2007-12-11

# A Corpus-based Approach to Determining Standard American English

Delys Ann Waite Snyder
*Brigham Young University - Provo*

Follow this and additional works at: http://scholarsarchive.byu.edu/etd

Part of the Educational Psychology Commons

A CORPUS-BASED APPROACH TO DETERMINING

STANDARD WRITTEN AMERICAN USAGE

by

Delys Waite Snyder

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Instructional Psychology and Technology

Brigham Young University

December 2007

BRIGHAM YOUNG UNIVERSITY


GRADUATE COMMITTEE APPROVAL


of a dissertation submitted by

Delys Waite Snyder


This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.


_____         _____
Date                                    David D. Williams, Chair


_____         _____
Date                                    Mark E. Davies


_____         _____
Date                                    Deborah Dean


_____         _____
Date                                    Kristine Hansen


_____         _____
Date                                    Richard R. Sudweeks

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Delys Waite Snyder in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____     _____
Date                                                     David D. Williams
                                                              Chair, Graduate Committee


Accepted for the Department

_____     _____
Date                                                     Andy Gibbons
                                                              Department Chair, Instructional
                                                                 Psychology and Technology


Accepted for the College

_____     _____
Date                                                     K. Richard Young
                                                              Dean, David O. McKay School of
                                                                 Education

ABSTRACT


A CORPUS-BASED APPROACH TO DETERMINING

STANDARD WRITTEN AMERICAN USAGE

Delys Waite Snyder

Department of Instructional Psychology and Technology

Doctor of Philosophy

Many teachers, test designers, textbook writers, and instructional designers turn to books written by usage experts to determine what is correct standard written American English. Unfortunately, though, experts often disagree about what is correct and what is incorrect, and this disagreement can create problems with validity when people create and assess instruction about usage. One way to discover the rules of standard English usage is to describe what writers actually do in printed, edited English. Researchers can access large collections of standard English through digital text archives, which can be searched electronically. The text archives for this study were taken from EBSCO and ProQuest digital libraries and divided into three different registers: (a) newspapers, (b) magazines, and (c) scholarly journals.

This study examines 30 representative items of controversial usage; such as *a lot* or *alot, between you and I* or *between you and me, had proved* or *had proven*; to determine the actual occurrence in these three registers of standard written American English.  The results list the percentage of use in each register, as well as the total averaged percentage of use in all three registers.  Items showing 90% to 100% usage in the total averaged percentages are considered standard English, but items showing 90% to 95% usage are borderline cases that should be monitored for future use.  If a variant form is used more than 10% of the time, then it should be considered a possible alternative usage in dictionaries, in text books, and in tests.  This study shows the results of using corpus linguistics to answer questions about usage in standard American English.

# ACKNOWLEDGEMENTS

Table of Contents

List of Tables

List of Tables (continued)

List of Figures

List of Figures (continued)

Chapter 1: Introduction

Most people think that there are rules for using standard American English, and they think the experts agree on what these rules are.  In reality, there is much disagreement about what is correct and what is not correct.  Furthermore, language changes, and these changes affect what people consider correct and incorrect.  This dissertation explores the problem of developing instruction and testing in a field where experts agree about what is correct in some areas, but disagree about what is correct in other areas.

No one authority decides the rules of language.  People actually negotiate the rules with each other, and the rules change with the people and circumstances.   If the language is always changing, and people disagree about what is correct and what is not, then instructional designers, English teachers, test makers, curriculum evaluators, and even students will have a difficult task in deciding what to teach, test, and study. This dissertation presents one way to empirically find consensus in an instructional field that currently relies on conflicting expert opinion to determine what is correct and what is not. The study goes right to the source of English language arts instruction:  the English language itself.  This project shows a way to begin to describe some aspects of language usage so that future instruction and testing will draw from an accurate account of the language.  This dissertation, therefore, addresses the problems of instructing and testing students in a field that is always in flux and a field that has areas of expert disagreement.

*Standard Written American English*

Educators in the United States expect their students to learn to write standard American English in school. Language curriculum from kindergarten to college focuses on teaching this standard written American English, and many tests try to assess students' knowledge of standard written English. Indeed, most comprehensive tests, from the California Standards Tests for kindergarten through 11th grade to the GMAT entrance exam for business graduate school, test students' knowledge of standard written English.

Test makers, instructional designers, and teachers, including many English teachers, test and grade as if there is absolute agreement about usage, mechanics, grammar, and spelling; and they consider the ways they themselves use the language are obviously the rules for everyone. Instructional designers create texts and courses, test makers create tests, and teachers grade papers with great confidence about the rules of the language; but a closer look shows areas of standard written English where there is actually a lot of disagreement among experts about what the rules really are. Instructional designers cannot create instruction that is helpful, and measurement experts cannot create valid tests if experts themselves disagree about what is correct and incorrect.  If there is no correct answer in actual usage, how can there be a correct answer on a high stakes exam? How can anyone evaluate the effectiveness of instructional materials in the language arts if the experts do not agree on what the rules are?

How do we currently know what constitutes standard written American English? We have no official language authority that decides what we should do, but we do have dozens of books that offer up expert opinion. *The Chicago Manual of Style* (2003)*, Garner's Modern American Usage* (Garner, 2003)*,* Evans and Evans (1981), Bernstein (1977), Fowler (Burchfield 2000), and many handbooks all present their own versions of correctness. William Safire (1984), Edwin Newman (1980), and others write newspaper columns pointing out the foolishness of many writers who do not follow the rules as these pundits see them. These experts, though, are often only prescribing what they think people should do, and they do not necessarily describe what educated writers actually do. They draw from their experience to determine what is right and wrong, and for examples they draw from their file collections.

In many cases, though, usage experts agree on correct usage. For example, most educated English speakers will label as *nonstandard* or *wrong* these sentences: (a) "He done it last week," (b) "They was lookin' out the window," and (c) "It don't matter none." Indeed, all usage experts consider these subject-verb combinations as colloquial, spoken English, but not standard written English. In other areas of usage, though, there is actual disagreement and more variation. Many test writers, instructional designers, teachers, students, and the general educated public think there is a right way and a wrong way to spell or use words, but in reality there are sometimes several possibilities in edited, educated written English. Consequently, a college English handbook says one usage is correct, and the GMAT considers a different usage to be standard. High school English teachers teach punctuation and usage one way, and college teachers and editors often expect something else. This

disagreement confuses students because many have the sense that there are rules out there, but these students have no idea what the rules are because they change from class to class.

Is there a way to accurately describe written American English so that what we teach and test in elementary school, junior high, senior high, and college matches what good writers actually do in published and edited written English?  To begin a search for an accurate description, language researchers can look at the place most educators go to find the rules of standard English.  Currently, many teachers rely on writing handbooks to supply the usage rules for students.  There are many handbooks in use, though, and they do not always agree with each other.

Usage books go into much greater detail about words than writing handbooks do.  The people who write usage books draw on (a) logic; (b) historical evidence; (c) the influence of other languages, especially Latin and Greek; (d) usage panel votes; and/or  (e) their own preferences to proclaim how we should use English. The editors of usage books keep extensive files of actual instances of particular usages, and they draw examples from these files to bolster their arguments in favor of particular usages. The literature review section discusses the most prominent usage books and examines the methods these books use to determine usage.  By searching in prefaces and introductions for usage books' methodology, one can uncover how each book claims to have authority to determine what correct standard written American English is.

Writing handbooks and usage books have gained much clout in representing standard English to the general public. This clout probably increased because people

are insecure in their knowledge of the usage rules of standard English (Schuster, 2003). How have rules of correctness proliferated in modern English? Many early rules, such as (a) "Do not use multiple negatives," (b) "Do not split infinitives," and (c) "Do not end a sentence with a preposition," began in the 18th century as scholars tried to upgrade English to be a more respectable language like Latin. Early grammarians, such as the British bishop Robert Lowth (1762) and the American businessman Lindley Murray (1795), published best-selling grammars that heavily influenced educational practices and ideas about language. These writers, as well as many others, proclaimed rules about writing, and they often chastised great writers from the past because these writers did not follow the newly proclaimed rules.

Authority has also come through dictionaries. Samuel Johnson published *A Dictionary of the English Language* (1755) with the intent of describing the language and not necessarily prescribing how the language should be used. His dictionary, though, began a standardization of spelling and definitions of many words in English. In the United States, Noah Webster (1828) wrote the first influential American dictionary, and his work also had the effect of standardizing many words, spellings, and definitions in American English. Modern writing handbooks, usage books, and dictionaries stem from this early tradition. Writing handbooks are created for writing students who need instruction on standard English. Many of these handbooks, usage books, and dictionaries continue the practice of prescribing correct usage for writers, and people continue to regard these books as language authorities on standard English.

Since the 1950s and 1960s, though, some linguists and dictionary makers have concentrated more on describing what language users actually do, as opposed to prescribing what language users should do. *Webster's Third New International Dictionary* (1961) was the first major dictionary to be published in this descriptivist tradition. Although many linguists praised the dictionary's new approach, many people reacted negatively to its publication because they expected the dictionary to be the authority that told people how to speak and write correctly, not just a record of the language (Feris, 1970).

Today, language experts are still split between the prescriptivists, who list and sometimes create the rules of the language and expect people to follow them, and the descriptivists, who want to describe how people actually use the language. Prescriptivists see no value in merely describing a language and not adding value judgments to certain usages. Some prescriptivists argue that the fact that more people use a word in a certain way does not make it correct. Popular, they say, does not necessarily mean correct or good (Safire, 1984). Descriptivists, on the other hand, question the authority of prescriptivists to proclaim rules, and they argue that speakers and writers create and negotiate the rules as they communicate with others (Pinker, 1994). Descriptivists claim the authority rests with the language users (Milroy, 1999).

The English language, because it is a live language with native speakers, is constantly changing, and this variability complicates the issue of capturing standard English. Schuster claims that the *Harbrace Handbook of English* presents 280 usage errors in its 1941 edition, compared with 156 errors in its 1998 edition: "For the most

part the 'errors' of 1941 have become standard English in 2003" (2003, p. 67).

Surprisingly, 91 of the 156 errors in 1998 are new errors not documented in the 1941

edition.  This study shows how people's perceptions of error change as the language

changes.

All of these factors, the entrenchment of rules in language instruction, the

conflict between prescriptivism and descriptivism, the insecurity of writers who want

to know rules from handbooks, and the variability of language, complicate language

education and affect language testing.  In the past there seemed to be no way to

accurately describe standard written American English with any kind of consensus.

Past descriptions of usage were really prescriptive pronouncements about how

language should be.  In order to create a better description of standard English,

researchers should rely more on empirical studies of the language.

Empirical studies of how people use the language were difficult before

computers were developed. With the rise of the computer and quick digital analysis,

though, many linguists have turned to corpus linguistics to research and document

actual language use. Corpus linguistics studies a sample of actual texts to determine

what happens in the entire register.  A register is the way people speak or talk in a

particular genre or situation, such as a newspaper register, a legal register, or a

baseball fan register.

Some people have used corpus linguistics to study language usage to create

curriculum for second language speakers (Aston, Bernardini, & Stewart, 2004;

Burnard & McEnery, 2000).  Others have studied the syntax and structure of native

English.  The 1985 publication of Quirk, Greenbaum, Leech, and Svartvik's

*Comprehensive Grammar of the English Language* and the 1999 publication of Biber, Johansson, Leech, Conrad, and Finegan's *Longman Grammar of Spoken and Written English* show how corpus linguistics can affect the description of the structure of English. The *Comprehensive Grammar* bases its descriptions on the information in a one-million-word corpus of actual text, and the *Longman Grammar* uses a 40-million-word corpus of both spoken and written English. These two books focus on the details of the structure of English, but they only briefly mention items of disputed usage.

Researchers can also use corpus linguistics to discover actual usage in standard written American English. If we can limit a corpus to edited, published American English, then we can research the practices of people who write, edit, and publish standard English. This dissertation seeks to use corpus linguistics techniques to discover actual expert usage, and this discovery could lead to better descriptions for standard written American English. Accurate descriptions can lead to better guidelines to teach students how to write like experts. Test makers can have accurate descriptions of how expert writers actually use the language. As language changes, descriptions can change, and the guidelines for writers, textbooks, handbooks, and exams can then change, too.

For this dissertation I have created a kind of corpus by using text archives of three registers: newspapers, magazines, and scholarly journals. A text archive is an unstructured corpus that is a searchable database of text, and the archive usually has limited searching capacity. I am limited by the technology and corpora available for American English in the summer of 2007, and I write this dissertation knowing that

within a year or two, many more corpora options will be available. Even as the technology and corpora options change, the research questions will be portable to new ways of doing research. Answering questions about the details of usage will only become easier.

*Research Questions*

This dissertation first compiles categories of common writing errors, based partly on a study of the most common student errors by Connors and Lunsford (1988). Then the study examines (a) what kinds of questions can be answered with a quantitative analysis using text archives, (b) which kinds of questions will be able to be studied when a more advanced corpus becomes available, and (c) which kinds of questions can be answered only with a qualitative analysis. This dissertation concentrates only on questions that can be answered with a quantitative analysis of existing text archives. I have chosen sample usage items from each category of common writing errors. These items are disputed usages that are commonly found in usage books. I have researched each item in text archives of American English published between January 2000 and April 2007. Because the text archives are so large, it is possible to look at extremely large samples of newspaper, magazine, and scholarly journal writing.

Beyond the scope of this dissertation lies the comprehensive question "What is standard written American English usage?" This dissertation will begin to answer that question by focusing on a few of the points of debate and disagreement. Future research can continue to fill out our description of standard written American English usage.

The research questions for this dissertation are as follows:

1. How can we use an empirical method to determine what is standard written American English in various linguistic registers?

2. What does empirical research using corpus linguistics tell us about certain controversial usage questions that are representative of the many questions still requiring research?

    a. What is the percentage usage of each item?

    b. Does the usage differ among the registers of newspapers, magazines, and scholarly journals?

    c. What usage or usages should be considered standard written American English?

3. How should this knowledge of standard written American English affect what we teach and test in schools?

Chapter 2: Literature Review

First, this literature review discusses the philosophy behind the research in this dissertation, situating the idea that the speakers of a language, rather than experts, can decide what is correct in the language. The next section explains the underlying philosophies and practices of traditional prescriptivists and linguistic descriptivists and then discusses the rise of standardization. This longstanding debate over philosophies and attitudes about language continues in the 21st century, and the philosophy of a usage book writer will greatly affect the advice and pronouncements in that book. Next, this dissertation reviews all of the usage books from prominent presses and writers, and it lists the primary style books used in the academic and publishing world. Each entry explores how that usage book claims its authority and what method the book uses to declare what is correct.

The next section explains several studies and articles that explore what teachers should instruct students about usage. Several researchers have described errors in writing and the effects of these errors in writing. Although these studies are more error based than usage based, I will draw on one study's list of errors to form categories of study for this dissertation.

Next, this literature review assesses what linguistic corpora are available for study. This section examines many of the existing English language corpora to find one that will work well with this dissertation's research questions.

Finally, the literature review discusses content validity and testing in the language arts. Most testing companies rely on subject experts to determine if the test content is valid or not, yet experts disagree on some points. If test makers use more

empirical studies to determine what published writers actually do, then these test makers can write test items that will match well with actual practice rather than just with expert opinion.

*Research Philosophy*

The main debate in usage is an argument between prescriptivists and descriptivists. Prescriptivists believe that experts should decide correctness of language and prescribe what writers should do. Then writers should obey these rules or be judged uneducated or uninformed. Descriptivists, on the other hand, believe that linguists should describe the way people use language and not attach judgments (especially moral judgments) to those descriptions. Descriptivists let the actual speakers and writers decide what is correct for a particular situation.

This division resembles the differences in social research philosophy. As we ask questions about the world, we need to ask a very basic epistemological question: How do we know what we know? Objectivism posits that there is a truth to be known separate from the entities involved: "Objectivism is the epistemological view that things exist as *meaningful* entities independently of consciousness and experience, that they have truth and meaning residing in them as objects . . . and that careful . . . research can attain that objective truth and meaning" (Crotty, 1998, p. 6). Verified by experience, objective knowledge is not tied to feelings, spirituality, or ethics. Meaning is discovered by what is inherent in the object, which leads to absolutism. Rational thought leads us to truth. Someone who thinks in positivist terms about usage will be comfortable with an expert who knows all the answers, and that expert expects people to follow his (and occasionally her) pronouncements.

Post-positivism, as proposed by Feyerabend (1987), questions the scientific method. Feyerabend situates all inquiry in a culture and questions whether science is value-neutral. He posits that "scientific truths are no less cultural in character, and no less socio-political in origin, than any other of the beliefs we hold" (Crotty, 1998, p. 40). His radical ideas can bring humility to those who favor a positivist epistemology.

Another kind of epistemology besides objectivism is constructivism. Constructivists do not claim that knowledge lies within an object, but that knowledge is constructed in a community. We must pay attention to the culture that shapes minds (Bruner, 1996). Merleau-Ponty (1962) argues that things are indeterminate, and objects do not have meaning until they interact with consciousness. According to Crotty (1998), because each person sees and interprets differently, constructivists do not look for one true meaning:

> What constructionism drives home unambiguously is that there is *no* true or valid interpretation. There are useful interpretations, to be sure, and these stand over against interpretations that appear to serve no useful purpose. . . . "Useful," "liberating," "fulfilling," "rewarding" interpretations, yes. "True" or "valid" interpretations, no. (Crotty, 1998, pp. 47-48)

Vygotsky (1978) and other social constructionists contend that we socially construct meaning. Social reality and meaning exist only as we create them. Social constructionism also shows that what some see as fact is actually relative. Crotty summaries this philosophy:

> What is said to be "the way things are" is really just "the sense we make of them." Once this standpoint is embraced, we will obviously hold our understandings much more lightly and tentatively and far less dogmatically, seeing them as historically and culturally effected interpretations rather than eternal truths of some kind. Historical and cross-cultural comparisons should make us very aware that, at different times and in different places, there have been and are very divergent interpretations of the same phenomena. (Crotty, 1998, p. 64)

Critical theorists, such as Benjamin (1969), argue that the way we construct meaning

can oppress others and continue hegemony, so we need to beware of our

constructions, as well as the ideologies behind them.

It seems, then, that if people are bent on pronouncing the one correct way to

use the language, they are leaning toward a positivist view of the universe. Someone

who sees usage and language as a cultural construction will approach usage

differently. In terms of usage, constructionists would value the ways people currently

use the language. Instead of looking for one self-professed authority to pronounce

correct usage, a constructionist would take a consensus of expert users. This

dissertation fits comfortably in this constructionist tradition.

*Prescriptivism, Descriptivism, and Standardization*

English developed in Great Britain with many different dialects.  During

Middle English times and by the end of the 15th century, the East Midlands dialect

became the standard dialect because residents of London, immigrants to London,

printers, academics at Oxford and Cambridge, members of parliament, and influential

merchants in London all used the East Midlands dialect (Keene, 2000). In the 16th and

17th centuries, this standardized language became the language of the King James

translation of the Bible, and it was the main dialect of Shakespeare and other

Renaissance playwrights.

Roger Lass (2000) defines standardization as cutting down on varieties and

codifying the language:

> It is usually assumed that standardization typically involves at least two major
> operations: elimination of variation, and codification (in dictionaries,

grammars, orthoepic treatises, and other 'authorities') of the trimmed-down and 'authorised' version. (p. 219)

By the end of the 17<sup>th</sup> century, writers such as Swift and Dryden, as well as others, tried to upgrade English by adding to its vocabulary and creating rules for the language. In the 18<sup>th</sup> century, Samuel Johnson created his dictionary (1755), and people such as Robert Lowth (1762) wrote grammars for the study of English. The prestige language in the European world at the time was Latin, and scholars tried to model English after Latin in its vocabulary and grammar. Standard English also became an important issue in the United States. Noah Webster made a step toward standardized spellings with his 1828 dictionary, and Lindley Murray (1795) sold many copies of his American grammar book.

Jim Milroy (2000) lists three characteristics of language standardization: (a) the language becomes more uniform (p. 13), (b) the standard language is "implemented and promoted by the written language" (p. 14), and (c) "standardization inhibits linguistic change and variability" (p. 14). Standardization does not mean, however, that all variation ceases. Milroy explains that even standardized languages have some change and variation:

> Standardisation inhibits linguistic change, but it does not prevent it totally: there is a constant tension between the forces of language maintenance and the acceptance of change. Thus, to borrow a term from Edward Sapir, standardization 'leaks'. In historical interpretation it is necessary to bear in mind this slow acceptance of change into the written language in particular, because even when the written forms are not fully standardized, they are still less variable than speech is. (p. 14)

Standard English is also tied to class because the standard is usually based on the language of the most prestigious speakers. Edgar Schuster (2003) defines modern standard American English in his book *Breaking the Rules*:

> Standard English, then, in America, would be the English of Well-Edited American Prose (WEAP); that is, the English of editors at such institutions as Alfred A. Knopf and *The New York Times*.  Especially when we are referring to the spoken standard, it is the English of our major news broadcasters.  Broadly, it is the dialect of educated speakers, of those who "run things." (p. 54)

In the United States, citizens expect students to learn standard English in the schools so these students will be able to fully participate in influential aspects of American culture.  In the last several decades, though, linguists have cautioned that educators should honor the language students speak at home, even if it is not standard English, and at the same time students should learn to communicate in standard English (Weaver, 2007).

Prescriptivism followed the rise of standard English (Hope, 2000).  Prescriptivism, according to David Crystal in his *Cambridge Encyclopedia of Language* (1987), can be characterized as the following:

> Prescriptivism is the view that one variety of language has an inherently higher value than others, and that this ought to be imposed on the whole of the speech community.  This view is propounded especially in relation to grammar and vocabulary, and frequently with reference to pronunciation. . . . Adherents to this variety are said to speak or write 'correctly'; deviations from it are said to be 'incorrect'. (p. 2)

Prescriptivists follow what they consider to be the rules of the language; these rules are generally based on written English, rather than spoken English.  The general attitude of prescriptivists is that they want to keep the language from changing, and they base their decisions of right and wrong usage on logic and on the patterns of other languages, such as Latin.

Crystal further explains about prescriptivists:

> The authoritarian nature of the approach is best characterized by its reliance on 'rules' of grammar.  Some usages are 'prescribed', to be learnt and

followed accurately; others are 'proscribed', to be avoided. (p. 2)

William Safire ( 1984) writes in his book on English usage *I Stand Corrected* that he himself is the authority who decides what is right and what is wrong. He looks around him to see what other language experts say, but he relies on his own good taste to decide correctness. Many prescriptivists equate correctness of language with morality, and if the language changes, then morals must be changing too.

Descriptivists, on the other hand, are more interested in describing what people are doing with language than with pronouncing what is correct. Most linguists see their role with language as descriptivists. Crystal discusses descriptivism:

> There is an alternative point of view that is concerned less with 'standards' than with the *facts* of linguistic usage. This approach is summarized in the statement that it is the task of the grammarian to *describe*, not *prescribe*—to record the facts of linguistic diversity, and not to attempt the impossible tasks of evaluating language variation or halting language change. . . . Linguistic issues, it is argued, cannot be solved by logic and legislation. (p. 2)

This dissertation is based on the philosophy that speakers and writers of a language determine what is correct for a particular register by their language practices. Researchers of language can infer what the language rules are by studying the actual practice of language users. This philosophy borrows from descriptivism by looking empirically at actual usage in actual texts. This dissertation also borrows the idea from prescriptivism that we can determine whether or not a particular usage is part of standard English.

Teachers, textbook writers, and instructional designers need a good description of standard English usage so they will know what to teach and what to test. By describing the language according to actual usage, we can articulate that actual usage for others. All live languages change, so when standard American

English changes, we can document it empirically, then we can change what we teach and what we test. In this way educators and test makers can have a good match between what currently exists as standard English and what we currently test in standard English.

*Usage*

What are the rules of usage and who decides them? Until now, we have had to rely on experts who use their intuition and experience. Bookstores and libraries have shelves of books about usage rules written by language experts. These experts take on most issues of spelling, usage, and punctuation, and they proclaim and prescribe what is correct.

This section will examine the organization, philosophy, and method of some usage books, dictionaries, and handbooks. Their methods and philosophies are revealed mostly in the prefaces and introductions. The main questions for these books are "How do the authors find answers to usage questions," "What is the source of their authority," and "Why should we pay attention to these people's proclamations?"

*Usage Books by Publishing Houses*

The first group of usage books has been published by the same presses that produce the best-known dictionaries: (a) Webster, (b) American Heritage, (c) Harper, (d) Penguin, (e) Longman, (f) Oxford, and (g) Cambridge. Of these, Webster, American Heritage, Harper, and Penguin are strictly American. Longman, Oxford, and Cambridge are highly respected British publications that include American English. Since each of these books carries the publisher's name in the title, these works carry the authority of a publishing house.

*Webster's Dictionary of English Usage.*  The editors at Merriam-Webster brought out in 1989 (reprinted in 1994) a revolutionary dictionary of usage. Instead of presenting one individual's view of correctness, the editors looked at the history of usage problems over the years. The dictionary compiles what different usage experts say about each topic, and then it shows what the current usage is according to its files of examples. The editors' methods, then, are to survey the history of what usage experts have said, and then rely on their files for a collection of actual usage.

The more extensive articles are organized as follows:  "origin and development of the usage with examples, origin and development of criticism of the usage, the contemporary status of the usage with examples, review of alternatives, summary and recommendation" (p. 5a). In the "Preface," Gilman, the editor, states,

> *Webster's Dictionary of English Usage* examines and evaluates common problems of confused or disputed English usage from two perspectives: that of historical background, especially as shown in the great historical dictionaries, and that of present-day usage, chiefly as shown by evidence in the Merriam-Webster files. Most of the topics treated have been selected from existing books on usage, primarily those published in the second half of the 20th century. (p. 4a)

This method of summarizing the positions in the debate and relying on extensive files creates a reliable dictionary of usage. A reader can expect Webster to give a consensus opinion.

*The American Heritage Book of English Usage.*  The *American Heritage Dictionary* (2000) and *The American Heritage Book of English Usage* (1996) are both built on the idea of a panel of experts deciding upon correct usage. This usage book claims to help people make decisions about questionable usage. The editors list a number of questions that people may have about usage:

Has this usage been criticized for some reason in the past? If so, are these criticisms substantial? What are the linguistic and social issues involved? Have people frequently applied this usage in the past, and for how long? What do well-respected writers think of the usage today? (p. ix)

To help answer these questions, the publishers formed the American Heritage Usage Panel, "a group of successful people whose work involved writing or speaking effectively" (p. ix). The panel for this usage book consists of 158 people, and the surveys for this edition started in 1987. The publishers have collected many examples of usage in their files. They send out ballots with controversial instances from their files to usage panel members. The editors use the word in question in several different environments, and the panel members write their preferences. The panel's decision, therefore, consists of 158 different opinions. The editors tabulate the responses and include the consensus decisions in the usage book. The publishers ask the panelists if they consider the word or usage to violate "some notion of propriety that they consider inherent to formal Standard English" (p. x). At times they ask the panelists if the usage would fit into informal English. They admit that questions of usage are questions of social acceptance, not just of grammar rules: "Acceptability is thus not really a matter of grammaticality but rather a broader notion of appropriateness. It can entail a sense of aesthetics, . . . a concern about pretentiousness," and "issues of social justice" (p. x). The editors sometimes recommend usage that has a long history of use by fine writers, even though the panel rejects the use today.

*The American Heritage Book of English Usage* divides English into several levels of discourse: (a) standard English, the language of public discourse in news, education, law, and government; (b) nonstandard English, expressions of people from "less prestigious social groups" (p. xi); (c) formal English, the language required by

serious public discourse, giving "full treatment to all the elements that are required for grammatical sentences" (p. xi); and (d) informal English, language that is more casual, used when it is not necessary "or even desirable, to use the conventions of formal discourse" (p. xii). The differences between formal and informal are not a matter of correctness because each has its own rules for correctness.

The book is organized differently from other usage books: The editors group by topic rather than by word. The chapter categories are as follows: (a) grammar, (b) style, (c) word choice, (d) science terms, (e) gender, (f) names and labels, (g) pronunciation challenges, (h) word formation, and (i) e-mail. Even with all of these categories, though, the largest by far is the word choice section, which looks much like the other usage books. Entries show the panel's response to each question. Each entry shows what percentage of the panel considers the usage unacceptable and what percentage considers it informal or acceptable.

The formation of the usage panel in the 1960s was in direct response to the publication of *Webster's Third New International Dictionary*. *Webster's Third* was shockingly descriptive, so the *American Heritage Dictionary* was formed in response to this perceived looseness in dictionary making. The *American Heritage Dictionary* editors formed a usage panel in the 1960s to restore prescriptivism to dictionary making. The panel was notoriously conservative in the 1960s and has become less conservative in the 1980s (see the "hopefully" entry in the *American Heritage Book of English Usage* book for further explanation). One valuable aspect of this *English Usage* book is the use of percentages to show divided opinions about usage. This

practice acknowledges the complexity of usage decisions and shows that even great writers and editors do not agree on many items.

*The American Heritage Guide to Contemporary Usage and Style.* The latest usage book from American Heritage, *The American Heritage Guide to Contemporary Usage and Style* (2005), has changed format. It has returned to the typical dictionary style, with each word listed alphabetically. The usage panel now has 200 people. The American Heritage company still sends out ballots showing actual usage of a word in several settings, and panel members vote on each one of these usages. This method tries to capture the nuances of usage for each word: In some settings it is not acceptable, and in some settings it is acceptable. When a word is deemed acceptable, the term means "the usage does not violate the propriety that the Panelists consider inherent to formal Standard English" (p. xiii). Showing that usage can change, they state, "When an overwhelming percentage of the Panel accepts a usage, this indicates that it has become standard and that it is likely to remain so" (pp. xiii-xvi). The panelists are mostly writers, editors, and professors of English or linguistics, but there are also scientists and government officials. These are the experts in language, and the panelists supposedly give us insight into the minds of some of the best writers around.

*Harper Dictionary of Contemporary Usage.* Morris and Morris had a newspaper column for many years, and this *Harper Dictionary of Contemporary Usage* (1985) is unique in that it contains answers to many questions brought up by readers. *Harper Dictionary* also uses a panel of 165 writers, editors, and public speakers. This panel ensures that *Harper* will receive advice from some of the best

language users in the country. This dictionary tries to call attention to inaccuracies in the language so readers can correct them. The editors also try to show "the standards of linguistic usage adhered to by those who use the language well" (p. xix). They point out that the usage panel members have divergent opinions, and on few items do they agree entirely. Morris and Morris note that the days when one person could dictate usage to all are gone, and we must get a wider accounting. The dictionary is organized alphabetically, and the entries that are more controversial list the comments from members of the usage panel.

*The Penguin Dictionary of American English Usage and Style.* This usage book, *The Penguin Dictionary of American English Usage* (Lovinger, 2000), focuses on American English. Penguin publishes many specialized dictionaries, but not a comprehensive desk dictionary; nevertheless, it publishes a usage dictionary. Lovinger draws on examples from "the popular press, broadcasting, books, and a variety of other sources, mostly in the latter eighties and the nineties" (p. vii). He sets out to "'illuminate many traps and pitfalls in English usage'" (p. vii), according to Lovinger's editor. Lovinger wants to correct and advise and not let all usage be correct usage. He wants no justification for bad English, such as the majority decides usage, or English is a living, changing language. Lovinger criticizes linguists who describe language and do not condemn bad language. He acknowledges that the differences in usage may just be personal choice and not a choice between good and bad, but he offers to give advice so we can make sound choices.

He takes as his sources mostly newspapers and magazines, drawing primarily from *The New York Times* for errors. He also uses 120 books and newspapers in the

San Francisco area. Almost half of his 2,000 errors come from newspapers and magazines. He seems to have collected errors wherever he could find them, and then he uses those examples to school the reader in correct English. He sees words that have a newer meaning which makes the word ambiguous as a "wounded word." He decries the "watering-down of distinctive words that we already have, the creation of ambiguity and fuzziness, the breakdown of grace and grammar, and irrational verbal fads" (p. x). His usage book is presented alphabetically with general topics dispersed among the word entries.

*Longman Guide to English Usage.* Greenbaum and Whitcut, authors of *Longman Guide to English Usage* (1988), write in their usage book "clear recommendations in plain English to those who look for guidance on specific points of pronunciation, spelling, punctuation, vocabulary, grammar, and style" (p. xiii). They base their recommendations on British usage, with some explanations of American usage. Greenbaum also co-wrote *A Comprehensive Grammar of the English Language,* a book which has highly influenced the usage guide. Whitcut was an assistant editor for the *Oxford Dictionary and Usage Guide*, another book that also influenced the Longman usage guide. They also acknowledge the influence of the *Survey of English Usage* at University College in London. Like many of the other usage books, *Longman Guide* is (a) published by a dictionary publishing house; (b) is British based; and (c) is organized alphabetically, word by word.

*The Oxford Dictionary and Usage Guide to the English Language. The Oxford Dictionary and Usage Guide to the English Language* (1995) is comprised of a usage guide and a dictionary. The usage guide contains sections on (a) word

formation, (b) pronunciation, (c) vocabulary, and (d) grammar. The prefaces do not give a clue about how the editors chose certain usages over others. The usage guide's aim is "to give guidance in as clear, concise, and systematic a manner as possible on matters of pronunciation, spelling, meaning, and grammar about which there is controversy or uncertainty" ("Introduction"). The guide gives examples mostly from 20th-century writers, and informal or substandard usage draws on examples from speeches in novels. The editors make up many examples, and they draw a lot on novelists.

Interestingly, the only usage books and dictionaries the editors quote are from (a) Oxford Press: *Hart's Rules for Compositors and Readers* (1983), (b) Fowler's *A Dictionary of Modern English Usage* (1965), (c) *The Oxford English Dictionary* (1989), and (c) *The Oxford Writers' Dictionary* (1990). If many variations of spelling exist, the usage guide chooses the house style of Oxford University Press. The Oxford guide gives recommendations, not always distinguishing between right and wrong, but rather between formal and informal use. Of course, it has a British bias (library is pronounced in only two syllables), but it does show some American differences.

*The Cambridge Guide to English Usage. The Cambridge Guide to English Usage* (Peters, 2004) tries to reach a global audience by considering not just British and American usage, but also Canadian, Australian, and New Zealand English. This usage guide is the first to use large corpora for its sources. It uses the 100 million words from the British National Corpus, and 140 million words of American English from the Cambridge International Corpus. The editors make the case for using large corpora:

> Corpus data allow us to look more neutrally at the distributions of words and constructions, to view the range of styles across which they operate. On this basis we can see what is really 'standard,' i.e. usable in many kinds of discourse, as opposed to the formal or informal (p. vii).

These corpora include speech as well as writing.

This usage book also includes information from population surveys sent all over the world. These surveys ask questions about "doubtful or disputed usage in spelling, punctuation, the use of capital letters and certain points of grammar" (p. viii). The editors also rely on the two great unabridged dictionaries: (a) *Oxford English Dictionary* (2nd ed., 1989), and (b) *Webster's Third New International Dictionary* (3rd ed., 1961). This usage book also draws on editorial style guides, such as (a) *Chicago Manual of Style* (2003) and (b) *Oxford Guide to Style* (2002). The *Cambridge Guide* draws on corpus-based grammars, such as *Longman Grammar of Spoken and Written English* (1999). Although many of the other usage books stay far away from linguists, this guide embraces linguistic research. It comes as no surprise that *The Cambridge Guide* is British based because the British have a longer history with corpus linguistics than the Americans. The usage guide is organized alphabetically like a dictionary.

*Usage Books by Individuals*

The next group of usage books contains works written by individuals. The titles do not include the name of the publishing house. These works are included because they are influential and are the books most likely to be found on a writer's reference shelf. Some of the works are rather old, but they are classics and are highly regarded in editing circles. These are the books found on most "recommended" lists informally written by teachers and editors.

*Garner's Modern American Usage.* Garner has a background in legal writing, so his books and articles contain straightforward pronouncements. *Garner's Modern American Usage* (2003) is published by Oxford University Press, but, unlike the books from the other publishing houses, this book bears Garner's name.  In his "Preface," Garner sets forth his principles for deciding the best usage. One principle is about "word-judging" (p. xii):

> A word or phrase is somewhat undesirable if it has any one of the following characteristics, and is worse if it has two or more:
> (a)      it sounds newfangled;
> (b)      it defies logic;
> (c)      it threatens to displace an established expression (but hasn't yet done so);
> (d)      it originated in a misunderstanding of a word or its etymology;
> (e)      it blurs a useful distinction. (p. xii)

One of Garner's principles is about actual usage: "In the end, the actual usage of educated speakers and writers is the overarching criterion for correctness. But while actual usage can trump the other factors, it isn't the only consideration" (p. xii). Garner is a conservative prescriptivist, and he knows that editors and writers need direction in how to use the language. He disparages modern linguists who think that actual usage—with no nod to educated English—is the only kind of usage worth considering. He considers actual educated usage to be very important, tempered by the principles mentioned earlier, as well as some others. He says that a good usage dictionary will help writers make decisions about their language.

Garner reveals his methods for his decisions about usage. First, he gathers instances from his own reading; second, he has friends from the H. W. Fowler Society send in clippings. He supplements his articles with a search of the online databases Nexis and Westlaw. Nexis collects news, law, and business articles.

Westlaw, not surprisingly, contains articles about the law. Also, Garner has written

*The Elements of Legal Style* and *The Dictionary of Modern Legal Usage.* These books

reveal a gigantic bias in Garner's writings: He bases his decisions about usage on the

practices of journalists and legal writers. Although journalists and legal writers have

much to offer discussions about usage, we can also learn much through writers of

academic journals in many fields. We can also look at magazines, not just

newspapers. We can examine writing in the humanities, social sciences, and the

sciences, not just business.

*The New Fowler's Modern English Usage.* Fowler published his first edition

of *Modern English Usage* in 1926 after working on it for years. According to the

"Preface to the Third Edition" of *The New Fowler's Modern English Usage,* revised

by Burchfield (2000, p. vii), Fowler and his brother lived on an isolated part of an

isolated island (Guernsey) while they both used the Oxford English Dictionary, as

well as newspapers and classic literature for their examples. Later, Fowler lived in

Somerset, but still in a village. Fowler never knew anything about American English,

so he completely disregarded it. Fowler's *Modern English Usage* has remained

popular throughout the 20th century. In 1965 Ernest Gowers created a second edition,

and now Burchfield has created a revised third edition.

Burchfield made an electronic database to collect items for his usage book. He

admits that his database is small compared to other databases, but he says he carefully

chose each item, and his book "contains no garbage" (p. x). Admitting that a usage

book reflects its sources, he lists his sources as (a) British and American newspapers,

(b) fiction from the 1980s and 1990s, and (c) periodicals. He also drew on a few non-

British, non-American English sources. Because his bias is about documenting

historical change, he admits he prefers historical usages rather than descriptive

pronouncements. He has little good to say about modern linguists. He draws on many

different usage books, mostly British, and some American. Burchfield does not seem

to distinguish between registers, such as formal and informal, but he is careful to

distinguish what is particularly American.

*American Usage and Style: The Consensus.* In *American Usage and Style:*

*The Consensus* (1980), Copperud tries to take a consensus of many different usage

books and general dictionaries on usage questions. He researches points of contention

in eight usage books and seven dictionaries. Copperud explains the disagreement in

usage from these many respected sources:

> Dictionaries of usage often disagree, but they have one quality in common: presumption. It could not be otherwise, for the authors are saying to the reader, "I know best." Yet correct usage, whatever that may be, is not a matter of revealed truth, but oftener than not reflects taste or opinion. Such books cast a wide net. Their judgments cover common errors in grammar, misapprehensions of the meanings of words, and the acceptability of changed meanings, to name their principal concerns. The implication is that the critic is reflecting the preponderance of educated practice. But this is not necessarily so, or there would be more agreement among the authorities. (p. vi)

> Copperud decided to overcome this tendency toward presumption by taking a

consensus of other usage experts. He muses,

> But since authorities on usage (all self-appointed, and regarded as authorities only because they have found publishers and audiences for their views) differ, like panelists it is a fair question whether a consensus of their opinions may not also be suspect. (p. vi)

Copperud's usage book is organized like a dictionary, and his entries list the

differences in pronouncements from the various usage books.

What follows is the list of usage books Copperud consults:

1. *The Careful Writer* by Bernstein
2. *Current American Usage* by Margaret Bryant
3. *A Dictionary of Contemporary American Usage* by Evans and Evans
4. *Current American Usage* by Copperud
5. *The ABC of Style* by Rudolf Flesch
6. *Modern American Usage* by Follett
7. *A Dictionary of Modern English Usage* by Fowler, 2$^{nd}$ ed. revised by Gowers
8. *Encyclopedic Dictionary of English Usage* by Mager and Mager
9. *Harper Dictionary of Contemporary Usage* by William and Mary Morris.

*Index to English.* Ebbitt and Ebbitt wrote *Index to English* (1990) to help college students write better, and it includes a lot of valuable usage information. It is organized like a dictionary with entries arranged alphabetically, like most usage books. Ebbitt and Ebbitt categorize usage into formal, informal, and general, with most educated writing and speaking being general American English. They are not quick to call usages wrong and right, but to show these different levels of usage:

> We refuse to condemn as "wrong" failure to observe a rule that has never been observed consistently by gifted writers, and we think it a waste of time to keep deploring practices that (though anyone may choose not to adopt them) have long been accepted in widely esteemed periodicals. When a locution is considered wrong, or illiterate, or merely distasteful by a sizable number of educated readers, we say so, even if the location is firmly established as majority usage—but we also report that it is majority usage. When the same locution appears regularly in publications edited by men and women of skill, taste, and intelligence, we offer an example. The writer can then decide whether or not to use it in a particular rhetorical situation. (p. vi)

This *Index to English* is actually the latest edition in a long line of editions, starting with Porter Perrin's original index in 1939; in fact, in scholarly literature, this book is referred to as *Perrin's Writer's Guide and Index*.  Wilma Ebbitt co-wrote the book in 1965, and by 1990, the latest edition, she is the primary author. Ebbitt is a well-known editor, and she includes examples from "books, learned journals, popular

magazines, and newspapers" (p. v). The guide is short (281 pages), and is a good reference for students. It contains entries about writing, not just about usage.

 *A Dictionary of Contemporary American Usage.* Evans and Evans wrote *A Dictionary of Contemporary American Usage* in 1957, and the book was reprinted in 1981. The authors say in the preface that language changes a lot, and we must adjust our expectations to those changes: "Since language changes this much, no one can say how a word 'ought' to be used. The best that anyone can do is to say how it *is* being used, and this is what a grammar should tell us" (p. vi). Evans and Evans, a brother-sister team, were both born in the United States, but spent their childhood in northern England. This upbringing gives them a sensitivity to the differences between American and British English. Their book emphasizes American English. Evans and Evans acknowledge that many of our earlier rules came from Latin, and these rules are not good descriptions of language today. They also describe how some linguists have made a scientific study of the language that shows us how educated people actually use the language.

 Evans and Evans acknowledge "standard English allows a certain amount of variation" (p. vii). They also acknowledge that modern English is less formal than in past years. They draw their information from (a) the *Oxford English Dictionary*, (b) Otto Jespersen's grammar, and (c) works by Charles Fries. They also use other dictionaries, such as (a) *A Dictionary of American English*, (b) *A Dictionary of Americanisms* by Matthews, (c) *The American College Dictionary*, and (d) *The American Language* written by H. L. Mencken. Furthermore, they draw on articles in *American Speech*, as well as from the writing of other linguistic scholars. Clearly,

they value the opinions of linguistics scholars and dictionaries. The examples they give are created by the authors, and they rarely quote actual usages.

*Modern American Usage: A Guide.* Follett's *Modern American Usage* (1966) is a classic usage book. Follett began writing the book in 1958, and Jacques Barzun and others finished the book after Follett's death in 1963 (p. xii). Follett was keen to write a book on American usage, as opposed to Fowler's book, which is British based. Follett argues that writers need tact; writers need to know what is correct or acceptable for the situation. He is hesitant to use the word "correct" because many linguists consider the idea of correctness to be "a hangover from aristocratic and oppressive times" (p. 4). Linguists, he proposes, believe whatever a native speaker says is correct. Follett, on the other hand, believes that "good usage is what the people who think and care about words believe good usage to be" (p. 6). He proposes that statistics of usage cannot always tell the whole story: "How can science know who the cultivated are and what number suffices to make them many?" (p. 7)

Follett disagrees with linguists who think that whatever people say is always right. He feels there is a right and wrong and a better and worse in usage, and people want to know the best way to say something. Using logic and history, he proposes that learning Latin is the best way to understand English grammar, and he demonstrates that only the knowledge of grammar will help people understand certain constructions. He ends his introduction with the following admonition:

> It is the duty to maintain the continuity of speech that makes the thought of our ancestors easily understood, to conquer Babel every day against the illiterate and the heedless, and to resist the pernicious and lulling dogma that in language—contrary to what obtains in all other human affairs—whatever is is right and doing nothing is for the best. (p. 30)

Follett's book is organized like a dictionary, with many short entries about particular points of usage interspersed with long essays on topics such as "adverbs, vexatious." He seems to base his opinions on his own experience and good taste.

*The Careful Writer: A Modern Guide to English Usage.* Theodore M. Bernstein was an editor at the *New York Times*. He also had a column for many years about language, and he compiled his ideas into *The Careful Writer: A Modern Guide to English Usage* (1977). He begins his introduction defending the idea of a usage book. He argues against linguists who say the "only authority in this field is the unclear, imprecise, and often vulgar voice of the masses" (p. viii). He claims his authority (a) from the "practices of reputable writers, past and present," (b) from "observations and discoveries of linguistic scholars," (c) from "the predilections of teachers of English wherever . . . these predilections have become deeply ingrained in the language itself," (d) from "observation of what makes for clarity, precision, and logical presentation," (e) from personal preferences of the author, and (f) from "experience in critical examination of the written word as an editor *The New York Times*" (pp. viii-ix).

Bernstein criticizes linguists who set out to describe the language but make no moral evaluation about that language, and he criticizes linguists for championing speech over written language. Bernstein also makes a case for the experts to decide what is wrong and what is right with written language. Interestingly, in his acknowledgements, Bernstein thanked Mario A. Pei for his help, and he added "It is not to be assumed that he is in agreement with every position taken in these pages— probably no authority on usage would be" (p. xvi). Bernstein, then, favors the written

word, favors the language authority, and acknowledges that there is disagreement even among language authorities. Bernstein's book is organized like a dictionary with alphabetized entries. He draws most of his examples from newspapers.

*Other language commentators.* Other writers draw from their newspaper columns on language. Edwin Newman wrote *Strictly Speaking and A Civil Tongue* (1980), which are compilations of his language columns. Newman's book consists of long essays about his complaints against errors in language. Because items are not listed individually, this book is not really a reference book but a book to be read front to back.

William Safire also wrote columns about language which he organized into books, such as *I Stand Corrected* (1984). His book is arranged alphabetically, like a dictionary, but he discusses mostly jargon and slang, rather than an entire range of usage questions. Safire is a political columnist, and he writes language columns. In his introduction to *I Stand Corrected*, he addresses his claim to authority in pronouncing usage bad or good:

> Who decides who wins? The answer is not universal, but personal. For myself, I decide. I look at the challenging usage, check around to see what the other usage mavens have said, apply my own standards, put it on a back burner, stir it now and then, and come forth with my decision. In the eternal tug-of(o')-war, I grab the rope in the position of round-heeled prescriptivist, happy to stand on the burning deck but not after all the others have fled. I am my own Final Authority; the Academie americaine, *c'est moi.* (p. ix)

*Conclusions about Usage Books*

The large number of different usage books shows that there is a market for these books. This literature review of usage books shows that most of these books take a very prescriptivist approach. Few usage book editors regard linguists' opinions

as valuable because the editors consider linguists to be too descriptivist and too allowing of constructions that may be popular but not traditionally correct. Most books rely on collections of sample usages in the publishing company's files, and many editors look to newspapers for their examples. Many of these usage books consider British English to be the primary English, and American English is a secondary variety.

Only two have used linguistic corpora at all to find out what published writers actually do: *Garner's Modern American Usage* (2003) and Peter's *Cambridge's Guide to English Usage* (2004). Garner discusses American English as the primary variety for his usage book, and he uses linguistic corpora to back up his pronouncements, but he uses a limited corpus of only newspapers, legal writing, and business writing. This limited subject matter skews his data. In *Cambridge,* Peters uses the British National Survey and the Cambridge International Corpus to research usage items. This usage book specializes in British usage, though, and adds information about different Englishes around the world, like American. No usage book concentrates on American English and bases its authority on actual practices in edited, published American English as found in different registers; therefore, there is a need to study corpora of American English writing in order to determine what rules writers and editors follow.

*Style Guides*

Up to this point, this review has focused on usage dictionaries and books. Another place people look to for answers about usage questions is a style guide. Different fields and publishing houses have their own style guides, which declare

how all people using this style should spell, use words, and cite information. Style guides, then, are usually specialized for particular fields, such as (a) social sciences, (b) sciences, (c) humanities, (d) newspapers, (e) magazines, and (f) the United States government. Many people rely on the *Chicago Manual of Style* as an all-purpose publishing style guide. Most newspapers follow the *Associated Press Stylebook.* These two style guides greatly affect the texts discussed in this dissertation, so this section will review these two guides. There are many other style guides covering many other fields, and these will just be listed.

 *The Chicago Manual of Style.* Many consider *The Chicago Manual of Style* (2003) to be the bible of the editing world. The newest edition of *The Chicago Manual of Style* is the result of much wider consultation than in past years. The editors "enlisted scholars, publishing professionals, and writers familiar with book and journal publishing, journalism, and—particularly valuable—electronic publication" (p. xi). This newest edition features a section on grammar and usage written by Garner (as in *Garner's Modern American Usage)*. This section includes a glossary of troublesome expressions, which looks much like the usage dictionaries previously reviewed in this dissertation.

 *The Chicago Manual of Style* sets out the rules of this publishing house's style. As the preface states, "it has retained its occasionally arbitrary character, for it reflects Chicago's house style" (p. xiii). It states it has a "conservative approach," which is "tempered by pragmatism" (p. xii). It sets out the correct punctuation and spellings, and then it discusses names, terms, numbers, foreign languages, quotations, and dialogues. At the end it shows how to document information. *Chicago* does not

present several different positions of a debate; it settles the debate by choosing one way to punctuate or spell. House styles must do this to set a consistent style for all who use this style.

*The Associated Press Stylebook and Briefing on Media Law.* Goldstein's *The Associated Press Stylebook and Briefing on Media Law* (2002) is a dictionary-like book which lists the preferences for usage and spelling for the print media. It lists the reference books that informed its choices, and it advises writers to consult these reference works to ask further questions. The most interesting choice is the "first reference for spelling, style, usage and foreign geographic names," (p. xiii) which is *Webster's New World College Dictionary* (4th Ed., New York: Hungry Minds, Inc.) The second reference for spelling, style and usage is *Webster's Third New International Dictionary* (Springfield, MA: Merriam-Webster). This is the only usage book that considers *Webster's New World College Dictionary* to be the best reference. Most other usage books and style guides choose Merriam-Webster's dictionary.

The following is a list of other important style guides that concentrate on format and put forth some usage pronouncements.

1. *MLA Handbook for Writers of Research Papers* . The *MLA Handbook* (Gibaldi, 2003) is the format bible for the humanities, with some usage items.

2. *Publication Manual of the American Psychological Association* (2001). Most of the social sciences and education fields use APA style.

3. *Style Manual 2000.* The United States Government Printing Office has its own style manual for all documents it prints.

4. *Scientific Style and Format.* The Council of Science Editors (2006) publishes a guide for scientific fields. This style guide originally came from the field of biology, but the council has expanded to include all of science.

5. *AMA Manual of Style. JAMA* and the *Archives Journal* (2007) publish a style manual for the American Medical Association (AMA). The AMA style manual is for the medical field and is created by the American Medical Association.

6*. IEEE Editorial Style Manual.* The IEEE style guide (2007) began as a guide for electrical engineers, but now it serves the engineering, computer science, and technology fields.

<div align="center">*Usage Teaching*</div>

Usage is taught in English classes from kindergarten to college. Many people call this *grammar* instead of usage. In fact, the word *grammar* confuses many people because it can mean many things. Instructional designers need to understand the many definitions of the word *grammar* if they are to design language instruction. Evaluators and assessors also need to understand the differences in definition if they are to evaluate educational material and design tests in language arts.

Patrick Hartwell (1985), in his article "Grammar, Grammars, and the Teaching of Grammar," builds on an older definition of grammar proposed by W. Nelson Francis (1954). Francis proposes that *Grammar One* is "the set of formal patterns in which the words of a language are arranged in order to convey larger meanings" (p. 300). This grammar is the one that we all use to communicate with each other. It is the native speaker's knowledge of how the language works. This knowledge is not conscious knowledge.

*Grammar Two* is "the branch of linguistic science which is concerned with the description, analysis, and formulization of formal language patterns" (p. 300). This is the grammar that linguists create. Grammar Two describes in great detail the rules and patterns of the language. This is not the innate knowledge of a native speaker, but a conscious study of the language and a conscious analysis of the patterns. Hartwell calls this "scientific grammar" (p. 110).

*Grammar Three*, according to Francis, is "linguistic etiquette" (p. 300). Hartwell calls this not grammar, but usage (p. 110). Grammar Three covers the concerns of this dissertation. Hartwell goes on to create Grammar Four, which is school book grammar, or the grammar that school books explain to students. It is based on the Latin grammars of the 18$^{th}$ century, and it has been used for years to explain the English language to students from elementary school to college. This is the grammar that scholars debate when they claim that teaching grammar does nothing to help students write better. Grammar Five, according to Hartwell, is "stylistic grammar" (p. 110), which is grammar terms used to teach people to write better. This kind of grammar is in the tradition of Strunk and White (*The Elements of Style)*, Williams (*Style: Ten Lessons in Clarity and Grace*), and Lanham (*Revising Prose)*.

Obviously, definitions of grammar vary from scholar to scholar. When some teachers discuss grammar errors, they are often speaking about usage errors. Connors and Lunsford (1988) conducted a study of the most common errors students make in their writing. First, they discovered that what is considered to be an error changes over time:

> What seem to us the most common and permanent of terms and definitions are likely to be newer and far more transient than we know. . . . Teachers have always marked different phenomena as errors, called them different things, given them different weights. Error-pattern study is essentially the examination of an ever-shifting pattern of skills judged by an ever-shifting pattern of prejudices. (p. 399)

As part of their research, Connors and Lunsford had 50 college teachers analyze

teachers' marks compared with actual errors in 3,000 student papers. From this study,

they make several generalizations:

> First, teachers' ideas about what constitutes a serious markable error vary widely. . . . Second, teachers do not seem to mark as many errors as we often think they do. . . . Third, the reasons teachers mark any given error seem to result from a complex formula that takes into account . . . how serious or annoying the error is perceived to be at a given time for both teacher and student, and how difficult it is to mark or explain. . . . Fourth, error patterns in student writing are shifting in certain ways, at least partially as a result of changing media trends within the culture. (p. 404)

The number one problem Connors and Lunsford found was spelling. Students made

three times more spelling errors than any other errors. This study was conducted

before spell check was widely used, though. The top 20 occurring errors or error

patterns besides spelling are as follows:

1. No comma after introductory element
2. Vague pronoun reference
3. No comma in compound sentence
4. Wrong word
5. No comma in non-restrictive element
6. Wrong/missing inflected endings
7. Wrong or missing preposition
8. Comma splice
9. Possessive apostrophe error
10. Tense shift
11. Unnecessary shift in person
12. Sentence fragment
13. Wrong tense or verb form
14. Subject-verb agreement
15. Lack of comma in series
16. Pronoun agreement error

17. Unnecessary comma with restrictive element
18. Run-on or fused sentence
19. Dangling or misplaced modifier
20. *Its/it's* error  (extracted from Table 1 on p. 403)

Connors and Lunsford show that what teachers consider an error is changing. They also show that students are not making any more errors now than in the earlier parts of the 20th century, just different ones.  Lunsford has recently repeated this study, but she has not published the results yet.

In this dissertation I consolidate the Connors and Lunsford list into categories of common errors.  I use this list of categories to organize the usage items I research using text archives.

Maxine Hairston (1981) takes a different slant on studying errors. She muses that "professional writers frequently violate handbook rules and their writing does not necessarily suffer" (p. 795). What should we teach students, then?  To answer this, she gave a survey to 101 professional people to find out what errors bothered them and what errors did not. Unfortunately, her methods of choosing the professionals were based more on who she knew rather than a principle-based scientific selection. She found that women were much more bothered by errors than men. The constructions that bothered these professionals the most were phrases that were clearly considered substandard, such as *we was* and *has went*.  On the next level down were sentences that had multiple negatives or objective pronouns used as a subject (*Him and me*). At the very bottom on the list were the least offensive errors.

> These included using a qualifier before "unique" ("that is the most unique city"), writing "different than" instead of "different from," using a singular verb with "data," using a colon after a linking verb ("Three causes of inflation are:"), and omitting the apostrophe in the contraction "it's." (p. 797)

Hairston was surprised that many respondents were passionate about certain usage choices. She also found it interesting that these professionals as a whole valued clarity and economy as much as correct usage and mechanics. Some usage mistakes did not bother readers much, and yet teachers often mark the items as errors. The following mistakes (in order as they appear on the survey on pp. 800-806) were rated mostly "bothers me a lot":

1. Fused sentence (two sentences put into one sentence with no conjunction)
2. Lack of parallelism in a list
3. Wrong word choice
4. Dangling participle
5. *Set* instead of *sit*
6. No capitals for *French* and *German*
7. Lack of apostrophe in a possessive
8. *Brung* for *brought*
9. A plural pronoun *them* with a singular noun *a person*
10. Lack of commas in a series
11. *My husband and I* instead of *my husband and me* as direct object
12. Sentence fragments—an appositive as separate sentence
13. *Bad* instead of *badly* in an adverbial position
14. A question mark after an implied question
15. No commas around *however*
16. Misplaced modifier
17. An objective pronoun, *Him and Richard,* as a subject
18. Multiple negatives in *There has never been no one*
19. *These kind* instead of *these kinds*
20. Inconsistent tense in a compound sentence
21. *If I was* instead of *if I were*
22. *Would of been* rather than *would have been*
23. *Has went* instead of *has gone*
24. No closing parenthesis symbol
25. Bad wording
26. *Can't hardly* rather than *can hardly*
27. Proper nouns not capitalized
28. Dangling participle combined with the wrong verb tense
29. A plural subject with *is*
30. Subject of *president or vice-president* takes a plural verb
31. Plural pronoun *they* with singular noun *person*
32. Lack of parallelism in a list after a colon
33. *Effect* for *affect*

34. Proper nouns in lower case
35. *That is her* instead of *that is she*
36. Subject and object separated with a comma
37. *We was* instead of *we were*
38. *It's* instead of *its*
39. Subordinate clause written as a sentence
40. *Don't* with a singular subject (third person singular)

Also interesting are the errors on the survey (pp. 800-806) that professionals are not bothered by much. The following mistakes were rated mostly "bothers me a little." Sometimes the survey takers marked one instance of an error as "bothers me a little," and they marked another instance of the same kind of error as "bothers me a lot" or "doesn't bother me at all." Survey takers are not always consistent in their evaluations, and the context of a phrase or word can change its meaning. This inconsistency accounts for errors listed in two different categories.

1. *For whoever* instead of *for whomever*
2. Too much repetition of a word in one sentence
3. Lack of quotation marks around a quote
4. *Between* for *among*
5. A comma instead of a semi-colon or a dash
6. *Everyone . . . they* instead of *everyone . . . he*
7. *Different than* rather than *different from*
8. *And* repeated with each item in a list
9. Lack of comma with an appositive
10. *Its* instead of *it's*
11. As accusative pronoun instead of possessive pronoun with a gerund
12. No commas around an inserted phrase
13. Comma splice (two sentences connected with a comma and no conjunction)
14. *Everyone of them are* rather than *everyone of them is*
15. A comma rather than a semi-colon

An even more interesting list is the one that includes errors that professionals are not bothered by (pp. 800-806):

1. *Most unique*
2. *Everyone . . . they*
3. *Different than* rather than *different from*

4. A colon coming after a *to be* verb
5. *The . . . is when*
6. *Data* with a singular verb
7. *That* instead of *who*
8. A lack of comma after an introductory clause

Curiously enough, usage books argue over these very issues that show up as bothering professionals a little or not bothering at all. Many of the issues that teachers mark a lot and students miss appear on the "bothers me a little" list, such as comma splices, *its* instead of *it's,* lack of quotation marks around a quote, etc. This means that teachers are more sensitive to some usage questions than professionals are, and, therefore, teachers are bothered more by those errors. Hairston's list makes it look like some distinctions are not as important as they once were. These items that professionals do not care as much about will probably show up with mixed usage in a corpus study of actual writing. Unfortunately, I have no good way of searching American punctuation until an American corpus is completed. Other usage, though, we can check.

Joseph M. Williams, in "The Phenomenology of Error" (1981a), attempts to show that not all errors are considered equal. He defines an error not as an "isolated item on a page," but "as a flawed verbal transaction between a writer and a reader" (p. 153). Some violations we notice, and some we don't notice. Williams is mostly concerned with the violations we notice when we are not looking for violations.

> We need not believe that just because a rule of grammar finds its way into some handbook of usage, we have to honor it. Which we honor and which we do not is a problem of research. We have to determine in some unobtrusive way which rules of grammar the significant majority of careful readers notice and which they do not. (p. 164)

Williams also asserts that when teachers read for error, they cannot devote as much attention to content, and when they read for content, they cannot devote as much energy to error. Williams argues that different people notice different errors:

> Well, it is all very puzzling: Great variation in our definition of error, great variation in our emotional investment in defining and condemning error, great variation in the perceived seriousness of individual errors. The categories of error all seem like they should be yes-no, but the feelings associated with the categories seem much more complex. (p. 155)

He goes on to show that some of the great writers about language and the errors in language, namely, E. B. White, George Orwell, and Jacques Barzun, all violate many of the rules they set forth. They say to avoid the passive, but they often write in passive, and few people notice it.

Williams categorizes error into several groups. The first group of errors is those whose violation we notice, but we don't notice when the rules are observed. These are errors such as "double negatives, incorrect verb forms, many incorrect pronoun forms, pleonastic subjects, double comparatives and superlatives, most subject-verb disagreements, certain faulty parallelisms, certain dangling modifiers, etc." (p. 161). The next category is a group of errors that most people do not notice when the rule is violated. "They constitute a kind of folklore of usage, rules which we can find in some handbook somewhere, but which have, for the most part, lost their force with our readers" (p. 161). Williams presents a partial list of errors that people do not notice when others violate them:

1. Beginning sentences with *and* or *but*;
2. Beginning sentences with *because* (a rule that appears in no handbook that I know of, but that seems to have a popular currency);
3. *Which/that* in regard to restrictive relative clauses;
4. *Each other* for two, *one another* for more than two;
5. *Which* to refer to a whole clause (when not obviously ambiguous);

6. *Between* for two, *among* for more than two. (p. 168)

He then lists errors that some readers feel very strongly about, but which he asserts occur frequently and are not noticed:

1. *Less* for *fewer*;
2. *Due to* for *because*;
3. The strict placement of *only*;
4. The strict placement of *not only, neither,* etc. before only that phrase or clause that perfectly balances the *nor*. (p. 168)

He also suggests that current usage also puts several other disputed words into this category: "*disinterested/uninterested, continuous/continual, alternative* for more than two" (p. 168).

Williams has a third category: those rules that, when violated, people do not notice, but when observed, people notice. This is an odd category with very few instances; nevertheless, there are some constructions that stand out when people use them and do not stand out when people violate the rules:

1. *Shall/will*
2. *Who/whom*
3. Unsplit infinitives
4. Fronted prepositions
5. Subjunctive form of *be,*
6. *Whose/of which* as possessives for inanimate nouns,
7. Repeated *one* instead of referring pronoun *he/his/him*,
8. Plural *data* and *media*, singular verb after *none*. (p. 168)

Williams argues that most people do not notice when these usages are violated, but we notice when the usages are observed because they stand out as not normal. Williams cleverly puts about 100 errors throughout his paper and challenges his readers at the end to notice which errors they picked up with a first reading. His point is that we do not notice some kinds of errors. Since we do not notice them, maybe they are not really errors.

Edgar Schuster also discusses errors that are not really errors. In *Breaking the Rules: Liberating Writers Through Innovative Grammar Instruction* (2003), he lists four kinds of rules. First, there are the intuitive rules that all native speakers have. Next, there are the bedrock rules of English syntax that most everyone agrees on. Then there is Well-Edited American Prose (WEAP), which includes the nuances of usage that only editors and a few others know. Finally, there are usage rules that are debatable: "Rules that Do Not Rule." These particular rules have no universally agreed upon standard. Schuster advocates looking to actual writing of writers we admire to find out the true rules of good writing. If a rule states "Don't start a sentence with *and* or *but,*" we should look through the works of a writer we admire and see if that person starts a sentence with *and* or *but*. We should look to the practice of good writers to discover how they use language. This philosophy influences the methodology of this dissertation because Williams encourages students to discover good usage from actual texts, not just from rules set out by experts.

Another study shows how academics and business executives perceive writing errors. Leonard and Gilsdorf (1990) researched impressions about writing errors in two specific groups: post-secondary business communication teachers and executive vice presidents in very large firms. Leonard and Gilsdorf assert that writers and teachers must make "judgment calls" about many usage items. "Many of the most hotly contested elements are matters of taste or personal opinion. With each published authority implicitly claiming to reflect a majority of educated persons' practice, more descriptive data are in order" (p. 137). This dissertation tries to

describe usage in educated, edited English, just as Leonard and Gilsdorf suggest we need to do.

In English composition, teachers are becoming less prescriptive, but leaders in business education are becoming more and more prescriptive:

> While most writers in composition journals now view contested elements of usage as a matter of situation, purpose, and other variables, most journals in business communication and business education have evinced less change in what they consider correct usage. Similarly, many business communication courses and texts have retained a list of shoulds and shouldn'ts generated decades ago. ( p. 199)

Leonard and Gilsdorf tested 58 usage elements with the two different groups to see who was bothered most by what errors. Leonard and Gilsdorf decided that we can choose what to emphasize and what not to emphasize in writing classes by examining how bothered future readers will be by certain mistakes. They also assume that if what some consider an erroneous practice does not bother other readers, then that erroneous practice may be in flux, and opinion about it may be changing. It may not be considered an error for much longer.

Greenbaum and Taylor (1981) also set out to test certain usages. They investigated ten usages and how writing teachers perceive these usage errors. They found much agreement among teachers identifying "clearly unacceptable" items, and they found just as much agreement in the "clearly acceptable" category. Surprisingly, many teachers could not identify any error in the "clearly unacceptable" items. In the middle category, "in divided usage," Greenbaum and Taylor found that teachers were split in identifying error. Some teachers considered the following items correct, and some considered them incorrect:

1. Plural verb with quasi-coordinator as subject

The instructor, as well as his students, are happy with the textbook.
2. *Between you and I* ("in the sense of confidentially")
    Between you and I, his grading is often unfair.
3. *Who* or *whoever* as direct object
    He can choose whoever he wants.
4. *Badly* after linking verb *feel*
    We felt badly when we saw the extent of the damage. (p. 170)

Larry Beason (2001) researched how people established in the business field react to error. He researched how certain errors negatively impact an author's ethos. By questionnaire and interview, he uncovered readers' reactions to the following error types: (a) fragments, (b) misspellings, (c) word-ending errors, (d) fused sentences, and (e) quotation mark errors. Beason suggests "that the extent to which errors harm the writer's image is more serious and far-reaching than many students and teachers might realize. . . . Errors affect a person's credibility as a writer or employee" (p. 48). Beason suggests researchers spend time discovering what errors will affect a writer's credibility, and educators should make these usage items part of the curriculum.

Obviously, many composition instructors must concern themselves with usage instruction. Teachers want students to learn to write clearly, and they want students to learn to use standard written English. Authors differ in what they consider to be the most important rules to teach, but at least they are asking employers and other teachers to find out what is most important to learn about writing and which errors annoy others the most. Researchers in rhetoric and composition are also aware of the perils of assuming there is only one correct way to use the language.

*Corpus Linguistics*

There may be another way to discover the working rules of a language besides relying on the opinions of experts. Corpus linguistics can help educators discover the actual practice of published writers. Researchers can inquire what actual practice is in standard written American English, and corpus linguistics can help researchers uncover this actual practice.

*Creation and Use of Linguistic Corpora*

Linguists have assembled corpora to study aspects of language. A corpus is a body of written and/or spoken language. Meyer (2002) defines a linguistic corpus as "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (p. xi). Biber, Conrad, and Reppen (1998) define a corpus as "a large and principled collection of natural texts" (p. 12), and McEnery and Wilson (2001) describe corpus linguistics as "the study of language based on examples of 'real life' language use" (p. 1). Currently, the useful corpora are electronic, although corpora existed before computers. Of course, computers have made the creation and use of corpora much easier. Corpora are created carefully to make each corpus a good sampling of what it represents. Some corpora represent spoken language, others written language, and some both. The corpus creator sets the proportions for each register to be included in the corpus. The linguistic corpora are carefully assembled, then each word is tagged with linguistic labels so researchers can ask many questions and search for many items.

Corpora can be researched to answer questions about the "prosody, lexis, grammar, discourse patterns or pragmatics of the language" (Kennedy, 1998, pp. 3-4).

Biber (1998) shows that corpus linguistics can aid (a) in creating dictionaries, (b) in describing grammar, (c) in studying discourse characteristics, (d) in documenting register variation, (e) in understanding language acquisition and development, and (f) in investigating historical and stylistic changes.

He ends his book suggesting that other areas of linguistic study would benefit from corpus linguistics: "As with other areas, analysis of a large body of authentic language can show the actual language patterns being used—rather than having to rely on intuition or anecdotes" (p. 236). Kennedy states that in the past, evidence for linguistic theories came from "intuition or introspection, from experimentation or elicitation, and from descriptions based on observations of occurrence in spoken or written texts" (1998, p. 7). These are the ways many usage experts have decided issues of usage in the past.

According to Kennedy, "In the case of corpus-based research, the evidence is derived directly from texts. In this sense corpus linguistics differs from approaches to language which depend on introspection for evidence" (pp. 7-8). One example of the use of intuition is Noam Chomsky's (1957) approach to language study, which uses constructed sentences and introspection. Theoretical linguists like Chomsky do not usually draw on actual examples, but create examples themselves. Corpus linguists, on the other hand, prefer to study actual texts and draw from real, not contrived, language.

Descriptive linguists need a body of text so they can accurately and validly describe language. In using corpora studies, descriptive linguists can research not only what is possible in a language, but also what is probable (Kennedy, 1998, p. 8).

In the past, usage experts have relied on their own observations, but now we have a chance to base usage guidelines on actual practice, on actual texts.

Corpus linguistics brings an entirely new level of research to usage debates. Because of increasing computer capabilities and databases, now we can rely on empirical data and not just on expert opinion when we are inquiring about many individual points of usage. In the field of corpus linguistics, a researcher creates or finds a corpus of sample documents that are representative of the language the researcher wants to study. Corpora can include written language and transcripts of spoken language. A corpus is usually a sampling of texts, so linguists must carefully set the proportions of what registers to include in the corpus so that the results of the research will reflect what is happening in the entire register. Of course, this will increase the validity of the findings. Meyers comments on the creation of corpora:

> If corpus linguists understand the methodological assumptions underlying both the creation and subsequent analysis of a corpus, not only will they be able to create better corpora but they will be better able to judge whether the corpora they choose to analyze are valid for the particular linguistic analysis they wish to conduct. (2002, p. xiv)

Linguists need to choose corpora that represent the language the researchers want to study. If they are interested in studying newspaper language, they need to use a newspaper corpus; if they are interested in the language of fiction, then they should use a fiction corpus; and if they are interested in the language of written scholarship, then they should study articles in scholarly journals. Most linguists wanting to make generalizations about a language seek to use a balanced corpus for the purpose of describing actual usage across several registers. A balanced corpus contains samples

from many different kinds of sources, and those samples are in a proportion

predetermined by the linguist. If no corpus exists, linguists can create one.

When linguists put together corpora, they often tag each word with linguistic

information, such as the part of speech. Tagging makes the corpus searchable and

very useful for linguistic purposes. My first choice would be to use a tagged, fully-

assembled corpus. This choice would allow me to research (a) parts of speech, (b)

collocates (words that occur near the word in question), (c) verb forms, (d)

combinations of parts of speech, and (e) punctuation.

*Examples of Existing Linguistic Corpora*

There are many well-designed and usable corpora already in existence.  This

next section will review the already-assembled corpora in English to find one that

will work for this study.

The original widely-used corpus is *Brown Corpus* (1961) (available on

ICAME CD ROM found at http://icame.uib.no/cd/readme.htm), compiled by Kucera

and Francis at Brown University. It contains one million words of written American

English. The first editions of *American Heritage Dictionary* drew examples of word

use from the Brown corpus.

More recently, Mair has overseen the creation of the *Freiberg-Brown Corpus*

*of American English* (*FROWN*) (available on ICAME CD-ROM).  This corpus

attempts to apply the Brown corpus model (1961 American English) to 1990s

English.  The *FROWN* is comprised of one million words from American English. As

a parallel corpus to the Brown Corpus, *FROWN* allows researchers to compare

language from different decades.

One of the first corpora of British English, *The Survey of English Usage Corpus* (*SEU*) (http://www.ucl.ac.uk/english-usage/), compiled by Quirk and Greenbaum, was originally compiled on paper and then later made electronic. Of the one million words, half are spoken British English and half are written British English collected from 1953 to 1987. Jan Svartvik directed the computerization of the *London-Lund Corpus of Spoken English* (on ICAME CD-ROM) which is the spoken part of the *Survey of English Usage*.

The *Lancaster-Oslo/Bergen Corpus* (*LOB*) (ICAME CD-ROM) is the British version of Brown's corpus. It consists of one million words from 1961, and it draws exclusively from British English.  Later, the *Frieberg-Lancaster-Oslo-Bergan Corpus* (*FLOB*), a parallel corpus, was created, and it mirrors the LOB corpus, but uses 1991 texts.

The *International Corpus of English* (*ICE*) has collected one million words from many of the different Englishes throughout the world, such as British and South African.

These corpora are valuable, but, at one million words each, they are too small for this study.  The next corpora listed are from large publishing houses.

Collins publishers assembled a corpus, the *Bank of English* (http://www.titania.bham.ac.uk/docs/svenguide.html), for their own dictionaries. The Collins website says that subscribers can access the 56 million word subcorpora, but only under special conditions can anyone access the large corpora (Collins, 2006). The current word count is about 525 million words. All but 10 million of these words are British English.

The *Longman Written American Corpus* (http://www.pearsonlongman.com/ dictionaries/corpus/written-american.html) contains 100 million words from newspapers, journals, novels, scientific works, and technical writing in American English. This corpus provides the examples and information for the Longman dictionaries. This corpus seems to be a perfect fit for this dissertation. Unfortunately, this corpus is not available to the public, but it is proprietary to Longman publishers.

The *Cambridge International Corpus* (http://www.cambridge.org/ etl/corpus/default.htm) claims to have one billion words in its corpus of British and American writing and speech. The American section has 250 million words of written American English, four million of American academic English, and 30 million of American business writing. Unfortunately, this corpus can be used only by authors writing for Cambridge University Press.

The next two corpora contain the written record of spoken American English. The *Corpus of Spoken Professional American-English* (http://www.atel.com/cpsa.html) contains English spoken in academic settings and English spoken in question and answer White House press conferences. Each sub-corpus contains about one million words. Spoken English is, of course, worth studying, but this dissertation needs a corpus of written American English.

The *Santa Barbara Corpus of Spoken American English* (*CSAE*) (http://www.inguistics.ucsb.edu/research/sbcorpus.html), prepared by Chafe, DuBois, and Thompson at the University of California at Santa Barbara, contains transcripts for 60 vocal segments. This corpus contains transcripts of mostly face-to-face

American Speech. This corpus doubles as the spoken segment of the American English section of the *ICE*.

The previously reviewed corpora will not work well for this dissertation because (a) they are either too small, (b) they contain only British English, (c) they are not open to the public, or (d) they contain only spoken American English. The best English corpus available to the public is the *British National Corpus* (*BNC*) (http://www.natcorp.ox.ac.uk/), which, unfortunately for this study, includes only British writing and speaking, but it serves as a great model of how a corpus can aid linguistic study. This corpus contains around 100 million words of British English from the 1990s. Mark Davies of Brigham Young University has created an interface with the corpus that makes the *BNC* searchable by register. His interface is found at http://corpus.byu.edu/bnc. The following are large categories within the *BNC*, and these categories can also break down into more specific categories: (a) spoken, (b) fiction, (c) news, (d) academic, (e) non-fiction miscellaneous , and (e) other miscellaneous. The *BNC* can also find collocates, words that are found close to the word being researched. If this dissertation were focusing on British English, then this corpus would be the perfect one to use.

Unfortunately, there is no large usable corpus for American English at this time. Chomsky, the American linguist who developed transformational grammar, has dismissed exemplar-based studies as incomplete and therefore uninformative, so consequently, American linguists have not pursued corpus-based studies with much passion.  The British and other Europeans, on the other hand, have been developing corpora for decades, and nothing in American English equals the current BNC.

This situation will soon change. The *BYU Corpus of American English* is being constructed as this dissertation is being written. It should be ready in the winter of 2008, after this dissertation is finished. It will open up corpus research into American English the way that the *British National Corpus* has opened up research into British English. The *BYU Corpus,* starting with 360 million words from 1990 to 2007, will also be updated every three months, so it will always have instances of current American English. It will include the registers of fiction, newspapers, scholarly journals, magazines, as well as transcripts from spoken English. Unfortunately, it is not yet available for use.

Instead of using a corpus assembled by linguists, I will use text archives that are searchable by word, if not by part of speech. In these databases, I can search certain words, phrases, and spellings, but not punctuation. Other usage questions (for example, how often writers end a sentence with a preposition) will have to wait for the tagged linguistic corpus, too.

*Testing and Content Validity*

Many tests proclaim one thing they are testing is knowledge of standard written English, meaning American English. How do test writers know what standard written American English is? Test makers try to achieve content validity partially by having a panel of experts review the material. Every national standardized test has been examined for validity for particular usages of that test. Often these tests are used to help decide if a person should be advanced a grade or if the person should be admitted to an undergraduate or graduate program of study.

Most of the validity studies on the Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT), and SAT that are published on the official test websites are studies of predictive validity. In other words, they are studying the test to see if it will accurately predict how a student will perform in the first year of college or the first year of graduate studies (Talento-Miller & Rudner, 2005). In fact, the Graduate Management Admission Council (GMAC) assists individual schools in performing a predictive validity study for just that particular school. The GMAC (at GMAC.com) provides a Validity Study Service (VSS) to schools who wish to participate.

To test for content validity, researchers take a different tactic. The College Board, which publishes the SAT, College Level Examination Program (CLEP), and Advanced Placement (AP) tests, defines content validity as follows:

> Content validity addresses the match between test questions and the content or subject area they are intended to assess. This concept of *match* is sometimes referred to as *alignment*, while the content or subject area of the test may be referred to as a *performance domain*.
> Experts in a given performance domain generally judge content validity. For example, the content of the *SAT II: Subject Tests* is evaluated by committees made up of experts who ensure that the test covers content that matches all relevant subject matter in each of the academic disciplines. Both a face validity and a curricular validity study may be used to establish the content validity of a test. (College Board)

Rubio, Berg-Weger, Tebb, Lee, and Rauch (2003) state that "a content validity study can provide information on the representativeness and clarity of each item" (p. 95). They continue to warn about the subjectivity of the experts. They state that "experts' feedback is subjective; thus, the study is subjected to bias that may exist among the experts" (pp. 95-96). Since experts are biased and disagree with each other, we can call content validity of language arts tests into question. If the experts

disagree on certain points, then it is impossible to design instruction and tests effectively for those particular points.

Constance Weaver has analyzed the ACT for what kind of questions it asks, so English teachers can determine what kind of teaching they need to do. She asserts that the makers of the ACT English tests have been highly influenced by the Connors and Lunsford study, and may have been too influenced by it, since some teachers claim the study emphasizes comma use too much (p. 63). Weaver warns that tests like the ACT are testing very conservative grammar rules:

> The tests are extremely conservative with regard to grammar. They test 'rules' that many or most published writers don't follow, 'errors' that aren't considered errors by most publishers, and 'no-nos' that never should have found their way into English grammar books in the first place, since they were based on the structure of Latin rather than English—or simply made up by the books' writers. (p. 65)

She then goes on to advocate teaching students to write like "published authors" and then to know the grammar rules required by standardized tests.

Another problem with content validity is that not all experts agree what the right answer should be. Weaver states: "Always keep firmly in mind that even excellent writers, even excellent editors of their own writing, may not be able to do well on such multiple-choice tests of writing skills" (p. 65). According to one expert who analyzes standardized tests for content validity studies, test makers and teachers often do not agree amongst themselves on the best usage or construction (Dean, 2006).

If test makers and teachers do not all agree about correctness or what is better writing, then how are teachers supposed to teach that particular point, and how are tests going to give any information about knowledge of these points? No one seems

to publish much about the lack of content validity in standardized tests about English. Clearly, though, there is disagreement about what is correct and what is not, and this problem should be discussed. Too many non-experts still think there is one right way to (a) spell, (b) use commas, (c) use pronouns, (d) place prepositions, (e) use modifiers, etc. If there are, in fact, several accepted ways to spell, use commas, use pronouns, place prepositions, use modifiers, etc. in published American English, test makers and people who interpret the results should pause and ask just what they are testing. They are not necessarily testing students' abilities to understand correct usage in standard American English.

In a content validity study, information empirically gathered would possibly lessen the total reliance on subjectivity. This is especially true for usage questions about which opinions vary widely. Some of the answers to usage questions on standardized tests could be disputed based on empirical evidence. If we can check usage through a corpus study, then we can discover which items clearly have one usage and not another, and we can check which items are of divided usage. Those items that have divided usage should not be items that are tested.

Corpus linguistics, then, offers teachers, textbook writers, instructional designers, test writers, and course evaluators the tools to research standard written American English usage empirically. In the end, these empirical searches can affect instruction and assessments by aligning them with the actual practice of published authors. Studies using corpus linguistics can add to a test's content validity.

Chapter 3: Method

This methodology section discusses the following topics: (a) researching usage items in text archives, (b) choosing usage items to study, (c) deciding upon a frequency of occurrence for items in standard written American English, and (d) describing the limitations of this research model.

*Research Process*

Research Question 1 asks How can we use an empirical method to determine what is standard written American English in various linguistic registers? In the past, writers of usage books used their intuition, logic, historical rules, and examples from their files to determine what is standard and what is not standard. Some, like the *American Heritage Dictionary,* still consult usage panels to determine the status of a particular usage. Now, with electronic corpora available, we can research actual usage of a particular group of people. For this dissertation, I narrowed the users to writers who published in edited newspapers, magazines, and scholarly journals.

For each research item I consulted what the experts say about its usage. The history of the word is presented, as found in the *Oxford English Dictionary* (Gove, 1989). I used (a) the *Chicago Manual of Style* (2003)*,* the stylebook used most by editors; (b) *Webster's Dictionary of English Usage* (1989)*,* one of the most respected and comprehensive usage books; as well as (c) *Garner's Modern American Usage* (2003), (d) *The American Heritage Dictionary* (2000), (e) the *Associated Press Stylebook* (2002), and (f) *Merriam-Webster's Collegiate Dictionary* (2004). Next, I researched these items in different registers: (a) newspapers, (b) magazines, and (c)

scholarly journals.  Magazines and newspapers are less formal, and scholarly journals represent formal English.

This dissertation was written in the summer of 2007, right before a major American corpus, the *BYU Corpus of American English,* was scheduled to be released.  As discussed in the literature review, many English corpora exist, but no extensive corpora of written and published American English are currently available to the public.  Ideally, I would research every usage item in a linguistic corpus that has a balanced selection of writing, searchable by register and tagged by part of speech.  Because I have no access to such a corpus at this moment, I have had to create a more primitive corpus from existing databases and archives. Text archives are searchable by word and phrase, but the searches cannot be limited by part of speech.  Collocates, words that occur within a few places of the word in question, cannot be searched in a text archive. Even so, text archives can be researched for some usage questions.

Since there are many archives available on the Internet, which ones would best answer usage questions?  Project Gutenberg (http://www.gutenberg.org/wiki/Main_Page) has thousands of online books, but it will not deliver statistics on anything over 1,000 hits. Lexis/Nexis was another obvious choice, but, again, it has a limited scope because it concentrates on law, government, business, and news. Google.com searches all web sites and gives the number of hits, but Google is not selective. Sometimes information from Google searches is helpful, but Google is not limited to only standard English or American English because it includes everything on the Internet.

Google Book Search (http://books.google.com/), though, is different and possibly more valuable. Publishers give a copy of their books to Google, which then scans the full text. If someone searches a particular word in this database, Google Book will then list the number of times that word shows up. It will then show a snippet from each text, which means the window shows the sentence this term is found in. On the positive side, this database is comprehensive because it contains many books. Unfortunately, though, the database is not selective. Anyone can send in a book to be included. Also, many of the books are at least 75 years old because Google Book will not display full text unless the copyright has run out. The statistical count from Google Book is interesting, but we cannot count on it representing today's standard written American English. Also, Google Book does not collect journal articles, just books. These databases are interesting, but they were not the best choices for this dissertation.

I selected three excellent text archives to serve as the corpora for this usage study. Because this study examined general usage across the United States, the corpora needed to contain items from many different publications so I could study many different editorial styles. Academic Search Premier provided a text archive for formal English found in scholarly journals. ProQuest was the text archive for magazines, and Newspaper Source Publications provided newspaper English.

Academic Search Premier (http://www.epnet.com/thisTopic.php?marketID =4&topicID=1) can search full text of scholarly articles from all subjects of academic study: (a) social sciences, (b) sciences including medicine and engineering, (c) humanities, (d) business, (e) fine arts, and (f) law. As a result, Academic Search

Premier is a wonderful text archive of formal English found in academic journals. The database, managed by EBSCO and available to full-time faculty and students at universities that hold a license, contains articles from 4,650 serials, most of which are peer-reviewed titles. Because this database is widely available, other scholars can duplicate this research and perform many other searches.  This archive can be searched by word or phrase, and the database gives a count of each publication that uses the particular word or phrase. For this dissertation, I researched the full text of scholarly journals from January 2000 to April 2007.

ProQuest (http://www.proquest.com/) makes it possible for people to search magazines by year, so I chose to search January 1, 2000-April 30, 2007.  Including magazines such as (a) *Sports Illustrated*, (b) *Time*, (c) *Good Housekeeping*, and (c) *Ebony*, the publications in this corpus are much less formal and more conversational than scholarly journals.  ProQuest searched 477,086 magazines articles printed from January 1, 2000 to April 30, 2007, and it provided the number of publications that include a particular usage. A very small percentage of the magazines are British, and the rest are American. For some of the items I used the magazine database of EBSCO's Academic Search Premier.  The items in this database are also from January 2000 through April 2007. A large variety of magazine titles represents an equally large variety of editorial styles, so searching magazines can give us a sweeping look at usage in a middle level of formality in edited, published English.

Newspaper Source Publications, managed by EBSCO (http://www2.ebsco.com/en-us/Pages/index.aspx), can be searched by date, so I searched from January 2000 to April 2007.  This database has full text for 25 large

American and international newspapers, as well as 260 American regional newspapers, so I personally selected only the American newspapers. This large collection of newspaper articles represents many different editorial styles throughout the nation. This database gives a count for each publication that shows a particular usage. The Newspaper Source Publications is an excellent corpus of newspaper language.

Another interesting register of English is fiction, as found in short stories. I searched a corpus of fiction published between January 2000 and April 2007, and this corpus is comprised of 2,630 full text short stories found on Wilson Web Short Story Index (http://www.hwwilson.com/Databases/storeindec.htm). Because this is a rather small corpus, some word searches have results and others have no results, so some words will have information from the fiction register and others will not. The fiction searches will provide extra information for each item, but the results will not be included in the averaged percentage of usage.

Garner (2003) uses corpora to report usage in his dictionary, but he uses only business, government, newspaper, and legal writing to determine usage. This list is extremely limited because these four kinds of writing are specialized genres. The three corpora used in this dissertation cover a much larger range of writing. The scholarly sources cover many fields, such as (a) social sciences, (b) humanities, (c) fine arts, (d) medicine, (e) math, (f) science, (g) computers, and (h) engineering. The newspapers are strictly American newspapers from every region in the United States, and the magazine corpus includes magazines about a range of topics from sports, health, fashion, politics, news, popular sciences, religion, and many more. These

three corpora cover much of the published writing in the United States found in

periodicals, journals, and newspapers from January 2000 to April 2007.

*Usage Items to Study*

Research Question 2 asks What does empirical research using corpus

linguistics tell us about certain controversial usage questions that are representative of

the many questions still requiring research?

a. What is the percent usage of each item?

b. Does the usage differ among the registers of newspapers,

magazines, and scholarly journals?

c. What usage or usages should be considered standard written

American English?

Every usage book contains hundreds or even thousands of words and phrases

of controversial usage.  In order to choose items to research for this dissertation, I

took the Connors and Lunsford (1988) list of commonly occurring errors in students'

writing, and I categorized the errors into large categories:

1. Spelling
2. Comma use
3. Inflected endings
4. Preposition use
5. Verb forms
6. Subject/verb agreement
7. Pronoun agreement and reference
8. Apostrophe use
9. Shifts in person or tense
10. Sentence recognition
11. Placement of modifiers
12. Wrong word

Errors from some of these categories cannot be researched in a text archive

but eventually can be researched quantitatively in a corpus when a corpus is

available: (a) comma use; (b) apostrophe use; and (c) sentence recognition; as well as

(d) individual items from other categories, such as spelling.  Items from other

categories will need to be researched qualitatively (checking each word in its context)

using corpora in the future:  (a) shifts in person or tense; (b) placement of modifiers;

as well as (c) individual items from other categories, such as wrong word.

This dissertation presents research of items from the following categories:

1. Spelling
2. Inflected endings
3. Preposition use
4. Verb forms
5. Subject/verb agreement
6. Pronoun agreement
7. Wrong word

The individual items are usually words and phrases labeled as controversial

and discussed in usage books.  Each item represents a type of item that can be

researched quantitatively by text archive.

*Spelling*

Early in the history of English, spelling varied greatly, not only among

writers but within one writer's manuscript. Printing and, later, the rise of the

dictionary helped solidify and standardize spellings. Many people recognize the

differences between typical British spellings, such as *theatre* and *colour*, and typical

American spellings, such as *theater* and *color*. Surprisingly, though, variation exists

within American English.  The items represent several kinds of variation in spelling:

(a) letters used, (b) spacing, and (c) hyphen use.

The following spelling variants are compared:

1. *Through/thru*
2. *Catalog/catalogue*

3. *Technique/technic*
4. *A lot/alot*
5. *Cannot/can not*
6. *All right/alright*
7. *E-mail/email*

Some variant forms, such as *judgment/judgement*, are recognized as two optional

spellings in the text archives, so the archives combine both spellings into one search.

Consequently, the text archives cannot distinguish between the two spellings.  We

will have to wait for a corpus to fully research spelling variants.

*Inflected Endings*

English was at one time a highly inflected language, but during the Middle

Ages, many inflections dropped off.  Now English speakers have only a handful of

regular inflections. This section will focus on the noun and adjective/adverb

inflections, while verb inflections will be discussed in a different section.  Many

nouns can add the plural /s/ and the possessive *s*, and many adjectives and adverbs

can add the comparative *–er* or superlative *–est*. Text archives will not search for

apostrophes, so we cannot search for the spellings of possessives until we can use a

corpus.  This section, then, will discuss irregular plural endings for nouns and explore

the rules for the comparative and superlative endings for adjectives.

*Plurals.* English speakers borrowed many words from Latin, Greek, and other

languages. English speakers often kept the plural endings from these foreign

languages. Yet, as these words were used by English speakers, gradually some of the

Latinized endings changed to the more regularized plurals of /s/, /z/, and /əz/. The

question arises, then, what is the correct plural of current English words that have

traditionally had Latin plural endings?

The following plural options are compared:

1. *appendixes/appendices*
2. *indexes/indices*
3 *syllabuses/syllabi*
4. *formulas/formulae*

*Comparatives and superlatives.* Most people follow the rule that one-syllable words take the *-er* ending for the comparative and the *-est* ending for the superlative. Two-syllable words usually take "more" and "most," and three-or-more-syllable words always take "more" and "most." In some words, though, there is variance from the standard way. Usage books rarely discuss the details of comparatives and superlatives, but dictionaries carefully list the inflections *–er* and *–est* with the appropriate adjectives and adverbs.

I will research the following word pair to test the usage of the comparative and superlative construction in one-syllable words: *more proud/prouder*; *most proud/proudest*. These words represent a one-syllable adjective that is controversial and has divided usage in making comparatives and superlatives, although dictionaries do not show divided usage.

The following two words represent two-syllable words that have divided usage, although dictionaries usually show one usage:

1. *more risky/riskier*; *most risky/riskiest*
2. *more lovely/lovelier; most lovely/loveliest*

*Preposition Use*

Prepositions are hard to research in text archives because they are so common that search engines will skip the most common prepositions and not search for them. EBSCO will not search (a) *of*, (b) *by*, (c) *is*, (d) *for*, (e) *in*, or (f) *to*. This inability to

search for the most common prepositions greatly hinders language research. The following usage issues cannot be searched with text archives and will have to be searched later with a corpus: (a) a preposition at the end of a setence, (b) the number of prepositional phrases per sentence, (c) phrasal verbs coupled with particular prepositions, (d) prepositions that commonly appear with certain verbs, and (e) use of *should of*, *could of*, and *might of* instead of *should have*, *could have*, and *might have.*

Because prepositional searching is so limited at this time, this section has only one example of controversial prepositional use: *different from/different than.*

*Verb Forms*

*Past tense*. Old English had many ways to make past tense. In Old English there were seven types of strong verbs, meaning a word that changes the vowel to make past tense, such as in *sing, sang, sung*. Eventually the weak construction prevailed, and now we all use this weak construction as our regular past tense: add a /t/, /d/, or /əd/, all usually spelled *-ed*, such as in *walk, walked; live, lived;* and *mate, mated*. Over the last thousand years, most strong verbs have changed from a strong construction of past tense to a weak construction. Words changed by analogy with other words. It should also be noted that a handful of words changed from a weak construction to a strong construction.

Most words now use the weak construction (regular) for past tense, but some words still use the strong construction (irregular—changing the vowel for past tense). Some words are in flux and are changing from one construction to the other, and so some verbs have both a weak past tense and a strong past tense. This fluctuation can

also occur in the past participle. This study presents research on the following past tense constructions in flux:

1. *snuck/sneaked*
2. *dove/dived*
3. *crept/creeped.*

Many usage books discuss these three verbs because their past tense seems to be changing.

*Past participle.* Some words show variation in the past participle, with one form being the regular *–ed* ending and another form adding *–en* or changing the vowel. The past participle form of a verb is found by putting *have, had,* or *has* in front of the verb. This study will present research on the following items to determine the standard usage of the past participle form:

1. *sped/speeded*
2. *proven/proved*
3. *kneeled/knelt*

Other verb forms cannot be researched until a corpus is created: researching the subjunctive and tense problems will probably take qualitative and quantitative evidence, and researching the past participle as an attributive adjective will require a quantitative corpus study.

*Subject-Verb Agreement*

Most issues of subject-verb agreement cannot be researched by text archives, such as a separated subject and verb matching in number, and a compound subject taking a plural or singular verb. This section, therefore, will focus on a problem of two words which are considered singular nouns taking a singular verb by some, and a

plural noun taking a plural verb by others. People who consider *data* and *criteria* to be plural are agitated by people who pair the words with singular verbs.

1. *this data/these data*
2. *this criteria/these criteria.*

*Pronoun Agreement*

All nouns in Old English were declined according to the function in the sentence: (a) subject, (b) object, (c) object of a preposition, or (d) possessive (also called (a) nominative, (b) accusative, (c) dative, and (d) genitive). Each noun had four possible endings in the singular and four possible endings in the plural, making eight possible endings for all nouns. Now, we have only three endings for most nouns: (a) no ending, (b) -*s* for plural, and (c) *'s* for possessive. The pronouns, on the other hand, have kept their various endings, probably because they are some of the most used words in the language. Choosing the correct pronoun case can be very complicated, especially since only the pronouns are declined this extensively and not any other nouns.

I will research pronoun use in the following phrases: *between you and me/between you and I.* This phrase *between you and ____* showcases the controversy of writers using the subjective or objective pronoun case with a preposition.

Other pronoun issues cannot be researched with just text archives. The issues of (a) *their* being used as a singular 3rd person pronoun, (b) gender and pronouns, and (c) *who* versus *whom* in the objective case can all be studied in a corpus when one is available.

*Wrong Word*

Many controversial words need to be researched in a corpus rather than a text archive: (a) *hopefully* as a sentence modifier, (b) *infer/imply*, (c) *accept/except*, and (d) *less/fewer*. All of these need to be researched qualitatively as well as quantitatively.

On the other hand, some words can be studied in a text archive. The following are two examples:

1. *toward/towards*
2. *regardless/irregardless*

*Research Statistics*

Research Question 2a asks What is the percent usage of each item? This study's innovative way of searching for usage in large corpora can yield reliable results because the searches can be easily replicated. The results from this kind of search also differ from typical corpus searches. The frequency count of a particular usage is not a word count like most corpora search results. Most linguistic corpora count a word every time it appears, even when the word appears many times in one document. When researchers are working with word searches in linguistic corpora, they often present the results as 10 occurrences per 10,000 words, or some similarly normed frequency method. The problem with word counts in a smaller corpus is that one document may use the word 100 times, and another only 10 times. The document with the most entries has the most influence on the results in traditional corpus studies. The three chosen text archives, on the other hand, count only how many texts contain a particular usage. This way, each text has an equal chance of influencing the results, no matter how many times the word in question is used in that text. Each text

gets weighted the same as the next. This way of counting gives a more accurate description of how widespread a particular usage is.

Next, the results show percentages of use, rather than merely a general pronouncement of what is right and what is wrong. This innovative way of presenting actual usage (innovative at least among usage experts) vividly demonstrates what percentage of writers and editors uses one usage or spelling, and what percentage uses another. A 95% confidence interval is provided for each item so we can estimate the true value of usage in the population of all edited and printed writing in the United States in magazines, newspapers, and scholarly journals from January 1, 2000 to April 30, 2007.

Research Question 2b asks Does the usage differ among the registers of newspapers, magazines, and scholarly journals? This study examines the variation within the different registers. A chi-square test ($\chi^2$) on the various registers' total use of a word will show if the registers significantly differ from each other. The null hypothesis is that the percentage of usage is the same throughout all registers, but a rejection of the null hypothesis at $p < 0.05$ supports the hypothesis that there is a difference in the registers. If the difference is statistically significant, then we can be at least 95% sure that the differences among the registers are because of actual differences in the registeres and not because of a sampling error.

A confidence interval will be calculated for each averaged total percentage for each word. This interval shows the range of percentages which will include the true percentage of use, at the 95% confidence level. The formula is as follows, with $\hat{p}$ representing the average of the three registers' percentage of use.

$$\hat{p} = \frac{p_1 + p_2 + p_3}{3}$$

$$\text{where} \quad p_1 = \text{proportion in newspapers} \qquad (1)$$

$$p_2 = \text{proportion in magazines}$$

$$p_3 = \text{proportion in academic journals}$$

$$\hat{p} \pm \frac{1.96}{3} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} + \frac{p_3(1-p_3)}{n_3}}$$

Instructional designers, teachers, test makers, and usage experts can now determine how common one usage is over another. This empirically determined knowledge can contribute to the content validity of tests that are used to identify students' knowledge of correct usage. It can help evaluators determine the content validity of some of the material taught in English and writing programs.

If the item in question has divided usage, then instructional designers, teachers, and textbook makers need to note that students can confidently choose either usage and still be correct. Exams should not test for correctness on items that have divided usage. On the other hand, if the usage is predominantly one way and not any other, then that item is one that can be taught as the correct way, and that item can be confidently tested. This dissertation will use 95% as the cutoff point of percent of usage to be considered standard written American English. If the usage of one item is 90% to 95%, then researchers should carefully consider if the usage is

standard or not. Any usage under 90% is not dominant enough to be considered the only correct usage in standard written American English.

*Cutoff for Considering an Item as Standard English*

Research Question 2c asks What usage or usages should be considered standard written American English? If we empirically research controversial items, then we can be more confident about what usage items to teach and what items to test. If the usage is divided between several possibilities, then educators and testers should be aware of the several possible accepted usages and not consider one correct and one incorrect. On the other hand, if published writers choose one usage 99% of the time, and other usage 1% of the time, then educators and testers can confidently say one way is clearly accepted in standard written American English and one way is not.

Most usages and spellings in edited standard English will have 100% compliance. For example, the only time an author in a scholarly journal uses *we was* is in quoting dialogue or using an obvious dialect; otherwise, *we was* has a 0% occurrence rate in scholarly journals according to this dissertation's text archive search. This is an example of a usage that is common in non-standard speech but never occurs in writing in standard English, except in quotations of dialogue or examples of dialect. Another example is the non-standard word *talkin'*. This word is spelled *talking*, of course, in standard English, but people representing dialects or casual dialogue will use *talkin'*. The word *talkin'* occurs often in representations of speech, but when authors using standard English are not quoting speakers or imitating dialect, they use *talking* 100% of the time in written standard English. These two

examples represent the thousands of spellings and usages that have 100% agreement in written, published standard English.

Some spellings and usage items, though, have several forms that occur in magazines, newspapers, and scholarly journals. Most native speakers and writers learn standard English by hearing or reading many examples of the language. People can also read usage handbooks or textbooks to learn about certain controversial items in standard English, but most people learn standard English by example. Native speakers learn language as it is culturally transmitted from person to person.

For people who learn a first or second language by example, frequency of occurrence can be crucial. As Cunningham and Stanovich have shown in three separate studies, children's and adults' spelling abilities are affected by print exposure (Cunningham & Stanovich, 1991, 1993; Stanovich & Cunningham, 1993). People who are exposed to more print material can spell better. If people are affected by how much they see a word spelled in printed material, then they will probably be affected by seeing variations in that spelling.

Hauk et al. (2006) have tested the brain activity of people who read words that vary in their "orthographic typicality, that is, in the frequency of their component letter pairs (bigrams) and triplets (trigrams)" (p. 818). They found that when a person reads a word spelled with atypical orthography, a different part of the brain has more activity (left anterior inferior temporal cortex) than the part that is activated when a person sees the word spelled with typical orthography (left perisylvian cortex). In other words, our brains process typical and atypical spellings in different places, and the atypical spellings elicit stronger brain activity. In comparing actual words and

pseudowords (made up words that still follow English orthographic patterns), these researchers found that pseudowords elicit stronger brain activity in readers than actual words elicit. This research shows that readers' brains especially notice words that are atypical and words that are spelled in atypical ways.

Nick Ellis has also studied how frequency affects the way we learn language. He claims that our acquisition of language is exemplar-based, and frequency of occurrences shapes how we learn the language:

> Frequency is thus a key determinant of acquisition because "rules" of language, at all levels of analysis (from phonology, through syntax, to discourse), are structural regularities that emerge from learner's lifetime analysis of the distributional characteristics of the language input. Learners have to *figure* language out. (p. 144)

Ellis explains that learners must have sufficient examples of linguistic constructions to learn the rules of the language:

> For language learners to be accurate and fluent in their generalizations they need to have processed sufficient exemplars that their accidental and finite experience is truly representative of the total population of language of the speech community in terms of its overall content, the overall frequencies of that content, and the mappings of form to functional interpretation. The enormity of the lexical pool, the range of frequencies from 60,000 per million down to 1 per million and below, and the wide range of different linguistic constructions, when considered from the point of view of sampling theory, makes it clear that the necessary representative experience for fluency must be vast indeed. (p. 167)

These studies discuss the frequency of words and the effect of this frequency on readers. If frequency of occurrence affects how well we know a spelling or a usage, then we must pay attention to the frequency of occurrence to find out what usages and spellings people learn from reading edited, published English. If a person comes across usage or spelling in print material that is used or spelled in different ways, then there will not necessarily be one clear way to spell or use words.

Frequency of use becomes an issue when considering what is standardized English. If one use or spelling is found 10% of the time in edited published English, then a person will see this use or spelling in 1 out of 10 texts that have instances of the word or words in question. This could be a high enough frequency to confuse people so they are not absolutely certain what usage or spelling of the word is standard English. If, on the other hand, a particular usage or spelling is found only 1% of the time in edited, published English, then a reader will come across the particular usage or spelling in 1 out of 100 texts that have instances of the word or words in question. This is a much more isolated usage that has a much smaller effect on what an English speaker/writer considers to be standard. For these reasons I have set a high level of 90% for declaring a usage standard.

*Effects on Teaching and Testing*

Research Question 3 asks How should this knowledge of standard written American English affect what we teach and test in schools? Test makers must be very careful to test standard English as people actually see it and use it, not just how a few people say it should be used. A spelling or usage that occurs in 95% to 100 % of the texts that include any instances of the word or words in question should be considered standard English. Spellings or usages that occur between 90% and 95% of the time should be looked at carefully as being on the border of standard English, but the variants might have to be considered as alternatives to the standard. Words or spellings that occur below 90% of the time should probably not be considered the only choice of standard English, and because the frequency of the alternate form is much higher, neither usage should be tested as the only form of the word.

Instructional designers, teachers, test makers, and usage experts can now determine how common one usage is over another. This empirically determined knowledge can contribute to the content validity of tests that are used to identify students' knowledge of correct usage. It can help evaluators determine the validity of some of the content taught in English and writing programs.

*Limitations*

The method of using text archives to discover usage has great advantages, but it also has limitations. First, the technology is not developed enough to research every question. Many of the previously cited studies about errors in college students' writing list wrong punctuation as a common mistake students make, but text archives cannot search punctuation at this time. Text archives cannot search for parts of speech or collocates more than one word away, as linguistic corpora can. Homonyms complicate searching; for example, "dove" is not only a past tense of "dive," but it is also a bird and a soap brand. Searching with collocates that appear immediately next to the word eliminates some of the problems, but not all. I use "he dove" to avoid the problem. I am, therefore, currently limited in what questions about usage I can ask.

The *BYU Corpus of American English* will be available in 2008, and this corpus will be able to search parts of speech, collocates within ten words in either direction of the target word, and punctuation, so in the future many of the limitations of researching text archives will be eliminated.

Another limitation of performing a quantitative study of usage in a text archive is that some controversial usages must be studied qualitatively, such as when do people use *farther* and when do they use *further*. A researcher must look at each

instance to see which meaning an author is using, or the researcher must search for a collocate that will give a clue about the author's usage.

Some prescriptivists would object to the methodology that considers the majority usage to be any kind of standard. These people think only experts who are sensitive to word usage should decide questions of usage. This dissertation compromises by not considering all written language, but by considering only edited and published writing from three different genres. This limits the texts to standard English and expands the expert pool to people whose writing has been edited and published in (a) newspapers, (b) magazines, or (b) scholarly journals. Most people would consider the language in these genres to be standard English, although in three different registers.

Chapter 4: Results

This chapter will show the results from text archive searches of standard written American English in three different registers. The searches are limited to edited and printed works published between January 1, 2000 and April 30, 2007. These searches will provide the results of actual usage in the sample texts, and the statistics will show how reliable the results are for the entire population of all standard written American English. By using this method of researching standard English, we can begin to describe actual usage. Most usage items have also been researched in a fiction register that contains 2,630 short stories published from 2000-2007. Because this is a rather small archive, the results will be presented separately and not added into the composite average.

The entries will include a history of each word and a survey of the opinions of writers of different usage books. The entry will first present a history of the word or phrase, as found in the *Oxford English Dictionary* (2nd ed., Gove, 1989, abbreviated *OED* and accessed online so no page numbers included), since it is the most comprehensive record of the English language. Each section also reports what *The Chicago Manual of Style* (15th ed., 2003, abbreviated *Chicago*) or *Garner's Modern American Usage* (2003, abbreviated *Garner*) says about each item. *Chicago* has included a usage chapter in its new edition, and the usage chapter is written by Garner, so *Garner* and *Chicago* represent mostly one view of usage, but they do not always comment on the same usage items.

Each entry also mentions either *Webster's Dictionary of English Usage* (1989, abbreviated *Webster*) or *Merriam-Webster's Collegiate Dictionary* (11th ed., 2004,

abbreviated *Merriam-Webster*).  Since these are from the same publishing company,
they offer mostly the same opinion, but some items appear only in dictionaries and
not in usage books.  *The American Heritage Dictionary of the English Language* (4th
ed., 2000, abbreviated *American Heritage* and accessed electronically, so no page
numbers are given) represents a different dictionary philosophy, so it is referenced.
Since newspapers generally follow *The Associated Press Stylebook and Briefing on
Media Law* (2002, abbreviated *AP*), this dissertation comments on AP rulings.

     The method chapter explained how usage items were chosen for this study.
The results chapter includes for each entry a table, which lists the raw data and the
percentage of usage in each of the three registers, (a) newspapers, (b) magazines, and
(c) scholarly journals.  Then the entry includes a figure, which shows the average of
the combined usages.  Each average of combined usages also has a confidence
interval, which shows the range of percentages that we are 95% sure contains the true
percentage of usage in all edited and published works in the three registers. A chi-
square statistic for each entry shows if the difference among the registers is
statistically significant, with $p < 0.05$.

     For the purposes of this dissertation, the cutoff for standard English is 95%
usage and above, and anything that occurs 90% to 95% should be examined carefully.
Any item that occurs in more than 10% of the texts should be considered as a
plausible variant of standard English.

     The usages that are clearly in standard written American English can be
included in textbooks, handbooks, and dictionaries for students.  Other words or
phrases that include divided use should be presented as words that have a variant

form.  Teachers, textbook writers, instructional designers, and test makers should allow for the variety of standard English and not require uniform usage when a variety exists in actual written standard English.

<div align="center">*Spelling*</div>

Most words have only one correct spelling; some words, though, have several possibilities.  Dictionaries will list the most preferred spelling first and an alternate spelling next.  Usage books will sometimes comment on a spelling, but dictionaries are the authority.  A search of the text archives shows us what spellings people really use and in what proportion. The first group of words, *through* and *thru, catalogue* and *catalog,* and *technic* and *technique,* vary in the letters used.  The next group of words, *a lot* and *alot, cannot* and *can not,* and *all right* and *alright* vary with spacing.  The last words, *e-mail* and *email,* vary in hyphen use.

"*Through" and "thru."*  The *OED* shows the spelling of this word as having a –*gh* or –*ch* ending most of the time since it was first used.  These letters represent a fricative sound that has dropped out of English but is still represented in the spelling. Nevertheless, the *OED* shows one spelling variant as *thru* in the 14th century.  The *OED* states that *thru* is "now used informally as a reformed spelling and abbreviation (chiefly) in N. Amer."

According to *Webster* (p. 906), *thru* was a proposed spelling from several organizations that tried to reform spelling:  (a) the American Philological Association in 1876, (b)) the National Education Association in 1898, (c) the Simplified Spelling Board in the early 1900s, and (d) the *Chicago Tribune* from 1935-1975.  Because of the spelling reform movement, some newspapers and magazines used the simplified

spelling of *thru* in the past. Neither *AP* nor *Chicago* mentions the issue, and

*Merriam-Webster* mentions only that *thru* is a variant of *through*. *American Heritage*

also mentions *thru* as a variant of *through*. *Garner*, though, mentions that *thru*

"should be shunned," and that this variant spelling "oddly . . . appears in parts of the

Internal Revenue Code" (p. 289).

The text archive search reveals that, indeed, the use of *thru* in standard written

American English is extremely uncommon. Table 1 shows how thoroughly *through*

dominates the spelling. The resulting $\chi^2$ (2, $N$ = 1,054,670) = 1731.52,

$p < 0.0001$ indicates that there is a significant difference among the registers, even

Table 1

*Number of Instances of "Through" and "Thru"*

| Register | through | thru | Total |
|---|---|---|---|
| Newspapers | 26,593 (99.87%) | 336 (0.13%) | 266,929 |
| Magazines | 390,058 (99.71%) | 1,122 (0.29%) | 391,180 |
| Scholarly Journals | 635,802 (99.88%) | 759 (0.12%) | 636,561 |

though the percentages of use are very similar. Since the sample size is so large, the

small observed difference is statistically significant, but probably is not important for

practical purposes.

Figure 1 shows the domination of *through* over *thru*. The 95% confidence

interval is extremely small because there are so many instances of the words: 99.81%

to 99.83% for *through* and 0.17% to 0.19% for *thru*. Fiction also has a

predominance of *through* spellings (2,427, 99.67%) compared to *thru* (8, 0.33%).

Fiction and magazines show that less formal written language has a slightly greater

chance of containing *thru* than more formal written language; nevertheless, *thru* is

```
                                                              |
           through 99.82%                                     |   thru 00.18%
                                                              |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                              50%                              100%
```
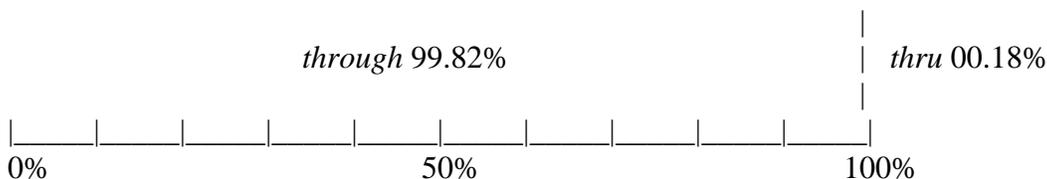
*Figure 1.*  Total percent usage of *through* and *thru*.

uncommon and used mostly in advertising and business names, such as a *drive-thru*

(probably because *thru* looks faster than *through* ).  Obviously the spelling reform of

*through* as *thru* did not catch on in printed standard English.  Dictionaries,

instructional designers, teachers, writers of textbooks, and test makers should

consider *through* the standard.

> *"Catalog" and "catalogue."*  According to the *OED,* the spellings *catalogue*

and *catalog* have both existed since the 16th century; indeed, even as early as the 15th

century, we can find *–ge and –gue* spellings.  Surprisingly, after the 16th century,

*catalog* does not appear again until the 19th century, so *catalogue* has a longer

continuous history in English spelling.   The word itself is borrowed from the French

word *catalogue*, which comes from the Latin word, *catalogus.*  Obviously, the *-gu*

spelling existed long before the English language borrowed it.  Now in the late 20th

and early 21st century, what do the experts say the correct spelling is?

*Merriam-Webster* and *American Heritage* put *catalog* as the first spelling and *catalogue* as the second spelling, meaning they prefer *catalog*. *Chicago* does not mention this word separately, but it spells it only one way in its book: *catalog*. On the other hand, *Garner* prefers *catalogue* as "the better form" (p. 133), especially for traditionalists, and he notes in the *–agog(ue)* entry that people like William Safire have identified the dropping of the *–ue* in several words, such as *dialogue* and *demagogue* (p. 30), so the dropping of the *–gue* in *catalogue* will probably continue. *AP* specifies *catalog*, so we expect the newspapers to use this spelling.

This text archive study shows interesting results for this word. Table 2 displays the data from newspapers and scholarly journals. ProQuest, the search engine for the magazine corpus, will not distinguish between *catalogue* and *catalog* because it groups the two spellings as one spelling, so we are left with only newspapers and journals. Since the *AP* style designates the spelling as *catalog*, the

Table 2

*Number of Instances of "Catalog" and "Catalogue"*

| Register | catalog | catalogue | Total |
|---|---|---|---|
| Newspapers | 4,063 (86.87%) | 614 (13.13%) | 4,677 |
| Scholarly Journals | 14,581 (55.78%) | 11,559 (44.22%) | 26,140 |

newspaper usage, not surprisingly, favors *catalog*; the scholarly journals, though, have an almost equally divided usage. More formal writing, then, has more instances

of the more formal spelling of *catalogue*. The $\chi^2$ (1, $N = 20{,}817$) = 1,604.76, $p <$ 0.0001 indicates that statistically the spelling is significantly different between categories, meaning that the spelling choices in newspapers and in scholarly journals are significantly different.

Figure 2 shows the averaged total percentages of the different spellings. The 95% confidence interval for these words is from 70.73% to 71.93% for *catalog* and from 28.07% to 29.27% for *catalogue*, which are both very small confidence

```
                                        |
        catalog 71.33%                  |   catalogue 28.67%
                                        |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                              50%                              100%
```
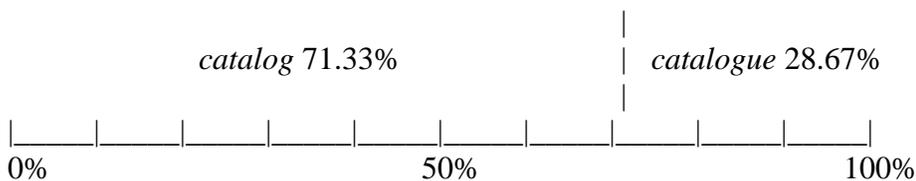
*Figure 2*. Total percent usage of *catalog* and *catalogue*.

intervals. The shorter spelling of *catalog* is indeed preferred more than *catalogue*, but in scholarly, formal writing, either spelling is acceptable, according to these data. The fiction search shows the spelling equally divided with 44 (48.35%) of fiction writers choosing *catalog* and 47 (51.65%) choosing *catalogue*. Writers of textbooks, dictionaries, and tests should not prefer *catalog* to *catalogue,* since respected texts use either spelling.

*"Technic" and "technique."* The *OED* gives each a word a separate entry. The *–ic* ending comes from Latin and Greek, and the *–ique* ending comes from the French. Both spellings were used for the noun version of the words in the 1800s. The adjective use of the word *technic* has existed since 1612, but it is of rare use

today; most people use just the noun version. The two spellings are not mentioned in *Webster, AP,* or *Chicago*, but *Garner* claims that *technique* is standard, and *technic* is a "variant spelling to be avoided" (p. 774). *Merriam-Webster* gives the words separate entries, but defines *technic* as *technique*, and not vice versa. *American Heritage* lists *technique* as the main spelling, but it also lists *technic* in the same entry.

Table 3 shows that *technic* is almost non-existent in (a) newspapers, (b) magazines, and (c) scholarly journals. Occasionally someone will use *technic*, but a heavy majority of the writers choose *technique*. The $\chi^2$ (2, $N = 409,261$) = 65,330,772, $p < 0.0001$ indicates that there is a statistically significant difference among the registers, even though the difference among the percentages is very small.

Table 3

*Number of Instances of "Technic" and "Technique"*

| Register | technic | technique | Total |
| --- | --- | --- | --- |
| Newspapers | 29 (0.27%) | 10,607 (99.72%) | 10,636 |
| Magazines | 22 (0.04%) | 60,053 (99.96%) | 60,075 |
| Scholarly Journals | 1,171 (0.35%) | 337,379 (99.65%) | 338,550 |

Figure 3 shows the average percentage of use for each word, and, again, *technique* is heavily favored. The 95% confidence interval for *technic* is 0.00% to 0.59%, and the confidence interval for *technique* is 98.07% to 100.00%. These

```
                                                        |
            technique 99.78%                             |   technic 0.22%
                                                        |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                              50%                              100%
```
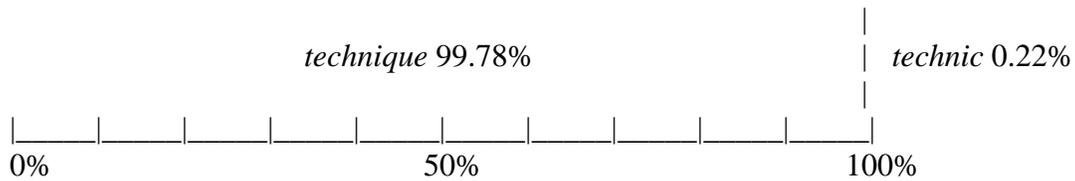
*Figure 3.*  Total percent usage of *technique* and *technic*.

are small confidence intervals because there are so many instances of the words.  The

fiction archive also favors *technique* with 44 (100.0%) instances and *technic* with 0

(0.0%) instances.  All indicators show that *technique* is favored so heavily that

*technic* is a rarity in standard written American English.  Writers of textbooks,

dictionaries, handbooks, and tests should consider *technique* to be standard.

     *"A lot" and "alot."*  *Webster* and *Garner* claim that the one-word version in

printed works is merely careless and the two-word version is standard.  The *OED,*

*Merriam-Webster, American Heritage, Chicago,* and *AP* do not even mention the

usage.  Table 4 shows the results of the text archive searches.  Once again, the

newspapers and magazines are more standardized, and the scholarly journals show a

little more variety with 0.40% using *alot.*  The $\chi^2$ (2, $N = 429,860) = 603.47, p <$

0.0001 indicates that there is a statistically significant difference among the registers,

although they are very close in percentage of use.

     Figure 4 shows the total percentages for each usage.  The 95% confidence

interval shows the estimated true value of the total usage of *a lot* to be between

Table 4

*Number of Instances of "A Lot" and "Alot"*

| Register | a lot | alot | Total |
|---|---|---|---|
| Newspapers | 167,333 (99.98%) | 30 (0.02%) | 167,363 |
| Magazines | 173,528 (99.87%) | 218 (0.13%) | 173,746 |
| Scholarly Journals | 88,407 (99.60%) | 353 (0.40%) | 88,760 |

99.80% and 99.83%, while the estimated true value of *alot* to be between 0.17% and

0.19%. This small interval is to be expected when the sample size is so huge.  The

```
                                                      |
         a lot 99.82%                                 |   alot 00.18%
                                                      |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                            100%
```
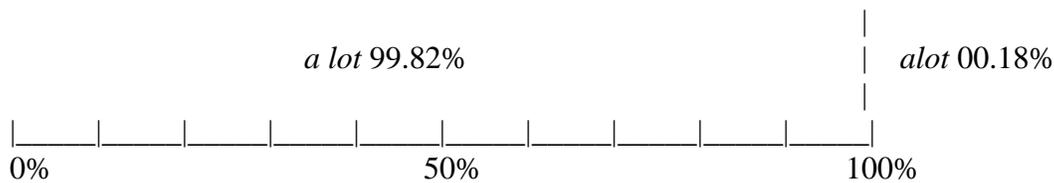
*Figure 4.*  Total percent usage of *a lot* and *alot.*

totals show that in standard written American English, *a lot* is the standard way of

spelling the word.  The fiction archive has 1,185 (100%) writers using *a lot,* and 0

(0%) using *alot.*  Because the text archives demonstrate such a large disparity

between the two words, teachers; writers of textbooks, handbooks, and dictionaries;

and test makers can consider *a lot* to be standard and to be an item that can be taught

and tested.

     *"Cannot" and "can not."*  The *OED* shows in the *can* entry that *cannot* and

*can not* spellings have both existed since the 14th century.  *Webster* states that "both

spellings are acceptable, but *cannot* is more frequent in current use" (p. 219). The two-word form can be used to place special emphasis on *not*. A study of the usage books shows that (a) *Merriam-Webster* chooses *cannot* over *can not*, (b) *American Heritage* does not mention it, (c) *AP* designates *cannot* as the correct form, (d) *Chicago* does not comment on the word, and (e) *Garner* says *cannot* is preferable.

The results from the archive search reveal how entrenched *cannot* is in standard written American English. Table 5 shows that all registers—(a) newspapers, (b) magazines, and (c) scholarly journals—heavily favor *cannot.* The $\chi^2$ (2, *N* = 462,898) = 46,240,548, *p* < 0.0001 indicates that there is a significant difference among the three registers. Fiction also shows that most writers choose *cannot* 614 (97.15%) as opposed to *can not* 18 (3.85%). Surprisingly, in the most formal writing and the least formal writing, we find instances of *can not.*

Table 5

*Number of Instances of "Cannot" and "Can Not"*

| Register | cannot | can not | Total |
|---|---|---|---|
| Newspapers | 22,725 (97.04%) | 694 (2.96%) | 23,419 |
| Magazines | 91,040 (97.92%) | 1,938 (2.08%) | 92,978 |
| Scholarly Journals | 323,191 (93.27%) | 23,310 (6.73%) | 346,501 |

Figure 5 shows the total averaged percentages for the three registers. The 95% confidence interval is quite small with *cannot* at 95.98% to 96.16%, and *can not* at 3.83% to 4.02%. *Cannot* usage is obviously dominant in standard written

```
                                                     |
           cannot 96.07%                             |   can not 03.93%
                                                     |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                            100%
```
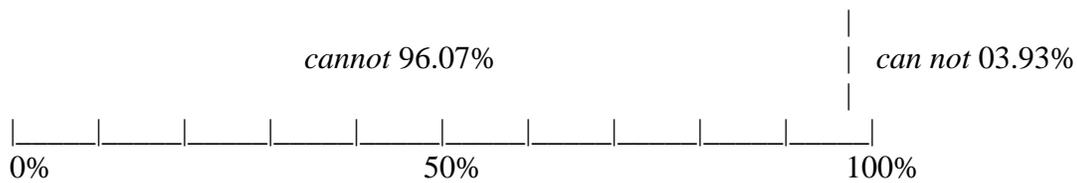
*Figure 5.* Total percent usage of *cannot* and *can not.*

American English, but *can not* still exists in a small percentage of texts. Since *cannot*

is so dominant, most people consider it to be the correct form. Anyone who uses *can*

*not* must realize that the spelling is not common enough to be considered standard,

but *can not* seems more acceptable when the *not* needs to be emphasized.

Instructional designers, textbook writers, teachers, and test writers should probably

consider *cannot* to be standard written American English, unless a person is

emphasizing the *not*.

   *"All right" and "alright."* The *OED* lists *alright* as a variant form of *all right*

and shows it being in use since 1893, with very early usages during Middle English.

*Webster* presents the extensive history of the debate over the spelling. The editors

state that both usages are acceptable, but *all right* is found more often in print. They

also warn the reader that most usage handbooks say that *alright* is wrong. The

controversy over the correct spelling has been strong since the early 20th century.

Some people think that *alright* is formed by analogy with *already, altogether,* and

*although. Webster* states that *alright* shows up in newspapers and business

publications most often. Usage handbooks, according to *Webster*, have labeled the

spelling of *alright* as colloquial or illiterate, yet "no very cogent reasons are presented

for its being considered wrong" (p. 79).

Merriam-Webster lists *alright* as a variant of *all right* and notes that many

consider *alright* to be wrong, but the spelling does occur sometimes in "fictional

dialogue" (p. 34) and other writing. *AP* states that the word must always be spelled *all

right* and never *alright*. *Chicago* says to use only the two-word version, and *Garner*

says the one-word version may becoming more acceptable in British English, but in

American English the one-word version "cannot yet be considered standard—or even

colloquially all right" (p. 35). *American Heritage* admits that *alright* is used by many

well-known authors, but in a usage note the editors proclaim that *alright* "has never

been accepted as a standard variant."

Since these two spellings have brought on so much controversy, a search of

text archives will bring a more empirical view of how these two words are used. In

the electronic search, I noticed that many usages of *all right* were actually part of the

phrase *all rights reserved*. The number of *all rights reserved* usages was subtracted

from the total usages of *all right* for a more accurate count. Table 6 shows the total

number of usages among the three different registers. The $\chi^2$ (2, $N =$ 31,339) =

1,621.99, $p < 0.0001$ indicates that the difference among the registers is

statistically significant. Obviously, scholarly journals have a much higher use of

*alright* than either newspapers or magazines, contrary to what *Webster* indicates.

Ironically, the most formal register has the highest percentage of usages of the

spelling that most dictionaries and usage books consider non-standard. Figure 6

shows the average percentage use of all the registers. The 95% confidence interval for

Table 6

*Number of Instances of "All Right" and "Alright"*

| Register | all right | alright | Total |
|---|---|---|---|
| Newspapers | 10,532 (94.87%) | 570 (5.13%) | 11,102 |
| Magazines | 11,409 (93.98%) | 731 (6.41%) | 12,140 |
| Scholarly Journals | 3,968 (73.73%) | 1,414 (26.27%) | 5,382 |

*all right* is 87.08% to 87.96%, and *alright* is 12.16% to 13.04%. In fiction, *all right* is

used in 922 (93.51%) instances, and *alright* is used in 64 (6.49%) instances.

```
                                       |
            all right  87.52%          |  alright 12.60%
                                       |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                          100%
```
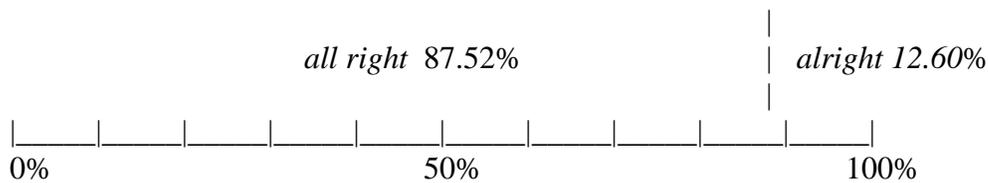
*Figure 6.* Total percent usage of *all right* and *alright.*

*Webster* claims that *alright* is replaced by *all right* by copy editors constantly.

*Webster's* editors conjecture that if copy editors did not make that change, we would

see *alright* many more times. If 95% usage is considered standard, and 90% to 95%

considered borderline, then *alright* is a standard variant, not a non-standard variant in

standard written American English. Consequently, (a) teachers, (b) dictionaries, (c)

handbooks, and (d) textbooks should allow the use of *alright*, and tests should not

consider this a wrong spelling.

*"Email" and "e-mail."* The *OED* shows both *email* and *e-mail* have been used since 1982. *Merriam-Webster, AP,* and *Chicago* use *e-mail.* *Webster* was written before there was *e-mail,* so it has no comment. *American Heritage* lists the spellings in this order: (a) e-mail, (b) email, and (c) E-mail. *Garner* comments that *e-mail, E-mail,* and *email* all exist, but "*e-mail* is five times as common as *email*" (p. 204). He predicts that the hyphen will drop off eventually. The archive search shows that *e-mail* is much more common than *email*. Table 7 shows the results of the searches. The $\chi^2 (2, N = 577,307) = 11,393.99$, $p < 0.0001$ indicates that the null hypothesis that the registers are all the same is rejected, and, consequently, the registers show difference in usage. Newspapers and magazines are very regular in using *e-mail.* Scholarly journals seem to allow more variation, and so the results show that 15.37% of journals that include the word spell it *email.*

Table 7

*Number of Instances of "E-mail" and "Email"*

| Register | e-mail | email | Total |
|---|---|---|---|
| Newspapers | 31,401 (96.26%) | 1,220 (3.74%) | 32,621 |
| Magazines | 65,801 (97.02%) | 1,606 (2.38%) | 67,407 |
| Scholarly Journals | 403,931 (84.63%) | 73,348 (15.37%) | 477,279 |

Figure 7 shows the averaged percentages of each of the registers. The 95% confidence interval shows that the estimated true value of the proportion of *e-mail* is

92.64% to 93.04%, and the 95% confidence interval for *email* is 6.96% to 7.36%.

Even though the majority of articles in magazines, newspapers, and scholarly journals

```
                                                          |
                    e-mail 92.84%                         |   email 07.16%
                                                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                            100%
```

*Figure 7.* Total percent usage of *e-mail* and *email.*

use *e-mail*, the alternate spelling of *email* has a fairly good usage.  The fiction register

is the most divided so far with 105 (79.54%) instances of *e-mail* and 27 (20.45%)

instances of short story writers using *email.*

Many compound words begin with hyphens and over time lose the hyphen.

This historical trend, combined with the fact that *email* is faster to type than *e-mail*,

lead me to the conclusion that the more formal *e-mail* is temporarily preferred.  This

word is only about twenty years old, so spelling is not yet settled.  The percentage use

of *e-mail* is close to being standard, but the use is changing enough that people should

track its use closely. Writers of textbooks and tests should consider both spellings as

valid in standard written American English.

*Inflected Endings*

*Latin Plurals*

During the Renaissance, English-speaking scholars tried to upgrade English so

they could write their scholarly writing in English, rather than traditional Latin.  In

order to improve the English language, many people imported wholesale many words

from Latin.  Latin scholars enjoyed showing off their knowledge of Latin, so it is not

surprising that people imported not only the Latin words but also their Latin plural

endings.  Most Latin words added to English became assimilated into the language

and added the typical plural morphemes of *s* and *–es.*  Some words that kept Latin

plurals are in flux between the English plural and the Latin plural.  One way to

determine which plural should be used in standard written American English is to

check what plurals writers and editors use.

   *"Appendices" and "appendixes."*  The *OED* shows both plurals extant from

the 16$^{th}$ century, and it lists *–ices* as the preferred plural.  *Merriam-Webster* lists

*appendixes* as the preferred plural spelling, *Chicago* uses only *appendixes, American*

*Heritage* lists first *appendixes* and then *appendices,* and *AP* does not mention it.

*Webster* claims that both usages can be found in the United Kingdom and the United

States, and one is not preferred over the other. *Garner* states that either spelling is

acceptable, but "*appendixes* is preferable outside scientific contexts" (p. 54).

The text archive search shows how people actually use these two spellings today.

Table 8 shows that *appendixes* is actually not the preferred form in standard written

American English.  Scholarly journals heavily prefer *appendices*, which is expected,

since the writing is more formal.  Magazines are not represented because ProQuest

does not differentiate between the two spellings in searches.  Newspapers, a less

formal register, use *appendices* more than *appendixes*, but the newspaper register

showed far few instances of either word than the scholarly journal register.  The $\chi^2$ (1,

$N = 4,601) = 11.64$, $p = 0.0006$ indicates that the registers do not have unified usage,

but the two registers, newspapers and scholarly journals, have different usages.

Table 8

*Number of Instances of "Appendices" and "Appendixes"*

| Register | appendices | appendixes | Total |
|---|---|---|---|
| Newspapers | 17 (60.71%) | 11 (39.29%) | 28 |
| Scholarly Journals | 3,856 (84.32%) | 717 (15.68%) | 4,573 |

Figure 8 shows the total averaged percentages for the two words.  The 95%

confidence interval that shows the estimated true value of the usage of the two words

is appendixes, 18.38% to 36.58% and appendices, 63.42% to 81.62%.  The

```
                                        |
          appendices 72.52%             |   appendixes 27.48%
                                        |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                            100%
```

*Figure 8*.  Total percent usage of *appendices* and *appendixes*.

large confidence intervals are the result of low numbers of examples in the text

archive.  The fiction register has too few examples to be helpful.

The usage books prefer the word *appendixes*, and they say that this plural

ending is used much more than *–ices* except in scientific settings.  In this archive

search, though, we find that *appendices* is used more, even in newspapers.  The usage

books are wrong about prevalent usage.  Obviously, teachers, instructional designers,

and writers of textbooks, handbooks, and dictionaries can list both spellings, but they

must not consider one right and another wrong because of the divided usage.

*"Indexes" and "indices."* The *OED* lists *indices* as the preferred spelling, but it clarifies that *indexes* is preferred when referring to the list of topics at the back of a document. When the subject matter is mathematics or computers, the spelling is generally *indices*. *Merriam-Webster* lists both spellings, but prefers *indexes*; *American Heritage* also lists both spellings, but prefers indexes; *AP* lists only indexes; and *Chicago* does not list anything but uses only *indexes*. *Webster* does not even list it as a controversial item. *Garner* heavily prefers *indexes* and calls *indices* pretentious, except in mathematics and the sciences.

Table 9 shows the results of the text archive search. Again, ProQuest does not differentiate between the two spellings, so no numbers are available from the magazine archive. The newspaper register heavily uses *indexes*, but the scholarly journal register heavily uses *indices*. These data divide right down informal-formal lines. The $\chi^2$ (1, $N$ = 81,480) = 22,418.79 $p$ <0.0001 indicates that the null hypothesis that the registers have the same percentage usage is rejected, and we

Table 9

*Number of Instances of "Indexes" and "Indices"*

| Register | indexes | indices | Total |
|---|---|---|---|
| Newspapers | 9,799 (97.37%) | 265 (2.63%) | 10,064 |
| Scholarly Journals | 16,373 (22.92%) | 55,043 (77.07%) | 71,416 |

can accept the hypothesis that the registers significantly differ in usage percentages.

In order to research magazines more, which are not tied to one style like newspapers

are, I searched two other magazine text archives and found the following usages: in a

magazine search of Questia.com, *indexes* were 48.17%, and *indices* were 51.83%. In

magport.com, which has more conversation-like magazines, *indexes* were 68.33%,

and *indices* were 31.67%. This search of other archives demonstrates that the usage

is not as polarized as the initial archives show.

Figure 9 shows the total percentage of usage of the registers. The 95%

confidence interval for *indexes* shows that the estimated true value of the percent of

usage is between 59.95% and 60.31%. The 95% confidence interval for *indices* is

39.65% to 40.05%. The fiction corpus does not have enough examples of the words

to add any information to this study.

```
                             |
        indexes 60.15%       |   indices 39.85%
                             |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                          100%
```
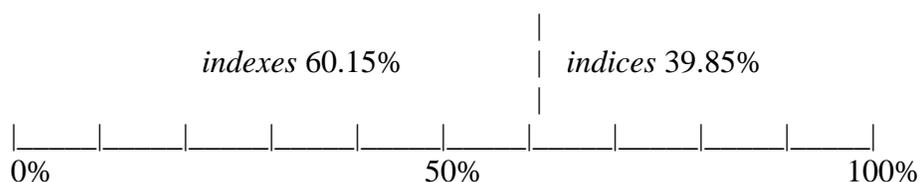
*Figure 9.*  Total percent usage of *indexes* and *indices.*

This study demonstrates that *indices* is heavily used in these three registers. In

standard written American English, *indexes* is used more often, but the usage is more

evenly divided than the usage books indicate. *Indices* is still very much alive in

formal and less formal usage, but in a newspaper register, *indexes* is the clear choice.

Writers of textbooks and tests should note that both usages are acceptable in standard

written American English, and both can be taught, but neither can be tested as the one

correct spelling. People should be sensitive to the register and context of use.

*"Syllabi" and "syllabuses."* The *OED* lists two plurals for *syllabus*: *syllabi*

and *syllabuses.* It shows no other variation in spelling over the years, and the two

examples of plural from 1881 and 1972 show only *syllabi* as the plural. *Merriam-*

*Webster* lists first *syllabi* and second *syllabuses* with no explanation. The item does

not appear in *Webster. American Heritage* lists first *syllabuses* and then *syllabi.* The

*AP* style lists only *syllabuses* as the plural. *Garner* declares that in American English,

*syllabuses* is preferred to *syllabi* twice as often., but in legal writing *syllabi* is used 10

times as often as *syllabuses.* The dictionaries and usage books clearly disagree about

preferred or standard usage.

The text archive search shows different results. Table 10 lists the percent

usage of the different plurals in all three registers. The $\chi^2 (2, N = 3{,}511) = 10.89$, $p =$

0.0043 indicates that there is a statistically significant difference among the

Table 10

*Number of Instances of "Syllabi" and "Syllabuses"*

| Register | syllabi | syllabuses | Total |
|---|---|---|---|
| Newspapers | 44 (58.67%) | 31 (41.33%) | 75 |
| Magazines | 499 (71.19%) | 202 (28.81%) | 701 |
| Scholarly Journals | 2,129 (75.10%) | 706 (24.90%) | 2,835 |

three registers.  Interestingly, even the newspaper register, which usually follows *AP* writing style, has a majority usage of *syllabi,* even though *AP* dictates *syllabuses* as the only plural.  Obviously, each register has a significant majority of users choosing *syllabi* over *syllabuses.*

Figure 10 shows the total average of all the registers.  A 95% confidence interval, showing the estimated true value of the usages in the full registers, is 64.4% to 72.2% for *syllabi* and 27.8% to 35.6% for *syllabuses.*  The fiction register has no instances of these words. Obviously, *Garner* is wrong.  The use of *syllabi* is overall about twice that of *syllabuses,* and this is especially true for magazines and scholarly journal registers.  The newspaper register has more equal usage.

```
                                        |
        syllabi 68.32%                  |   syllabuses 31.68%
                                        |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                               100%
```
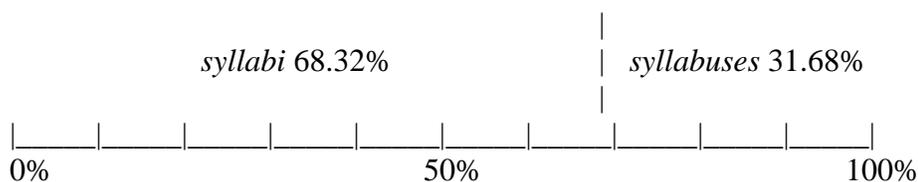
*Figure 10.*  Total percent usage of *syllabi* and *syllabuses.*

Since the usage is so divided, there is no clear single usage in standard written American English.  Teachers, textbook writers, and dictionary makers should be aware that more people use *syllabi* than use *syllabuses* in printed, edited English newspapers, magazines, and journals.  Test makers cannot designate one usage as correct and the other as incorrect.  It seems that the Latin plural is still strong in English usage, but the anglicized plural also has a strong presence.

*"Formulas" and "formulae."* When English speakers borrowed *formula* from the Latin, some writers borrowed the Latin ending, and some anglicized the plural of the word. The *OED* lists the first plural as *formulae* and the second as *formulas. Webster* acknowledges that both plurals exist in English, and the editors can find no pattern of use. *Merriam-Webster* lists *formulas* first, and then *formulae* second. *Chicago* does not list the word, and the *AP* style book lists only *formulas* as acceptable. *Garner* lists *formulas* as the acceptable plural everywhere but in scientific writing.

The text archive search shows how the plural is used in three different registers. This study limited the use of the two plurals to *formulas are* and *formulae are* to limit the samples to just the plural form. Table 11 shows the percent usage of the two plurals in newspapers, magazines, and scholarly journals. The $\chi^2$ (2, $N$ = 15,703) = 2,097.94, $p$ <0.0001 indicates that there is a statistically significant difference among the three registers. The newspaper register has the largest percentage of

Table 11

*Number of Instances of "Formulas" and "Formulae"*

| Register | formulas | formulae |
|---|---|---|
| Newspapers | 3,254 (99.06%) | 31 (0.94%) |
| Magazines | 5,308 (83.15%) | 1,076 (16.85%) |
| Scholarly Journals | 3,595 (68.47%) | 2,439 (31.53%) |

regular plurals.  Magazines use more Latinized endings than newspapers, and

scholarly journals use *formulae* 31.53% of the time.  As the writing gets more formal,

more people use *formulae.*  Although most people use *formulas, formulae* is not dead.

Figure 11 shows the average of the use of the two plural forms in all three

registers.  A 95% confidence interval, showing the estimated true value of the

percentage of use in the full registers, is 83.16% to 83.96% for *formulas* and 16.04%

to 16.84% for *formulae.*  The fiction register contains no instances.


```
                                                  |
            formulas 83.55%                       | formulae 16.44%
                                                  |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                              100%
```
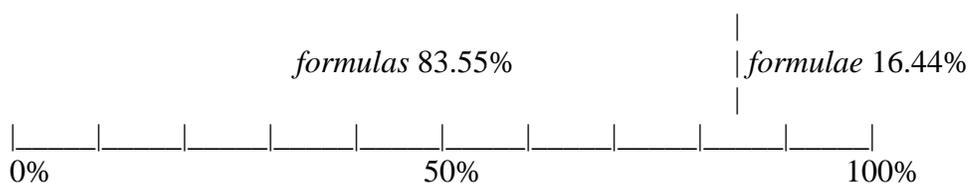
*Figure 11.*  Total percent usage of *formulas* and *formulae.*


Obviously, there is divided usage for the plural, and the Latin plural,

*formulae*, does not seem to be dying.  In the three registers, newspapers rarely use

*formulae,* and in scholarly writing, *formulae* appears more often.  Further research

can focus on which scholarly writing uses *formulae.*  This may be a plural found more

often in the sciences than in the arts and humanities, in which case the spelling should

be determined by context. The preferred plural is *formulas,* but *formulae* as a plural

has a presence.  Teachers, textbook writers, and dictionaries should list both plurals as

possibilities, noting that *formulae* can be found in more formal writing.  This is not an

item that should be included in tests because there is too much variance in the use.

*Comparatives and Superlatives*

In Old English and Middle English, people used inflections for the comparatives and superlatives. Today, these are the *–er* and the *–est* inflections. The use of *more* and *most* as periphrastic comparatives (adding a word rather than an ending) began in Middle English. The *OED* says the modern rule is that we use *more* or *most* with most words of two syllables and all words that have more than two syllables; otherwise, we use the inflectional endings *–er* and *–est.*

*"Prouder" and "more proud."* Not all words follow the general rules about when to add an inflectional ending and when to use a periphrastic comparative. Only the dictionaries mention the word *proud*: (a) The *OED* records the *–er* and *–est* inflections since Middle English; (b) *American Heritage* lists *-er* and *–est* as the proper inflections; and (c) *Merriam-Webster* lists no inflections, which means the comparative can be made with the word *more*. Because *proud* has only one syllable, it is obviously inflected, according to the standard rules, but a search of text archives shows that there is more variety than the dictionaries present.

Table 12 shows the results of the search by register. The $\chi^2\,(2, N = 1{,}947) = 10.50, p = 0.0052$ indicates that there is a statistically significant difference among the registers. Surprisingly, most registers show a close split in actual usage. *More proud* is used half the time in newspapers, and it is used more than 40% of the time in magazines and newspapers.

Figure 12 shows that a higher percentage of people use *prouder,* but a substantial percentage uses *more proud.* The confidence interval for the two different options is 51.45% to 57.19% for *prouder* and 42.81% to 48.55% for *more*

Table 12

*Number of Instances of "Prouder" and "More Proud"*

| Register | prouder | more proud | Total |
|---|---|---|---|
| Newspapers | 609 (49.84%) | 613 (50.16%) | 1,222 |
| Magazines | 307 (58.03%) | 222 (41.97%) | 529 |
| Scholarly Journals | 108 (55.10%) | 88 (44.90%) | 196 |

*proud.* The fiction register has a small number of instances, but the numbers are

similar to the other registers: 4 (40%) for *prouder* and 6 (60%) for *more proud.*

Dictionaries differ in their recommendations, and the text archive search shows that,

indeed, usage is closely divided between the two forms of the comparative.

Obviously usage handbooks and dictionaries should acknowledge the two variant

forms.

```
                              |
        prouder 54.32%        |        more proud 45.68%
                              |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                           50%                           100%
```
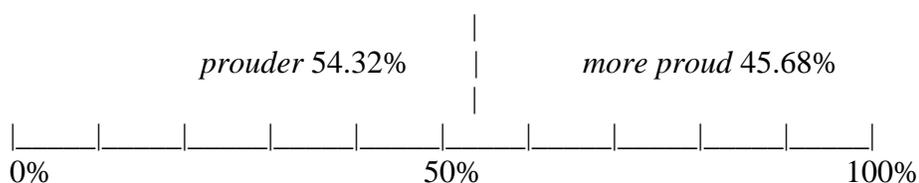
*Figure 12.* Total percent usage of *prouder* and *more proud*.

*"Proudest" and "most proud."* If the usage of the two different forms are

fairly equally divided in the comparative, does that trend carry over to the

superlative? The *OED* mentions only *–est*, *Merriam-Webster* does not mention any

inflections, and *American Heritage* lists only *–est*.

Table 13 shows the number and percentage of instances in the three registers. The $\chi^2$ (2, $N$ = 4,705) = 26.50, $p < 0.0001$ indicates that there is a statistically significant difference between the registers. The use of *proudest* predominates in every register. Interestingly, *proudest/most proud* appear in each of

Table 13

*Number of Instances of "Proudest" and "Most Proud"*

| Register | proudest | most proud | Total |
|---|---|---|---|
| Newspapers | 1,236 (53.09%) | 1,092 (46.91%) | 2,328 |
| Magazines | 1,013 (59.34%) | 694 (40.66%) | 1,707 |
| Scholarly Journals | 419 (62.54%) | 251 (37.46%) | 670 |

the registers two to three times more often than *prouder/more proud.* This choice may be because people are more comfortable using the superlative when writing about *pride.*

Figure 13 shows the averaged total usage amongst all three registers. The confidence intervals for the two usages are as follows: 56.72% to 59.92% for *proudest*, and 40.08% to 43.28% for *most proud. Proudest* seems to dominate *most*

```
                                    |
        proudest 58.32%             |      most proud 41.68%
                                    |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                                50%                               100%
```
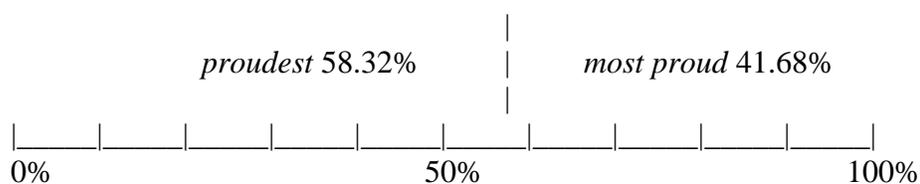
*Figure 13.* Total percent usage of *proudest* and *most proud.*

*proud* to a higher degree than *prouder* dominates *more proud*. The fiction register favors *proudest* with 8 (100%) instances, compared to *most proud* with 0 (0%) instances. This archive search shows that people favor *proudest*, but *most proud* is used often, too. Because of this high usage of *most proud*, dictionaries and handbooks should list both possibilities as the superlative.

　　*"Riskier" and "more risky."* Whether a writer should use the inflection *–er* or the periphrastic comparative *more* with a two-syllable adjective does not seem to be a controversy with certain words. This study focuses on two different two-syllable adjectives to test the idea that the way for making the comparative and superlative is standard across three different registers.

　　In the case of *risky*, the only books that mention the comparative and superlative are dictionaries. Both *Merriam-Webster* and *American Heritage* list *riskier* and *riskiest* as the only alternatives for the comparative and superlative. Since the comparative form of two-syllable adjectives is sometimes in flux, a study of the text archives can reveal what edited writers actually do.

　　Table 14 shows that, indeed, people use both the inflected comparative and the periphrastic comparative, with most people using the inflected comparative. The $\chi^2$ (2, $N$ = 4,606) = 482.45, $p$ <0.0001 indicates that the registers differ in their percent usage by a statistically significant amount. Considering that both dictionaries do not acknowledge the possibility of *more risky*, these statistics are rather surprising. *Riskier* is the more common form in all registers, but *more risky* appears in scholarly journals in 44.46% of the instances. This discrepancy represents a divided usage that dictionaries have not yet recorded.

Table 14

*Number of Instances of "Riskier" and "More Risky"*

| Register | riskier | more risky | Total |
|---|---|---|---|
| Newspapers | 571 (87.31%) | 83 (12.69%) | 654 |
| Magazines | 1,735 (83.94%) | 332 (16.06%) | 2,067 |
| Scholarly Journals | 1,047 (55.54%) | 838 (44.46%) | 1,885 |

Figure 14 shows the total percentages for all the registers. The 95% confidence interval, showing the estimated true value of the percent of usage in the complete registers of newspapers, magazines, and scholarly journals, is 74.35% to 76.85% for *riskier* and 23.19% to 25.69% for *more risky.* Clearly, both usages exist

```
                                    |
         riskier 75.60%             |    more risky 24.40%
                                    |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                            100%
```
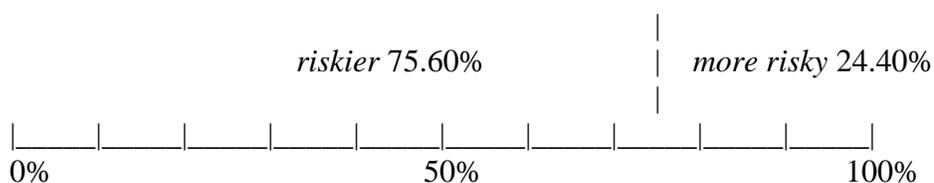
*Figure 14.* Total percent usage of *riskier* and *more risky.*

in all three registers of edited, written English. The fiction corpus shows 100% of the five instances use *riskier,* and 0% of the writers use *more risky.* Across the three registers, there is no definitive usage. Teachers, instructional designers, and textbook makers should not teach one usage as correct and the other as incorrect, and test makers and dictionaries should consider both usages as standard.

*"Riskiest" and "most risky."* When *risky* is made a superlative, do writers of edited, printed English use the inflected word or the periphrastic phrase? Again, only the dictionaries mention anything about the word, and both dictionaries list *–er* and *-est* as the only possibilities for the comparative and superlative. The text archives show a different story.

Table 15 shows that *riskiest* is used more often, but *most risky* is also in use. The $\chi^2$ (2, $N = 1,072$) = 207.09, $p < 0.0001$ indicates that there is a significant difference among the three registers. The scholarly journals show an almost equally the divided usage between the two superlatives, whereas the magazine and newspaper

Table 15

*Number of Instances of "Riskiest" and "Most Risky"*

| Register | riskiest | most risky | Total |
|---|---|---|---|
| Newspapers | 175 (91.62%) | 16  (8.38%) | 191 |
| Magazines | 501 (89.62%) | 58 (10.38%) | 559 |
| Scholarly Journals | 163 (50.62%) | 159 (49.38%) | 322 |

registers use *riskiest* much more often than *most risky*. It seems quite different that the scholarly journal register uses *most risky* almost as much as *riskiest*. This might be because *more risky* sounds slightly more formal, so writers might choose the more formal phrase in their formal writing. Curiously, *riskier* has three times more occurrences than *riskiest*. People seem to be more comfortable using comparative

inflection rather than the superlative one, or they compare only two things, not three

or more.

Figure 15 shows the total averaged percentages of usage in all three registers.

*Riskiest* is the clear favorite, but *most risky* is used substantially, too.  The 95%

confidence interval, showing the estimated true value of the usage in the three

```
                                          |
            riskiest 77.29%               |   most risky 22.71%
                                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                          100%
```
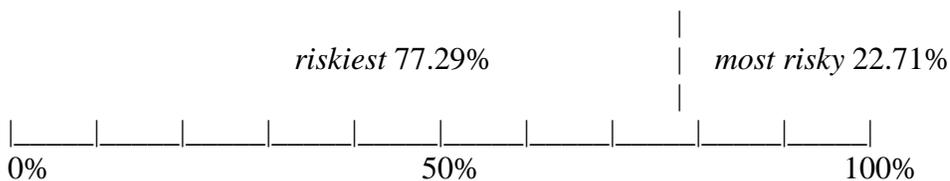
*Figure 15*.  Total percent usage of *riskiest* and *most risky*.

registers, is 74.89% to 79.69% for *riskiest* and 20.31% to 25.11% for *most risky*.  The

fiction corpus has only three recorded uses, and each use is *riskiest.*  Teachers,

instructional designers, textbook writers, and dictionary creators should note that

people use both *riskiest* and *most risky* in edited, written American English.

Dictionaries do not list all the possible ways of making this word a comparative or a

superlative, and this item does not follow the OED's guideline that two-syllable

words should generally add *most* rather than *–est.*   Dictionaries should adjust the

entries, and teachers, textbook writers, and test makers should consider both usages

correct.

*"Lovelier" and "more lovely."*  This study will search for one more two-

syllable adjective made into the comparative and superlative.  The question with the

two-syllable words is if the words are consistently made comparative and superlative

with adding the words *more* or *most* or if the forms have a lot of divided usage.

The *OED* lists the inflected forms only. *American Heritage* and *Merriam-*

*Webster* list only *–er* and *–est* as the possible inflections for *lovely*. Table 16 shows

Table 16

*Number of Instances of "Lovelier" and "More Lovely"*

| Register | lovelier | more lovely | Total |
|---|---|---|---|
| Newspapers | 97 (79.51%) | 25 (20.49) | 122 |
| Magazines | 190 (73.93%) | 67 (26.07%) | 257 |
| Scholarly Journals | 73 (62.93%) | 43 (37.07%) | 116 |

the results of the text archive search. The $\chi^2 (2, N = 495) = 8.63, p = 0.0134$

indicates that there is a statistically significant difference among the three registers.

Although there are not as many instances as other words studied in this dissertation,

the numbers still reveal a varied usage. All three registers show a preference for

*lovelier,* but the more formal register of scholarly journals shows over one-third of all

usages are *more lovely*.

Figure 16 shows the total average of the different registers. The 95%

confidence interval for these two options is as follows: 67.94% to 76.20% for *lovelier*

and 23.70% to 32.06% for *more lovely*. The fiction register shows 6 (50%) instances

of *lovelier* and 6 (50%) instances of *more lovely*. Clearly, the

```
                                         |
             lovelier 72.12%             |    more lovely 27.88%
                                         |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                             50%                              100%
```
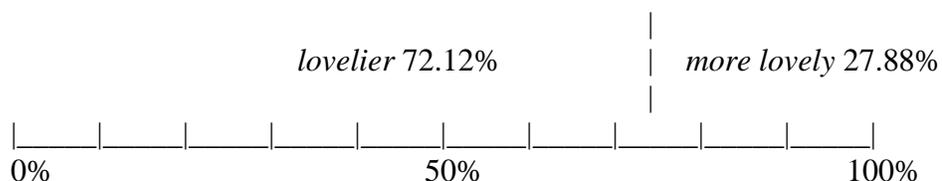
*Figure 16*.  Total percent usage of *lovelier* and *more lovely.*

usage of *lovelier* is preferred, but *more lovely* also occurs a little over one-fourth of

the time.  This text archive search shows that actual printed and edited texts use two

forms of this comparative.  Obviously no one should claim one form is correct and

the other incorrect.

　　　　*"Loveliest" and "most lovely."*  Does the same division hold for the

superlative versions of *lovely*?  This word is not considered controversial by anyone,

yet writers do not use a uniform comparative.  This item represents the many other

two-syllable words in English, and some of these words are sometimes inflected for

the comparative and superlative, and some of the words are used with *more* or *most*.

　　　　The *OED, American Heritage,* and *Merriam-Webster* list only the inflected

ending *–est* as the correct superlative.  No other possibility is mentioned.  Table 17

shows the results from the text archive search.  The $\chi^2 (2, N = 1{,}058) = 13.64$, $p =$

0.0011 indicates that the three registers vary from each other significantly.

There are about twice as many instances of *loveliest/most lovely* compared with

*lovelier/more lovely*.  Writers must be more comfortable with the superlative

rather than the comparative.  Once again, the formal register, scholarly journals,

contains more instances of the two-word superlative than the less formal registers

Table 17

*Number of Instances of "Loveliest" and "Most Lovely"*

| Register | loveliest | most lovely | Total |
|---|---|---|---|
| Newspapers | 188 (90.38%) | 20  (9.62%)) | 208 |
| Magazines | 525 (90.05%) | 58  (9.95%) | 583 |
| Scholarly Journals | 218 (81.65%) | 49 (18.35%) | 267 |

.

Figure 17 shows the averaged totals among the three registers.  The 95% confidence intervals are as follows: 85.16% to 87.56% for *loveliest*, and 10.44% to 14.84% for *most lovely.*  The fiction register has 10 (90.91%) instances compared

```
                                          |
          loveliest 87.36%                |   most lovely 12.64%
                                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                              100%
```
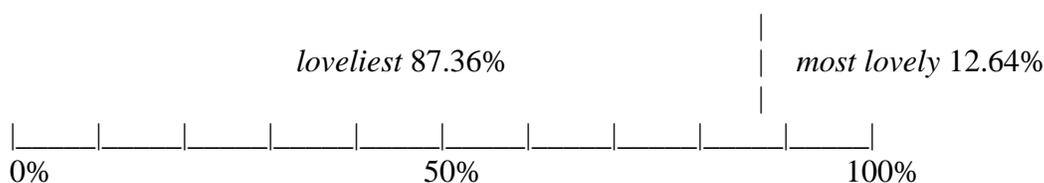
*Figure 17.*  Total percent usage of *loveliest* and *most lovely.*

with 1 (9.09%) instance for *most lovely.* Clearly, *loveliest* is the preferred form for the superlative, but enough *most lovely* forms exist that both should be considered part of standard written American English. Handbooks and dictionaries should include both forms as possibilities, and tests and graders cannot consider one form correct and the other incorrect.

*Double comparatives and superlatives.*  Even though usage is divided in the forms of comparatives and superlatives presented in this dissertation, one kind of usage is 100% consistent:  People use single comparatives and superlatives, and no double comparatives or superlatives exist for any of the above words and phrases in any of the corpora.  In Early Modern English, writers used double comparatives and superlatives often, such as *more riskier*, or *most riskiest*.  Because of early 18th century grammarians' efforts, English speakers dropped all double comparatives and superlatives (except for a few phrases found in the Bible, such as God as the *most highest*).

A test of *more riskier* and *most riskiest* in all the corpora shows that this rule of no double comparatives or superlatives has successfully changed English, for no examples of double comparatives and superlatives exist for these words in any of the three text archives of edited, published, written English.  Using single comparatives and superlatives rather than double is certainly something that teachers, instructional designers, and textbook writers can teach, and test makers can test.

*Preposition Use*

As stated in the method chapter, prepositions are hard to research in text archives because they are such common words that some search engines do not search for them.  Most preposition research will have to be delayed until an American corpus is ready.  This dissertation presents one example of controversial prepositional use.

In the phrases *different from* and *different than,* everyone will agree that *from* is a preposition, but most grammarians consider *than* a conjunction and not a

preposition. Common use of *than*, though, shows that many speakers use *than* as a

preposition, even if it is not officially a preposition. What is the correct preposition

to follow *different*? *AP* states that *different from* should be used rather than *different*

*than*. *Chicago* also states that *different from* is preferable to *different than*. *Garner*

explains that *than* implies a comparative, and *different* is not a comparative but

*different than* may occasionally be used if it avoids an awkward construction.

*Webster* states that "*different than* is standard in American and British usage,

especially when a clause follows *than*, but is more frequent in American" (p. 341).

A search of the text archives reveals what people actually do in edited,

published English. Table 18 shows the results of the usage in different registers. The

Table 18

*Number of Instances of "Different From" and "Different Than"*

| Register | different from | different than | Total |
|---|---|---|---|
| Newspapers | 4,615 (65.17%) | 2,466 (34.83%) | 7,081 |
| Magazines | 9,562 (85.87%) | 1,115 (14.13%) | 10,767 |
| Scholarly Journals | 131,346 (92.39%) | 10,815 (7.60%) | 142,161 |

$\chi^2$ (2, $N$ = 176,193) = 6,419.01, $p$ <0.0001 indicates that there is a statistically

significant difference among the three registers. Table 18 shows that newspapers use

*different than* a third of the time, but in scholarly writing, *different than* is rare.

Magazines are in between. The most formal writing then, uses *different from*; in less

formal writing, people use *different from* the majority of the time, but *different than* shows up, too.

Figure 18 shows the total of all the usages. The confidence intervals, showing the estimated true value of the percent usage in all the registers, is 79.17% to 79.97% for *different from* and 20.03% to 20.82% for *different than.* Clearly, *different from* is not an exclusive usage; enough usages of *different than* exist that they both are part of standard published American English. A search of the fiction corpus shows the usage of *different from* at 81.44%, and the usage of *different than* at 18.56%. Teachers, instructional designers, and textbook writers can indicate a preference for *different from*, but test makers cannot test for the correctness of *different from* over *different than* because so many edited, published authors use *different than*.

```
                                    |
        different from  79.57%      |   different than  20.43%
                                    |

|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                           50%                           100%
```

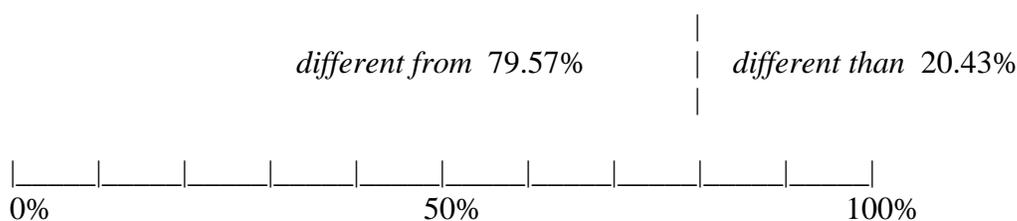*Figure 18.* Total percent usage of *different from* and *different than.*

## *Verb Forms*

### *Past Tense*

Old English speakers made verbs into past tense in several ways. One way was to add a /t/, /d/, or /əd/ to a verb, while other words formed past tense by changing the vowels. Over the centuries, most strong verbs (words that change the vowel for past tense) have changed to weak verbs (words that add a dental sound for

past tense), but English still has plenty of strong verbs, such as (a) *speak*, (b) *run*, (c) *sing*, (d) *see,* etc.  Some words are in flux and have some usages that are like a weak verb and some like a strong verb.  Sometimes there are two versions of a weak verb. A text archive search shows us what current usage is for these verbs in certain registers.  This search will show us how writers of standard English make the past tense in the years 2000 to 2007.

 *"Snuck" and "sneaked."*  The *OED* states the connection between this 17th-century word and the Old and Middle English *snican* is tenuous, and the origins of this word are in question; nevertheless, according to *Webster, snuck* is more of a United States invention of the 19th century.  *Webster* shows that *snuck* began as a dialectical word, then it became mainstream, and with younger speakers *snuck* is a common usage.  *AP* does not allow the use of *snuck*, and *Garner* says *snuck* is non-standard, but it is used half as much as *sneaked.  American Heritage* states that 67% of their panel in 1988 disapproved of *snuck*, and it points out the origins of *snuck* are in non-standard English.  The editors also acknowledge that the usage of *snuck* has increased by about 20% from 1985 to 1995.

 This interesting word has gone against the current of language change because speakers have added to the standard weak form a strong form.  Most other words go the other way around.  The results of a text archive search are shown in Table 19 below. Obviously, *snuck* is used a lot in all three registers.  *Sneaked* is still used more often, but *snuck* is used so much that it must be considered part of standard written American English.  The $\chi^2$ (2, $N = 3{,}514) = 41.02$, $p < 0.0001$ indicates that there is a

statistically significant difference in the usage among (a) newspapers, (b) magazines,

and (c) scholarly journals.  Magazines have the largest percentage of

Table 19

*Numbers of Instances of "Snuck" and "Sneaked"*

| Register | snuck | sneaked | Total |
|---|---|---|---|
| Newspapers | 320 (34.22%) | 615 (65.77%) | 935 |
| Magazines | 152 (45.89%) | 240 (54.11%) | 392 |
| Scholarly Journals | 121 (36.67%) | 209 (63.33%) | 330 |

usage, which makes sense, since they are less formal, but *snuck* still is used one-third

the time in scholarly journals, a very formal kind of writing.

Figure 19 shows the total percentages of the usage.  A 95% confidence

interval of the estimated true percentage of use shows *snuck* at 36.83% to 41.03% and

*sneaked* at 58.97% to 63.17%. *Sneaked* is still clearly favored, but *snuck* has a

substantial number of authors who use it.  The fiction archive does not have any

instances of these words.  People who plan curriculum, write textbooks, and write

```
                        |
   snuck 38.93%         |   sneaked 61.07%
                        |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                         50%                        100%
```
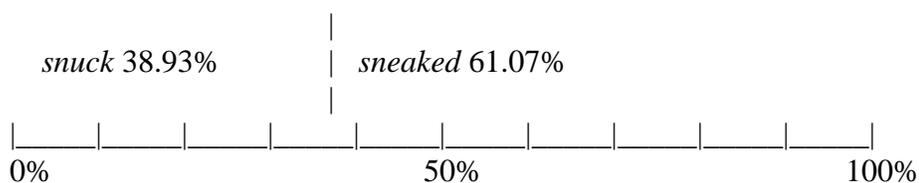
*Figure 19.*  Total percent usage of snuck and sneaked.

tests should not favor one usage over the other; neither can one be considered incorrect. Although its use has been condemned in the past, *snuck* has a large presence in standard American English.  If 37% of the occurrences of the past tense of *sneak* are represented as *snuck* even in scholarly journals, then *snuck* has entered standard English, although *sneaked* is used more often.

　　*"He dove" and "he dived."*  According to the *OED, dive* had a strong and weak form in Old English, but the strong form dropped out and left just the weak form, *dived.* The modern form *dove* was formed, probably by analogy with *drive/drove* and *weave/wove*.  This is another unusual case of a weak verb taking on a strong past tense.  *Webster* claims that the newer form *dove* began in the 19th century and is found in the northern United States and Canada, but its usage is moving down and across the country.  *Webster* says editors of the *New York Times* over the years have forbidden the use of *dove*, but *Webster* says either usage is acceptable.  *AP* says to use only *dived* and not *dove*, and *Garner* says that *dove* is common, but *dived* is preferable and more common. *American Heritage* lists either *dived* or *dove* as acceptable, and mentions that in the North, *dove* has been more accepted, while in the Midlands, *dived* is more common.

　　Table 20 shows the results of the text archives search.  Because the word *dove* has so many more meanings than the past tense of dive, such as the bird and the soap, I had to include another word that would limit its use to the verb.  This limitation has cut the number of instances down substantially.  In the future, a corpus will be able to search for the verb *dove,* and I will not have to limit the phrase to *he dove.*  The $\chi^2$ (2, $N = 439) = 23.08$, $p < 0.0001$ indicates that the differences among the three registers

of (a) newspapers, (b) magazines, and (c) scholarly journals are statistically

significant.  Although several usage books consider *dived* to be the most

Table 20

*Number of Instances of "He Dived" and "He Dove"*

| Register | he dived | he dove | Total |
|---|---|---|---|
| Newspapers | 64 (50.00%) | 64 (50.00%) | 128 |
| Magazines | 102 (45.95%) | 120 (54.05%) | 222 |
| Scholarly Journals | 8 (32.00%) | 17 (68.00%) | 25 |

prevalent, and even though *AP* style recommends the use of only *dived*, half of the

newspapers used *he dove*.  Scholarly journals have the greatest percentage usage of *he

dove* with 68.00%.

Figure 20 summarizes the percentages of all the uses. The 95% confidence

interval, which estimates the true percentage of use in the three registers, is  50.25%

to 64.45% for *he dove* and 35.55% to 49.75% for *he dived*.  Clearly, the usage book

proclamations do not match what the text archive search shows:  *he dove*

```
                          |
   he dived 42.65%        |   he dove 57.35%
                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                         100%
```
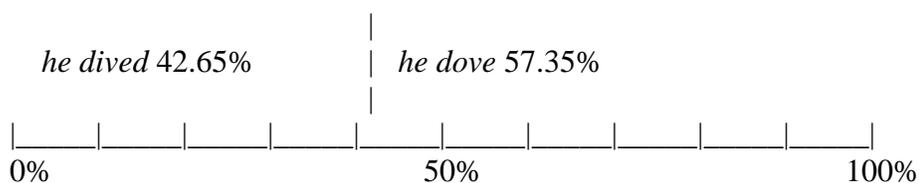
*Figure 20.*  Total percent usage of *he dived* and *he dove*.

is used more often than *he dived* in edited, printed American English.  The fiction

corpus has 100% (of 11 instances) use *he dove* and 0% use *he dived*. All registers,

from scholarly journals to magazines to fiction use *he dove* more than *he dived*.

Designers of curriculum, textbooks, and tests cannot choose one usage as correct and

one as incorrect because both usages are strongly entrenched in standard written

American English. Dictionary and handbook makers should note that usage has

changed.

　　　*"Crept" and "creeped."*  The *OED* states that *creep* was originally a strong

verb, changing its vowel for past tense.  In the 14th century, the weak forms *creeped*

and *crept* appeared, and since the 16th century *crept* has dominated.  *Webster* claims

that *creeped* is the more recent invention, and it does not appear in print very much,

but it will probably increase in usage.  *AP* does not mention the issue.  *American*

*Heritage* acknowledges only *crept. Garner* says *crept* is the standard form, but

*creeped* is infiltrating the language.  Table 21 shows the results of the archive search.

The $\chi^2$ (2, $N$ = 5,073) = 35.53, $p$ < 0.0001 indicates that there is a statistically

Table 21

*Number of Instances of "Crept" and "Creeped"*

| Register | crept | creeped | Total |
|---|---|---|---|
| Newspapers | 1,051 (95.29%) | 52 (4.71%) | 1,103 |
| Magazines | 2,000 (94.77%) | 149 (5.23%) | 2,849 |
| Scholarly Journals | 1,109 (98.93%) | 12 (1.07%) | 1,121 |

significant difference among the three registers. Obviously, *crept* is the dominant

usage in newspapers, magazines, and scholarly journals, and *creeped* is indeed rare.

The more formal register of scholarly journals has only 1.07% usage of *creeped*.

Obviously, in formal written English, *crept* is the standard form.  In less formal

writing, *creeped* is used a little bit more often.  Since *creeped* shows up more in

informal English, and since informal forms often make their way to formal English,

*creeped* may indeed be on the rise.

Figure 21 shows the totals for usage of the two forms.  The 95 % confidence

interval, showing the estimated true value of the percent of usage in the three

registers, is 95.83% to 96.83% for *crept* and 3.17% to 4.17% for *creeped.  Crept* is so

dominant a usage that most writers of standard English choose it, yet the usage of

*creeped* is probably going up.  In the fiction corpus, the usage is 100% (of a

```
                                                         |
              crept 96.33%                               |   creeped 3.67%
                                                         |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                              100%
```

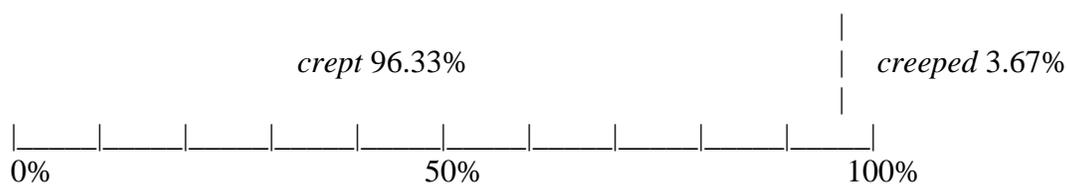*Figure 21.*  Total percent usage of *crept* and *creeped.*

total of 187 instances) usage as *crept* and 0% usage as *creeped.* Since this word is

changing, writers of curriculum, textbooks, and tests should probably not consider

*creeped* as wrong, but allow it a minority status in English spelling. Clearly, though,

users prefer *crept* at an overwhelming rate, and it is the standard way of using the

word.

*Past Participle*

Most past participles are formed the same way the past tense is formed: by adding an *–ed* to the verb. However, past participles of irregular verbs are often formed by adding an *–en* to the verb. Sometimes the past tense and past participle of a verb come in several different forms. This next section will take a few words that have variants in the past participle and search for the standard usage.

*"Sped" and "speeded."* The two forms *sped* and *speeded* appear in American English in the past participle. This study used the phrases *has speeded, have speeded, had speeded, has sped, have sped,* and *had sped* to search for the past participle form in the text archives. This search limits the results to the past participle used as a verb rather than as an attributive adjective. The *OED* shows that forms of *sped* existed since the 13$^{th}$ century, but *speeded* did not appear until the 18$^{th}$ century. *Webster, Chicago,* and *AP* do not list the item, but *Merriam Webster* and *American Heritage* list *sped* first and then *speeded* as the past tense, and therefore the past participle forms. *Garner* states that the best past tense and past participle form is *sped.*

The archive search shows what people actually do in three different registers. Table 22 shows the results of the archive search in (a) newspapers, (b) magazines, and scholarly journals. The $\chi^2$ (2, $N$ = 5,190) = 320.85, $p < 0.0001$ indicates that there is a statistically significant difference between the registers. Obviously, the less formal registers use *sped*, and the more formal register, scholarly journals, uses *speeded* more often.

Table 22

*Number of Instances of "Sped" and "Speeded" in the Past Participle*

| Register | sped | speeded | Total |
|---|---|---|---|
| Newspapers | 3,419 (63.74%) | 1,945 (36.26%) | 5,364 |
| Magazines | 208 (50.61%) | 203 (49.39%) | 411 |
| Scholarly Journals | 120 (38.10%) | 195 (61.90%) | 315 |

Figure 22 shows the average of the totals across the three registers. The 95% confidence interval, showing the estimate of the true value of the percent usage in the complete registers, is 48.41% to 53.21% for *sped* and 46.79% to 51.59% for *speeded.* Usage seems to be fairly divided between the two forms. The fiction corpus does not have enough examples to use. The less formal registers use *sped* more, and the more formal register uses *speeded* more. Teachers, textbooks, and

```
                          |
     sped 50.81%          |    speeded 49.19%
                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                      50%                      100%
```
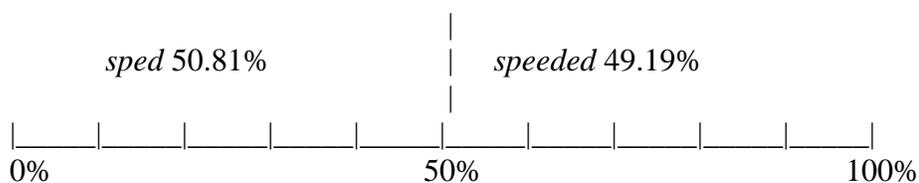
*Figure 22.* Total percent usage of *sped* and *speeded* in the past participle.

dictionaries should acknowledge both forms of the past participle, but they cannot make a blanket statement about which form is preferred because it depends on the register. Obviously, both forms exist in standard American English, and test makers cannot test one form as correct and another form as incorrect.

*"Proven" and "proved."*    The past tense of *prove* is *proved,* but the past

participle can be either *proven* or *proved.*  This study has used *has proved, have*

*proved, had proved, has proven, have proven,* and *had proven* as phrases to search for

the past participle use. Like the search in *sped/speeded,* this search concentrates on

the word *prove* being used as a perfect construction (*have + -en* verb) and not as an

attributive adjective. The *OED* lists *proved* as a past participle form from Middle

English, and *proven* as a past participle form later in the 15$^{th}$ century, and it lists

modern past participles as *proved,* first, then *proven,* second. The version of *proven*

came from Scottish English, and the word probably changed by analogy with *cloven*

and *woven.*

The *OED* says today that American English uses *proven* and *proved* at about

the same rate.  The *AP* style guide says that the past participle is *proved,* and *proven*

should only be used before a noun.  This search isolates its usage not as an attributive

adjective (coming before the noun), but as a past participle verb.  *Webster* says the

controversy of which form to use started in the 19$^{th}$ century when *proved* was used

four times more often.  Now, the editors say, the use of *proven* has caught up, and

*Webster* declares both possibilities as equally good.  *Merriam-Webster* and *American*

*Heritage* both list *proved* first and *proven* second, but they both state that, when used

as an attributive adjective, writers prefer *proven.*  *Chicago* and *Garner* both say the

past participle should be *proved* with *proven* used only as an adjective.  *Garner* calls

the past participle form, *proven,* as "ill-advised" (p. 65).

This text archive study compares the usage of *proved* and *proven* in the past

participle by looking at the phrases *has proved, have proved, had proved, has proven,*

*have proven,* and *had proven.* Table 23 shows the results of this search. The $\chi^2$ (2, *N* = 99,233) = 202.00, *p* < 0.0001 indicates that the three registers differ significantly from one another. All three show quite evenly divided usage between *proven* and *proved.*

Table 23

*Number of Instances of "Proven" and "Proved" in the Past Participle*

_____

| Register | proven | proved | Total |
|---|---|---|---|
| Newspapers | 8,280 (48.77%) | 8,696 (51.23%) | 16,976 |
| Magazines | 10,140 (43.01%) | 13,437 (56.99%) | 23,577 |
| Scholarly Journals | 28,250 (48.14%) | 30,430 (51.86%) | 58,680 |

_____

Figure 23 shows the averaged total of all the registers. A 95% confidence interval, showing the estimate of the true value of the percentage usage in all the registers, is 46.24% to 47.04% for *proven* and 52.96% to 53.76% for *proved.* Obviously, the usage is split fairly evenly in the past participle, and this trend is found in every register. In the fiction register, *proven* is used in 6 instances (40.00%), and *proved* is used in 9 instances (60.00%).

```
                                |
        proven 46.64%           |       proved 53.36%
                                |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                             50%                              100%
```
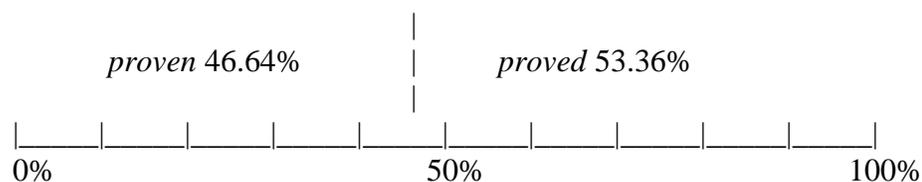
*Figure 23.* Total percent usage of *proven* and *proved* in the past participle.

This usage is split so evenly that it is hard to recommend which is the best usage. In American English, both forms exist in standard English. Several of the usage books say that *proven* should be used only as an adjective. This archive study shows that almost half of all usages as a past participle are the form *proven.* Its use is more prevalent than the usage books say. Teachers, textbooks, and dictionaries should teach both usages as possibilities. Test makers cannot test this item because there is no right or wrong.

As we have seen, some words can have one form in the past tense and another form as the past participle. Randolph Quirk has researched some of these words in British and American English. He found that the corpora could not produce enough examples, so he used elicitation techniques to get subjects to supply words in blank spots of the sentence (1995, p. 129). Quirk's study showed that the British and Americans alike prefer the *–t* spelling for the perfect form of verbs like "*spoil*, *dream*, *spill*, *learn*, *spell*, *leap*, *kneel*, and *smell*" (1995, p. 195). A perfect is a verb form that uses a version of *have* plus the past participle. Quirk found that the British and Americans also preferred *–ed* in the preterit (past tense) (p. 195). This archive study will see what the preferences are in one verb in modern American English.

*"Kneeled" and "knelt."* The word *kneel* has two possible past tense forms. They both add a dental sound, and one word, *knelt,* changes the vowel. This study looks first at the past tense and then at the past participle forms of the verb. The *OED* lists *kneeled* first and *knelt* second as past tense and past participles. The entry notes

that *knelt* entered English in the 19th century, and the –*t* form began in southern England. The other version, *kneeled,* has deeper roots. *Webster* claims that, although it is a newer form, *knelt* occurs more often than *kneeled,* although either is acceptable. *Merriam-Webster* and *American Heritage* both list *knelt* first and *kneeled* second. *Garner* claims that *knelt* is the only form to properly use as past tense and past participle, and he claims that *knelt* occurs five times more often than *kneeled.*

The archive search shows that *kneeled* and *knelt* have a presence in modern American English. Table 24 shows the percentage of the use of each word across the three registers. A $\chi^2 (2, N = 20,710) = 34.09, p < 0.0001$ indicates that there is a statistically significant difference among the registers. The percentage numbers are very close in each register, and the usage does not seem to differ from one register to the next.

Table 24

*Number of Instances of "Kneeled" and "Knelt" in the Past Tense*

| Register | kneeled | knelt | Total |
|---|---|---|---|
| Newspapers | 3,633 (19.09%) | 15,397 (80.91%) | 19,030 |
| Magazines | 816 (12.07%) | 112 (87.93%) | 928 |
| Scholarly Journals | 636 (15.43%) | 116 (84.57%) | 752 |

Figure 24 shows the average of the totals in each of the registers. A 95% confidence interval, which shows the estimated true value of the percentage of

use in the registers, is 83.37% to 85.57% for *knelt* and 14.42% to 16.63% for *kneeled.*

Obviously, *knelt* is the form used most often, but in the past tense, *kneeled* does

occur.  In the fiction corpus, *knelt* occurs in 218 instances (77.85%), and *kneeled*

occurs in 62 instances (22.14%).  Quirk claimed from his elicitation studies that

Americans prefer the *–ed* form in the past tense, but this text archive search shows

that Americans prefer the *–t* version for the past tense. Teachers, dictionary makers,

and textbook makers can teach *knelt* as the favored form, but it is not the only form of

the past tense. *Kneeled* has enough of a presence that it also is a variant form in

standard written American English.

```
                                       |
            knelt 84.47%               |    kneeled 15.53%
                                       |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                           50%                           100%
```
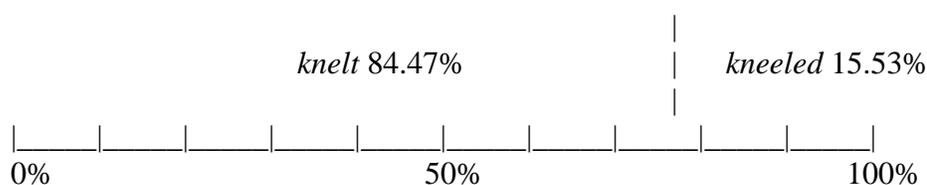
*Figure 24.*  Total percent usage of *knelt* and *kneeled* as the past tense.

The past few paragraphs show the results for *kneel* in its past tense form

(excluding any perfect aspect).  Now, the question is whether the past tense verb form

is the same as the past participle, or whether, as Quirk posits, the preferred perfect is a

different form.

This study focused on the phrases *have kneeled, has kneeled, have knelt,* and

*has knelt* to discover the past participle use in the text archives.  Table 25 shows the

usages of *knelt* and *kneeled* in the three different registers.  The $\chi^2$ (2, $N = 553$) =

7.21, $p = 0.0272$ indicates that the registers are significantly different from each other. Writers choose *kneeled* as the past participle form in less formal writing more often than in formal writing. *Knelt* is still used as the primary form, but *kneeled* in the perfect aspect is used about a third of the time in newspaper writing. The actual number of instances is much lower in the perfect than in the past tense, which shows that people do not use the perfect construction of the verb *kneel* nearly as much as they use the past tense.

Table 25

*Number of Instances of "Kneeled" and "Knelt" in the Past Participle*

| Register | kneeled | knelt | Total |
|---|---|---|---|
| Newspapers | 60 (31.58%) | 130 (68.42%) | 190 |
| Magazines | 0 (0.00%) | 10 (100.00%) | 10 |
| Scholarly Journals | 85 (6.25%) | 268 (93.75%) | 353 |

Figure 25 shows the average of the total usage in all three registers. The 95% confidence interval, showing the estimated true value of the percent usage in the complete registers, is 82.89% to 91.89% for *knelt* and 8.11% to 17.11% for *kneeled*. The fiction register does not have any examples of these words. Clearly, *knelt* is preferred, but *kneeled* also exists in standard English. *Knelt* is the preferred form in both past tense and past participle. The usage between the past tense and the past participle is not much different. Quirk's idea that Americans prefer the *–t* ending in

the past tense and the *–ed* ending in the perfect is wrong:  Americans prefer the *–t*

ending in both the perfect and the past tense.  Teachers, dictionary writers, and text

book writers should teach *knelt* as the past tense and past participle, but they need to

allow for *kneeled* also, which is a variant.  Test makers cannot test for a right and a

wrong usage in standard English, since both forms are present.


```
                                                  |
                    knelt 87.39%                  |       kneeled 12.61%
                                                  |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                          50%                               100%
```

*Figure 25.*  Total percent usage of *knelt* and *kneeled* in the past participle.


*Subject-Verb Agreement*

A text archive cannot search most subject-verb agreement problems, so we

will have to wait for an American corpus to search these.  There are two nouns,

though, that writers have problems agreeing with the verb because not everyone

agrees about whether the nouns should take a singular verb or a plural verb.  This

section will explore the plural and singular use of the words *data* and *criteria*.

*Data* and *criteria* were borrowed as plural words, but their singular forms,

*datum* and *criterion*, are not as heavily used.  To the general public, words like *data*

and *criteria* have become singular and take a singular verb.  Some people, though,

insist that *data* and *criteria* are clearly plural and take a plural verb and nothing else.

These people consider the plural *data* or *criteria* with a singular verb to be a mistake.

A text archive search reveals how these words are used in different registers.

*"This data" and "these data."* The *OED* gives instances of *data* and *datum* from the 17th century. The use of *data* as a plural form with a singular construction, though, is mostly a 20th-century phenomenon. *Webster* devotes an entire page to the history of *data* as a singular and a plural. This usage book shows that *data* has two constructions: a plural with plural determiners, pronouns and verbs; and a singular used as a mass noun with singular determiners, pronouns, and verbs. *AP* states that *data* is plural, but it does allow one usage of a mass noun, which is considered singular: If *data* is considered one unit, as in "the data is sound" (p. 52), then it is a mass noun, but if *data* means individual items, as in "the data have been carefully collected" (p. 52), then the word is a plural.

*American Heritage* argues that *data* is being accepted more and more as a singular and a plural. In their usage survey, 60% of the *American Heritage* usage panel accepted *data* as a word that can take a singular verb and a singular pronoun. *Garner* states that if you have to choose a plural or a singular verb with data, either way you are going to call attention to yourself and bother someone. *Garner* argues that since the 1940s, *data* has been used as a singular more and more; today, though, *data* as plural is more formal, and *data* as a singular mass noun is also considered correct.

In order to separate *data* plural from *data* singular, I coupled the words with the singular and plural determiners, *this* and *these,* since the text archives will not all search for the *to be* verb because it is such a common verb. Table 26 shows the results of the text archive search to determine if people do use *data* as singular at all. The plural use of the word *data* is strong in scholarly journals, and scholarly journals

have the most formal writing.  Magazines are split almost evenly, and newspapers use

twice as many singulars as plurals.  Probably because of the nature of the word, both

Table 26

*Number of Instances of "This Data" and "These Data"*

| Register | this data | these data | Total |
|---|---|---|---|
| Newspapers | 163 (70.87%) | 67 (29.13%) | 230 |
| Magazines | 1,126 (51.36%) | 1,066 (48.61%) | 2,192 |
| Scholarly Journals | 16,402 (13.97%) | 100,969 (86.03%) | 117,371 |

*this data* and *these data* appear in scholarly journals much more than in magazines

and newspapers. The $\chi^2$ (2, $N = 119,793$) = 2,966.71, $p < 0.0001$ indicates that there is

a significant difference among the registers.   There is no easy way to research which

of these uses of singular *data* are collective nouns at this time, but the numbers do

show that the use of *data* as a singular is common.

Figure 26 shows the total usage of *data* to be closely split between plural and

singular. The 95% confidence interval of the estimated true value of the percent of

usage of *this data* is 43.30% to 47.50%, and the confidence interval for *these data* is

52.50% to 56.70%.  The fiction corpus has too few instances to give us much

information.  Although some strict grammarians consider *data* to be plural in all

instances, text archive searches show that usage is divided.  Writers of curriculum,
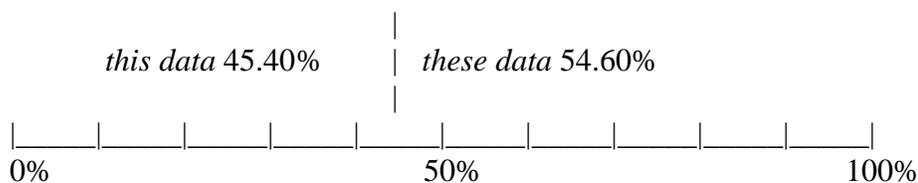
```
                          |
        this data 45.40%  |  these data 54.60%
                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                           50%                           100%
```

*Figure 26.* Total percent usage of *this data* and *these data*.

textbooks, dictionaries, handbooks, and tests should capture the differences in usages

of *data* to teach students, but, because of the divided usage in standard written

American English, these professionals cannot call one way correct and one way

incorrect.  Writers should also be sensitive to the usage within a particular register.

 *"These criteria" and "this criteria."*  *Criteria*, like *data*, is a controversial

word.  Some people think it should be plural at all times, and some use it as a

singular.  The *OED* shows that *criterion*, with plural *criteria*, is a Greek borrowing

from the 17th century.  *AP* designates *criteria* as plural and *criterion* as singular.

According to *Webster*, *criteria* as a singular began to be more commonly used in the

1960s.  It is sometimes used as a count noun (and therefore takes a singular *this*), but

this usage is criticized.  *Webster* conjectures that some day the plural form *criteria*

used as a singular count noun will be as common as *agenda* is as a singular noun.

*American Heritage* lists *criteria* as a plural and acknowledges that some use the word

as a singular, but this usage is not accepted yet.  *Garner* calls *criteria* a plural noun,

but acknowledges that some people use it incorrectly as a singular.

 A search of the text archives shows what writers actually do, and Table 27

displays the results.  The $\chi^2 (2, N = 15,085) = 646.43, p < 0.0001$ indicates that there is

a statistically significant difference among the three registers.  The majority of writers

in all registers use *criteria* as a plural; indeed, in scholarly journals, almost all writers use *criteria* as a plural. In newspapers and magazines, though, about a quarter of all writers use *criteria* as a singular. The scholarly journal register has far more examples of *criteria* than the other two registers, probably because *criteria* is more of a scholarly word. The newspaper register, on the other hand, has only a handful of instances. Even so, the samplings are sufficient for generalizing to the entire population, and the confidence interval takes into consideration the small number of samples.

Table 27

*Number of Instances of "These Criteria" and "This Criteria"*

| Register | these criteria | this criteria | Total |
|---|---|---|---|
| Newspapers | 17 (77.27%) | 5 (22.73%) | 22 |
| Magazines | 528 (74.05%) | 185 (25.95%) | 713 |
| Scholarly Journals | 13,737 (95.73) | 613 (04.27%) | 14,350 |

Figure 27 shows the final results of this text archive search. The 95% confidence interval, which shows the estimated true value of the percent of usage in all three registers, is 11.75% to 23.33% for *this criteria* and 76.45% to 88.25% for *these criteria*. A large majority of writers in all three registers use *criteria* as a plural, but this usage is not exclusive. These words do not show up in the fiction corpus, so there is no data from the short stories.
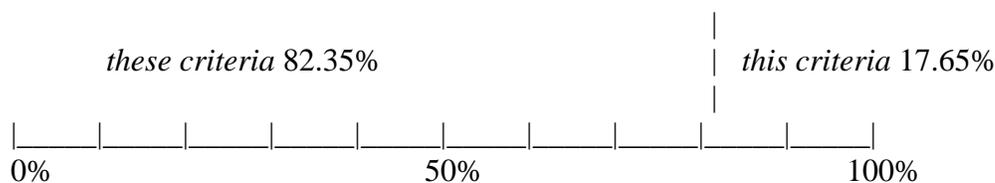
```
                                           |
        these criteria 82.35%              |  this criteria 17.65%
                                           |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                          100%
```

*Figure 27.* Total percent usage of *these criteria* and *this criteria*.


Since formal English often changes over time to become more like informal English, educators should watch this usage to see if the patterns in formal English change.  Writers of curriculum, textbooks, and tests should bring up the issue of plurality and the word *criteria*, but, with 18% of texts in the archive of standard written American English showing *criteria* as a singular, educators cannot consider the singular usage as wrong. This may be a usage that is in flux, and researchers should continue to watch how people use *criteria*, and writers need to be sensitive to the register.

### Pronoun Agreement

All Old English nouns had up to eight possible endings:  The pronouns had a different form in the singular and plural for the nominative (subject), accusative (object), genitive (possessive), and dative (object of a preposition or indirect object). The nouns have dropped all of the endings except for the genitive and general plural, but the pronouns have kept most of the Old English variety.  Because only the pronouns have an accusative case today, English speakers do not have to decide the case very often, and they sometimes have problems deciding which pronoun to use in different settings. This section will concentrate on the problem with choosing an accusative or a nominative pronoun with a preposition.

One phrase with a controversial pronoun is *between you and me* or *between you and I.* Because *between* is a preposition, the word or words that follow should be in the accusative (objective) case. The grammatically correct usage, then, is *between you and me.* Many people are uneasy about the phrase *you and me*, and many educated people are used to correcting most *you and me* phrases to *you and I* automatically. Because of this tendency to correct, many people overcorrect some phrases, changing *between you and me* to *between you and I,* even though they would never say *between we* for *between us.*

*Webster* follows the use of *between you and I* over the centuries, and the editors found that even Shakespeare used the phrase. The editors try out several theories trying to explain the usage: For example, some people think that maybe *you and I* is a set phrase people insert into many places. *Webster* suggests that people can say the phrase *between you and I*, but they will be judged uneducated if they write it. *Garner* shows differing points of view about the subject, but he disparages any descriptive linguists who say the rules of English should merely describe the language. He advocates only the use of *between you and me.* Is *between you and me* becoming antiquated, or is it still the prevailing usage in standard written American English?

A text archive search reveals what current usage is in three different registers. Table 28 shows the results by register. The $\chi^2$ (2, $N = 364$) $= 0.92$, $p = 0.6296$ indicates that there is not a statistically significant difference among the registers. Table 28 shows a rather shocking statistic: scholarly journals have a 10.59% usage of the phrase *between you and I.* The most formal writing of the three registers actually

has a larger occurrence of the usage many experts consider uneducated. This is less

shocking when we realize that the people using the phrase *between you and I* are

often, as *Garner* puts is, "educated speakers trying a little too hard to sound refined

and stumbling badly" (p. 100).

Table 28

*Number of Instances of "Between You and Me" and "Between You and I"*

| Register | between you and me | between you and I | Total |
|---|---|---|---|
| Newspapers | 27 (93.10%) | 2 (6.90%) | 29 |
| Magazines | 152 (92.12%) | 13 (7.88%) | 165 |
| Scholarly Journals | 152 (89.41%) | 18 (10.59%) | 170 |

Figure 28 shows the total percentages of the two usages in all the registers.

Clearly, *between you and me* is the standard, but the usage of *between you and I* is

quite prevalent at 8.45%. A 95% confidence interval shows the estimated true value

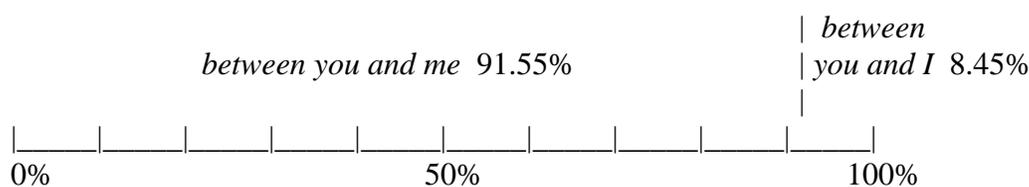of the percent of usage in these three registers is 87.45% to 94.85% for *between you*

```
                                                          |  between
        between you and me  91.55%                        | you and I  8.45%
                                                          |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                           50%                          100%
```

*Figure 28.* Total percent usage of *between you and me* and *between you and I*.

*and me* and 4.75% to 12.15% for *between you and I.* The fiction corpus shows 13

instances (100.0%) of *between you and me* as opposed to 0 instances (0.0%) of

*between you and I.*

Does the rather large percentage of use of *between you and I*, especially in the

scholarly journals, mean that the phrase using *I* is gaining acceptability? This is a

hard question because this one phrase is a linguistic shibboleth for many language

experts; people are often judged to be uneducated if they use *between you and I.*

Because neither of these phrases shows up in print very often, as shown by the small

number of items found in the corpora, this item might be one to follow over the

decade rather than make a pronouncement based on too few items. Writers of

curriculum, textbooks, and tests should teach *between you and me* to students, but

they should watch for changing usage if they plan to test students on a right or wrong

form of this item.

<div align="center">

*Wrong Word*

</div>

Some people consider one version of a word to be correct and another to be

incorrect. In the future with a linguistic corpus of American English, we will be able

to test more word pairs. Some word pairs require qualitative studies to see how each

word is actually used. For this dissertation, I chose two controversial word pairs to

see how each variant is used in the three different registers and to determine the

standard English usage.

*"Toward" and "towards." American Heritage, Garner*, and *Webster* all state

that there is no semantic difference between these two words. They state that *toward*

is predominantly American usage, and *towards* is predominantly British usage. *AP*

allows *toward* only.  The three text archives show that the usage is not so cleanly

divided.  The text archives contain written American English, so we expect *toward*

and not *towards*.  Table 29 shows the results.  *Toward* is the preferred spelling in all

three registers, but it is not the exclusive spelling.  The $\chi^2$ (2, $N$ = 746,821) =

84,940,784, *p* <0.0001 indicates that there is a statistically significant

Table 29

*Number of Instances of "Toward" and "Towards"*

| Register | toward | towards | Total |
|---|---|---|---|
| Newspapers | 65,661 (93.39%) | 4,651   (6.61%) | 70,312 |
| Magazines | 112,784 (79.48%) | 29,110 (20.52%) | 141,894 |
| Scholarly Journals | 345,302 (64.59%) | 189,313 (35.41%) | 534,615 |

difference in percentages among the three registers.  Newspapers, for the most part,

follow *AP,* so, not surprisingly, the newspaper corpus search shows that newspapers

use *toward* most of the time.  Magazines and scholarly journals, though, show that

*towards* is used in edited, printed American English, although it is not the

predominant form.

Figure 29 shows the total results of the text archive searches.  The 95%

confidence interval, which shows the estimated true value of the percentage of usage

in all the registers, shows *toward* as 78.94% to 79.14% and *towards* as 20.85% to

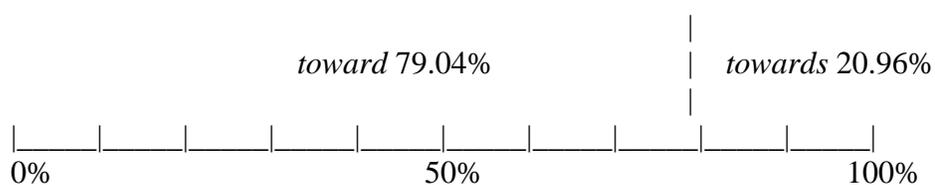21.05%.  Although usage books claim that *toward* is the spelling to use in American

```
                                      |
        toward 79.04%                 |    towards 20.96%
                                      |
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%                            50%                            100%
```

*Figure 29.*  Total percent usage of *toward* and *towards.*

writing, the text archive searches show that *towards* is also used in edited, printed

American English.  A search of the fiction corpus shows a heavy usage of  *toward* at

1,644 instances (81.35%) over *towards* at 377 instances (18.65%), but still both

usages exist.  Both usages are alive in all registers, except the newspaper register.

Teachers, instructional designers, and textbook writers can suggest one usage over

another, but they must remember that this is not an exclusive usage.  Test makers

cannot test *toward* or *towards* as correct or incorrect because each usage has a large

presence in published American English.

      *"Regardless" and "irregardless."*  The word *irregardless* is a shibboleth for

many language watchers.  *Webster* and *Garner* acknowledge that occasionally

*irregardless* can be found in published writing, and more often it can be found in

speech; nevertheless, both warn that it is not standard.  *Garner* calls it a "nonword,"

and suggests "careful users of language must continually swat it when they encounter

it" (p. 466).  *AP* says to use only *regardless*, and *American Heritage* states that some

people mistakenly think that *irregardless* is the more formal word.  Usage gurus

reject the word because it is a double negative:  *ir-* and *–less* in one word.

The text archive search shows that *irregardless* shows up very rarely in all the registers. Table 29 shows the results. The $\chi^2$ (2, $N$ = 131,930) = 9.57, $p$ = 0.008 indicates that there is a statistically significant difference among the three registers. In every register, *regardless* is clearly the standard usage.

Table 30

*Number of Instances of "Regardless" and "Irregardless"*

| Register | regardless | irregardless | Total |
|---|---|---|---|
| Newspapers | 7,565 (99.80%) | 15 (0.19%) | 7,580 |
| Magazines | 23,163 (99.88%) | 16 (0.12%) | 23,179 |
| Scholarly Journals | 101,171 (99.90%) | 100 (0.10%) | 101,271 |

Figure 30 shows the totals for all the registers with *regardless* and *irregardless*. The 95% confidence interval, showing the estimated true value of the percentage of use in the three registers, is 99.83% to 99.89% for *regardless* and 0.11% to 0.17% for *irregardless*. The totals are so heavily weighted toward
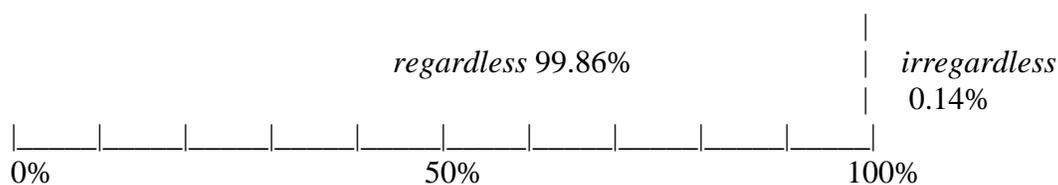
*regardless* 99.86%　　　　　　　　　　| *irregardless*
　　　　　　　　　　　　　　　　　　　|　0.14%

|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|
0%　　　　　　　　　　　50%　　　　　　　　　100%

*Figure 30.* Total percent usage of *regardless* and *irregardless*.

*regardless* that it can be considered standard written American English. A search of

the fiction corpus shows *regardless* at 89 instances (98.89%) is extremely favored

over *irregardless* at 1 instance (1.11%). In less formal and formal English,

*irregardless* is rarely used. Teachers, instructional designers, textbook writers, and

test makers can teach and test *regardless* as standard written English.

Chapter 5: Conclusions

Since the 18<sup>th</sup> century, the usage rules of English have emerged as prominent guidelines about how to write and edit. These rules arose as people tried to upgrade English and make it into a more respectable and ordered language. Experts promote the rules they favor by publishing usage books and dictionaries, which further codify the guidelines. And yet, not all experts agree about all the rules, and, further, the language changes, and the rules often do not.

In schools at all levels, teachers try to teach the rules of standard English to their students. English classrooms often have dictionaries and writing handbooks that contain many of these rules, and assessments for every grade check for students' knowledge of standard English., whether the tests require writing samples or answers to multiple choice questions. Many writers, though, have a general unease about knowing the rules of standard English, often because they were taught rules, but they do not see those rules executed in standard English, or, in other words, the rules do not always match what they see in standard written English. Writing practices also change over the years, but the rules often do not. Also, some rules are particularly British, but they do not apply to American English.

For these reasons, there is a need to pinpoint actual practices in standard written American English so we can write guidelines, rules, dictionaries, and usage books that will teach students how to write effectively and correctly in standard written American English. In the past we relied on authorities who wrote usage books to delineate the actual practices of written English, but with authorities disagreeing and the language changing, we can learn a lot about the written language by studying

it empirically. This dissertation shows one way of studying language use empirically by answering the following research questions.

*Research Question 1*

How can we use an empirical method to determine what is standard written American English in various linguistic registers? The rise of corpus linguistics has already heavily affected the ways people describe the syntax of English and the ways people write accurate definitions for dictionaries. Now that corpora are being created for American, not just British English, we can use these massive collections of texts to study actual practices in standard written American English. If we use edited, published English in particular registers as our examples of standard English, we can research these texts to discover what writers in standard English actually do.

*Research Questions 2a, 2b*

What does empirical research using corpus linguistics tell us about certain controversial usage questions that are representative of the many questions still requiring research?

a. What is the percent usage of each item?

b. Does the usage differ among the registers of newspapers,

magazines, and scholarly journals?

This dissertation listed categories of common errors condensed from Connors and Lunsford (1988) to form a framework of usage questions that would address the problems of errors. Then it listed disputed items from dictionaries and usage books that were appropriate for those categories. Each entry for each word presented the history of the usage problem and a survey of what usage books and dictionaries

proclaim about the item. Then the results of a text archive search were presented, including the percentages of use in each register, the total averaged percentages along with confidence intervals, and a chi-square statistic, which shows if there is a statistically significant difference among the registers.

No good corpora of contemporary standard written American English exist presently, although one will be completed in 2008. Because of this restraint, I researched electronic text archives of newspapers, magazines, and scholarly journals from January 2000 to April 2007 to find out the actual usage in each register.

*Research Question 2 c*

What usage or usages should be considered standard written American English? For this dissertation, I considered items used 95% of the time or higher to be standard English and words and phrases with usages between 90% and 95% as items to also consider. The following items, with their percentage usage across the registers, are considered standard written American English.

1. *Through* 99.82% rather than *thru* 0.18%
2. *Technique* 99.78% rather than *technic* 0.22%
3. *A lot* 99.82% rather than *alot* 0.18%
4. *Cannot* 96.07% rather than *can not* 3.93%
5. *E-mail* 92.84% rather than *email* 7.17%
6. *Crept* 96.33% rather than *creeped* 3.67% as a past tense
7. *Between you and me* 91.55% rather than *between you and I* 8.45%
8. *Regardless* 99.86% rather than *irregardless* 0.14%

The following items are of divided use. One of the two usages might be preferred, but both should be acceptable in standard written American English. Teachers, instructional designers, test makers, dictionary makers, and evaluators should also be sensitive to differences among the registers.

1.  *Catalog* 71.33% and *catalogue* 28.67%
2.  *All right* 87.52% and *alright* 12.60%
3.  *Appendices* 72.52% and *appendixes* 27.48%
4.  *Indexes* 60.15% and *indices* 39.85%
5.  *Syllabuses* 68.32% and *syllabi* 31.68%
6.  *Formulas* 83.55% and *formulae* 16.44%
7.  *More proud* 54.32% and *prouder* 45.68%
8.  *Most proud* 58.32% and *proudest* 41.68%
9.  *More risky* 75.60% and *riskier* 24.40%
10  *Most risky* 77.29% and *riskiest* 22.71%
11.  *More lovely* 72.12% and *lovelier* 27.88%
12.  *Most lovely* 87.36% and *loveliest* 12.64%
13.  *Different from* 79.57% and *different than* 20.43%
14.  *Sneaked* 61.07% and *snuck* 38.93% as past tense
15.  *Dived* 42.65% and *dove* 57.35% as past tense
16.  *Sped* 50.81% and *speeded* 49.19% as past participles
17.  *Proved* 53.36% and *proven* 46.64% as past participles
18.  *Knelt* 84.47% and *kneeled* 15.53% as past tense and past participles
19.  *Knelt* 87.39% and *kneeled* 12.61% as a past participle
20.  *These data* 54.60% and *this data* 45.40%
21.  *These criteria* 82.35% and *this criteria* 17.65%
22.  *Toward* 79.04% and *towards* 20.96%

*Research Question 3*

How should this knowledge of standard written American English affect what we teach and test in schools?  If we can describe standard written American English with more accuracy, then we can develop instructional material and train teachers to teach standard English better.  Our usage books and writing handbooks will not just be promoting outdated rules, but they can give up-do-date guidelines about writing in standard ways.  Teachers will be able to promote practices that are part of standard English, and teachers and texts can teach more standardized practices. Our tests can align with our practices, and we can achieve better content validity in our testing.

Searching text archives cannot answer all usage questions.  Many questions will have to wait until the *BYU Corpus of American English* or another American

corpus is assembled, and even then, some questions cannot be answered by a quantitative study. Some will require qualitative searches.

This dissertation is written before any corpora of standard written American English have been assembled. As the technology improves, we can search for a bigger range of usage items, and we can probe even deeper into standard English. Because the technology will improve very quickly, the use of text archives will be replaced by actual corpora, and the research practices of this dissertation will quickly be outdated. The research questions, though, will remain much the same as the ways of searching the language technologically improve.

In the future, we will not have to rely on experts debating with experts about the nature of the English language. We will also add searches of real text that tell us what writers of edited, published American English actually do.

References

*The American heritage book of English usage.* (1996). Boston: Houghton Mifflin.

*The American heritage dictionary* (4th ed.). (2000). Boston: Houghton Mifflin. Retrieved June 18-22, 2007 from http://www.bartleby.com/61/

*The American heritage guide to contemporary usage and style.* (2005). Boston: Houghton Mifflin.

Aston, G., Bernardini, S., & Stewart, D. (Eds.). (2004). *Corpora and language learners.* Amsterdam: John Benjamins.

Beason, L. (2001). Ethos and error: How business people react to errors. *College Composition and Communication, 53,* 33-64.

Benjamin, W. (1969). *Illuminations.* New York: Schocken.

Bernstein, T. (1977). *The careful writer: A modern guide to English usage.* New York: Atheneum.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge, England: Cambridge University Press.

Biber, D., Johansson, S., Leech, G. Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, England: Longman, 1999.

Brunner J. S. (1996). T*he culture of education.* Cambridge, MA: Harvard University Press.

Burchfield, R. W. (2000). *The new Fowler's modern English usage* (3rd ed. rev.). Oxford: Oxford University Press.

Burnard, L. & McEnery, T. (2000). *Rethinking language pedagogy from a corpus perspective.* Bern: Peter Lang.

*The Chicago manual of style* (15th ed.). (2003). Chicago: University of Chicago.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

College Board. Types of validity evidence. In *ACES: Validity Handbook.* Retrieved November 3, 2006, from http://www.collegeboard.com/highered/apr/aces/vhandbook/evidence

Collins Publishers. (2006). *Accessing the corpora*. Retrieved November 21, 2006, from http://www.collins.co.uk/books.aspx?group=155

Connors, R. J., & Lunsford, A. A. (1988). Frequency of formal errors in current college writing, or Ma and Pa Kettle do research. *College Composition and Communication, 39,* 395-409.

Copperud, R. H. (1980). *American usage and style: The consensus*. New York: Van Nostrand Reinhold Company.

Council of Science Editors. (2006). *Scientific style and format: The CSE manual for authors, editors, and publishers* ( 7th ed.). Reston, VA: Council of Science Editors.

Crotty, M. (1998). *The foundations of social research*. London: Sage.

Crystal, D. (1987). *The Cambridge encyclopedia of language.* Cambridge, England: Cambridge University Press.

Cunningham, A. E. & Stanovich, K. E. (1991). Tracking the unique effects of print exposure in children: Associations with vocabulary, general knowledge and spelling. *Journal of Educational Psychology, 83,* 264-274.

Cunningham, A. E. & Stanovich, K. E. (1993). Children's literacy environments and early word recognition subskills. *Reading and Writing: An Interdisciplinary Journal, 5,* 193-204.

Dean, D. (2006). Interview on November 29, 2006.

Ebbitt, W. R., & Ebbitt, D. R. (1990). *Index to English.* Oxford: Oxford University Press.

Evans, B., & Evans, C. (1981). *A dictionary of contemporary American usage.* New York: Galahad Books.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *SSLA, 24*, 143-188.

Feris, P. (1970). Two views of English usage: Webster's Third and the American Heritage panel. *College English, 31,* 836-844.

Feyerbend, P. (1987). *Farewell to reason*. London: Verso.

Follet, W. (1966). *Modern American usage: A guide.* New York: Hill Wang.

Francis, W. N. (1954). Revolution in grammar. *Quarterly Journal of Speech*, *40,* 299-312.

Garner, B. (2003). *Garner's modern American usage.* Oxford: Oxford University Press.

Gibaldi, J. (2003). *MLA handbook for writers of research papers* (6th ed.). New York: Modern Languages Association of America.

Goldstein, N. (2002). *The Associated Press stylebook and briefing on media law*. Cambridge, MA: Perseus Publishing.

Gove, P. B. (1961). *Webster's third new international dictionary, unabridged* (1961). Springfield, MA: G. and C. Merriam Co.

Greenbaum, S & Taylor, J. (1981). The recognition of usage errors by instructors of freshman composition. *College Composition and Communication, 32,* 169-174.

Greenbaum, S., & Whitcut, J. (1988). *Longman guide to English usage.* Essex, England: Longman.

Hairston, M. (1988). Not all errors are created equal: Nonacademic readers in the professions respond to lapses in usage. *College English, 43*, 794-806.

Hartwell, P. (1985). Grammar, grammars, and the teaching of grammar. *College English, 47*, 105-127.

Hauk, O., Patterson, K., Woollams, A., Watling, L., Pulvermüller, F., & Rogers, T. T. (2006). [Q:] When would you prefer a *sossage* to a *sausage*? [A:] At about 100 msec. ERP correlates of orthographic typicality and lexicality in written word recognition. *Journal of Cognitive Neuroscience, 18,* 818-832.

Hope, J. (2000). Rats, bats, sparrows and dogs: Biology, linguistics and the nature of Standard English. In L. Wright (Ed.) *The development of standard English 1300-1800: Theories, descriptions, conflicts* (pp. 49-56). Cambridge, England: Cambridge University Press.

*IEEE Editorial Style Manual.* http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/authors/ transjnl/stylemanual.pdf. Accessed October 27, 2007.

JAMA and the Archives Journal. (2007). *AMA manual of style* (10th ed.). New York: Oxford UP.

Johnson, S. (1755). *A dictionary of the English language: In which the words are deduced from their originals, and illustrated in their different significations by examples from the best writers.* London: Printed by W. Strahan, for J. and P. Knapton; T. and T. Longman; C. Hitch and L. Hawes; A. Millar; and R. and J. Dodsley.

Keene, D. (2000). Metropolitan values: Migration, mobility and cultural norms, London 1100-1700. In L. Wright (Ed.) *The development of standard English 1300-1800: Theories, descriptions, conflicts* (pp. 93-114). Cambridge, England: Cambridge University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Lanham, R. (1979). *Revising prose*. New York: Scribner's.

Lass, R. (2000). A branching *path*: low vowel lengthening and its friends in the emerging standard. In L. Wright (Ed.) *The development of standard English 1300-1800: Theories, descriptions, conflicts* (pp. 219-229). Cambridge, England: Cambridge University Press.

Leonard, D. J. & Gilsdorf, J. W. (1990). Language in change: Academics' and executives' perceptions of usage errors. *The Journal of Business Communication, 27,* 137-158.

Lovinger, P. W. (2000). *The Penguin dictionary of American English usage and style*. New York: Penguin.

Lowth, R. (1762). *A short introduction to English grammar.* London: A. Miller & R. & J. Dodsby.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2$^{nd}$ ed.). Edinburgh: Edinburgh University Press.

Merleau-Ponty, M. (1962). *Phenomenology of perception*. London: Routledge & Kegan Paul.

*Merriam-Webster's collegiate dictionary* (11$^{th}$ ed.). (2004). Springfield, MA: Merriam-Webster.

Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge, England: Cambridge University Press.

Milroy, J. (1999). The consequences of standardization in descriptive linguistics. In T. Bex and R. J. Watts (Eds.) *Standard English: The widening debate* (pp. 16-39). London: Routeledge.

Milroy, J. (2000). Historical description and the ideology of the standard language. In L. Wright (Ed.) *The development of standard English 1300-1800: Theories, descriptions, conflicts* (pp. 11-28). Cambridge, England: Cambridge University Press.

Morris, W., & Morris, M. (1985). *Harper dictionary of contemporary usage.* New York: Harper & Row.

Murray, L. (1795). *English grammar.* York, England: Wilson, Spence, and Mawman.

Newman, E. (1980). *Strictly speaking and a civil tongue.* New York: Warner Books.

*Oxford English dictionary* (2nd ed.). (1989). New York: Oxford University Press.

*The Oxford dictionary and usage guide to the English language*. (1995). Oxford: Oxford University Press.

Peters, P. (2004). *The Cambridge guide to English usage.* Cambridge, England: Cambridge University Press.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York: W. Morrow and Co.

*Publication manual of the American Psychological Association* (5th ed.). (2001). Washington, D. C.: American Psychological Association.

Quirk, R. (1995). *Grammatical & lexical variance in English.* London: Longman.

Quirk, R., Greenbaum, S., Leech, G., & Swartvik, J (1985). *A comprehensive grammar of the English language.* London: Longman.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*, 94-104.

Safire, W. (1984). *I stand corrected.* New York: Times Books.

Schuster, E. H. (2003). *Breaking the rules: Liberating writers through innovative grammar instruction.* Portsmouth, NH: Heinemann.

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20,* 51-68.

Strunk, W., & White, E. B. (1979). *The elements of style* (3rd ed.). New York: Macmillan.

Talento-Miller, E., & Rudner, L. M. (2005). GMAT validity study summary report for 1997 to 2004. In *GMAC Research Reports*. Retrieved November 3, 2006 from http://www.gmac.com/gmac/ResearchandTrends/Tools/RR0506_VSSSmmary 1997to2004.htm

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Weaver, C. (2007). The grammar book: A guide to smart thinking. Portsmouth, NH: Heinemann.

Webster, N. (1828). *An American dictionary of the English language.* New York: S. Converse.

*Webster's dictionary of English usage.* (1989). Springfield, MA: Merriam-Webster.

Williams, J. M. (1981a). The phenomenology of error. *College Composition and Communication, 32*, 152-168.

Williams, J. M. (1981b). *Style: Ten lessons in clarity and grace*. Glenview, IL: Scott, Foresman.