means "include all n-grams."	1
"Show n-grams"	Causes each list of extracted *n*-grams to be displayed in a new window.	[selected]
"Chars to sort"	Specified the number of significant characters to sort at the beginning of each *n*-gram.	128
"Not Case Sensitive"	Determines whether *n*-grams are considered the same or separate, depending on character case.	[selected]
"Keep internal"	Retains specific marks internal to each word found, keeping words from being split into separate words.	The period, comma, dash, and apostrophe.
"Frequency Sort"	Specifies whether the resulting list of extracted *n*-grams is sorted alphabetically or by *n*-gram frequency.	[selected]
"Change numerals to #"	Each numeric digit is replaced with a pound sign ("#").	[selected]

Table 5: The kfNgram processing options as selected for this study (with descriptions freely adapted from the online help documentation).

```
CREATE TABLE `n_grams_temp' (
        `n_gram'        varchar(500)    NOT NULL,
        `match_count'   int(11)         DEFAULT NULL,
        PRIMARY KEY (
                'n_gram'
                )
        )
```

Figure 19a: SQL used to create the **n_grams_temp** table.

```
CREATE TABLE `book_n_grams' (
        `Book_ID'       int(11)         NOT NULL,
        `n_gram_type'   varchar(1)      NOT NULL,
        `n_gram'        varchar(500)    NOT NULL,
        `match_count'   int(11)         DEFAULT NULL,
        PRIMARY KEY  (
                'Book_ID',
                'n_gram',
                'n_gram_type'
                )
        )
```

Figure 19b: SQL used to create the **book_n_grams** table.

```
delete from test.n_grams_temp;

load data local infile 'Pride and Prejudice.txt-03-ngrams-Freq.txt'
into table test.n_grams_temp
fields terminated by '\t'
lines terminated by '\n'
(n_gram, match_count);

insert into test.book_n_grams(Book_ID, n_gram_type, n_gram, match_count)
select 1, '3', n_gram, match_count from test.n_grams_temp;
```

Figure 19c: Sample SQL used to load *3*-grams extracted from *Pride and Prejudice* into the **book_n_grams** table, via the **n_grams_temp table**.

Document	*3*-grams	*4*-grams	*5*-grams	Totals
Book of Mormon	131,849	190,068	220,614	542,531
Chronicles of Eri	88,908	109,966	120,061	318,935
Common Sense	17,701	18,638	18,823	55,162
Ern Malley Poems	2,681	2,692	2,692	8,065
Frankenstein	67,385	73,633	74,724	215,742
The Metamorphosis (English)	19,801	21,542	21,896	63,239
The Metamorphosis (German)	19,978	20,944	21,091	62,013
Pride and Prejudice	102,709	118,023	120,803	341,535
Vortigern and Rowena	14,384	14,705	14,749	43,838
Totals	465,396	570,211	615,453	1,651,060

Table 6a: Distinct *n*-grams extracted from each document.

Document	*3*-grams %	*4*-grams %	*5*-grams %
Book of Mormon	24.30257442	35.03357412	40.66385147
Chronicles of Eri	27.87652656	34.47912584	37.64434759
Common Sense	32.08911932	33.78775244	34.12312824
Ern Malley Poems	33.24240546	33.37879727	33.37879727
Frankenstein	31.23406662	34.13011838	34.63581500
The Metamorphosis (English)	30.07275975	34.55663402	35.37060623
The Metamorphosis (German)	31.31137431	34.06442227	34.62420342
Pride and Prejudice	32.21582571	33.77356361	34.01061068
Vortigern and Rowena	32.81171586	33.54395730	33.64432684
Totals	28.03051137	34.56581209	37.40367654

Table 6b: Distinct *n*-grams percentages extracted from each document.

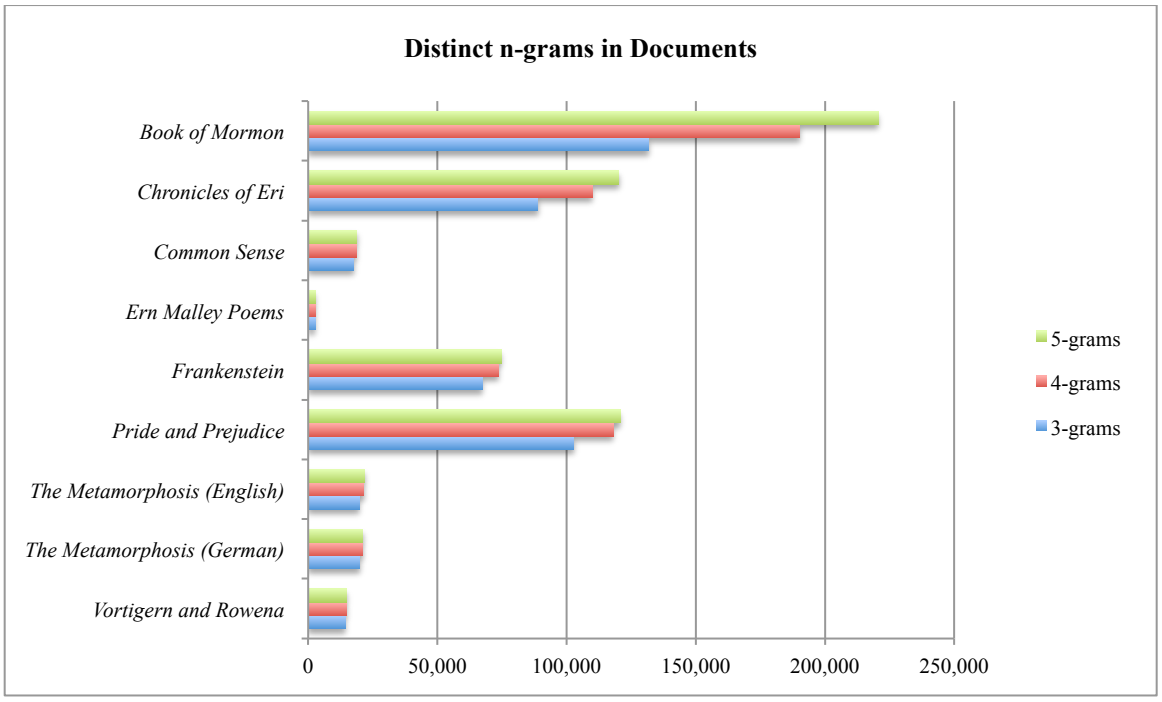


Figure 20a: Distinct *n*-grams in documents.

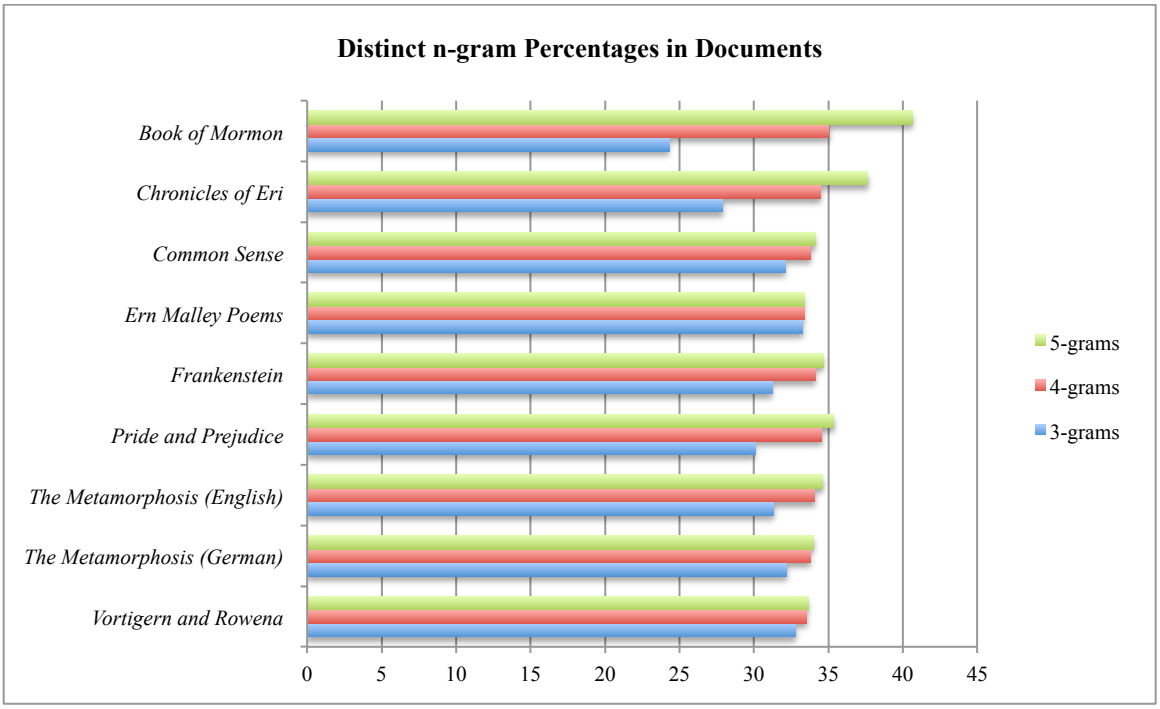


Figure 20b: Distinct *n*-gram percentages in documents.

5.4 Finding the Extracted *n*-grams in the Google database

After *n*-grams were extracted from the documents, the Google database was searched for matches, and each match was recorded. As stated earlier, due to disk space constraints on the local system, this searching process was modified to include the import phase of section 3.2. The method proposed in Chapter 4 describes the process of first loading the Google database into the local DBMS in its entirety, and then searching it for the *n*-grams that were extracted from the documents. Since extracting and importing the entire Google database prior to searching proved impractical, the process was modified so as to only extract and import one Google database CSV file at a time, search only within that specific file for *n*-gram matches, and then discard the file from the local DBMS (see Figure 21).

Given the large number of files to be processed, and the amount of time required to process each file, an automatic procedure was created to allow the process to execute unattended. This procedure was composed of two elements:

- An AppleScript routine
- A SQL script

A separate pair of these elements was written for each *n*-gram class. The AppleScript routine controlled the overall process loop, as well as the main steps of the process flow, such as copying a file to the work area, decompressing the file, renaming the file, executing the SQL script, and cleaning up the work area. The SQL script performed database-specific tasks, such as truncating the Google *n*-gram data temporary table, loading it with the current Google *n*-gram data from the work area, recording *n*-gram matches in the appropriate table, and logging the auditing information to a log table (see Figures 22a-f). Approximately 58 days were required to complete this process, which resulted in the location of over 26.6 million *n*-gram matches (see Table 7).

n-gram Class	Processing Time (days)	Total CSV Files Processed (English + German)	Total *n*-grams Extracted from Documents	Total *n*-gram Matches Found
3-grams	18	400	465,396	19,299,884
4-grams	23	800	570,211	5,999,081
5-grams	20	1,600	615,453	1,352,866
Totals	61	2,800	1,651,060	26,651,831

Table 7: Processing time required to process matches against *n*-gram classes, along with totals.

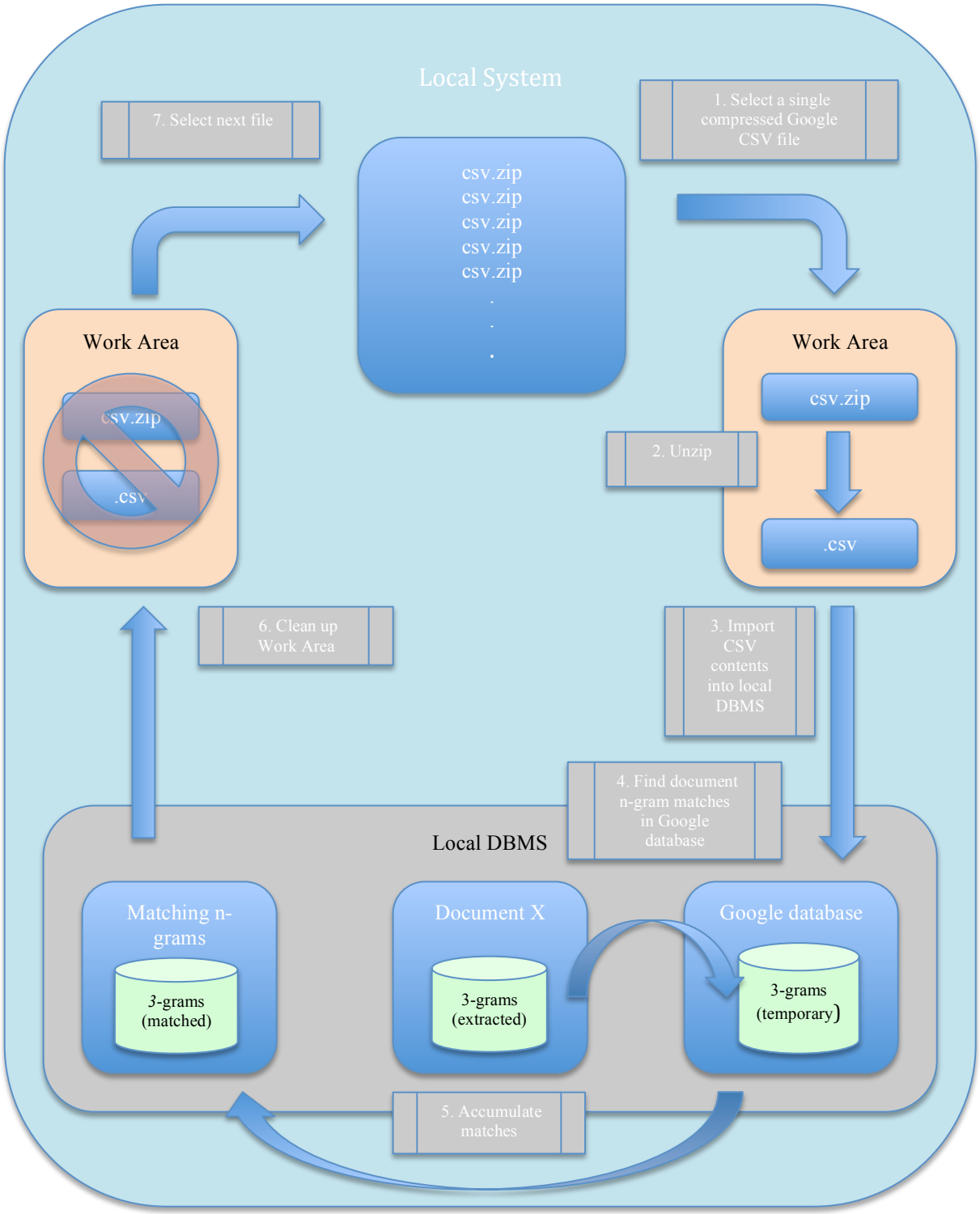


Figure 21: Sample modified *3*-gram search process.

```
# Variables
set ngramHome to "Macintosh HD:NGrams:Google databases:3-grams"

set fileWorkArea to "Macintosh HD:NGrams:Workarea"
set fileWorkArea2 to "/NGrams/Workarea/"
set fileScripts2 to "/NGrams/Scripts/"

set googleFileZip to "google_file.zip"
set googleFileCSV to "google_file.csv"

set fileNamePrefix to "googlebooks-eng-all-3gram-20090715-"

set fileIndex to 0

repeat 200 times

	set fileName to fileNamePrefix & fileIndex & ".csv"
	set fileNameZip to fileName & ".zip"

	tell application "Finder"
		copy file fileNameZip of folder ngramHome to folder fileWorkArea
		set name of document file fileNameZip of folder fileWorkArea to googleFileZip
		do shell script "/usr/bin/ditto -xk '" & fileWorkArea2 & googleFileZip & "'" & " '" &
				fileWorkArea2 & "'"
		set name of document file fileName of folder fileWorkArea to googleFileCSV
		delete file googleFileZip of folder fileWorkArea
	end tell

	do shell script "/usr/local/mysql/bin/mysql < '" & fileScripts2 & "Load3grams.sql" & "'"

	tell application "Finder"
		delete file googleFileCSV of folder fileWorkArea
	end tell

	set fileIndex to fileIndex + 1

end repeat
```

Figure 22a: AppleScript routine for finding *3*-gram matches in each of the 200 Google *3*-gram database files.

```
CREATE TABLE `books` (
	`Book_ID`		int(11)		NOT NULL,
	`Book_Name`		varchar(45)	DEFAULT NULL,
	`Year_Pub`		int(11)		DEFAULT NULL,
	`Author_Name`		varchar(45)	DEFAULT NULL,
	`Year_Written`		int(11)		DEFAULT NULL,
	PRIMARY KEY (
		`Book_ID`
		)
)
```

Figure 22b: SQL used to create the **books** table.

```
CREATE TABLE `log_3_grams` (
        `log_seq_no`     int(11)          NOT NULL AUTO_INCREMENT,
        `log_date_time`  datetime         DEFAULT NULL,
        `log_text`       varchar(500)     DEFAULT NULL,
        PRIMARY KEY (
                `log_seq_no`
                )
        )
```

Figure 22c: SQL used to create the **log_3_grams** table.

```
CREATE TABLE `Google_3_grams` (
         `n_gram`        varchar(100)   NOT NULL,
         `year`          int(11)        NOT NULL,
         `match_count`   int(11)        DEFAULT NULL,
         `page_count`    int(11)        DEFAULT NULL,
         `volume_count`  int(11)        DEFAULT NULL,
         PRIMARY KEY (
                `n_gram`,
                `year`
                )
        )
```

Figure 22d: SQL used to create the **Google_3_grams** table.

```
CREATE TABLE `Match_Details_3_grams` (
        `Book_ID`       int(11)          DEFAULT NULL,
        `n_gram`        varchar(500)     DEFAULT NULL,
        `year`          int(11)          DEFAULT NULL,
        `match_count`   int(11)          DEFAULT NULL,
        `page_count`    int(11)          DEFAULT NULL,
        `volume_count`  int(11)          DEFAULT NULL
)
```

Figure 22e: SQL used to create the **Match_Details_3_grams** table.

```
insert into test.log_3_grams(log_date_time,log_text) values( Now(), '** Start File Load **');

truncate table test.Google_3_grams;

insert into test.log_3_grams(log_date_time,log_text) values( Now(), 'Google_3_grams truncated');

load data local infile '/NGrams/Workarea/google_file.csv' into table test.Google_3_grams
fields terminated by '\t'
lines terminated by '\n'
(n_gram, year, match_count, page_count, volume_count);

insert into test.Match_Details_3_grams(Book_ID,n_gram,year,match_count,page_count,volume_count)
select
book_n_grams.Book_ID,Google_3_grams.n_gram,Google_3_grams.year,Google_3_grams.match_count,
Google_3_grams.page_count,Google_3_grams.volume_count
from test.Google_3_grams,test.book_n_grams,test.books
where Google_3_grams.n_gram=book_n_grams.n_gram
and book_n_grams.Book_ID=books.Book_ID
and book_n_grams.n_gram_type='3'
and book_n_grams.match_count > 0
and Google_3_grams.year < books.Year_Pub;

insert into test.log_3_grams(log_date_time,log_text) select Now(),ConCat('Total rows in
match_details_3_grams: ', Cast(count(*) as char(10)) ) from test.match_details_3_grams;

insert into test.log_3_grams(log_date_time,log_text) values( Now(), '** End File Load **');
```

Figure 22f: **Load3grams.sql** – a SQL script for finding *3*-gram matches in a specific Google database file, and logging progress details.

5.5 Determining Time Period of Greatest Cultural Influence

This section discusses the method that was employed to determine the period of cultural influence for the documents. The discussion will show the application of the method in detail, the results that were obtained, and the interpretation of those results.

5.5.1 Method Application

The method was designed to return two types of results for each document and *n*-gram class: a date range, and a mean date. In order to implement this method, a stored procedure was created and executed for each combination of document and *n*-gram class. The procedure was named **PerformMethod**(), and its full text is included in Appendix B. It was designed to accept two input parameters:

- Book Id
- *n*-gram Class

A unique "Book Id" was assigned to each book in the study, and the metadata for each book was entered manually into the **books** table (see Table 8). Application of the method, then, was simply a matter of executing the **PerformMethod**() procedure for each "Book Id" and *n*-gram class. For example, to apply the method to the *3*-gram matches collected for *Pride and Prejudice*, the following database stored procedure call was issued:

Call **PerformMethod**(1, 3);

Here, "1" is the "Book Id" for *Pride and Prejudice*, and "3" is the *n*-gram class desired. The output of **PerformMethod**() consisted of two main items of information: a date range, and a mean year.

Book_ID	Book_Name	Year_Pub	Author_Name	Year_Written
1	*Pride and Prejudice*	1813	Jane Austen	1798
2	*Metamorphosis* (English)	1915	Franz Kafka	1912
3	*Metamorphosis* (German)	1915	Franz Kafka	1912
4	*Frankenstein*	1818	Mary Shelley	1815
5	*Common Sense*	1776	Thomas Paine	1775
6	*Chronicles of Eri*	1822	Roger O'Connor	?
7	*Ern Malley Poems*	1945	James McAuley and Harold Stewart	1944
8	*Vortigern and Rowena*	1796	William Henry Ireland	1794
9	*Book of Mormon*	1830	Joseph Smith	?

Table 8: Contents of the **books** table.

5.5.2 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: a period of cultural influence, and a year of peak influence. These results are presented in Tables 9 and 10, along with other pertinent information about each document.

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

Table 9: Results – Period of Cultural Influence. Results were rounded to the nearest year.

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.

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

Table 10: Results – Peak Year of Cultural Influence. Results were rounded to the nearest year.

The remainder of this chapter is concerned with interpreting these results and evaluating the effectiveness of the method as it was applied.

5.5.3 Interpretation of Results

This section presents interpretations of the results, and will be organized into two sub-sections: the established documents and the questionable documents. Each document will be considered individually.

5.5.3.1 Established Documents

This section discusses the interpretation of the results from the group of "established" documents. More attention is given to the results from *The Metamorphosis*, due to the unexpectedly broad gap between its period of cultural influence and its period of composition.

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. These results are reasonable and consistent with our expectations (see Figure 23).

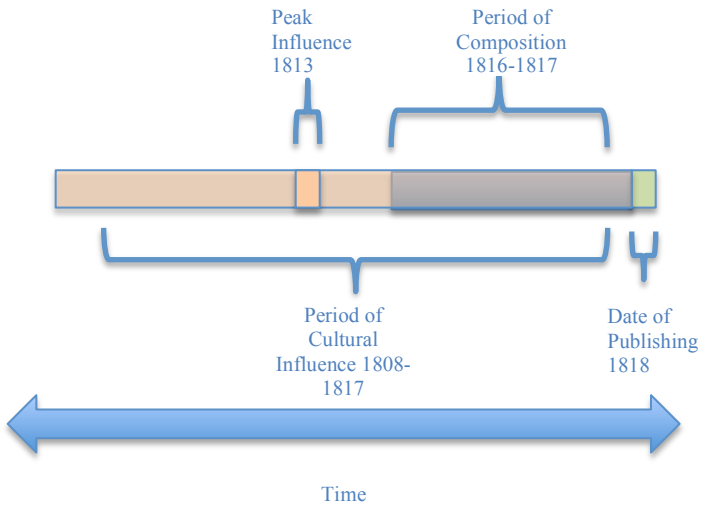


Figure 23: Method results for *Frankenstein*.

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. These results are reasonable and consistent with our expectations (see Figure 24).

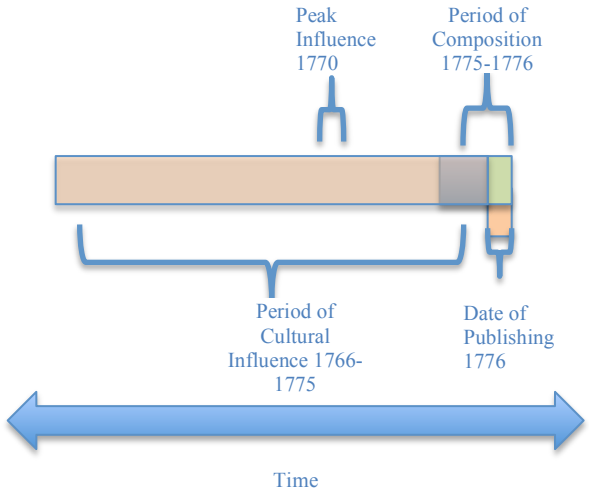


Figure 24: Method results for *Common Sense*.

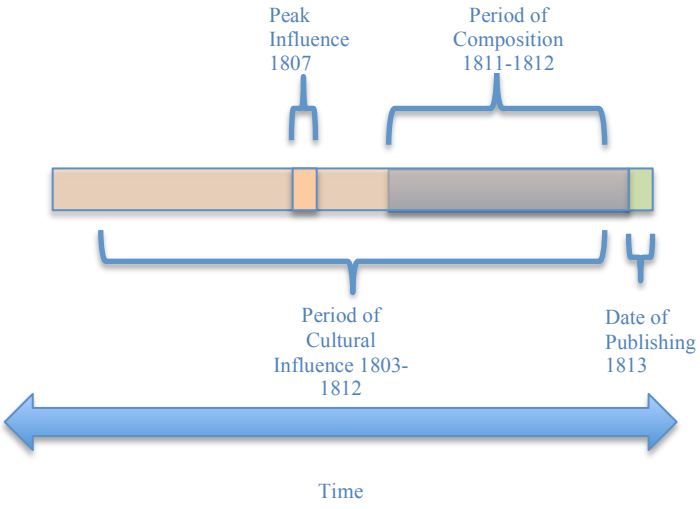


Figure 25: Method results for *Pride and Prejudice*.

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. These results are reasonable and consistent with our expectations (see Figure 25).

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, that 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 were quite similar, we concluded 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" [21]. Flaubert lived from 1821 to 1880, and produced most of his output during the mid-1800s [22]. 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 (see Figure 26).

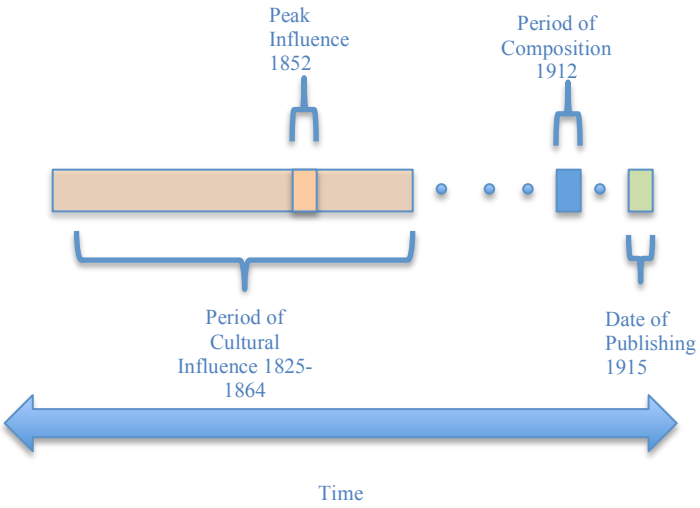


Figure 26: Method results for *The Metamorphosis* (German).

5.5.3.2 Questionable Documents

This section discusses the interpretation of the results from the group of "questionable" documents. More attention is given to the results from *Book of Mormon* and the *Ern Malley Poems*, due to the support this method lends to the results of other researchers concerned with time period of cultural influence.

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. These results are reasonable and consistent with our expectations (see Figure 27).

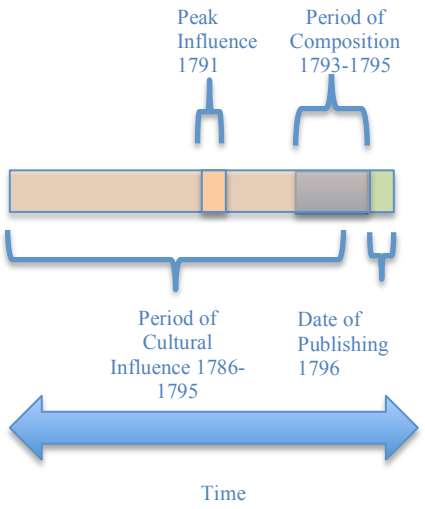


Figure 27: Method results for *Vortigern and Rowena*.

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 must be emphasized that the exact dates for the period of composition are not well established. We conclude that these results are reasonable and consistent with our expectations (see Figure 28).

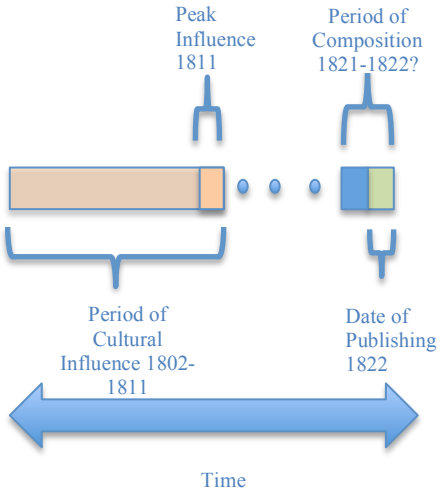


Figure 28: Method results for *Chronicles of Eri*.

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" [23]. 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 (see Figure 29).

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. [24]. 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 between any of them (see Figures 30a-d). The scatter plots for *4*- and *5*-grams are just as 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.

Identifying *Book of Mormon n*-gram data within this time period suggests other possibilities for authorship than those put forward by Smith. One such conjecture attributes authorship to Smith, and claims he was strongly influenced by another work on a similar topic entitled *View of the Hebrews*, published in 1825 by Ethan Smith (no relation). While further discussion of this theory is beyond the scope of this study, it should be noted that the time period of cultural influence suggested by this method is in agreement with the time period implied by this alternate theory of authorship [25, 26, 27].

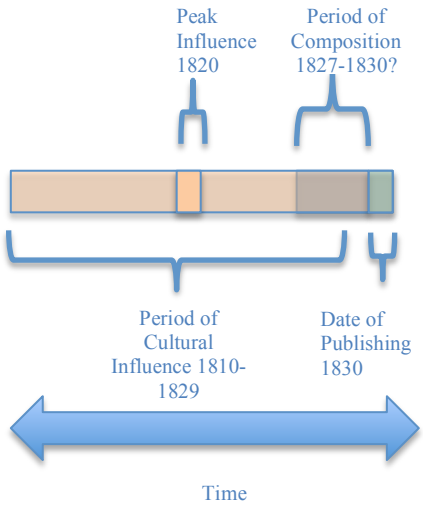


Figure 29: Method results for *Book of Mormon*.

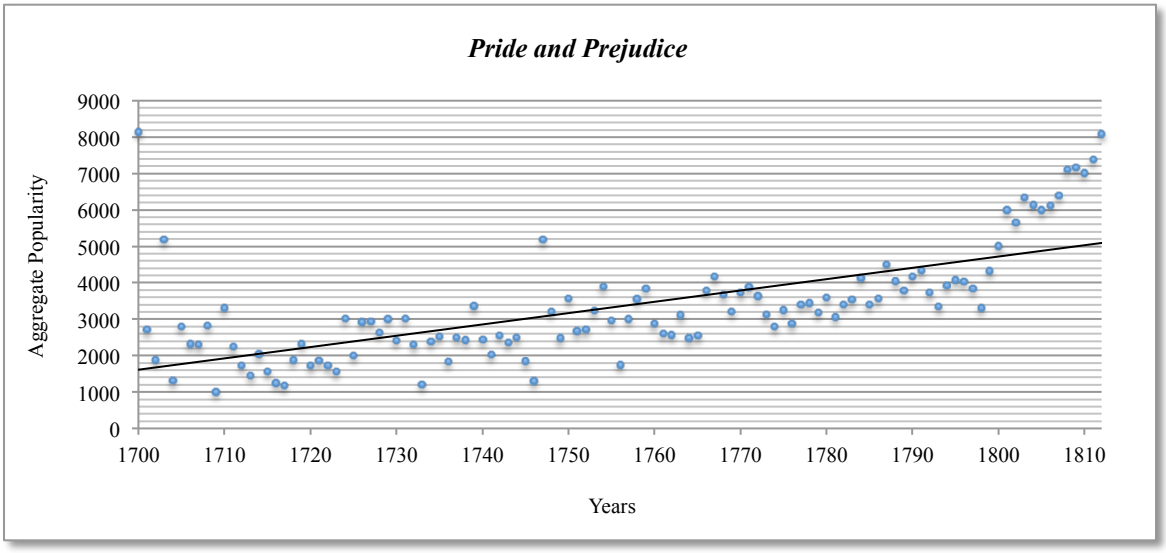


Figure 30a: Aggregate Yearly Popularity for *Pride and Prejudice* 3-gram data.

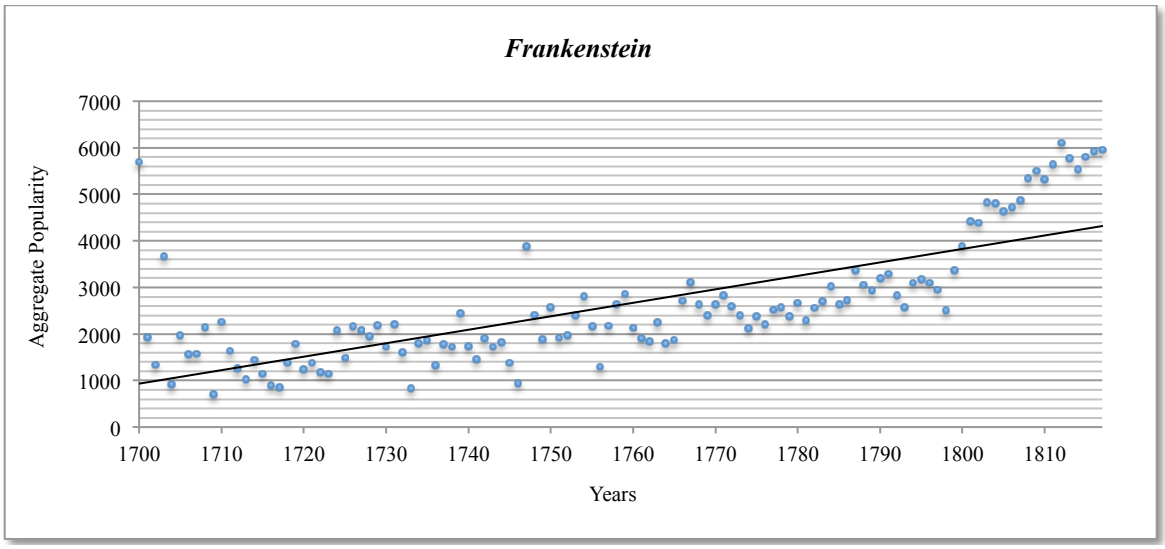


Figure 30b: Aggregate Yearly Popularity for *Frankenstein 3*-gram data.

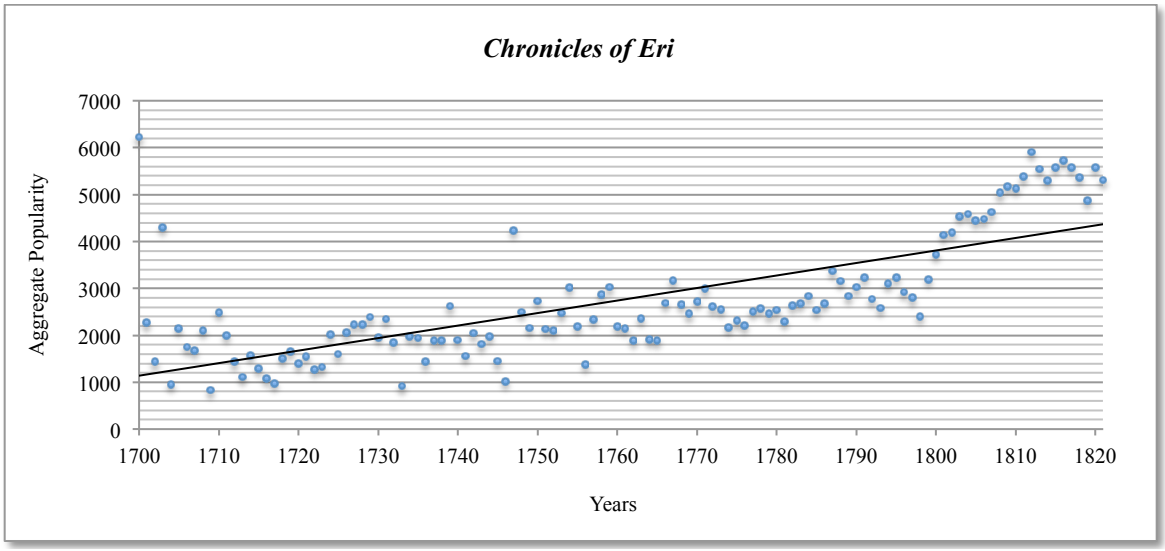


Figure 30c: Aggregate Yearly Popularity for *Chronicles of Eri 3*-gram data.

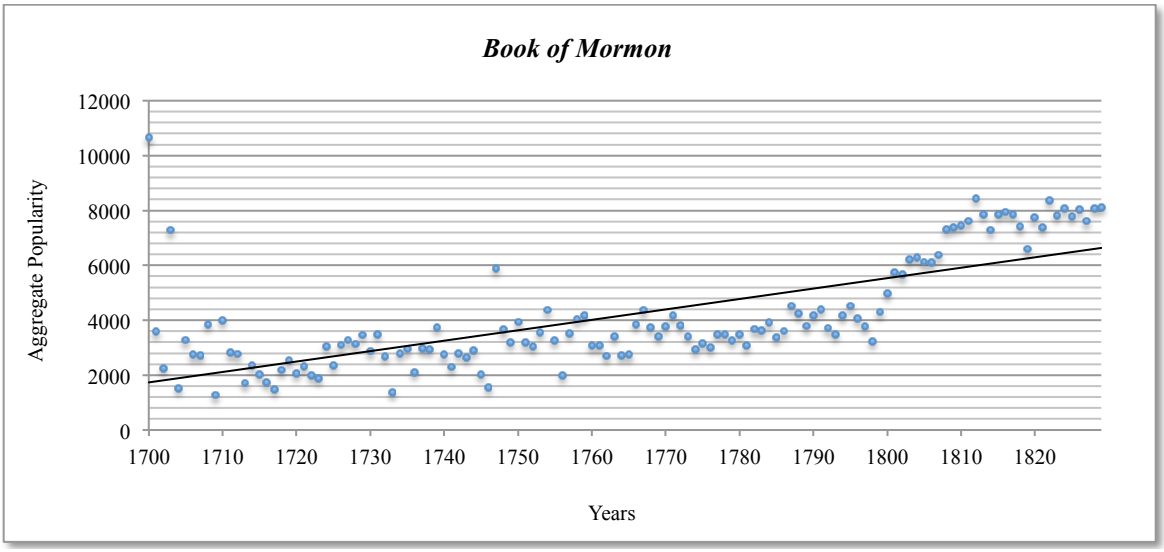


Figure 30d: Aggregate Yearly Popularity for *Book of Mormon* 3-gram data.

Ern Malley Poems

The interpretation of results from the *Ern Malley Poems* presented 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 Figures 31a-b).

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 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 [18]. 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" [28].

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 method [29]. 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.

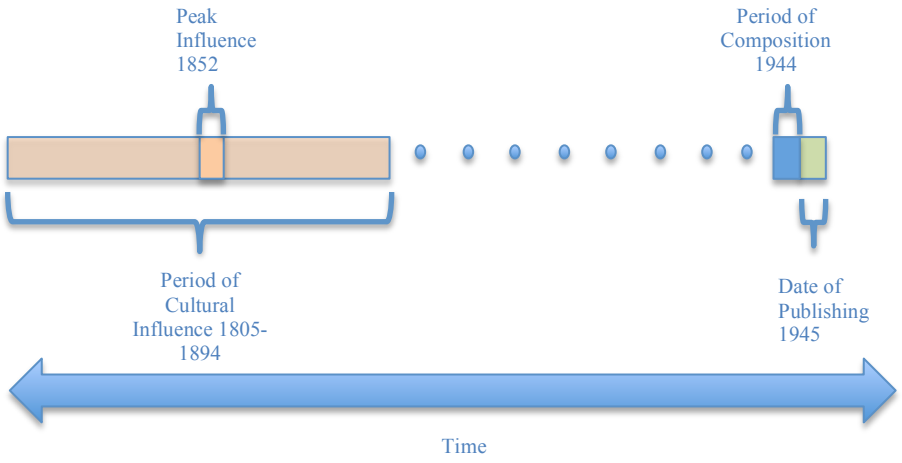


Figure 31a: Method results for *Ern Malley Poems*.

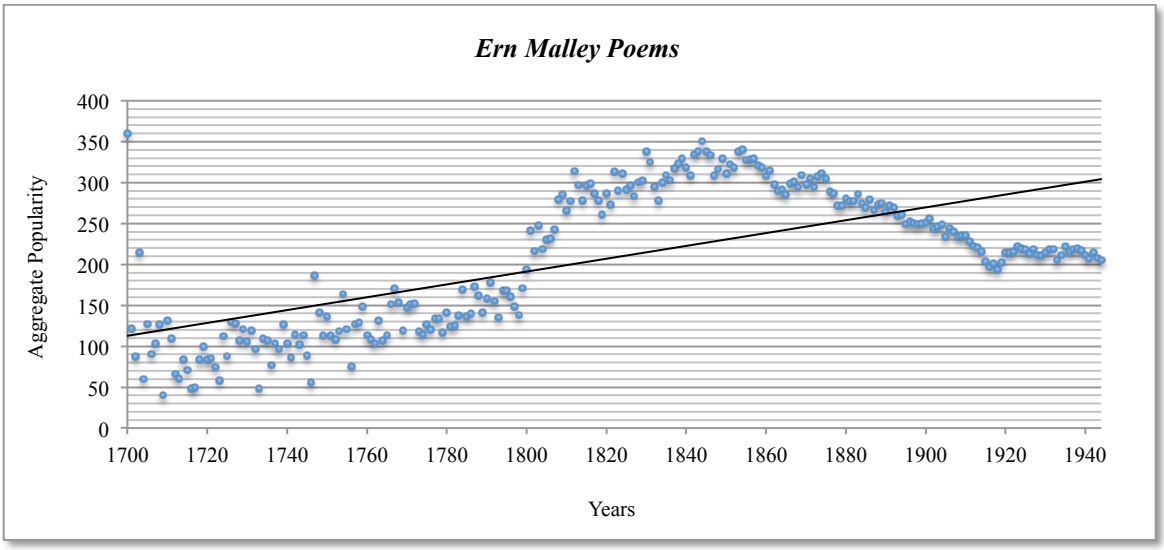


Figure 31b: Aggregate Yearly Popularity for *Ern Malley Poems* 3-gram data.

5.5.4 Method Evaluation

As discussed in Chapter 3, the method is evaluated against several criteria we would expect from a valid model. These are expressed in the form of six research hypotheses (see Table 2). Each of these will be tested in the sections that follow.

A number of these hypotheses (H_3 through H_6) depend on the interpretation of the Pearson Correlation Coefficient. The number of data pairs in our population is 9. Therefore, for each of the hypotheses tested below that involve the Pearson Correlation Coefficient, the critical value for $\alpha = 0.05$ is 0.666, and for $\alpha = 0.01$ is 0.798.

Testing H_1

Hypothesis H_1 is stated as follows:

$H_{1:}$ "The beginning of the period of cultural influence should either precede or be equal to the beginning of the period of composition."

This hypothesis can easily be tested and accepted by examining the results in Table 10. The method returned a predicted period of cultural influence with a beginning date that is less than the beginning date of the period of composition for every book in the study. This allows us to accept H_1.

Testing H_2

Hypothesis H_2 is stated as follows:

$H_{2:}$ "The ending of the period of cultural influence should either precede the beginning of the period of composition or fall within the period of composition."

Like H_1, this hypothesis can also be easily tested and accepted by examining the results in Table 10. A period of cultural influence with an ending year that either precedes or falls within the period of composition is predicted for every book in the study. This allows us to accept H_2.

Testing H_3

Hypothesis H_3 is stated as follows:

$H_{3:}$ "As the birth year of the author increases, the length of the Period of Cultural Influence should increase."

Since this hypothesis tests correlation of the population, our test is:

H_0: $\rho = 0$ (no linear correlation)

H_3: $\rho \neq 0$ (linear correlation)

The data used to test this hypothesis are displayed in Table 11, and their scatter plot is shown in Figure 32. The equation of the regression line through the data points is:

$$\hat{y} = 1769.3 + 1.9547x$$

This gives a Pearson Correlation Coefficient $\rho = 0.7706$. But we have $0.7706 > 0.6660$, allowing us to reject the null hypothesis in favor of H_3 with a 95% confidence level.

Document	Author Birth Year	Period of Cultural Influence	Period of Cultural Influence Length (years)
Pride and Prejudice	1775	1803-1812	10
The Metamorphosis (English)	1883	1825-1874	50
The Metamorphosis (German)	1883	1855-1864	10
Frankenstein	1797	1808-1817	10
Common Sense	1737	1766-1775	10
Chronicles of Eri	1762	1802-1811	10
Ern Malley Poems	1917	1825-1884	80
Vortigern and Rowena	1775	1786-1795	10
Book of Mormon	1805	1810-1829	20

Table 11: Period of Cultural Influence vs. Author Birth Year.

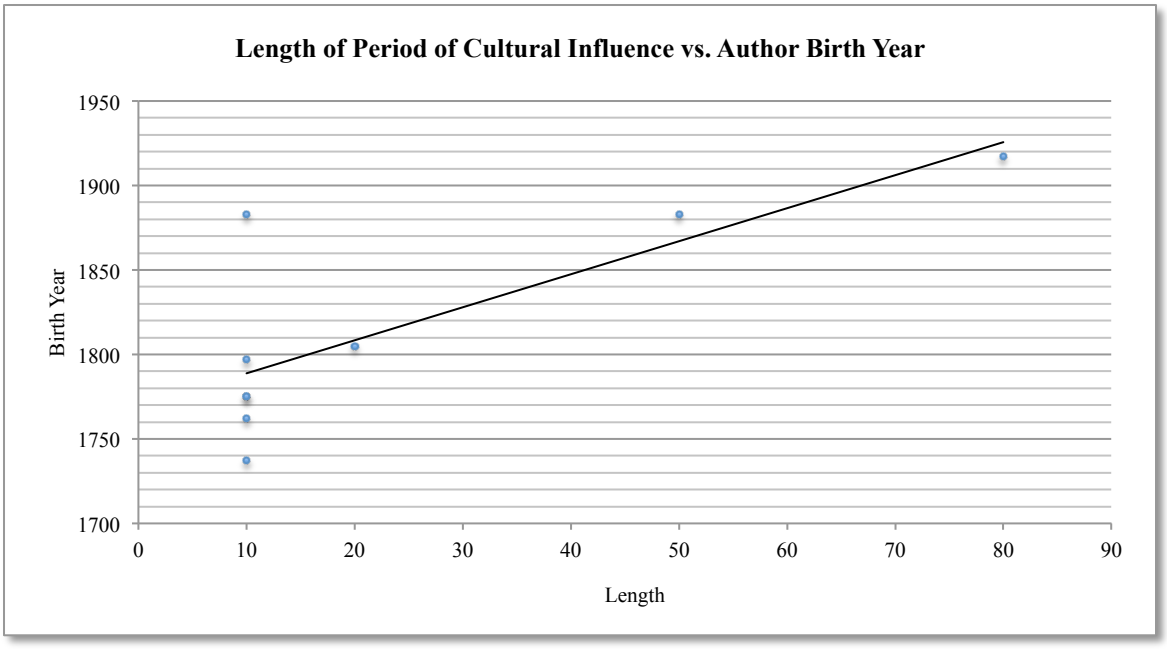


Figure 32: Scatter plot of Period of Cultural Influence Length vs. Author Birth Year.

Testing H_4

Hypothesis H_4 is stated as follows:

H_4: "As the birth year of the author increases, the start date of the Period of Cultural Influence should increase."

Since this hypothesis tests correlation of the population, our test is:

H_0: $\rho = 0$ (no linear correlation)

H_4: $\rho \neq 0$ (linear correlation)

The data used to test this hypothesis are displayed in Table 12, and their scatter plot is shown in Figure 33. The equation of the regression line through the data points is:

$$\hat{y} = 2000.1 + 2.109x$$

This gives a Pearson Correlation Coefficient $\rho = 0.8391$. But we have $0.8391 > 0.7980$, allowing us to reject the null hypothesis in favor of H_4 with a 99% confidence level.

Document	Author Birth Year	Period of Cultural Influence	Period of Cultural Influence Start Year
Pride and Prejudice	1775	1803-1812	1803
The Metamorphosis (English)	1883	1825-1874	1825
The Metamorphosis (German)	1883	1855-1864	1855
Frankenstein	1797	1808-1817	1808
Common Sense	1737	1766-1775	1766
Chronicles of Eri	1762	1802-1811	1802
Ern Malley Poems	1917	1825-1884	1825
Vortigern and Rowena	1775	1786-1795	1786
Book of Mormon	1805	1810-1829	1810

Table 12: Period of Cultural Influence Start Year vs. Author Birth Year.

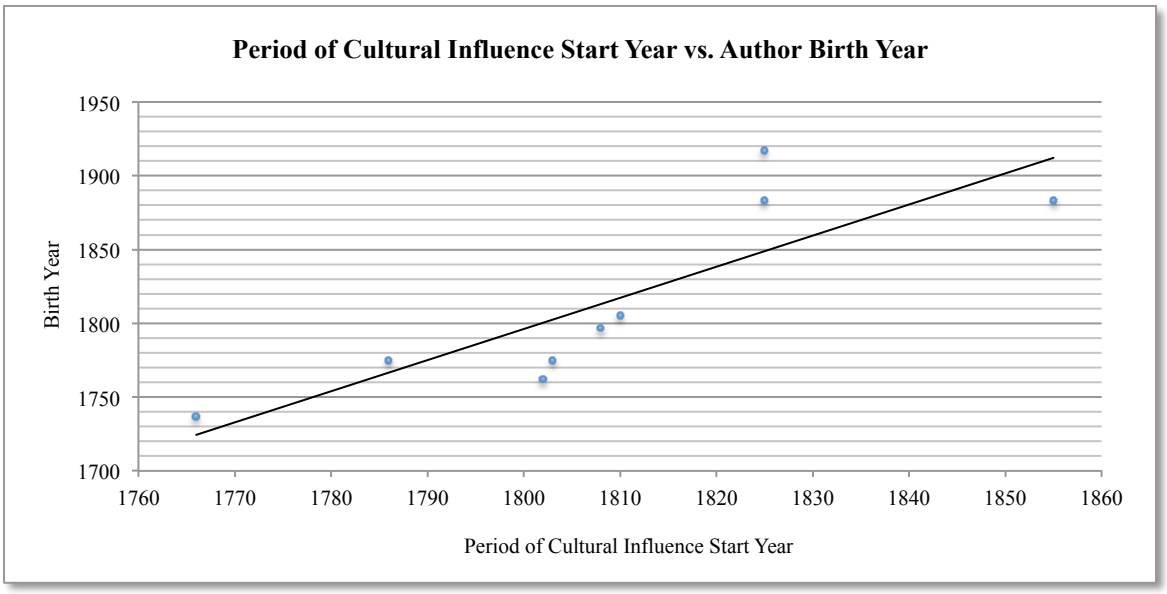


Figure 33: Scatter plot of Period of Cultural Influence Start Year vs. Author Birth Year.

Testing H_5

Hypothesis H_5 is stated as follows:

H_5: "As the age of the author increases, the length of the period from the Period of Cultural Influence Start Year to the Publish Year should increase."

Since this hypothesis tests correlation of the population, our test is:

H_0: $\rho = 0$ (no linear correlation)

H_5: $\rho \neq 0$ (linear correlation)

The data used to test this hypothesis are displayed in Table 13, and their scatter plot is shown in Figure 34. The equation of the regression line through the data points is:

$$\hat{y} = 34.199 - 0.0337x$$

This gives a Pearson Correlation Coefficient $\rho = 0.1153$. Since $0.1153 < 0.666$, we cannot reject the null hypothesis in favor of H_5.

Document	Author Birth Year	Period of Cultural Influence	Period of Cultural Influence Start Year	Publish Date	Age at Publish	Year Range
Pride and Prejudice	1775	1803-1812	1803	1813	38	10
The Metamorphosis (English)	1883	1825-1874	1825	1915	32	90
The Metamorphosis (German)	1883	1855-1864	1855	1915	32	60
Frankenstein	1797	1808-1817	1808	1818	21	10
Common Sense	1737	1766-1775	1766	1776	39	10
Chronicles of Eri	1762	1802-1811	1802	1822	60	20
Ern Malley Poems	1917	1825-1884	1825	1945	28	120
Vortigern and Rowena	1775	1786-1795	1786	1796	21	10
Book of Mormon	1805	1810-1829	1810	1830	25	20

Table 13: Year Range vs. Author Age.

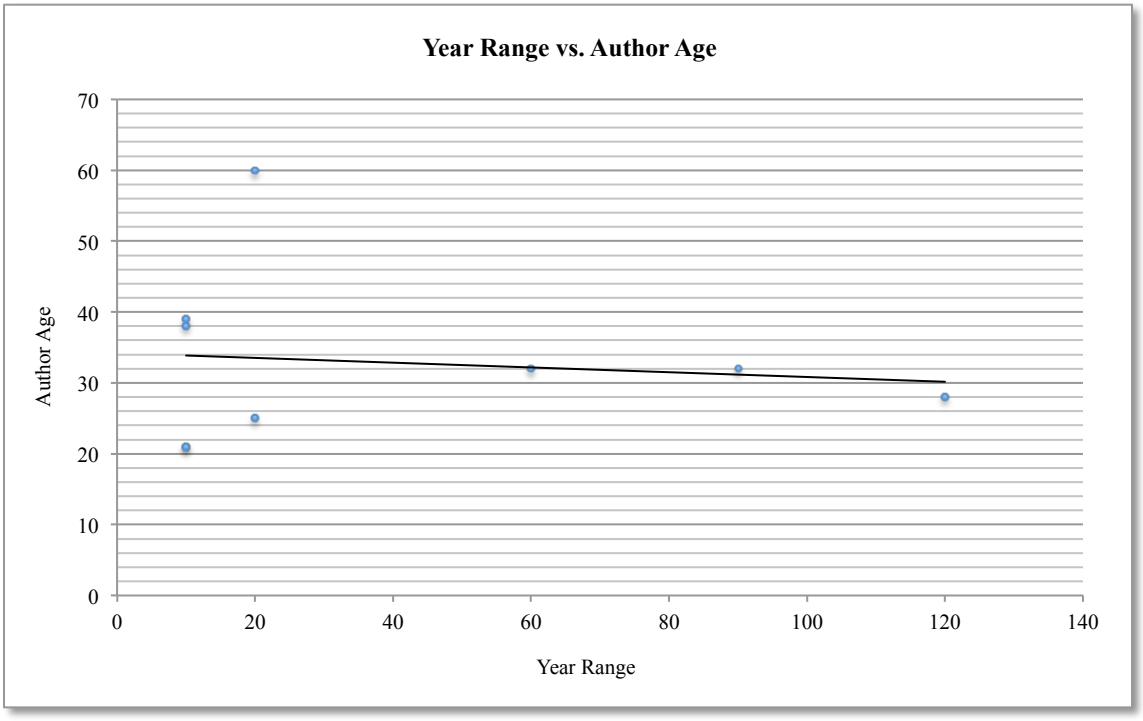


Figure 34: Scatter plot of Year Range vs. Author Age.

Testing H_6

Hypothesis H_6 is stated as follows:

H_6: "As the document file size increases, the variance between the results returned from different *n*-gram classes should decrease."

Since this hypothesis tests correlation of the population, our test is:

H_0: $\rho = 0$ (no linear correlation)

H_6: $\rho \neq 0$ (linear correlation)

In order to calculate the variance, two sets of data were considered. First, to calculate the variance of the time periods, the mean year was calculated for each time period, and then the standard deviation was calculated over these means. For the dataset of the mean years, we simply calculated the standard deviation directly. The data used to test this hypothesis are displayed in Table 14.

The scatter plot for the variance of the mean of the time periods is shown in Figure 35a. The equation of the regression line through the data points is:

$$\hat{y} = 3.5128 - 0.000004x$$

This gives a Pearson Correlation Coefficient $\rho = 0.5126$. Since $0.5126 < 0.666$, we cannot reject the null hypothesis in favor of H_6.

The scatter plot for the variance of the mean years is shown in Figure 35b. The equation of the regression line through the data points is:

$$\hat{y} = 3.7309 - 0.000004x$$

This gives a Pearson Correlation Coefficient $\rho = 0.553$. Since $0.553 < 0.666$, we cannot reject the null hypothesis in favor of H_6.

Document	File Size (bytes)	Variance (mean of time periods)	Variance (mean year)
Pride and Prejudice	682,851	0.0000	0.0000
The Metamorphosis (English)	118,582	5.7735	5.5076
The Metamorphosis (German)	125,989	7.6376	7.2111
Frankenstein	418,592	0.0000	0.0000
Common Sense	108,352	0.0000	0.0000
Chronicles of Eri	733,993	0.0000	0.0000
Ern Malley Poems	15,820	5.0000	6.1101
Vortigern and Rowena	88,154	0.0000	0.5774
Book of Mormon	1,350,985	0.0000	0.0000

Table 14: File Sizes and Result Variances.

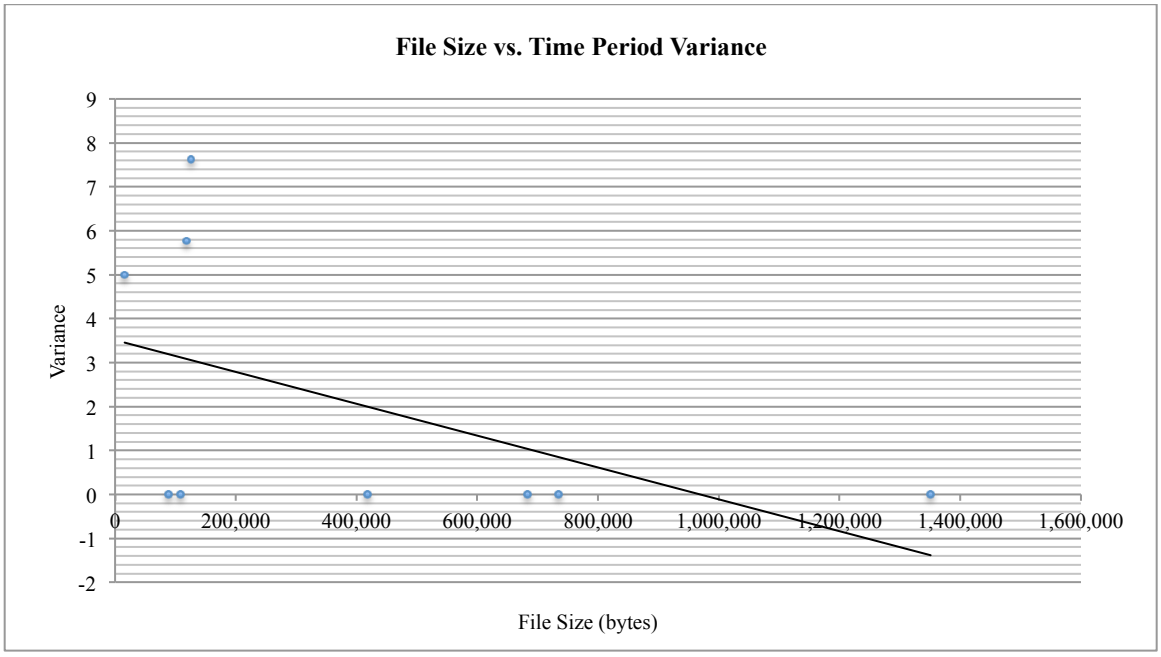


Figure 35a: Scatter plot of File Size vs. Time Period Variance.

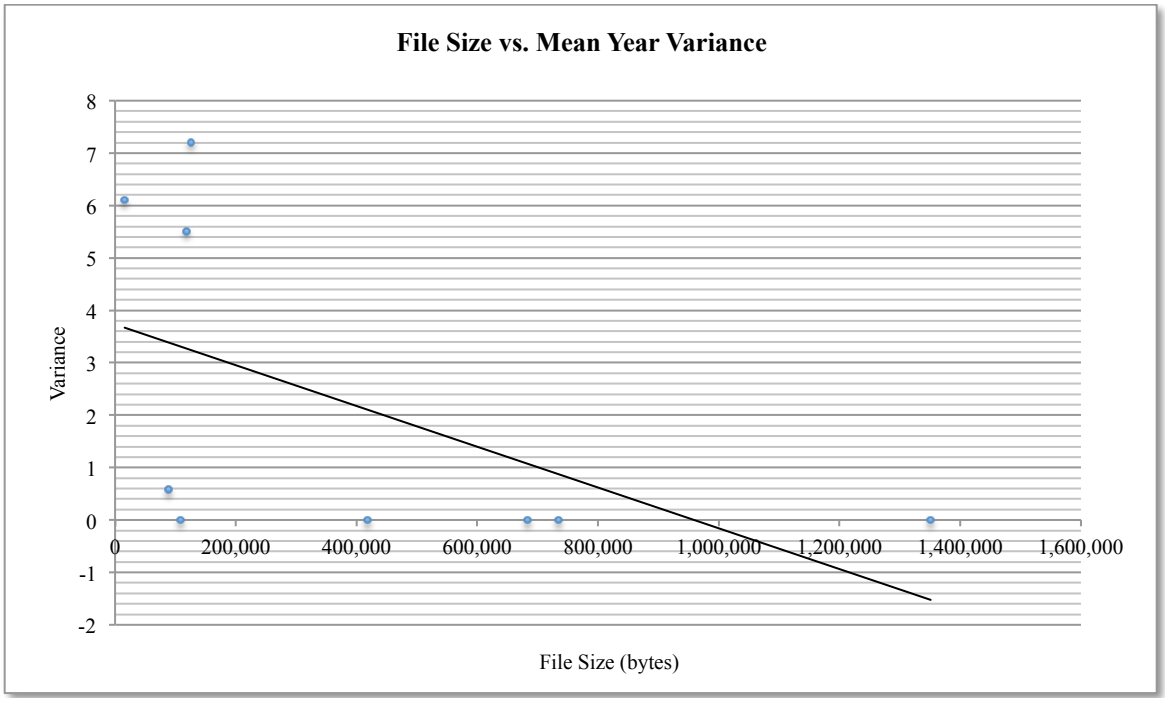


Figure 35b: Scatter plot of File Size vs. Mean Year Variance.

CHAPTER 6: CONCLUSION

6.1 Research Hypotheses

After applying the method to the selected documents and interpreting the results returned, we present our conclusions. The results of testing our research hypotheses are summarized in Table 15.

Hypothesis No.	Description	Results of Tests	Confidence Level
H_1	The period of cultural influence should begin before the beginning of the period of composition.	Accepted	100%
H_2	The period of cultural influence should end either before or during the period of composition.	Accepted	100%
H_3	As the birth year of the author increases, the length of the period of cultural influence should increase.	Accepted	95%
H_4	As the birth year of the author increases, the start date of the period of cultural influence should increase.	Accepted	99%
H_5	As the age of the author increases, the difference between the start of the period of cultural influence and the Publish Year should increase.	Rejected	
H_6	As the document file size increases, the variance between the results returned from different *n*-gram classes should decrease.	Rejected	

Table 15: Results of testing the research hypotheses.

These are discussed separately below.

H_1 – Accepted

H_1 simply imposes the restriction that the period of cultural influence must begin before the period of composition. This is a condition that is imposed by the timeline model proposed in Chapter 1. The method met this restriction in every case.

H_2 – Accepted

H_2 requires that the period of cultural influence must end either before the period of composition or within it. Again, the timeline model imposes this condition, and the method met it in every case.

H_3 – Accepted

H_3 goes beyond the implications of the timeline model and attempts to test for a specific phenomenon. The volume count information collected in the Google database demonstrates that, generally speaking, more volumes are made available with the passing of each year (see Figure 13). This suggests

that authors from more recent time periods have access to a greater historical breadth of literary works than did authors from earlier time periods. This implies that more modern authors may be influenced by the culture of earlier time periods, and that this may broaden their period of cultural influence. Placing this in more concrete terms, H_3 tests to see if there exists a correlation between the birth year of an author, and the length of the period of cultural influence predicted by the method.

Analysis of the data shows that there is indeed a positive correlation between these two values, and this correlation can be asserted with a 95% confidence level. We conclude that the method produces results that exhibit this expected phenomenon.

H_4 – Accepted

H_4 also goes beyond the implications of the timeline model and attempts to test for another specific phenomenon. Normally, one expects the output that an author produces to be representative of the time period in which the author lives. This means that, while the period of cultural influence must start before the period of composition, the gap between them would not be expected to be excessive. So, generally speaking, as the birth year of the author increases, so should the start of the period of cultural influence. Placing this in more concrete terms, H_4 tests to see if there exists a correlation between the birth year of an author, and the start of the period of cultural influence predicted by the method.

Analysis of the data shows that there is indeed a positive correlation between these two values, and this correlation can be asserted with a 99% confidence level. We conclude that the method produces results that exhibit this expected phenomenon.

H_5 – Rejected

H_5 suggests that older authors should exhibit broader periods of cultural influence. The rationale here is that, the longer one lives, the more one is exposed to cultural influences, and these extend over longer lengths of time. Also, one should have more time to study the works of earlier authors and be influenced by them. Placing this in concrete terms, H_5 tests to see if there exists a correlation between the age of an author, and the length of the gap between the start of the period of cultural influence and the publish date.

Analysis of the data shows that there is not enough evidence to conclude that such a correlation exists. We conclude that the method produces results that do not allow us to reach this conclusion. In effect, the results from the method suggest that the length of this time period is not influenced by the age of the author.

H_6 – Rejected

H_6 simply tests to see if the method returns better results for larger documents. It formulates this test by checking for a correlation between the file size of the document and the variance in the results returned from the method. The rationale here is that, the more data one has to analyze, the more accurate the analysis should be.

Interestingly, analysis of the data shows that there is not enough evidence to conclude that such a correlation exists. We conclude that the method produces results that do not allow us to reach this conclusion. In effect, the results from the method suggest that the length of the document is not a factor in determining the accuracy of the results produced by the method.

6.2 General

As mentioned previously, 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 DBMS, 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, along with the results of the research hypotheses, 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 provides 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.

CHAPTER 7: REFERENCES

[1] Robert McKee, *Story: Substance, Structure, Style, and the Principles of Screenwriting* (New York: HarperCollins, 1997), p. 135.

[2] Jack Grieve, "Quantitative Authorship Attribution: An Evaluation of Techniques," Literary and Linguistic Computing, Vol. 22, No. 3, 2007.

[3] Gregory Shalhoub, Robin Simon, Ramesh Iyer, Jayendra Tailor, Dr. Sandra Westcott, "Stylometry System – Use Cases and Feasibility Study," Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 7th, 2010.

[4] Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods," JASIST. 01/2009; 60:538-556.

[5] "About Google Books," available at books.google.com.

[6] Brian Hayes, "Bit Lit," *American Scientist*, Vol. 99, No. 3 (May-June 2011): 190. DOI: 10.1511/2011.90.190.

[7] 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, "Quantitative analysis of culture using millions of digitized books," *Science*, Vol. 331, No. 6014 (January 14, 2011): 176–182. DOI: 10.1126/science.1199644; John Bohannon, "Google Books, Wikipedia, and the Future of Culturomics," *Science*, Vol. 331, No. 6014 (January 14, 2011): 135. DOI: 10.1126/science.331.6014.135; Patricia Cohen, "In 500 Billion Words, New Window on Culture," *New York Times* (December 17, 2010): A3; Eric Hand, "Word Play," *Nature*, Vol. 474 (June 17, 2011): 436-440. DOI: 10.1038/474436a; Erez Lieberman Aiden, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak, "Quantifying the evolutionary dynamics of language," *Nature*, Vol. 449, No. 7163 (October 11, 2007): 713–716. DOI: 10.1038/nature06137.

[8] Thomas Paine, *Common Sense* (New York: Fall River Press, 1995); "Tom Paine: Plain Arguments for Independence," *The Annals of America, Volume 2 (1755-1783): Resistance and Revolution* (Chicago: Encyclopedia Britannica, Inc., 1968); David McCullough, *John Adams* (New York: Simon & Schuster, 2001), pp. 96-97; David McCullough, *1776* (New York: Simon & Schuster, 2005), pp. 250-251; Nicholas Hagger, *The Secret Founding of America* (London: Watkins Publishing, 2007), p. 149; Ron Chernow, *Alexander Hamilton* (New York: The Penguin Press, 2004), p. 70.

[9] Mary Shelley, *Frankenstein; or, the Modern Prometheus* (London: George Routledge and Sons, Limited, 1891), pp. v-xii.

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

[11] Jane Austen, *Pride and Prejudice: A Novel* (T. Egerton, Military Library, Whitehall: London, 1813); Jane Austen, *Pride and Prejudice: An Annotated Edition*, Patricia Meyer Spacks, ed. (Harvard, 2010), pp. 3-4; David Nokes, *Jane Austen: A Life* (University of California Press: Berkeley and Los Angeles, 1997), pp. 2, 51.

[12] Fawn M. Brodie, *No Man Knows My History: The Life of Joseph Smith* (New York: Vintage Books, 1945), p. 50.

[13] Joseph Smith, *History of the Church, Volume I* (Salt Lake City: Deseret Book Company, 1980), pp. 15-18.

[14] Donna Hill, *Joseph Smith: The First Mormon* (Salt Lake City: Signature Books, 1977), pp. 35, 410-416.

[15] Roger O'Connor, *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, 1822), pp. iii-xi; For sources that take the *Chronicles of Eri* seriously, see: F. R. A. Glover, *England the Remnant of Judah and the Israel of Ephraim*, 2nd ed. (Rivingtons: London, 1881), pp. 19, 76-77; E. Raymond Capt, *Jacob's Pillar - A Biblical Historical Study* (Artisan Sales, 1977); http://www.britam.org/traditions17.html; http://www.ensignmessage.com/archives/jeremiahinireland.html.

[16] R. A. S. Macalister, *Irish Historical Studies*, Volume 2, No. 7 (March 1941): p. 335.

[17] W. J. Fitzpatrick, Sidney Lee, ed., *Dictionary of National Biography, Volume 41* (MacMillan and Company: New York, 1895), pp. 407-408.

[18] David Brooks, *The Sons of Clovis: Ern Malley, Adoré Floupette and a Secret History of Australian Poetry* (University of Queensland Press: Brisbane, Australia, 2011).

[19] William Henry Ireland, *Vortigern, an Historical Tragedy, in Five Acts; Represented at the Theatre Royal, Drury Lane* (J. Barker: London, 1799); William Henry Ireland and Richard Grant White, *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, 1874), pp. vii-xxxi, 135; Jeffrey Kahan, *Reforging Shakespeare: The Story of a Theatrical Scandal* (Associated University Presses, Inc.: Cranbury, NJ, 1998), p. 11; Jeffrey Kahan, "Shakespeare and the forging of belief," *Critical Quarterly*, January 24, 2003, Volume 32, No. 2 (July 2001): p. 1. DOI: 10.1111/1467-8705.00352; John Mair, *The Fourth Forger: William Ireland and the Shakespeare Papers* (Ayer Publishing, 1938); Bernard D. N. Grebanier, *The Great Shakespeare Forgery* (W. W. Norton: New York, 1965); S. Schoenbaum, *Shakespeare's Lives* (Oxford University Press, 1991).

[20] Deirdre Le Faye, *Jane Austen: The World of Her Novels* (London: Frances Lincoln Limited, 2002), p. 178.

[21] Franz Kafka, *The Metamorphosis* (New Mexico: CSF Publishing, 2011). Available at books.google.com.

[22] Alexander Jessup, ed., George Burnham Ives, trans., *Little French Masterpieces* (New York and London: The Knickerbocker Press, 1903), ix-xxv.

[23] Joseph Smith, Jr., *Book of Mormon* (Palmyra, New York: E. B. Grandin, 1830), title page.

[24] LeGrand Richards, *A Marvelous Work and a Wonder* (Salt Lake City: Deseret Books, 1976), pp. 72-73.

[25] David Persuitte, *Joseph Smith and the Origins of the Book of Mormon, 2nd ed.* (London: McFarland & Company, Inc., 2000), pp. 268-280.

[26] Richard S. Van Wagoner, *Sidney Rigdon: A Portrait of Religious Excess* (Salt Lake City: Signature Books, 1994), 462-467.

[27] Brigham H. Roberts, "The Origin of the Book of Mormon," *American Historical Magazine*, No. 4 (March 1909): 179-196.

[28] Don Anderson, "Ern, it turns out, has a French cousin," *The Australian*, October 1, 2011. Available at http://www.theaustralian.com.au/arts/ern-it-turns-out-has-a-french-cousin/story-e6frg8n6-1226145808046.

[29] Henry Weinfield, trans., *Collected Poems / Stéphane Mallarmé* (Berkeley and Los Angeles: University of California Press, 1994), p. 179.

[30] William Henry Wilde, Joy W. Hooton, B. G. Andrews, eds., *The Oxford companion to Australian Literature* (Oxford University Press, 1994), p. 257.

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

[32] Philip Mead, *Networked Language: Culture & History in Australian Poetry* (Australian Scholarly Publishing, 2008), p. 89.

[33] Oscar James Campbell and Edward G. Quinn, eds., *The Reader's Encyclopedia of Shakespeare* (New York: Thomas Y. Crowell Company, 1966), pp. 386-387.

APPENDIX A: TRANSIENT *N*-GRAM METHOD

The first analytical method constructed in this study that returned acceptable results was based on the concept of a *transient* n-*gram*. During the time period prior to the date of publication, *n*-grams extracted from a given document have a "lifespan" – the length of time between their first and last appearance in the Google database. Those *n*-grams having very short lifespans are, by definition, restricted to usage within a narrower window of time. For the purposes of this method, these *n*-grams having very short lifespans were referred to as "transient" *n*-grams. Since the goal was to restrict or narrow down the time period of greatest cultural influence, it was believed that an analysis of the time periods associated with these transient *n*-grams could provide an indicator of the time period of greatest cultural influence.

By way of illustration, Figure 36 presents a hypothetical list of 10 *n*-grams extracted from a document published in 1800. Examining the lifespans of all these *n*-grams, one could tentatively conclude that the time period of influence lasted anywhere from around 1710 until about 1780 – a broad window. However, if one considers only transient *n*-grams (in red), we arrive at a narrower range of about 1740 to 1760.

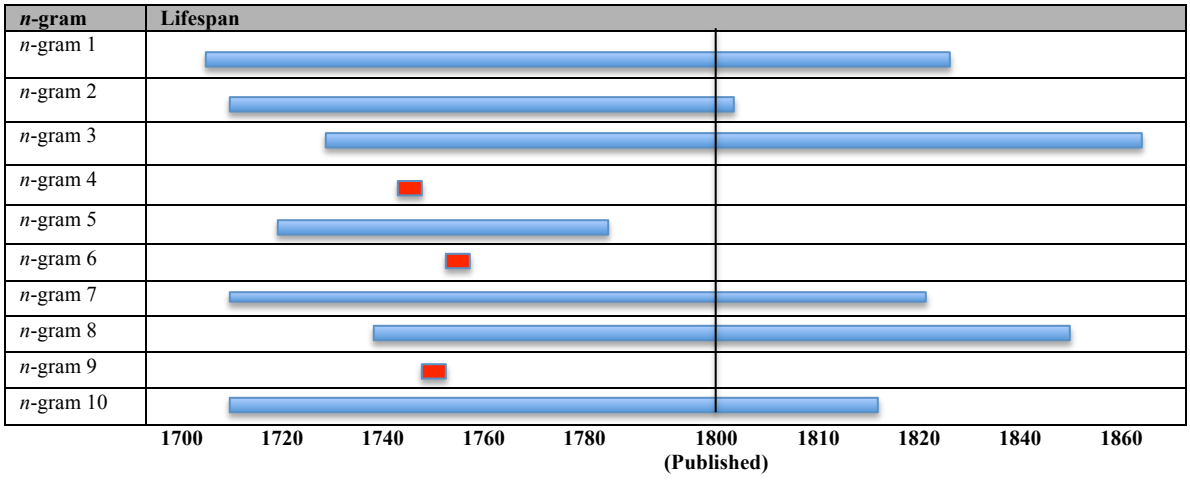


Figure 36: Transient *n*-grams (red) can help to narrow the window of cultural influence.

A transient *n*-gram, then, was defined as an *n*-gram having a very small difference between the year of its initial occurrence in the Google database, and the year of its final appearance. Specifically, this method

constrained the difference to 1 - 2 years. Stated more formally, the set of transient *n*-grams for a specific document is as follows:

> *The set of all n-grams in a document having 0 ≤ Lifespan ≤ 1, where*
>
> *Lifespan = {Year of last occurrence} – {Year of first occurrence}, and*
> *{Year of last occurrence} < {Year published}.*

Or, more succinctly,

> *The set of all* n-*grams in a document having*
> *{Year of last occurrence} ≤ {Year of first occurrence} + 1 ≤ {Year published}.*

This can be demonstrated with some concrete examples. In *Pride and Prejudice*, the *5*-gram "a ball of this kind" makes its first and final appearance (prior to the publish date of 1813) in the Google database in the year 1810. Since this *n*-gram has a "lifespan" of 1 year, it qualifies as a transient *n*-gram. Likewise, the *4*-gram "be mentioned and that" taken from *Common Sense* makes its first appearance in 1759 and its last appearance in 1760 (prior to 1776), so it has a lifespan of just 2 years, and therefore qualifies as transient. Conversely, the *3*-gram "pleasure such as" from *Frankenstein* appears first in 1800 and last in 1817 (prior to 1818) – a lifespan of 18 years, which is too long a lifespan to be considered transient.

The SQL used to apply this method is presented in Figure 37, and the results obtained are presented in Table 16. An obvious drawback of this method is that it only considers *n*-grams of a transient nature (ironically, its defining characteristic). If a document contains a small number of transient *n*-grams, the results will likely be poor. And it is possible that a document (especially a smaller one) may contain no transient *n*-grams at all. For example, the *Ern Malley Poems* contain no transient *5*-grams.

Document	Year Published	Est. Year: 3-grams	Est. Year: 4-grams	Est. Year: 5-grams	Mean Year	(Mean – Published)
Pride and Prejudice	1813	1799	1799	1800	1799	-14
The Metamorphosis (English)	1915	1892	1896	1895	1894	-21
The Metamorphosis (German)	1915	1885	1892	1903	1893	-22
Frankenstein	1818	1804	1804	1804	1804	-14
Common Sense	1776	1740	1747	1752	1746	-30
Chronicles of Eri	1822	1805	1804	1805	1805	-17
Ern Malley Poems	1945	1928	1918	N/A	1923	-22
Vortigern and Rowena	1796	1759	1761	1765	1762	-34
Book of Mormon	1830	1811	1811	1811	1811	-19

Table 16: Overall results obtained using the Transient *n*-grams Method.

```
select d.Book_Id,
b.book_name,
b.year_pub,
b.year_written,
AVG(d.Low_Year) as Low_Year_Mean,
AVG(d.High_Year) as High_Year_Mean
from
(
select m.book_id as Book_Id,
m.n_gram as n_gram,
MIN(m.year) as Low_Year,
MAX(m.year) as High_Year,
MAX(m.year)-MIN(m.year) as Range_Years
from match_details_3_grams m
group by m.book_id,m.n_gram
having MAX(m.year)-MIN(m.year) <= 1
) d, books b
where d.Book_Id=b.book_id
group by d.Book_Id;
```

Figure 37: SQL used to select transient *3*-grams and their mean years.

APPENDIX B: PerformMethod() SQL STORED PROCEDURE

```
DROP PROCEDURE IF EXISTS PerformMethod$$

CREATE PROCEDURE PerformMethod( IN theBookId INT, IN theNGramType INT)
BEGIN

DECLARE theYear INT;

DECLARE n     LONG;
DECLARE meanX FLOAT;
DECLARE sumX  LONG;
DECLARE sumXX LONG;
DECLARE meanY FLOAT;
DECLARE sumY  FLOAT;
DECLARE sumYY FLOAT;
DECLARE sumXY FLOAT;

DECLARE b      FLOAT;
DECLARE a      FLOAT;
DECLARE r      FLOAT;

DECLARE proc1_year    LONG;
DECLARE proc1_density FLOAT;
DECLARE max_density   FLOAT;
DECLARE max_year      LONG;
DECLARE min_year      LONG;
DECLARE search_done   INT;
DECLARE end_of_file   INT;

DECLARE theEndYear    LONG;
DECLARE thePeriod     INT;

DECLARE cur1 CURSOR FOR
   SELECT year,density FROM proc1_temp ORDER BY year DESC;
DECLARE CONTINUE HANDLER FOR NOT FOUND
   SET end_of_file = 1;

DELETE FROM proc2_temp;

CASE theNGramType

WHEN 3 THEN

INSERT INTO proc2_temp(year,match_factor)
   select m.year,sum(m.match_count)/t.volume_count as year_matches
   from match_details_3_grams m,books b,totals_1_grams t
   where m.book_id=b.book_id
   and m.year=t.year
   and m.book_id=theBookId
   and m.year >= 1700
   group by m.book_id,b.book_name,b.year_pub,b.year_written,m.year;

WHEN 4 THEN
```

```
INSERT INTO proc2_temp(year,match_factor)
   select m.year,sum(m.match_count)/t.volume_count as year_matches
   from match_details_4_grams m,books b,totals_1_grams t
   where m.book_id=b.book_id
   and m.year=t.year
   and m.book_id=theBookId
   and m.year >= 1700
  group by m.book_id,b.book_name,b.year_pub,b.year_written,m.year;

WHEN 5 THEN

INSERT INTO proc2_temp(year,match_factor)
   select m.year,sum(m.match_count)/t.volume_count as year_matches
   from match_details_5_grams m,books b,totals_1_grams t
   where m.book_id=b.book_id
   and m.year=t.year
   and m.book_id=theBookId
   and m.year >= 1700
   group by m.book_id,b.book_name,b.year_pub,b.year_written,m.year;

END CASE;

SELECT
@n := COUNT(match_factor),
@meanX := AVG(year),
@sumX := SUM(year),
@sumXX := SUM(year*year),
@meanY := AVG(match_factor),
@sumY := SUM(match_factor),
@sumYY := SUM(match_factor*match_factor),
@sumXY := SUM(year*match_factor),
@theStartYear := MIN(year),
@theEndYear := MAX(year)
FROM proc2_temp;

SET @b := (@n*@sumXY - @sumX*@sumY) / (@n*@sumXX - @sumX*@sumX);

SET @a := (@meanY - @b*@meanX);

SET @r := (@n*@sumXY - @sumX*@sumY) / SQRT((@n*@sumXX - @sumX*@sumX) *
(@n*@sumYY - @sumY*@sumY));

SELECT CONCAT('Y = ',@b,'X + ',@a,'; r = ',@r) AS 'least-squares regression with correlation
coefficient';

#----------------------------
# Calculate the starting year
# ----------------------------

SET @theStartYear := @theEndYear - FLOOR((@theEndYear - 1700)/10)*10 + 1;

SET theYear = @theStartYear;

DELETE FROM proc1_temp;

REPEAT
```

```
   IF (@theEndYear-theYear < 9) THEN

      SET @thePeriod := @theEndYear-theYear+1;

   ELSE

      SET @thePeriod := 10;

   END IF;

   INSERT INTO proc1_temp
   select theYear,(count(*)/@thePeriod)*100
   from
   (
   select year,match_factor
   from proc2_temp
   ) d1
   where theYear <= d1.year and d1.year < (theYear+10)
   and (@b*d1.year+@a) <= d1.match_factor;

   SET theYear = theYear + 10;

UNTIL theYear > @theEndYear
END REPEAT;

SELECT * FROM proc1_temp;

#------------------------------------------------
# Find start/end dates of most likely time period
# and insert into proc3_temp
#------------------------------------------------

SET max_year    = 0;
SET min_year    = 9999;
SET max_density = 0;

SET search_done = 0;

DELETE FROM proc3_temp;

OPEN cur1;

FETCH cur1 into proc1_year,proc1_density;

REPEAT

   IF proc1_density > max_density THEN

      SET max_density = proc1_density;
      SET max_year    = proc1_year;
      SET min_year    = proc1_year;
      SET search_done = 0;

   ELSE
```

```
      IF proc1_density = max_density AND search_done = 0 THEN

         SET min_year = proc1_year;

      ELSE

         IF proc1_density < max_density THEN

            SET search_done = 1;

         END IF;

      END IF;

    END IF;

FETCH cur1 into proc1_year,proc1_density;
UNTIL end_of_file = 1
END REPEAT;

SET theYear = min_year;

REPEAT

   INSERT INTO proc3_temp(year) values (theYear);

   SET theYear = theYear + 10;

UNTIL theYear > max_year
END REPEAT;

SELECT CONCAT('min_year = ',min_year,'; max_year= ',(min_year+10 - 1) ) AS 'Time Period';

SELECT * FROM proc3_temp order by year;

#-------------------------------------------
# Calculate weighted-Average Year
# over proc3_temp
#-------------------------------------------

SELECT SUM(p2.year*p2.match_factor)/SUM(p2.match_factor) AS 'Mean Year' from proc2_temp p2
 WHERE year >= (SELECT MIN(p3.year) from proc3_temp p3)
 AND   year <= (SELECT MAX(p3.year)+10 from proc3_temp p3);

END$$
```