CONSTRUCTION AND APPLICATION OF BOUNDED VIRTUAL CORPORA OF BRITISH AND AMERICAN ENGLISH

by

Garrison E. Bickerstaff, Jr.

(Under the Direction of William A. Kretzschmar, Jr.)

Abstract

This study proposes a systematic approach that capitalizes on the quantitative advantages of corpus linguistics and caters to researchers' needs for extremely large amounts of text that can address regional American speech. The Bounded Virtual Corpus (BVC), an online virtual corpus is constructed within finite boundaries and contains a ten-year span (1998 to 2007) of 25 American newspapers distributed over five different regions. The BVC thus reflects regional frequency of word use and supports lexicographic decision making. A separate British component reflects lexical activity in British English and contains the full text of 5 newspapers over the same ten-year span, which enables direct comparisons between British and American English. The BVC's virtual corpus methodology allows for calculation of rates of occurrence and other metrics for word use that cannot be achieved by using the Internet as a corpus. Estimated word counts for each component—5 billion words for the American regions and 1 billion words for the British section—provide a baseline for such calculations. The BVC is unusually large in comparison to other currently well-known corpus projects and provides more evidence than smaller corpora for the study of careers of words, especially lower frequency words. The *LexisNexis Academic* interface includes a robust downloading interface that allows BVC search results to be saved conveniently and uploaded to offline concordancing applications as needed for analyses. Through results from the BVC, this study shows some recent differences in use of American and British English in the phrase *go missing* as well how the BVC can be used to identify and describe the trajectory of a neologism, *carb*, before and after it has reached its peak in frequency. There is also a demonstration of how the BVC can be a useful lexicographic tool regarding potential revision of *A Dictionary of Americanisms on Historical Principles*. The BVC can be replicated using detailed instructions included in the dissertation, and can be used free of charge by any researcher who has legal access to *LexisNexis Academic*.

Index Words: American English, American-British Variation, Corpus linguistics, Virtual corpus, *Dictionary of Americanisms on Historical Principles*, Neology, Lexicography, *LexisNexis Academic*

CONSTRUCTION AND APPLICATION OF BOUNDED VIRTUAL CORPORA OF BRITISH AND AMERICAN ENGLISH

by

Garrison E. Bickerstaff, Jr.

B.A., Tennessee Technological University, 2000

M.A., Tennessee Technological University, 2002

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2010

CONSTRUCTION AND APPLICATION OF BOUNDED VIRTUAL CORPORA OF BRITISH AND AMERICAN ENGLISH

by

Garrison E. Bickerstaff, Jr.

Major Professor:	William A. Kretzschmar, Jr.
Committee:	Don R. McCreary Christy Desmet

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2010

Acknowledgements

In my career as a doctoral student, UGA's reference librarians have constantly supported me. This study could not have begun or finished without their support. Fellow students have been an important support as well. I have especially benefitted from the friendship and advice of Iyabo, Bess, Sasha, Ondra, Mark, Kyle, and Jenn.

The First-year Composition Program at UGA connected me with excellent students and colleagues and gave me the opportunity to be a part of an important community within UGA's Department of English. The Division of Academic Enhancement has also been a source of important friendships and learning experiences in connection with both students and colleagues.

I wish to acknowledge the committee members for this dissertation who have provided support and leadership over the years of my studies at UGA. Dr. Christy Desmet has not only supported me when I was a TA in the First-year Composition program at UGA, but she has also sacrificed time to read drafts, ask and answer questions, and provide valuable feedback.
Dr. Don McCreary has supported me with his leadership and knowledge of lexicography at UGA in class, in office visits, and even at out-of-town conferences. No matter the location, he has remained concerned, supportive, and interested in my progress.

Finally, this dissertation's director, Dr. William A. Kretzschmar, Jr., has provided ideas, support, direction, and motivation in times of my successes and difficulties. He has constantly returned e-mails and provided time in office visits. His sacrifices of time and concern have been remarkable, and I have truly appreciated all of his contributions and leadership.

Table of Contents

List of Tables

Page

List of Figures

Page

Chapter 1

Introduction

This study explains specific steps that are required to construct and use Bounded Virtual Corpora (BVCs) in the *LexisNexis Academic* database. These corpora, American and British, are online constructions of full-text American and British newspapers, respectively. The BVCs possess specific elements of construction, such as specific titles and a time frame that consists of the ten year period of 1998 to 2007. Each corpus is very large, and the study has systematically calculated estimated word counts for each. The combination of both constructed and virtual attributes makes BVCs highly innovative and powerful tools for the study of recent American and British English.

Corpus linguistics does not currently embrace a systematic approach that capitalizes on the quantitative advantages of corpus linguistics and caters to researchers' needs for extremely large amounts of texts that focus on regional speech. Current corpora frequently exceed the classic size of one million words and can reach into the hundreds of millions; however, the word counts for such a needed, new project could extend well into the billions. The number of files required for such a methodology would be extremely burdensome and time consuming to download and difficult store on a personal computer.

The significance of regionalisms has been known for some time, but linguists have not had the ability to quantitatively judge how regionalisms are actually being used in terms of frequency within large amounts of regional texts. So, clearly defined regional texts within BVCs make quantitative judgments about the use of regional speech possible.

BVCs can also assist the study of neology in American English. Frequently, new words are identified, which is important, but the careers of such forms are often unknown. The appearance of new forms, as well as their rises and falls, can be determined and monitored by region through the regionally-based text approach that is possible with a BVC. A BVC can isolate, by region, the frequency of new words and show that while a new word may be dying in one area, the same form may be current in another. Such regional information is valuable for the challenging of native speaker intuition that may be mistaken about actual usage (Stubbs 9).

British and American BVCs can also answer questions about British constructions that are fairly recent manifestations in American English. As noted above, in the case of neology, a large corpus of electronic texts could be used to identify regional patterns of a Briticism in American English and also provide evidence of the construction's career in American English.

Further, the BVC approach, which uses large numbers of regional texts, can assist with answering the question, "What is an Americanism?" Often dictionaries of American English do not acknowledge frequency, regional use, or show a comparison with British English in the determination of what an Americanism is. Similarly, dictionaries of American English often tend to include world English forms in large quantity with forms that are Americanisms. Therefore, BVCs can assist lexicographic decision making with empirical evidence of the lexical use that is current, by region, throughout the United States.

The BVC is a significant contribution to scholarship because its innovative construction and use of widely available online resources share with the scholarly community the fact that a very large and well organized corpus of recent newspaper texts can be constructed within the *LexisNexis Academic* database. The construction of the BVCs can be replicated and, more importantly, can be altered to fit the exact needs of another researcher. These needs could be

met through the flexibility of the online database and the user's own wishes for a specially designed BVC.

Further, beyond the feasibility of construction, this study shows that a word count estimation can be employed in connection with a BVC. Word count and virtual corpora are concepts that usually do not associate with each other; this study shows that a word count estimation is feasible for a BVC. The estimated word count will give significance to lexical frequencies within the corpora. This significance enables quantitative judgments, based on regional comparisons of American English, as well as comparisons of American English and British English.

LexisNexis Academic is available at many schools and libraries, so the scholarly community can access the database from a number of locations and replicate the methodology of this study for their own scholarly applications.

Overview of Methodology

As noted above, BVCs, which are the focus of this study, are both virtual and constructed. The Internet, via a search engine (such as Google) is a virtual corpus (Teubert and Cermakova 124). Teubert and Cermakova explain, "The Internet is a virtual corpus, and, like the discourse of any language community, we cannot expect to access it as a whole. Normally if someone wants to use the Internet as a source, they should, therefore, download all the texts they are working with, and compile them in a special corpus [. . .]" (125). In contrast to this explanation of a virtual corpus, all of the texts within the BVC s are accessible and contained in full-text format within the finite bounds of the *LexisNexis Academic* database. These texts

consist of the available full text of 25 American and 5 British newspaper titles from 1998 to 2007.

The virtual nature of the BVCs rests in the fact that the texts are reposed online, and they are not downloaded into a portable file or directory. The texts within the BVCs are accessible only within the online *LexisNexis Academic* database. Another finite, non-virtual attribute that these corpora possess is the estimation of the total word count for each corpus. After a specialized estimation process, the word count is estimated at 5 billion words for the American corpus and 1 billion words for the British corpus.

The use of newspapers is a natural choice for these corpora because of their convenient online availability and because newspapers reflect the speech of authentic social situations. Hockey explains, "that newspapers create a specific type of language from a fusion of literary language and some spoken and special purpose language. They represent the modern language as it is, and can be viewed as both homogeneous and heterogeneous" (21). Newspapers, therefore, can record volatility within language, such as new words that can disappear shortly after they appear as well as established high frequency forms. Further, newspapers in this study, because of their significant word count, should reflect the career of a Briticism that has entered American currency within recent years.

Methods of Analysis

On one hand, *LexisNexis Academic* is merely a repository; however, on the other hand, *LexisNexis Academic* can be focused to answer specific linguistic questions relating to electronic texts. For example, the American BVC is organized into 5 subcorpora; each subcorpora represents a region of the United States. The 25 titles can be searched as a group (complete

BVC) or by region (subcorpora), which would be 5 titles. Also, *LexisNexis Academic* displays search results by newspaper title which instantly gives the user information about use by region.

The American and British corpora will be used for comparison; for example, currency of a form in British English would be evidence that a form is not an Americanism. In the case of *go missing*, its recent career in American English can be compared to that of *go missing* in recent British English. The British BVC will also be useful in noting how American neologisms might manifest themselves in British English.

Organization of the Dissertation

In Chapter 2, literature relevant to the study is acknowledged and discussed. This literature relates to the historical and current progress of corpus linguistics. The use of databases as linguistic tools is also discussed. Some overview of the study of American English as well as the comparison of American and British English is discussed as well. Finally, the use of statistics in connection with corpus linguistics is discussed.

The methodology of this study is explained and rationalized in Chapter 3. This chapter explains BVC construction, estimated word count in relation to a BVC, and the use of the *LexisNexis Academic* interface as a linguistically useful tool.

The British construction, *go missing*, is the focus of the case study in Chapter 4. The form's recent careers in both British and American English is observed through the BVCs, and background on this and similar constructions is explained through the use of the BVCs. In this chapter the utility of the BVC approach for comparison of corpora is especially important because of the heritage of *go missing* as a British construction.

Chapter 5 is a case study of the American English neologism, *carb*, as it relates to the popular "low carb" diet movement. This form's career is analyzed in both the American and British BVCs. The question of the form's endurance beyond the diet's peak is also pursued.

Chapter 6 demonstrates how the American BVC can be used to support a definition of "What is an Americanism?" In this chapter the British BVC shows what American forms in the pilot study may have reached currency in British English. This chapter demonstrates that the BVC can be used for practical lexicographic applications. Chapter 6 focuses on a specific span of *A Dictionary of Americanisms on Historical Principles* (*DA*) (1951) as the core of a pilot study to update the *DA*.

Chapter 7 is the conclusion for this study in which the literature review is revisited and final thoughts are shared.

Chapter 2

Literature Review

Two of the longstanding goals of corpus linguistics are size (large amounts of text) and balance (a variety of text types). These attributes work together to yield a corpus that is useful to a researcher. Useful refers to the evidence a corpus can supply in response to questions about lexical use. In the case of low frequency forms, for example, a million word corpus might not provide any examples of use, but a hundred million word corpus might be able to give such useful examples. Such examples, real use in authentic texts, are important because they stand as evidence of how a form is used. Examples of real use have stood as an identifying characteristic of many important English dictionaries.

Evidence in Lexicography

Historically and currently, the practice of lexicography has not maintained a single system for using and providing evidence to the users of dictionaries. The role of evidence in lexicography has evolved and in lexicography, newer, advanced practices or older practices may be employed according to the methodologies of the person or publisher behind each dictionary. Lexicographers have employed evidence of lexical use to provide examples, real or invented, of a particular form to illustrate use and meaning.

Dictionary makers have always relied on introspection, to some degree, as evidence. In an effort to employ real use, lexicographers have created citation files which are paper or electronic collections of quotations that employ a form that the dictionary makers are interested

in. Citation files are useful because they provide examples of real use, but the organization behind a citation file, such as why some forms are included and others are not, depends heavily on human subjectivity. Therefore, important forms might be excluded from the citation file, or important variations of included forms may be excluded (Sinclair 100).

An antidote to the shortcomings of a citation file as the central source of evidence for a dictionary is the use of a balanced electronic corpus. The use of a corpus reverses the process of evidence gathering established by a citation file approach; a corpus approach allows a lexicographer to observe patterns within the corpus and reflect that evidence in the making of a dictionary (Sinclair 26). Some dictionaries, which relied on citation files, have provided important pioneering accomplishments for subsequent lexicographers, who had access to more advanced tools.

Dictionaries and citation files

Samuel Johnson's *Dictionary* (1755) united a corpus of headwords with illuminating copy (Landau 64). This work is significant not only for its enduring influence but also for its employment of evidence. Johnson used authentic texts to show the reader usage of dictionary entry forms. This illuminating copy lent authority to the *Dictionary* and presented evidence to its user. Such a corpus, both of dictionary entry forms and of text that illuminates use of the forms, continues as a practice in dictionary making, but the notion of what a corpus is and how it relates to dictionary making has evolved since 1755.

The *New English Dictionary* (*NED*) was led by James A. H. Murray and published between 1882 and 1928 (Mugglestone 1). One of the guiding principles of the *NED* was the inclusion of illustrative quotations for evidence of use, and the politics of text-type for

illustration was sometimes a matter of controversy for Murray because he included a wide variety of text types (Read (1986) 43-44). The *NED*, was later renamed and combined with a supplement and released as *The Oxford English Dictionary* (1933 *OED 1*); four volumes of supplements followed from 1972 to 1986; and in 1989 after another revision process, *The Oxford English Dictionary* Second Edition (*OED 2*) was published (Mugglestone 1). The *OED 2* is legendary for being the dictionary "[. . .] which defined a total of 615,100 words, and illustrated those definitions with 2,436,600 quotations" (Winchester 247). The *OED 2* created an extremely large collection of headwords with illustrative quotations for evidence, and the Oxford University Press had the foresight to expand users' access to their *Dictionary*.

In 1984, keyboarding began to convert the *OED* to an electronic format (Burchfield 26). This effort eventually resulted in a series of CD-ROM versions of the *Dictionary* that began in 1992 (Jackson 58). In 2000 the *OED Online* was launched, and it reflects the *OED 2*, and completed work toward the *OED 3* (Jackson 59). The compact disc and online formats gave users the ability to execute complex searches quickly with regard to many variables beyond a simple headword search.

Johnson and the *OED* editors compiled useful citation files. These quotations or segments of text, on paper, do not function as dynamically as a balanced, constructed corpus can. For example, a linguist might be curious about what forms collocate with a certain entry form. Instead of referring to citation files, in paper or electronic form, which contain some sentences that exemplify use of the form, the linguist could search an electronic corpus for the form in question (node) plus surrounding text (horizon), such as five words to the left and five words to the right. The corpus would contain more textual evidence and provide more information to the user. In the field of corpus linguistics today, a corpus must be able to provide support for

research questions that reach beyond the one or so sentences that illuminate dictionary entry forms. The text of a citation file is isolated and does not have the dynamic capability that a corpus does to provide evidence about lexical use. Traditions within British linguistics eventually led to an approach in dictionary making that relies on corpus linguistics rather than the traditional citation file.

British Linguistics

A tradition of British linguistics for decades has been an emphasis on the study of words in their real uses. This emphasis is similar to Johnson's and the *OED*'s use of authentic citations but different because the British tradition seeks large numbers of examples of use in order to make determinations about meaning. J. R. Firth explains that lexicographic decision making based on a traditional citation file, without considering the use of words on a large scale, creates a problematic approach: "When a lexicographer has arbitrarily decided how many 'meanings' he can conveniently recognize in the uses of a given word, he limits his entries accordingly and, after definitions of the 'meanings' in *shifted terms*, he supports them by citations, usually with literary authority" (Firth 11). Firth explains that an improved lexicographic process would be to invert the process and allow large amounts of text (real use) to testify of words' meanings. He admonishes, "The listing or preparation of written materials in the restricted language from which exhaustive collocations of the selected word are to be collected" (26). He explains the inversion of tradition: "Draft entries can now be made, one for each group, definitions can be given and from the *collocations* one or two may be chosen to become *citations* keyed to the definitions" (26). Therefore, Firth's notion is to collect uses of words, which show a significant trend, and from that collection of texts a quotation can be selected to stand as an illustration in a

dictionary. The foundation of this notion of an improved lexicographic approach rests in his classic line, "You shall know a word by the company it keeps!" (11). In 1957 Firth's vision of an improved process was just that—a vision. Technology was not in place to connect his vision to practical application; however, John Sinclair followed in the Firthian tradition and connected his vision with new advances in technology.

Sinclair notes in his 1991 book, *Corpus, Concordance, Collocation*: "Linguists have had to rely on their intuitions, their limited capacity for thorough textual analysis, and whatever has caught their eye or ear as they have encountered large extents of language behaviour, in their daily lives or in their professional work (100). Sinclair's comments highlight the need for corpus linguistics which can give researchers the ability to rely less on intuition and study real use of words on a large scale. In an echo of Firth's remarks on citation file-based lexicography, Sinclair addresses another problem: "The great dictionaries of English used human beings to evaluate their examples, and there is as yet no substitute. This method is likely to highlight the unusual in English and perhaps miss some of the regular, humdrum patterns" (100). Of corpus linguistics, Sinclair remarks, "The language looks rather different when you look at a lot of it at once" (100). Sinclair's focus on advancing the study of real use in lexicography yielded the *Collins Cobuild English Language Dictionary* (Moon 159).

The COBUILD project (Collins Corpus) began in 1981, and the *Collins Cobuild English Language Dictionary* (CCELD), which was constructed based on the Collins Corpus, was completed in 1986 (Moon 159-60). Moon notes of the CCELD, "As a text, it had a massive impact on the practice of commercial dictionary-making, on meta-lexicography and lexicographical theory, and beyond that on language description in general" (Moon 159). The CCELD marks the beginning of a new stage in lexicography because corpus methods were used

and evidence in the form of citations was abandoned (Moon 165). Moon comments about the position of the CCELD in the field of lexicography: "It took less than seven years—just sixty-four months, in fact, between the main start-up of the Cobuild project in 1981 and the completion of the CCELD dictionary text in 1986—for the lexicographical world to be transformed" (160). Johnson and Murray both produced fabulous achievements with the tools that were available to them, but the CCELD took advantage of new tools they did not have.

The Firthian tradition that led to the CCELD has focused on the study of large amounts of text that give evidence of authentic lexical use. The CCELD is more than just a dictionary with a new approach; the CCELD represents a new standard in the field. Lexicographers might not embrace a corpus-based approach, after the CCELD, but the projects that are leaders in the field will. The CCELD showed not only the feasibility of a corpus approach, but also the detail and precision that a corpus approach can give to a dictionary project. Such detail and precision come from the evidence that results from significant size and balanced construction that corpus linguistics offers researchers. The heritage in lexicography that would eventually lead to the CCELD began with the classic corpora.

Corpus Construction

The first of the classic corpora was the Brown Corpus, which was the first machine readable corpus of English, and it was created between 1963 and 1964 (Kucera and Francis xvii). The Brown Corpus " [. . .] consists of 1,014,312 words of running text of edited English prose printed in the United States during the calendar year 1961" (Hockey 16). One of the pioneering aspects of the Brown Corpus is the fact that later corpus compilers would follow its methodology in the pursuit of balance (Hockey 16). Hockey explains, "The Corpus is divided into 500

samples of 2000+ words each. Each sample begins at the beginning of a sentence and ends at the first sentence ending after 2,000 words (Hockey 16). Hockey notes that the Brown Corpus is both useful for current research purposes and represents a methodology of construction that is limited (16); for example, the Brown Corpus does not "[. . .] include verse, drama, or spoken texts" (16).

The Brown Corpus was constructed according to a focused and well-ordered plan. The Brown Corpus is organized into fifteen categories of genres or text types (Kucera and Francis xix). Within each of these categories is a number of texts, ranging from 6 to 80 (Kucera and Francis xix). So, the construction of the corpus is deliberate in terms of text type and the number of samples for each text type and the identification of each text sample; each text type was designated by a letter (A – R), and each text sample in each text type was assigned a numerical designation (i.e. 1, 2, 3 and so on) (Kucera and Francis xix). This methodology allows the corpus builders to explain how the balance of the corpus was achieved, and through the alpha numeric assignments, text samples could be traced to their point of origin. The text types within the Brown Corpus include: three types of press, religion, skills and hobbies, popular lore, Belles Lettres, miscellaneous, learned and scientific writings, five types of fiction, and humor (Kucera and Francis xix). Even though larger and more complex corpora followed the Brown Corpus, the standards of balance that it set forth have been upheld through many subsequent corpus projects.

The construction of the Lancaster-Oslo/Bergen Corpus (LOB) followed the model of the Brown Corpus and included one million words of British English texts from 1961 within fifteen categories of text types (Baker et al. 101). The presence of LOB gave linguists the opportunity to compare American and British English via constructed, downloadable corpora. The similar

construction methodologies of the Brown and LOB corpora and significant amount of text (for their time), constructed within a balanced plan, facilitated scientific inquiries into comparisons of British and American English.

The Freiburg-Brown Corpus (FROWN) consists of "One million words of edited American English published in 1991; divided into 2000-word samples in varying genres intended to replicate the Brown Corpus (Meyer 145). The Freiburg-Lancaster-Oslo-Bergen corpus (FLOB) consists of "One million words of British English published in 1991; divided into 2,000-word samples in varying genres intended to replicate the LOB Corpus" (Meyer 145).

The presence of FROWN and FLOB, along with the Brown and LOB corpora, gave linguists the opportunity to pursue both synchronic and diachronic studies and comparisons of American and British English. Further, analysis of these classic corpora could give linguists important starting points for linguistic investigations; that is, these corpora could not only answer questions about use of American and British English, but they could also contradict native speaker intuition (Stubbs 9). Such a contradiction could have been the motivation for important studies that, without these classic corpora, would not have appeared. The classic corpora are the first step not only toward larger, next generation corpora but also toward independently constructed corpora.

Next generation corpora

The compilers of next generation corpora capitalized on the need that researchers have for larger corpora, which would give new and increased evidence about words and their uses. Such new evidence could include evidence of low frequency forms, evidence of forms that are non-standard or slang in nature, and evidence of unusual uses of usual forms. Such evidence

could be helpful in a variety of ways, such as mainstream lexicography as well as learners' dictionaries. Authentic evidence for high frequency forms could be especially helpful for learners' dictionaries because of, for example, the complexities connected with delexicalization which might not challenge native speakers' understanding of their own language—but could well challenge learners of English (Sinclair 113). These unusual uses exemplify "[. . .] the regular, humdrum patterns [. . .]" that Sinclair refers to, and he notes that too much dependence on introspection can cause us to fail to recognize them (100). The next generation corpora also borrowed from the classic corpora the notion of a plan. The plan that the next generation corpora tend to follow focuses on the construction of a corpus that is much larger than the classic corpora and the inclusion of Web-based texts. Web-based texts, such as online postings are not only conveniently accessible, but they add a new, contemporary text type to the next generation corpus. Such an addition shows not only a focus on an increase in size but also the dynamic nature of the next generation corpora. That is, the ways in which communities communicate are changing, so the corpus makers are reflecting that change in their corpus building.

Most major projects in corpus linguistics, either current or recently completed, generally focus on the construction of corpora that are much larger than one million words. Such next generation corpora include The International Corpus of English, The British National Corpus, The American National Corpus, The Collins Corpus (which includes The Bank of English), The Cambridge International Corpus, and The Oxford English Corpus.

The International Corpus of English (ICE) is a different project from other large English corpus projects, which tend to focus on the representation of one type of English. Baker et al. explain that "The goal of the [ICE] project was to develop a series of comparable corpora of different Englishes" (92). The ICE website explains that "Each ICE corpus consists of one

million words of spoken and written English produced after 1989" ("International Corpus of English"). The website also indicates that currently corpora for these varieties of English have been completed for the project: Hong Kong, East Africa, Great Britain, India, New Zealand, Philippines, and Singapore ("International Corpus of English").

The British National Corpus (BNC) is "An approximately 100-million-word corpus of written (90 per cent) and spoken (10 per cent) British English" (Baker et al. 24). Also, "The 4,124 texts mainly originate from the late 1980s and 1990s, although about 5.5 million words were first published between 1960 and 1984" (Baker et al. 24). This corpus is balanced because it encompasses a variety of text types (spoken, newsprint, fiction, academic essays) and time spans of texts.

The American National Corpus (ANC) is an ongoing project that includes texts from 1990 forward ("American National Corpus"). The ANC is not yet complete, but when completed, will contain "[. . .] a core corpus of at least 100 million words, comparable across genres to the British National Corpus (BNC)" ("American National Corpus"). It is interesting to note that just as the construction methodologies of the classic corpora mirrored each other, the ANC mirrors the BNC's construction methodology. The ANC's embrace of the BNC's methodology will ensure that the project will achieve and maintain balance, and again, as we saw with the classic corpora, the congruent methodologies of the ANC and BNC will position them as great tools for the comparison of British and American English.

The size and scope of the Collins Corpus clearly identify it as a next generation corpus. On their website, Collins explains:

> The Collins Corpus is a 2.5-billion word analytical database of English. It contains written material from websites, newspapers, magazines and books published around the world, and spoken material from radio, TV and everyday conversations. New data is fed into the corpus every month, to help the Collins dictionary editors identify new words and meanings from the moment they are first used. ("About the Collins Corpus")

The Bank of English (BoE) is a part of the COBUILD project, and "The corpus contains 524 million words and it continues to grow with the constant addition of new material" ("The Bank of English"). Therefore the BoE is both a reference corpus and a monitor corpus (Baker et al. 18). In order to support the monitor aspect of the corpus "Material is downloaded from websites, and daily feeds are received from many newspapers" ("The Bank of English"). The BoE reflects a variety of Englishes including mostly British as well as some North American and other varieties (Baker et al. 19).

As discussed earlier, the Collins Corpus is the project that supported the publishing of the CCELD which stands as a landmark in the connection of corpus linguistics and lexicography. Note that the corpus did not stand still after the publication of the CCELD; rather, the corpus continues to grow, as other commercial next generation corpora do. This growth follows an orderly plan, so even though the size of the corpus increases, the balance of the corpus is not challenged.

The Cambridge International Corpus (CIC) consists of over one billion words and new texts are added each year ("Cambridge International Corpus"). The CIC includes a wide variety of text types, and the CIC is a consortium member of the American National Corpus, which will increase their holdings of American texts ("Cambridge International Corpus"). The Oxford

corpus share similarities in methodology of construction as well as motivation for construction with the CIC.

The Oxford English Corpus is comprised of texts from 2000 to 2006 ("Composition and Structure"). The texts in the corpus come from a variety of text types and include: academic papers, technical manuals, journals, and a variety of Internet texts ("Composition and Structure"). Oxford's website notes that the journal texts, which, most likely, come from journals that Oxford publishes, are useful (along with newspapers and magazines) for "[. . .] building a picture of norms and standards in English usage" ("Composition and Structure"). Oxford continues to explain that balance is enhanced with the inclusion of Web texts: "Weblogs and newsgroups, which are largely unedited, are a rich resource for examining non-standard language such as slang, regionalism, and neologisms" ("Composition and Structure"). The Oxford English Corpus is mostly constructed of texts of British and American English (80%), and the remainder of the corpus (over 200 million words) represents other varieties of English from around the world ("Composition and Structure").

Oxford and Cambridge both employ a plan that includes a variety of text types and the ongoing inclusion of new texts to achieve balance and significant size. These publishers need large corpora to support their dictionary as well as other language related projects. The large corpora can help projects that need to identify low frequency forms, and inasmuch as the Oxford and Cambridge corpora have an open-ended dimension, these corpora continue to acquire samples of contemporary English. These publishers could face the need to publish a dictionary of forms that are indicative of Web communication. The presence of Web texts in their corpora would be a rich source to find evidence of such forms. Further, the broad array of printed text types, within the corpora, could function as a contrast to the Web texts in the making of

decisions about which forms, for example, really are indicative of Web communication. The multiplicity of text types and large amounts of text within these corpora make them versatile and therefore useful in the support of a variety of publishing projects.

The next generation corpora have capitalized on the principles of construction that the classic corpora established. These principles have allowed the next generation corpora to be balanced and, in keeping with Firthian linguistics, the next generation corpora are an improvement because they are larger than the classic corpora. As the ways modern culture communicates has progressed, the next generation corpora have followed through adopting web-based texts. This adoption is one example of the open-ended nature of many of the next generation corpora. The next generation corpora use web-based texts to remain up-to-date with the culture around them. Some corpora focus on useful web-based texts, and, instead of simply absorbing some web-based texts, these corpora exist entirely on the Internet, and they are known as virtual corpora.

Virtual corpora

The classic and next generation corpora mentioned above all share the attributes of a plan behind their composition as well as construction itself. In this case, construction is the downloading and organizing of the electronic texts within the corpus. Some of the next generation corpora have a monitor feature that allows them to constantly expand by adding new texts. These new texts often are often acquired from the World Wide Web, which is the primary means of access to virtual corpora.

Kilgarriff and Grefenstette assert that the Web is a corpus (334). They also explain that even with the size of the next generation corpora, they are arguably not large enough—especially

in the case of low frequency forms (336). The Web can be a convenient source to access texts to increase corpus size. Kilgarriff and Grefenstette note some problems, from a linguistic point of view, with commercial search engines as a way to access the Web in pursuit of linguistic research. They identify several problems: the search engine results do not present enough preview text surrounding a search form; searches may not be performed according to linguistic specification, such as part of speech; and they also assert that commercial search engine statistics are unreliable (345). The authors conclude with "Our take on the web is that it is a fabulous linguists' playground" (345). The Web has strengths, such as large amounts of text, but it also has attributes that require a linguistic researcher to use it with caution.

In "Googleology is Bad Science," Kilgarriff continues to address and develop concerns about the use of Google for linguistic research. He notes that linguists educate themselves about Google's shortcomings, such as mentioned above, and he asserts: "The argument that the commercial search engines provide low-cost access to the Web fades as we realize how much of our time is devoted to working with and against the constraints that the search engines impose" (148). Kilgarriff also predicts: "Researchers will continue to use Google, Yahoo, and Altavista unless the NLP community's resources are 'Google-scale'" (149). The Web, via Google, can be very useful for locating evidence for the presence of a single form, but because we do not know exactly what the Internet contains, the Internet as corpus should only be used with caution (Teubert and Cermakova 125).

Fairon and Singler used the online application, GlossaNet, to find evidence of quotative *like* on the Internet (325). GlossaNet is a free service which allows the user to obtain concordances from over one hundred newspapers (Baker et al. 78). After a search has been established with GlossaNet, concordance results are e-mailed to the user daily (Baker et al. 78).

GlossaNet is a convenient way to access large amounts newspaper texts online. GlossaNet has boundaries; for example, "GlossaNet does not cover all of the Web, but only certain pre-defined newspaper website [sic] (users are welcome to suggest the addition of new sites)" (Fairon and Singler 333).

Further, Glossanet users cannot access the original text that their search results are based on; rather, users only have the search results. This lack of access prevents users from pursuing pilot studies, such as a search for a particular form, and then downloading the full-text of all the texts in which the form appears. Such texts could function as a sub-corpus for research purposes. Also, users do not have a knowledge of the size of the Glossanet corpus, and because of the detached way users interface with the corpus, they do not have the tools to estimate a word count for the corpus.

Users of virtual corpora take advantage of the Web as linguistic resource, but in doing so, they face a multiplicity of both positive and negative repercussions. For example, the Web does access an enormous amount of text, but the user cannot know what the texts actually are. That is Google cannot be used as a corpus with any understanding of balance as we know we have in the classic and next generation corpora. Glossanet, provides users access to a large amount of texts; however, again, users have no specific understanding of what is in the corpus. With Glossanet, users cannot access the specific texts that they are searching, so the corpus's utility is limited due to the detached nature of the interface.

Independently constructed corpora

McEnery et al. (2006) explain that there is no exact formula for what the size needs to be for a corpus that a user might wish to build; however, he does note the shift that has caused what

used to be large (around one million words) to now be regarded as small (71). They also note the need for a balance of text types to be considered in corpus construction to help the corpus most effectively serve the linguist's research needs (73). As the field continues to evolve, corpus linguists have a selection of construction models to follow, such as ICE, Brown, LOB and the BNC (73). Just as these famous corpora have differing goals, the corpus linguist may construct a corpus to meet specific, individual research needs.

The Norwegian Newspaper Corpus (NNC) is a dynamic (constantly growing) corpus, which was begun in 1998, and consists of newspaper texts collected from the Internet ("The Norwegian Newspaper Corpus"). The NNC website explains that "Approximately 200,000-250,000 running words are added per day. As of April 2008, the database consists of about 640 million words, and it is by far the greatest searchable corpus of Norwegian" ("The Norwegian Newspaper Corpus"). This corpus features advanced technologies for both the collection of texts and the analysis of the corpus. For example, as a text is collected, it is classified by which of the two types (*bokmal*, or *nynorsk*) of Norwegian language that it represents (The Norwegian Newspaper Corpus). Also, advertisements are excised, through a software application, and the texts are constantly automatically analyzed in search of developments, such as neologisms (The Norwegian Newspaper Corpus).

Cristiano Furiassi and Knut Hofland created a 20 million word corpus (HF Corpus) from online Italian newspaper texts (347). The motivation of the construction of this corpus was the identification of false anglicisms in Italian, and this corpus followed the design of the NNC for collection of newspaper texts from the Internet (361). The HF Corpus includes texts from three Italian newspapers, and the texts were gathered over the course of ten months (347).

Mark Davies created a corpus from the text of *TIME* magazine from 1923 to the present. This corpus is known as the TIME Magazine Corpus (TIME), and he placed the TIME corpus online for public use. This project is noteworthy for its inclusion of large amounts of full-text articles that span several decades. Davies also created the Corpus of Contemporary American English and placed it online for public use. He notes, "The corpus contains more than 385 million words of text, including 20 million words each year from 1990-2008, and it is equally divided among spoken, fiction, popular magazines, newspapers and academic texts" (COCA). The corpus can be searched in terms of the above divisions, or in terms of even more specific classification of texts (COCA). Such search possibilities enable a variety of comparisons to be conducted within the corpus (COCA).

The TIME corpus covers a long time span and consists of a large amount of text, but the fact that all of its text falls under one title creates a troublesome circumstance in terms of its function as a linguistic corpus. Magazines have in-house style sheets and other directives that must be followed. Because all of the texts in the TIME corpus are within a single title, a sense of balance cannot be achieved with the TIME corpus. In comparison to the TIME corpus, COCA has a more balanced construction. Still, COCA has some troublesome construction attributes. For example, the user does not know an exact or estimated word count for each year or for each text type (spoken, fiction, popular magazines, newspapers, and academic texts) (COCA). The COCA website does explain that the corpus is "equally divided" among these text types. Because specific estimated or exact word counts are not available with COCA, quantitative studies, such as those that require tests of statistical significance, would be extremely difficult to conduct based on search results from COCA.

Even though the TIME and COCA corpora provide the user with access to bibliographic information for the original unit of text that connects to each search result, they do not provide local access to the original full-text. For example, the corpora provide an expanded context that includes several lines of text before and after the search word. This is as much text as COCA will supply; the TIME corpus supplies a similar segment of text and a hyperlink to an external commercial website for access to the full text. The user of the TIME corpus does have access to the original full text, but the detached location of it is inconvenient.

Finally, COCA contains some regionally-based texts, such as newspapers, but the corpus does not control for geography. That is, search results can be refined by text type, but they cannot be refined by geography. For example, a search might provide results that include a mid-western newspaper, but a search of only mid-western sources is not possible. An element of specific, geographic control would make COCA a more powerful and useful linguistic corpus of American English.

Sebastian Hoffmann constructed a corpus of CNN transcripts to pursue research questions relative to recent spoken English (Hoffmann 69). Hoffmann notes that the Internet as a whole is a large corpus, but it is filled with a many different text types and is not organized (Hoffmann 69). The corpus of around 172 million words that Hoffmann created was annotated so that specific searches can be performed (Hoffmann 69). For example, one priority for Hoffmann was the ability to search by speaker; the study identified that Wolf Blitzer was responsible for the greatest number of spoken words at over 5 million, and George W. Bush accounted for over 2 million words (Hoffmann 77). This study shows that useful texts, which pertain to specific research questions, can be found and harvested on the Internet. Further, corpus annotation, after the downloading process, allows the researcher, who uses this corpus, to

focus on highly specific matters, such as speaker and related variables, such as political role or gender. Also, this corpus, as shown by the author, is also highly useful for identifying and tracking changes in English usage, such as the *so* (intensifier) + verb collocation (Hoffmann 79).

All of these independently constructed corpora take advantage of the Internet for harvesting or displaying text or both. The Norwegian Newspaper Corpus uses the Internet to obtain its texts and increase its size as this corpus continues to harvest online texts. The HF newspaper corpus is not open-ended, as the Norwegian Newspaper Corpus is, but its builders used the Web to obtain texts in following the methodology of the Norwegian Newspaper Corpus. Davies's corpora feature easy to use publicly available interfaces, and they feature large word counts. His COCA is open-ended and will continue to acquire more texts. Davies's COCA maintains a balanced approach, but the single title inclusion of the TIME corpus prevents it from achieving balanced construction. Hoffmann's CNN corpus responds to the notion that the Web itself is a corpus—but an unbalanced one. His corpus is especially useful for identifying the text attributed to a specific speaker, such as a particular politician or news anchor. The CNN corpus is also very useful for tracking conversational speech trends as the corpus texts are transcripts of speech.

Corpus Linguistics and the Law

When anyone wishes to build their own corpus, they should make themselves aware of the legal implications of their pursuits. McEnery et al. (2006) note, "You might think that you need not worry about copyright if you are not selling your corpus to make a profit. Sadly, this is not the case. Copyright holders may still take you to court" (77). They continues, "Copyright issues in corpus-building are complex and unavoidable. While they have been brought up

periodically for discussion by corpus linguists, there is as yet no satisfactory solution to the issue of copyright in corpus-building" (77). McEnery et al. note that corpus builders should harvest texts from the Internet cautiously (78). They advise, "For example, Cornish (1999:141) argues that probably all material available on the Web is copyrighted, and that digital publications should be treated in the same way as printed works" (78). Still, one can easily make and use large corpora; however, some guidelines are useful.

Corpus linguist Adam Kilgarriff explains, "To be unequivocally, completely, totally, in the clear you need to get copyright clearance from all copyright holders (publishers and/or authors, all speakers for spoken material) ("Legal Aspects"). Kilgarriff notes some factors that can be critical for corpus builders in terns of copyright matters: publishing, extract size, cooperation ("Legal Aspects"). In the area of publishing, he explains:

> The issue is heavier if you are going to publish/copy on the data than if you are not. If it's only for in-house use, then one simple issue is 'who will ever know,' and it is not clear that eg, downloading a report onto your PC's desktop is any different to downloading it into a corpus. Copyright law in general is about the case where someone makes money from selling intellectual property; if you are going to sell a corpus, the issues need taking very seriously, as people will be upset by you making money out of selling their text (unless you give them a share) ("Legal Aspects").

Even though so much text is available on the Web, corpus builders need to be very cautious about what texts they select and how they use them. Because of the easy access to large amounts of useful, copyrighted text, this study will show how to use such copyrighted text in a legal manner.

Online Newspaper Databases

The connection between corpus linguistics and newspaper texts is natural. Newspapers provide text that can be studied, and newspapers have specific geographic origins and

circulations that determine the scale of their distribution. Newspaper texts are available online in several different formats. Both obscure and well known newspapers have websites that present up to the day information as well as archive older texts. Frequently in conversation, "the demise of the (paper) newspaper" is a topic. For the linguist, "the rise of the newspaper" could be a more accurate and relevant conversation topic. For example, in addition to the many current newspaper titles that are available online, databases also archive historical newspapers. These electronic texts can be studied one title at a time, or complex searches of several titles can be constructed.

The *ProQuest Historical Newspapers* database, has several newspaper titles from the United States that can be searched from their earliest issues to recent decades, which can, in some instances, cover over a century. *ProQuest* utilizes optical character recognition to electronically search newspaper text (MacQueen 127). The condition of the newspaper at the time of scanning can be a factor as water spots, wrinkles, faint ink, and other factors can sometimes complicate searches (MacQueen 128). Advanced search features and an understanding of how to effectively search the texts can help users overcome the challenges of the texts that can sometimes be difficult for the interface to recognize. *ProQuest Historical Newspapers* has been especially helpful with antedating dictionary forms (Popik 114).

The *LexisNexis Academic* database is also helpful to linguists. Since the texts in *LexisNexis Academic* are generally more current than the texts in the *ProQuest Historical Newspapers* database, *LexisNexis Academic* assists with research questions that relate to recent neology and other matters of current language use. *LexisNexis Academic* stores hundreds of newspaper titles—some for spans of several years and some for decades. In addition to newspapers, *LexisNexis Academic* also reposes other text types, such as news transcripts and

magazine articles. *LexisNexis Academic* lends itself naturally to the linguist who needs to build a corpus; the database allows 500 articles at once to be merged into a single text (.txt) file. This extremely fast method of file downloading, a variety of text types, and a useful search interface make *LexisNexis Academic* a valuable database for linguists.

Statistics and Corpus Linguistics

Inherent in the notion of a balanced corpus is the presence of a quantitative accounting (word count) for the various elements of a corpus. Generally, these elements are text types, and the user of a balanced corpus should know a word count for each text type. Beyond a simple word count are more complex functions, such as tests of statistical significance.

Raw numbers as a method of comparing corpora can become problematic quickly. For example, two corpora may have different word counts, or a single corpus may have sub-corpora with varying word counts (McEnery and Wilson (2001) 82). These variations quickly make raw word counts problematic for quantitative judgments in corpus linguistics.

McEnery and Wilson overview proportions: "There are several ways of indicating proportion, but they all boil down to a ratio between the size of the sample and the number of occurrences of the type under investigation. The most basic involves simply calculating the ratio: Ratio = number of occurrences of the type / number of tokens in entire sample" (83). This ratio establishes a standardized rate, but it does not determine statistical significance.

Davis explains that when frequency distributions cannot be calculated, such as in the case of the number of instances of a certain form in a corpus, a proportion test may be used to compare the two percentages (number of occurrences per total corpus word count) (39). Davis

shows applications of the proportion test to measure the statistical difference of linguistic usage between groups, such as male and female speakers and speakers from differing social classes (39, 41). The proportion test is especially useful because it allows statistical difference to be measured between two samples of differing population sizes; in the case of corpus linguistics, the proportion test allows the occurrences of one form in two corpora (that have different total word counts) to be tested for statistical difference.

McEnery et al. (2006) explain that "The chi-square test compares the difference between the observed values (actual frequencies extracted from corpora) and the expected values (e.g. the frequencies that one would expect if no factor other than chance were affecting frequencies [. . .] (55). Oakes (1998) clarifies that "This test does not allow one to make cause and effect claims, but will allow an estimation of whether frequencies in a table differ significantly from each other" (24).

Mutual information is a methodology for the calculation of the likelihood of words to collocate with each other. A methodology for this determination is an important way to bypass native speaker intuition and rely on a corpus for results or real use (McEnery and Wilson (2001) 86). McEnery and Wilson show the applicability of mutual information: "Given a text corpus, it is possible to determine empirically which pairs of words have a substantial amount of glue between them and which are, hence, likely to constitute significant collocations in that variety rather than chance pairings (86). They continue, "The mutual information score between any pair of words—or indeed any pair of other items such as, for example, part-of-speech categories—compares the probability that the two items occur together as a joint event (i.e. because they belong together) with the probability that they occur individually and that their co-occurrences are simply a result of chance" (86). The importance of mutual information to

lexicography is the determination of words and their interactions (or careers) for the purpose of determining meaning and senses (McEnery and Wilson 86; Sinclair 100).

Corpus Comparison

Corpus comparison is a form of analysis that naturally follows the presence of corpora. The presence of the Brown Corpus gave linguistics an important, but unilateral, opportunity to study American English. The LOB Corpus created another dimension of analysis; American and British English could be compared through corpus studies. As corpora have evolved, the possibilities for comparison have evolved as well. Still, the notion of what is being compared remains vital to the notion of comparison. Brown and LOB gave linguists the opportunity to compare two corpora with extremely similar methodologies behind their construction. This similarity in terms of balance and construction makes the results of such a comparison significant.

Hofland and Johansson's *Word Frequencies in British and American English* shares information about frequency of forms from studies of the LOB corpus and the Brown Corpus (Hofland). Leech and Fallon note, "This book largely consists of word frequency lists for the British (LOB) Corpus, but contains in one section (Ch. 8) a parallel alphabetical frequency list of both the Brown and LOB corpora" (29). Leech and Fallon's study, "Computer Corpora—What do they tell us about culture?" uses Chapter 8 in Hofland and Johansson's 1982 study as "the chief starting point of this study" (29). Leech and Fallon used the Brown and LOB corpus to go beyond an investigation of frequencies and into a study of how corpora, composed of texts from different cultures, can inform us of cultural differences (Leech and Fallon 31). They constructed 15 domains (categories) to organize their study and have a basis for comparison of British and

American English (35). These domains include, for example, Sport, Transport and travel, Administration and politics, and Social hierarchy (35). They begin the concluding comments of their study with:

> Wrapping up the whole analysis of Section 4 in one wild generalization, we may propose a picture of US culture in 1961—masculine to the point of machismo, militaristic, dynamic, and actuated by high ideals, driven by technology, activity and enterprise—contrasting with one of British culture as more given to temporizing and talking, to benefitting from wealth rather than creating it, and to family and emotional life, less actuated by matters of substance than by considerations of outward status. (44-45)

Leech and Fallon continue with words of caution that this study is an initial work and is based on their belief that linguistic corpora can provide evidence of cultural differences (45).

Oakes (2003) followed the pattern of the 1982 Hofland and Johansson study, and he compared the FLOB and FROWN corpora in his study (215). Oakes embraced the domains that were used in Leech and Fallon's study, plus one more domain, forms of the auxiliary verbs *be* and *have*, to group forms, for comparison (215-16). Oakes noted forms that may have shifted from being classified as either British or American in Leech and Fallon's study to "more typical of the other type of the other type of English in the 1990s corpora" (215). This comparison is important because it capitalizes on the presence of the Brown, LOB, FROWN, and FLOB corpora to pursue complex questions of frequency, corpus comparison through both synchronic and diachronic approaches.

Kilgarriff's study, "Comparing Corpora" begins with several questions, such as: "[. . .] is a new corpus significantly different from available ones, to make it worth acquiring?" (98). Such questions motivated his study which pursues methods for measuring similarity of corpora. He points, through a discussion of several different statistical functions, to a determination he refers to as Known Similarity Corpora (KSC) (120). Kilgarriff asserts, "We argue that corpus

linguistics is in urgent need of such a measure: without one, it is very difficult to talk accurately about the relevance of findings based on one corpus, to another, or to predict the costs of porting an application to a new domain" (127).

In 2004, MacQueen used historical newspaper databases to compare American and British English. MacQueen's study focused on the use of electronic newspaper texts stored in databases. MacQueen devised a methodology to determine the total number of articles that were contained in the *ProQuest Historical Newspapers* available to him. He used special syntax recognized by *ProQuest Historical Newspapers* to search for all articles, per year, that contained between one and one million words (MacQueen 129). This syntax would resemble: WC (>1 AND <1000000) (Macqueen 129; "Search Tips"). MacQueen's determination is significant, but MacQueen's methodology does not determine or estimate word count.

Corpus comparison can be a powerful way to highlight differences between the texts and social situations of different cultures. As corpora have evolved the very definition of a corpus, in terms of construction and inclusion, has evolved as well. With the completion of the American National Corpus, which is modeled after the British national Corpus, linguists will have, again, a balanced corpora for the comparison of American and British English. The Web has brought researchers vast amounts of texts within easy reach that may be very useful for linguistic research; however, cross-cultural comparisons need to be conducted with corpora that have similar construction methodologies and rationales behind their approach to balance.

Conclusion

Corpus linguistics is a relatively new field. This study looks to the classic Brown Corpus as the field's birth. As the field has progressed from the classic corpora to the next generation corpora and the Web as corpus, new research questions continue to emerge and new ideas surface about how to construct and use linguistic corpora. Still the principle of balance, which the Brown Corpus brought to field, remains as valid today as it was in 1964. This principle can still be achieved through twenty-first century resources with extremely large amounts of text.

Chapter 3

Methodology

This study endeavored to create an extremely large corpus that would preserve the principle of balanced construction, established by the Brown Corpus, and provide solutions to some deficiencies that are inherent in the next generation corpora and the Web as corpus.

The next generation corpora, such as The Oxford Corpus, The Cambridge International Corpus, and The Collins Corpus, employ the principle of balance in their construction, but the corpus user is not necessarily able to access all of the texts in the corpus in their original full-text format. Also, the next generation corpora do not give the user a reflection of lexical use in terms of geographic communities. Web-based corpora also fail in these areas, but the Web as corpus has an additional deficiency; users cannot know what texts are being accessed when a search, such as with Google, is conducted.

The construction of the BVC positions itself in between the next generation corpora and the Web as corpus. The BVC is stored online, as the Web is, but a word count can be estimated for the BVC, and the BVC is organized under principles of balance. These attributes make the BVC similar to the next generation corpora. Finally, all of the texts in the BVC are full texts, and they are conveniently accessible. Further, the BVC presents a solution to the lingering problem of copyright complications.

Title Inclusion in the American BVC

The building blocks of the BVC are full-text newspapers in the *LexisNexis Academic* database. The study borrowed from *LexisNexis Academic* prefabricated geographic organizations of newspaper titles. The *LexisNexis Academic* database provides an organizational shortcut to newspaper titles through placing them within folders, inside its interface, that correspond to these U.S. regions: Northeast, Southeast, Midwest, and West. The study began the title selection process by considering and analyzing the titles that *LexisNexis Academic* had organized under these regional labels. This regional organization of titles proved very useful to the study, which used the very similar regions of: Northeast (NE), Southeast (SE), Midwest (MW), West (W), and Coastal west (CW).

These regions generally reflect cultural boundaries that Zelinsky identifies as bounded by "first order cultural boundaries" (119). He identifies these regions as: New England, The Midland, The South, The Middle West and The West (119). This study's Northeast region is essentially a combination of Zelinsky's New England and Midland regions. Also, this study identifies a Southeast rather than the South that Zelinsky does. Zelinsky's South would include thihs study's South and some Midwest and possibly even a small portion of the West. Finally, the notion of a Coastal west region was not a part of Zelinksy's cultural units of the U.S., but this study created it to reflect the western U.S. in the most specific way possible.

The implementation of a Coastal west region allowed the study to make two clearly delineated regions out of the single west category in *LexisNexis Academic*. The study's use of a West as well as a Coastal west region reflects U.S. culture and geography better. This organizational decision creates balance and allows regional differences in American English to be more accurately reflected in the BVC.

Each newspaper in the BVC is a regional newspaper. Some newspapers, such as *USA Today*, embrace the country's news, but they are not connected to a specific U.S. region as all the newspapers in the BVC are. Again, the inclusion of regionally-based newspapers allows regional differences in American English to be reflected in the BVC.

Decisions had to be made about the inclusion of titles within each region of the BVC. For each region, the study faced a variable number of available titles. In order to establish a standard that would lead to a balanced corpus, the study included five newspaper titles per region, which yields a total of twenty-five titles for the American BVC. The study established for inclusion in the American BVC that a newspaper title must have seven-day circulation over the period of January 1, 1998 to December 31, 2007 and full-text availability in *LexisNexis Academic*. The study also sought to select from each region, as closely as possible, one or two large newspapers with a circulation above two hundred thousand; one or two medium newspapers with a circulation between one hundred and two hundred thousand; and one or two small newspapers with a circulation below one hundred thousand. The study inevitably had to include some newspaper titles and reject others; the study included titles that would ensure that a sense of balance would be achieved in each region because this study is not merely creating a collection of newspapers; this study is systematically organizing a corpus of newspaper texts that work together to create a balanced linguistic corpus.

A single newspaper title will have in itself a variety of text types through the sections it encompasses. Similarly, a mixture of newspapers from a variety of circulation sizes creates a dimension of variety for the BVC. The audience of a newspaper, such as *The Herald-Sun*, from Durham, N.C. with an average daily (weekday) circulation of 29,000, will likely be different from the audience of *The Washington Post* with an average daily (weekday) circulation of

635,000. This difference in audiences should result in some linguistic contrast between their respective texts and, ultimately, should create a greater variety of text types within the BVC than would have been achieved with newspapers of a single circulation size.

The motivation of this study to include newspaper titles with a variety of circulation sizes from a variety of regions mirrors the methodology of the Brown Corpus that included fifteen categories of text types to achieve balance (Kucera and Francis xix).

LexisNexis Academic provides some background information on the newspaper titles it makes available, but the study could not find evidence of how recent the information, such as circulation size, was. Also, circulation information was not always available within *LexisNexis Academic*. The study used the *SRDS Media Solutions*, accessed through the UGA's GALILEO family of databases, to determine circulation information. This information was generally only several months old, at the most, and was reflected through the various newspaper publishers' official statements.

The Northeast region

The Northeast folder for U.S. newspapers in Browse Sources in *LexisNexis Academic* contains 64 sources, and of that number, 13 are newspaper titles that have full-text, seven day coverage over the period of 1998 to 2007. Of these 13, the largest, in terms of circulation, is *The New York Times* (New York, NY), and the smallest is the *Union Leader* (Manchester, NH). The study selected these for their respective circulation sizes. In order to include titles smaller than *The New York Times*, but in the range of a circulation of at least 200 thousand, the study included *The Boston Globe* (Boston, MA) and the *Pittsburgh Post-Gazette* (Pittsburgh, PA) which

represent the east and west margins of the region respectively. The study included the *Portland Press Herald* (Portland, ME), which has a circulation below 100 thousand, and represents the northern extremity of the region.

The *Boston Herald* and the *Post Standard* (Syracuse, NY) were excluded because no circulation information was available for them through SRDS. Other possible candidates for inclusion included the *Star Ledger* (Newark, NJ) and *The Record* (Bergen County, NJ) because their circulation sizes placed them in the same respective circulation categories as *The Boston Globe* and the *Pittsburgh Post-Gazette* that were included based on their size and geographic representation. The *Telegram & Gazette* (Worcester, MA) was excluded because the study had included two small titles, and because one of the large included titles, *The Boston Globe*, comes from Massachusetts, that state had representation in the study. The *Providence Journal Bulletin* (Providence, RI) could have been a good selection for a medium-sized title, but because the study had included the smaller *Union Leader* from nearby New Hampshire, the study excluded the *Providence Journal Bulletin*. The *Buffalo News* (Buffalo, NY) and *Daily News* (New York, NY) were both eliminated because of the inclusion of *The New York Times*.

Of the titles that were candidates for inclusion from the Northeast region, only three had a circulation below one hundred thousand. Two of these were included, which was useful, but the candidates in the Northeast region tended to have large circulations. *The New York Times* has the largest circulation of any newspaper in the American BVC. Therefore, a comparison of the circulation of the region's largest and smallest titles represent a large variation even though most of the candidate titles had large circulations. Table 3.1 displays the candidates for inclusion in the Northeast region.

Table 3.1: Candidates for inclusion in the Northeast region

Circulation in thousands	Title and City	Inclusion Status
323	*Boston Globe* *Boston, MA*	Included
Not available	*Boston Herald* Boston, MA	Not included
175	*Buffalo News* Buffalo, NY	Not included
632	*Daily News* New York, NY	Not included
1,000	*New York Times* New York, NY	Included
203	*Pittsburgh Post-Gazette* *Pittsburgh, PA*	Included
65	*Portland Press Herald* *Portland, ME*	Included
Not available	*Post Standard* Syracuse, NY	Not included
131	*Providence Journal Bulletin* Providence, RI	Not included
156	*The Record* Bergen Co., NJ	Not included
316	*Star Ledger* Newark, NJ	Not included
79	*Telegram & Gazette* Worcester, MA	Not included
51	*Union Leader* Manchester, NH	Included

The Southeast region

The Southeast folder for U.S. newspapers in Browse Sources in *LexisNexis Academic* contains 55 sources, and of that number, 23 have full-text, seven day coverage over the period of 1998 to 2007. The study initially gravitated toward the selection of the largest and smallest titles, in terms of circulation, as the study had in the case of title selection for the Northeast region. The largest title in the candidate group was *The Washington Post* (Washington, D.C.), and the smallest was *The Herald-Sun* (Durham, NC). Both of these titles were included. Next, in order to extend inclusion to the southern and western extremities of the region, *The Florida Times Union* (Jacksonville, FL) and *The Times-Picayune* (New Orleans, LA) were included. With the inclusion of these titles, the study had acquired a large title, a small title, and two medium titles, so the study needed another small title for the region. *The Charleston Gazette* (Charleston, WV) was included as the other small title for the region.

The study endeavored to include only one title from Florida, so the selection of *The Florida Times-Union* (Jacksonville, FL) meant that these titles: *The Tampa Tribune* (Tampa, FL), the *Sarasota Herald-Tribune* (Sarasota, FL), the *St. Petersburg Times* (St. Petersburg, FL), and *The Palm Beach Post* (West Palm Beach, FL) could not be included. Beyond the fact that the circulation size of the *The Florida Times-Union* was appropriate, the location of its city of publication, Jacksonville, was geographically appropriate as well. Because Jacksonville is in Florida's northern region, it is close to the state of Georgia, which does not have a title in the study, so this title is in a geographically useful location for the American BVC.

The *District of Columbia News*, *The Washington Times* and *The Capital*, all of which are published in the District of Columbia, were excluded because the *Washington Post* was included. Further, the circulation of the *District of Columbia News* could not be determined through SRDS,

so it had two reasons (location and lack of circulation information) for exclusion. *The Richmond Times Dispatch* (Richmond, VA), *The Roanoke Times* (Roanoke, VA) and *The Virginian-Pilot* (Norfolk, VA), which were small and medium-sized titles, were excluded because of the inclusion of a small title, *The Charleston Gazette* (Charleston, WV). *The Post and Courier* (Charleston, SC), a small title, was not included because of the inclusion of the nearby, smaller *The Herald-Sun*. Also, a number of North Carolina titles, both medium and small, were excluded because of the inclusion of *The Herald-Sun*. These titles included: *News & Record* (Greensboro, NC), *The News and Observer* (Raleigh, NC), and the *Chapel Hill Herald* (Chapel Hill, NC). No circulation information was available for the *Chapel Hill Herald*, so it was eliminated on the bases of geography and the lack of available information on its circulation.

Finally, the *Arkansas Democrat-Gazette* (Little Rock, AR), and the *The Advocate* (Baton Rouge, LA) were excluded because of the inclusion of *The Times-Picayune* (New Orleans, LA) which was selected to represent the western area of the Southeast region.

The Atlanta Journal-Constitution (Atlanta, GA) was excluded, but, as mentioned above, the Jacksonville, Florida title that was included is extremely close to Georgia. Further, the study endeavored to select titles from these extremities of the Southeast region: The District of Columbia, Florida, and Louisiana, so some possibly useful titles were not selected because of the need to include both a variety of newspaper circulation sizes and geographic representations. Table 3.2 displays the candidates for inclusion in the Southeast region.

Table 3.2: Candidates for inclusion in the Southeast region

Circulation in thousands	Title and city	Inclusion status
92	*The Advocate* Baton Rouge, LA	Not included
176	*Arkansas Democrat-Gazette* Little Rock, AR	Not included
274	*The Atlanta Journal-Constitution* Atlanta, GA	Not included
42	*The Capital* Annapolis, MD	Not included
Not available	*Chapel Hill Herald* Chapel Hill, NC	Not included
70	*The Charleston Gazette* Charleston, West Virginia	Included
Not available	*District of Columbia News* Washington, D.C.	Not included
127	*Florida Times-Union* Jacksonville, FL	Included
29	*The Herald-Sun* Durham, NC	Included
77	*News & Record* Greensboro, NC	Not included
158	*The News and Observer* Raleigh, NC	Not included
134	*The Palm Beach Post* West Palm Beach, FL	Not included
95	*The Post and Courier* Charleston, SC	Not included
160	*Richmond Times Dispatch* Richmond, VA	Not included
87	*The Roanoke Times* Roanoke, VA	Not included
84	*Sarasota Herald-Tribune* Sarasota, FL	Not included
268	*St. Petersburg Times* St. Petersburg, FL	Not included
187	*The Tampa Tribune* Tampa, FL	Not included
175	*The Times-Picayune* New Orleans, LA	Included
174	*The Virginian-Pilot* Norfolk, VA	Not included
622	*The Washington Post* Washington, D.C.	Included
92	*The Washington Times* Washington, D.C.	Not included

The Midwest region

The Midwest folder for U.S. newspapers in Browse Sources in *LexisNexis Academic* contained 41 sources. Of that number, 14 had full-text, seven day coverage over the period of 1998 to 2007. *The Chicago Sun-Times* (Chicago, IL) had the largest circulation and was included; on the other extreme of inclusion, the two smallest candidates, the *Telegraph Herald* (Dubuque, IA) and the *Topeka Capital-Journal* (Topeka, KS) were included. The *St. Paul Pioneer Press* (St. Paul, MN) was selected to represent the western geographic extremity of the Midwest region; the *Topeka Capital-Journal* (Topeka, KS) was selected to represent the southern extremity; the *Dayton Daily News* (Dayton, OH) was selected to represent the eastern extremity of the region.

For the Midwest region, three candidates had circulations near or below fifty thousand. So, for this region, title selection began with the *Chicago Sun-Times*, which had the largest circulation, and also included the titles with the smallest circulations, which were the *Telegraph Herald* (Dubuque, IA) and the *Topeka Capital Journal* (Topeka, KS). Two medium titles still were needed, and the inclusion of the *St. Paul Pioneer Press* and the *Dayton Daily News* achieved both medium circulation representation and variety in terms of geography.

Eight candidate titles were not selected; of these, circulation figures were not available for two, *The Plain Dealer* (Cleveland, OH) and the *Chicago Daily Herald* (Chicago, IL), so they were excluded. Further, *The Columbus Dispatch* (Columbus, OH) was excluded because the smaller circulation size of the nearby *Dayton Daily News* made it a better candidate for inclusion. *The Milwaukee Journal Sentinel* (Milwaukee, WI) was excluded in favor of the slightly smaller *St. Paul Pioneer Press* that is geographically farther to the west; also, the *Star Tribune*

(Minneapolis, MN) was excluded in favor of the geographically close *St. Paul Pioneer Press*, which had a smaller circulation size that was closer to the study's needs for a medium-sized title. Table 3.3 displays the candidates for inclusion in the Midwest region.

Table 3.3: Candidates for inclusion in the Midwest region

Circulation in thousands	Title and city	Inclusion status
Not available	*Chicago Daily Herald* Chicago, IL	Not included
575	*Chicago Sun-Times* Chicago, IL	Included
195	*The Columbus Dispatch* *Columbus, OH*	Not included
141	*Dayton Daily News* Dayton, OH	Included
212	*The Milwaukee Journal Sentinel* Milwaukee, WI	Not included
169	*Omaha World Herald* Omaha, NE	Not included
45	*The Pantagraph* Bloomington, IL	Not included
Not available	*The Plain Dealer* Cleveland, OH	Not included
240	*St. Louis Post-Dispatch* St. Louis, MO	Not included
184	*St. Paul Pioneer Press* St. Paul, MN	Included
322	*Star Tribune* Minneapolis, MN	Not included
27	*Telegraph Herald* Dubuque, IA	Included
42	*Topeka Capital-Journal* Topeka, KS	Included
104	*Wisconsin State Journal* Madison, WI	Not included

The West region

The West folder for U.S. newspapers in Browse Sources in *LexisNexis Academic* contained 98 sources, and of that number, 9 are newspaper titles, from the study's geographically defined West region, that have full-text, seven-day coverage over the period of 1998 to 2007. The second largest title, in terms of circulation, was *The Denver Post* (Denver, CO), and it was selected as the largest title for the region. The candidate with the smallest circulation, *The Santa Fe New Mexican* (Santa Fe, NM) was also selected.

Four candidates were in the one hundred thousand range of circulation, and two were included: *The Austin American-Statesman* (Austin, TX) and *The Salt Lake Tribune* (Salt Lake City, UT). Because of the inclusion of a title from Austin, Texas, the candidate with the largest circulation, *The Houston Chronicle* (Houston, TX) was not included for reasons of geography; also, the *San Antonio Express*, with a larger circulation than *The Austin American-Statesman*, was not included. *The Denver Post* actually has only a slightly smaller circulation than *The Houston Chronicle*. The *Deseret Morning News* (Salt Lake City, Utah) was excluded because *The Salt Lake Tribune*, a larger title from the same city, was included.

The *Albuquerque Journal* (Albuquerque, NM) was excluded because of the inclusion of the geographically close *Santa Fe New Mexican* which the study included initially as the smallest title from the region. Finally, *The Tulsa World* (Tulsa, OK) was included to reflect the eastern extremity of the West region, near the western margin of the Midwest region. To reflect the region's southern extremity, *The Austin American-Statesman*, which is near the western margin of the Southeast region, was included. Table 3.4 displays the candidates for inclusion in the West region.

Table 3.4: Candidates for inclusion in the West region

Circulation in thousands	Title and city	Inclusion status
102	*Albuquerque Journal* Albuquerque, NM	Not included
151	*The Austin American-Statesman* Austin, TX	Included
426	*The Denver Post* Denver, CO	Included
71	*Deseret Morning News* Salt Lake City, UT	Not included
448	*The Houston Chronicle* Houston, TX	Not included
119	*The Salt Lake Tribune* Salt Lake City, UT	Included
206	*San Antonio Express-News* San Antonio, TX	Not included
25	*The Santa Fe New Mexican* Santa Fe, NM	Included
110	*The Tulsa World* Tulsa, OK	Included

The Coastal west region

The West folder for U.S. newspapers in Browse Sources in *LexisNexis Academic* contained 9 newspaper titles, from the study's geographically defined Coastal west region, that have full-text, seven day coverage over the period of 1998 to 2007. These newspaper titles are published in California, Oregon, and Washington. *The San Francisco Chronicle* (San Francisco, CA) was the title with the largest circulation, and it was included. *The Spokesman Review* (Spokane, WA) was included as the region's smallest title. The study attempted to include *The Columbian* (Vancouver, WA), which had the smallest circulation of the candidates, but it

had no text available for a random sample date (for the study's word count estimation), so *The Columbian* was excluded. The *San Diego Union-Tribune* (San Diego, CA) was included as a title with a medium-sized circulation.

Two other titles: *The Press Enterprise* (Riverside, CA), and the *San Jose Mercury News* (San Jose, CA) were excluded because of the previous inclusion of two titles from California. The *Seattle Post-Intelligencer* (Seattle, WA), was excluded because its Sunday coverage, noted in visual checks of several Sundays over several years, frequently included only four to six articles. This very small number of articles disqualified this title because its Sunday representation in *LexisNexis Academic* was so far out of balance in comparison to its Monday through Saturday holdings. *The Oregonian* (Portland, OR), a large title, was also included. Table 3.5 displays the candidates for inclusion in the Coastal west region.

Table 3.5: Candidates for inclusion in the Coastal west region

Average daily circulation in thousands	Title and city	Inclusion status
42	*The Columbian* Vancouver, WA	Not included
129	*The Daily News of Los Angeles* Los Angeles, CA	Included
283	*The Oregonian* Portland, OR	Included
149	*The Press Enterprise* Riverside, CA	Not included
269	*San Diego Union-Tribune* San Diego, CA	Included
339	*The San Francisco Chronicle* San Francisco, CA	Included
234	*San Jose Mercury News* San Jose, CA	Not included
117	*Seattle Post-Intelligencer* Seattle, WA	Not included
87	*The Spokesman-Review* Spokane, WA	Included

Conclusion

For the American title selection process, *LexisNexis Academic* afforded the study a broad array of titles that allowed the inclusion of titles with variety in terms of circulation size and geographic representation. This inclusion process did not include any titles from Alaska or Hawaii. *LexisNexis Academic* did possess one title from Alaska, but because the study decided not to represent Alaska or Hawaii, that title was not regarded as a candidate. No newspapers from Hawaii were included in *LexisNexis Academic*. The combination of circulation sizes and titles with a variety of geographic representations have systematically afforded the study the opportunity to create a balanced BVC.

Title Inclusion in the British BVC

The first step in the title selection process for the British Titles was to indentify daily full-text British newspaper titles in *LexisNexis Academic* with coverage from 1997 to 2008. A total of 14 British titles had the ten year coverage, and of that number, five were titles that had seven day per week coverage; eight had six day coverage; and one had five day coverage. The titles were published in the following cities: Belfast, Ireland (1), London, England (6), Glasgow, Scotland (2), Edinburgh, Scotland (1), Newcastle, England (1), Darlington, England (1), Nottingham, England (1), and Bristol, England (1). The study consulted *The Europa World Year Book 2007* Volume II Kazakhstan-Zimbabwe to determine circulation amounts for the British titles in the study.

The Belfast Telegraph (Belfast, Ireland) was considered, but it was eliminated from consideration because after an inspection of the paper through a variety of weeks and years, some blackouts of coverage, for several days, were detected. The *Western Daily Press* (Bristol,

England) had a problem with blackouts of coverage as well, and it was eliminated from consideration. The *Times* (London, England) was considered, but two conditions caused it to be excluded. First, the *Times* is actually two titles in the seven-day sense—the *Times* (six day coverage) and *The Sunday Times* (one day coverage); these titles are stored separately in *LexisNexis Academic*, and different editors were listed in the research found for each title (*The Europa World Year Book* 2007 Volume II 4645-4648). This situation of two titles created a problem because the study had established a policy of not mixing titles—even sister titles—to expand coverage. Second, the study considered inclusion of *The Times* as a six-day title, but because, from the London area, a larger seven-day title and a smaller six day title were available, those were both considered in favor of *The Times*.

Because of the concentration of population in London, and availability of titles in *LexisNexis Academic* from London, the study included the available newspaper with the largest circulation, *The Daily Mail and Mail on Sunday*, from London, and the available newspaper with the smallest circulation from London, *The Independent*. The inclusion of these titles excluded the other available titles from London: *The Evening Standard*, *The Guardian*, *The Mirror*, as well as *The Times*, which is addressed above.

The Daily Record and Sunday Mail (Glasgow, Scotland) and *The Scotsman and Scotland on Sunday* (Edinburgh, Scotland) were not included because the study included the much smaller title, *The Herald* (Glasgow, Scotland) from Scotland. The *Evening Chronicle* (Newcastle, England) and *The Northern Echo* (Darlington, England) were included to establish representation from the northern portion of England. The *Nottingham Evening Post* was excluded because of the study's motivation to include titles from further north since two titles from the south (London) were included.

The five titles that are included in the British BVC reflect a variety of daily circulation sizes, from 50 thousand to 2.2 million and a variety of geographical locations in the U.K. These locations include: London (2 titles), Northeast England (2 titles), and Scotland (1 title).

Table 3.6: Candidates for inclusion in the British BVC

Circulation in thousands	Title and city	Inclusion status
Not available	*Belfast Telegraph* Belfast, Ireland	Not included
2200	*The Daily Mail and Mail on Sunday* London, England	Included
397	*Daily Record and Sunday Mail* Glasgow, Scotland	Not included
77	*Evening Chronicle* Newcastle, England	Included
266	*The Evening Standard* London, England	Not included
322	*The Guardian* London, England	Not included
71	*The Herald* Glasgow, Scotland	Included
205	*The Independent* London, England	Included
1400	*The Mirror* London, England	Not included
50	*The Northern Echo* Darlington, England	Included
62	*Nottingham Evening Post* Nottingham, England	Not included
57	*The Scotsman and Scotland on Sunday* Edinburgh, Scotland	Not included
600	*The Times* London, England	Not included
43	*Western Daily Press* Bristol, England	Not included

The title selection process for the British BVC presented some challenges that were not faced by the study in the title selection process for the American BVC. For example, of the 14 candidates for inclusion in the British BVC, only three had seven-day coverage. This situation is different from the American BVC, whose selection process generally always faced many seven-day titles. The British BVC also features some noteworthy distinctions; for example, *The Daily Mail and Mail on Sunday* has the largest daily circulation of any title in the study. Also, with the inclusion of a title from Scotland, the countries of England and Scotland have representation in the British BVC. As with the American BVC, *LexisNexis Academic* provides the tools to create a BVC that reflects a variety of newspaper circulation sizes and geographic regions in Great Britain.

Estimated Word Count

In order to obtain random dates to use as the building blocks for the estimated word count process, the study created two sets of numbered pieces of paper. The first set included numbers 1 through 12, and the second set included numbers 1 through 31. For each year that the BVC covers (1998-2007), the study drew one number from each set to determine a random pair that represented a day and month. From the outset, the study sought to include only one Saturday and one Sunday because the majority of each week consists of weekdays. After the Saturday and Sunday had been selected, the study would re-select if Saturday or Sunday were selected again. Since not all of the British titles were seven-day titles, the date of the Sunday sampling frame, which was used with the American BVC, had to be randomly re-selected for the word count estimation process for the British BVC. Table 3.7 lists the dates of the random samples for the American and British BVCs for the word count process.

Table 3.7: Random sampling dates

Year	American BVC	British BVC
1998	Wednesday January 28	Wednesday January 28
1999	Tuesday August 31	Tuesday August 31
2000	Saturday November 4	Saturday November 4
2001	Sunday 25 March	Monday 30 July
2002	Friday 27 September	Friday 27 September
2003	Tuesday 18 March	Tuesday 18 March
2004	Tuesday 11 May	Tuesday 11 May
2005	Friday 9 December	Friday 9 December
2006	Wednesday 31 May	Wednesday 31 May
2007	Wednesday 12 December	Wednesday 12 December

When the study had established the ten sampling dates (month, day, year), the metadata, in the form of Expanded List of results from *LexisNexis Academic*, was obtained for all of the articles in the 25 titles for the sampling dates. The bibliographic metadata includes information, such as the article title, the newspaper title, the word count for the article, and the author's name. Refer to Figure 3.12 for examples of metadata in the Expanded List of results.

After the complete list of Expanded Results for the randomly selected date was saved in the .txt format, the software application, WordSmith Tools was used to isolate the collocation, from the Expanded Results, that included *words* preceded by an integer. This process began with the selection of the concordance feature in the WordSmith Tools interface and the

uploading of the .txt file to WordSmith Tools. Next, the concordance was created. Figure 3.1 shows the selection of the search word, or node, to begin the concordance process.

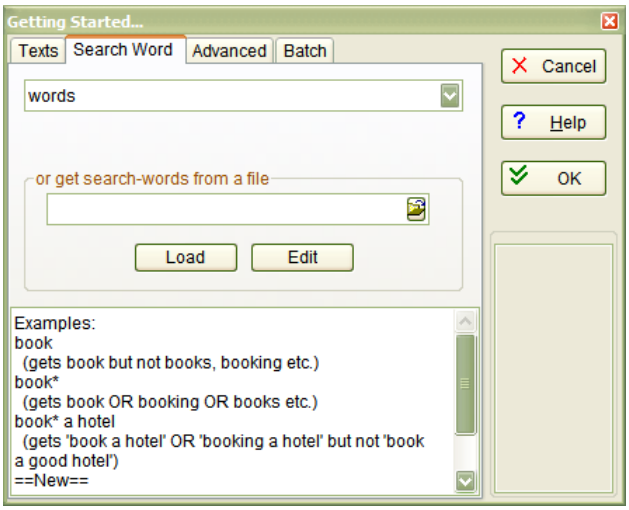


Figure 3.1: Search Word selection in WordSmith Tools

A specialized function within WordSmith Tools, a concordance cluster, was generated within the concordance results. Figure 3.2 shows the selection of cluster (collocation) length, minimum frequency, and the horizon. In this case the horizon is one word to the left of *words* and zero words to the right. WordSmith Tools recognizes integers as words, so this application worked very well for the study's needs.

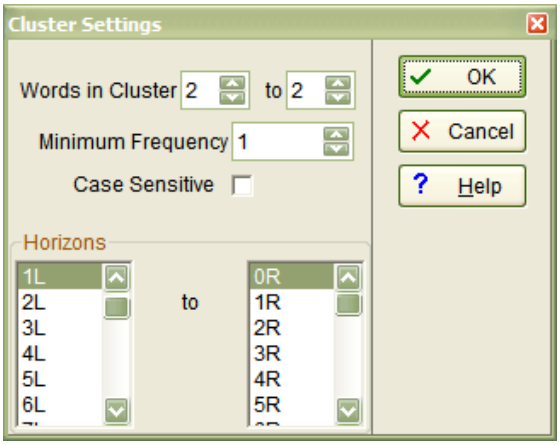


Figure 3.2: Cluster Settings in WordSmith Tools

Note, for example the first result in Figure 3.11: "Pg. A7, 270 words, By Judy Foreman, Globe Staff." WordSmith Tools was used to take the entire file of Expanded Results for one title, from a random sampling date, and yield only the collocation, such as *270 words*, for each result. One complication that the study encountered, but only rarely, was the presence of the form *words* in the Expanded List in a function other than word count. This manifestation was usually in a title, such as "Good Words." In the WordSmith Tools cluster output, a visual scan can quickly detect the presence of such an unwanted result. When the cluster output from WordSmith Tools is pasted into an Excel spreadsheet, an unwanted result, such as "Good Words" can be easily deleted. When the cluster results were generated, the study always compared the number of *words* collocations with the original number of Expanded Results from *LexisNexis Academic*. If these totals were not congruent, the output was investigated to determine the cause. Often a slight variation, such as 5, was regarded as acceptable. The sources of such as discrepancy can be many. For example, a publisher may send their files to *LexisNexis Academic* and include caption text with no word count; also, a few articles may have

no word count. Therefore, a small number of articles with no word count was consistently maintained as acceptable during the word count process.

Also, *LexisNexis Academic* can give its user some unusual challenges, such as the word count process. For example, when an article is retrieved in *LexisNexis Academic*, it is sometimes delivered twice in the results output. Some titles have this tendency more than others. The antidote for this problem was a slow and tedious process of visual verification of the expanded list for repeated results or duplication of article metadata. The study also performed word searches on the articles that repeated, and they would repeat in the word search results as well. The user should not be lulled into a sense of complacency by the size and power of *LexisNexis Academic*. The user must be cautious and monitor results judiciously because variables, such as duplication of results can occur.

WordSmith Tools maintains a running total (N) of the collocations. The presence of this count is a way to confirm that the number of search results from *LexisNexis Academic* equals the number of collocations in the WordSmith Tools calculation. This total should be checked in an effort to verify the total number of collocations equals the total number of search results. Figure 3.5 displays N1-N8 for the concordance of the 119 articles from *The Boston Globe* on 28 January 1998, and the collocations are listed in terms of descending frequency.

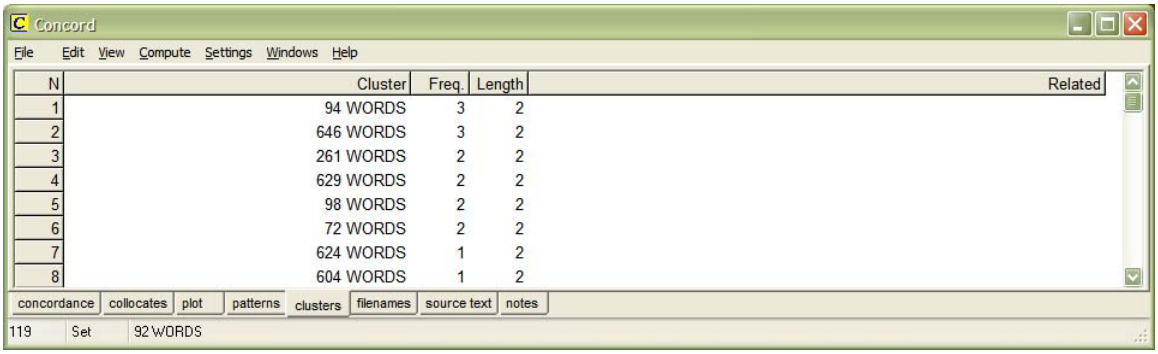


Figure 3.3: N1-N8 concordance cluster output for *words*

The concordance cluster output always displays the clusters, which have a frequency greater than 1, initially; the concordances with a frequency of 1 follow. Clusters with a frequency of 1 were by far the most commonly encountered in the word count functions for this study. Figure 3.4 displays the N105-N111 concordance cluster output for *words*.

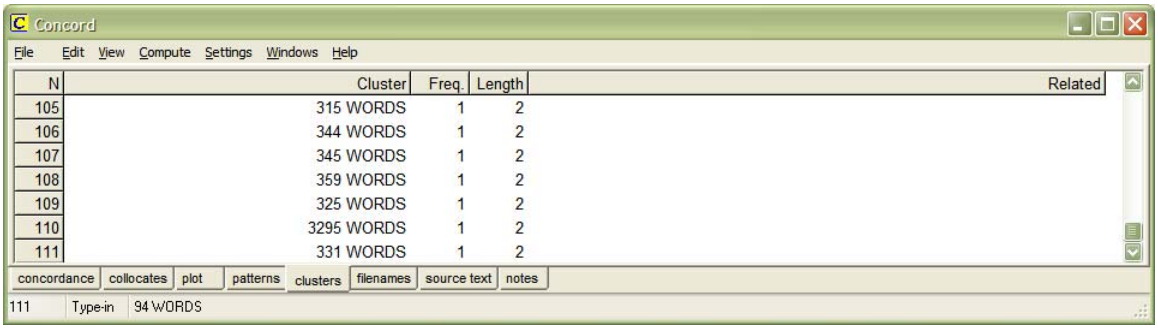
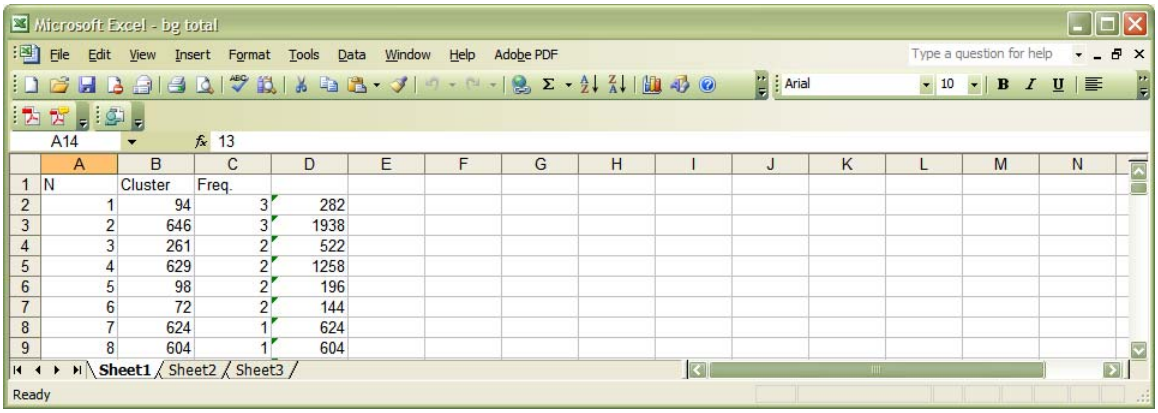


Figure 3.4: N105-N111 concordance cluster output for *words*

The concluding concordance output in Figure 3.4 indicates a total of 111 collocations (N=111); initially, this number may appear to reflect a malfunction because the total output of results from the LexisNexis interface was 119. The frequencies (number of occurrences) in Figure 3.3 shows that N1-N8 accounted for 14 collocations or an increase of 8 more collocations beyond 111. The

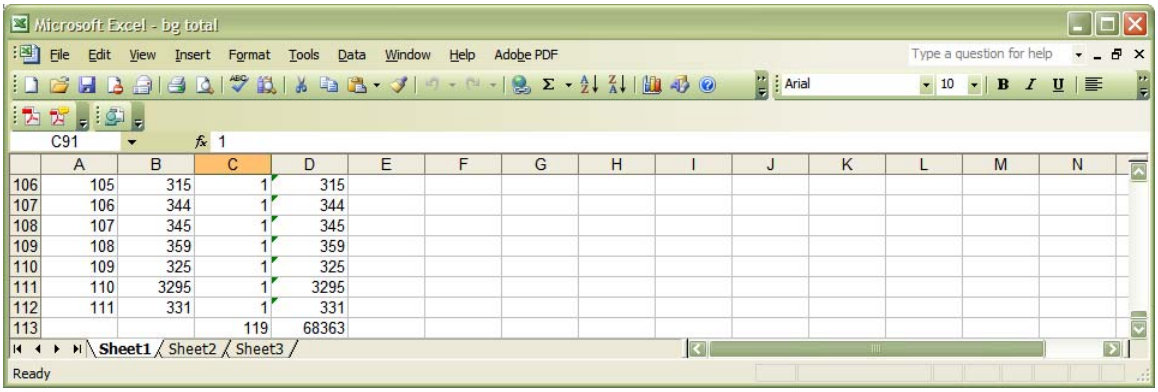
111 figure plus 8 yields 119, which equals the original results output from *LexisNexis Academic*. The study pasted the output from WordSmith Tools into a Microsoft Excel worksheet. Before calculations in Excel could be performed, the form *words* was removed from the file with the Find and Replace feature in Excel, so the spreadsheet would consist entirely of integers except for the labels at the top of the A, B, and C columns that remain from the WordSmith Tools output. Figure 3.5 displays the multiplication and addition processes for the data relating to *The Boston Globe* for 28 January 1998. Similar calculations led to a total word count for the randomly selected dates for each title in the BVC.



	A	B	C	D
1	N	Cluster	Freq.	
2	1	94	3	282
3	2	646	3	1938
4	3	261	2	522
5	4	629	2	1258
6	5	98	2	196
7	6	72	2	144
8	7	624	1	624
9	8	604	1	604

Figure 3.5: Multiplication processes for N1-N8 in Microsoft Excel

In Figures 3.5 and 3.6, note the N count in column A; the integer element from the *words* collocation in column B; the number of occurrences for the integer in column C; the product of the word count integer and its frequency and in column D. Therefore, in the case of N1, 3 occurrences of articles with 94 words yields a product of 282, and in N7, 624 occurs only once, so its product equals itself. The total number of collocations (119) and total number of words (68,363) are indicated at the end of columns C and D respectively in Figure 3.6.



	A	B	C	D
106	105	315	1	315
107	106	344	1	344
108	107	345	1	345
109	108	359	1	359
110	109	325	1	325
111	110	3295	1	3295
112	111	331	1	331
113			119	68363

Figure 3.6: Total word count for *The Boston Globe* on 28 January 1998

Table 3.8 and Table 3.9 show the total estimated word counts for the American and British BVCs, respectively. The estimated word count for the American BVC is 5.3 billion words. Two regions, the Northeast and Southeast, have estimated word counts of over 1 billion words (1.4 and 1.2 billion respectively), and the other three regions, Coastal west, West, and Midwest have estimated word counts of .93 billion, .82 billion, and .77 billion respectively.

The estimated word count for the British BVC is 1 billion words. Of the 5 newspapers in the British BVC, the newspaper with the lowest estimated word count is the *Northern Echo* with an estimated total of 68 million words. The *Daily Mail and Mail on Sunday* has the largest estimated word count with an estimated total of 312 million words. Appendix A contains step by step instructions for replication of the word count estimating process.

Variations of geography, circulation size, and word count contribute more than just convenient numbers to tables. These variables work to create, in essence, slices of American and British culture through newspaper texts that are accessible on a large scale.

Table 3.8: Estimated word count for the American BVC

Northeast region	**City**	**Estimated Word Count**
The Boston Globe	Boston, MA	312363715
The New York Times	New York, NY	550546290
Pittsburgh Post	Pittsburgh, PA	358727840
Portland Press Herald	Portland, ME	98183540
The Union Leader	Manchester, NH	103679710
Estimated regional total		1423501095
Southeast region		
The Charleston Gazette	Charleston, WV	207480235
Florida Times Union	Jacksonville, FL	151462225
The Herald-Sun	Durham, NC	99032895
The Times-Picayune	New Orleans, LA	281535450
The Washington Post	Washington, D.C.	478566830
Estimated regional total		1218077635
Midwest region		
Chicago Sun-Times	Chicago, IL	233375525
Dayton Daily News	Dayton, OH	106028485
Saint Paul Pioneer Press	St. Paul, MN	180159255
Telegraph Herald	Dubuque, IA	140623915
Topeka Capital-Journal	Topeka, KS	112895960
Estimated regional total		773083140
West region		
Austin American-Statesman	Austin, TX	123602505
The Denver Post	Denver, CO	179445315
The Salt Lake Tribune	Salt Lake City, UT	150383285
The Santa Fe New Mexican	Santa Fe, NM	94589750
The Tulsa World	Tulsa, OK	281767590
Estimated regional total		829788445
Coastal west region		
The Daily News of Los Angeles	Los Angeles, CA	182151790
The Oregonian	Portland, OR	195578680
San Diego Union Tribune	San Diego, CA	348787430
San Francisco Chronicle	San Francisco, CA	246733795
The Spokesman-Review	Spokane, WA	135022990
Estimated regional total		931359550
Estimated American BVC total		5352725000

Table 3.9: Estimated word count for the British BVC

Title	City	Estimated Word Count
The Daily Mail and Mail on Sunday	London, England	387276680
Evening Chronicle	Newcastle, England	103523811
Herald	Glasgow, Scotland	184388371
The Independent	London, England	296550598
Northern Echo	Darlington, England	68309398
Estimated British BVC total		1040048858

Standardized Rates

In Chapters 4, 5, and 6 the study will use rates (per ten million words) in the comparisons of forms both between regions in the American BVC and between the American BVC and the British BVC. The per ten million word rates allow standardized comparison.

In order to determine a form's status as an Americanism, the study will compare per ten million word rates of British and American forms over the ten-year coverage of the BVC in order to make conclusions about inclusion in the practical lexicographic application in Chapter 6.

Through random sampling of the titles in the American and British BVCs, the study estimated word counts for each corpora for the ten-year window that the corpora cover.

Use of the *LexisNexis Academic* Database

The Easy Search form in *LexisNexis Academic* can be accessed through links provided by the UGA Libraries or directly via http://lexisnexis.com/universe from a subscribing network. The Easy Search form gives a brief overview of source selections a user can make.

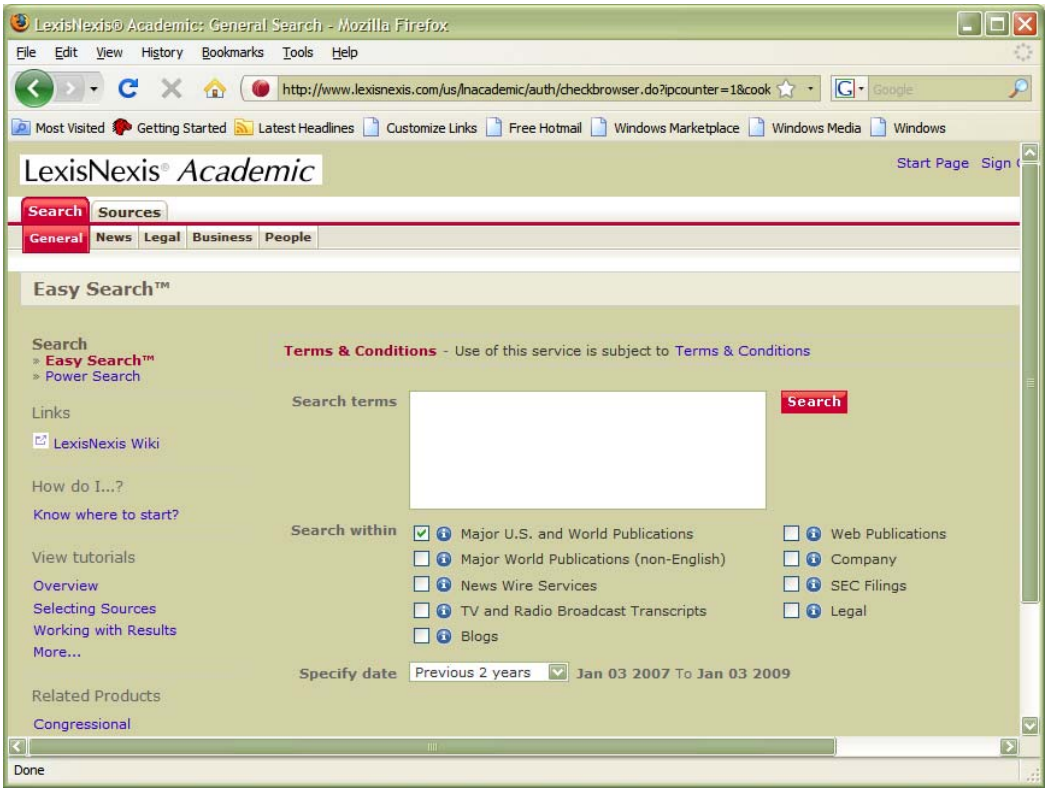


Figure 3.7: Easy Search form

The Power Search form, accessed via a link to the left of the search window in the Easy Search form, allows the user more flexibility and customization of searches. For example, if a user wishes to search for articles in *The Boston Globe* on 28 January 1998, a specific process must be followed from the Power Search form. First the title has to be selected. The More sources link below the Sources window opens a title search interface in which titles can be selected through a search window below Find a Source. Figure 3.9 shows the More sources form.

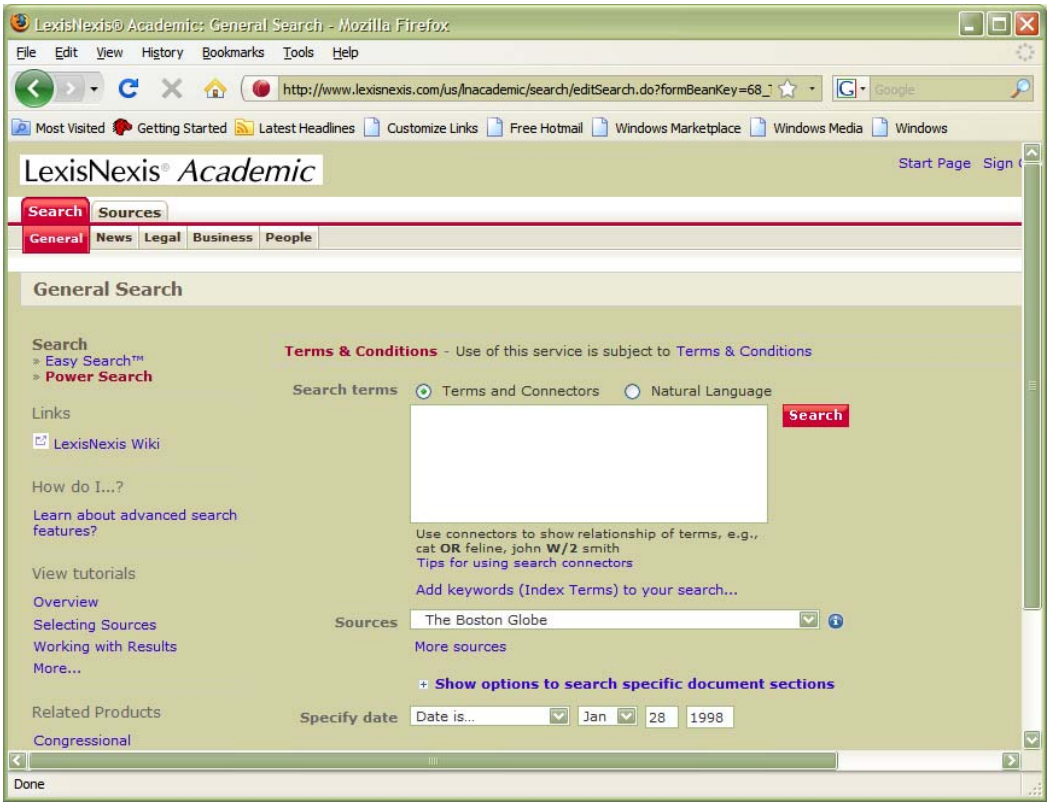


Figure 3.8: Power Search form

Access to Titles

The Find a Source application in the More sources form is most useful when a user knows all or a part of the title that needs to be selected. Search results for titles can be surprisingly large; a search for "Times" yields 280 results. Some of theses results, such as "*The New York Times* - Asia/Pacific Stories" are subdivisions of larger titles. Titles can also be browsed through folders that are accessible in the Browse Sources form.

The folders that can be browsed are an important way for a user to become educated about the holdings within *LexisNexis Academic*. The library staff at the University of Georgia has explained that one subscriber's holdings in *LexisNexis Academic* could be different from another's based on the specific details that each institution works out with LexisNexis Academic.

Subsequently, another library, which has a subscription to *LexisNexis Academic*, could have fewer or more newspaper titles according to the details of their particular agreement with *LexisNexis Academic*. Also, titles are frequently added and coverage of titles can be resumed or suspended at any time, so a user has no external master-list of the newspaper title holdings in their version of *LexisNexis Academic*. The user must independently find what holdings are available in the database.

The following selections can be made through drop-down boxes in the Browse Sources area in an effort to browse newspaper titles published in the United States: Country: United States; Topics: General News; then these folders can be selected: News and as a result, Newspapers. This sequence leads to the selection of possible regions for browsing: Midwest (41 sources), Northeast (64 sources), Southeast (55 sources), and West (98 sources).

The selecting (clicking the box to the left of the title) in the results of a Find a Source search, or the selecting of one of the titles, which is displayed within one of these regional folders, will make the title appear in the Selected Sources column. As sources are selected, they will appear in this column, but, when the OK – Continue button is selected, the titles will be placed in the Sources window in the Power Search form—which is displayed as a result of the selection of OK – Continue. If one needs to select many sources, then the need for caution in the selection process is acute because when the OK – Continue button is selected, that selection session is terminated. For example, if a user needs to select 50 titles and selects OK – Continue after selecting 20 titles, the selection process would have to be re-started from the first title in order to have all 50 titles combined. For convenience, strategies can be employed to make the title selection process faster and more convenient. For example, the American BVC involves 25 titles from 5 regions; these five regions can be selected independently, in five title units, as

needed from the More sources form. Such an approach is convenient and useful as long as searches do not have to be conducted on all 25 titles at one time. Still, all 25 titles in the American BVC can be searched at once if the user chooses to do so.

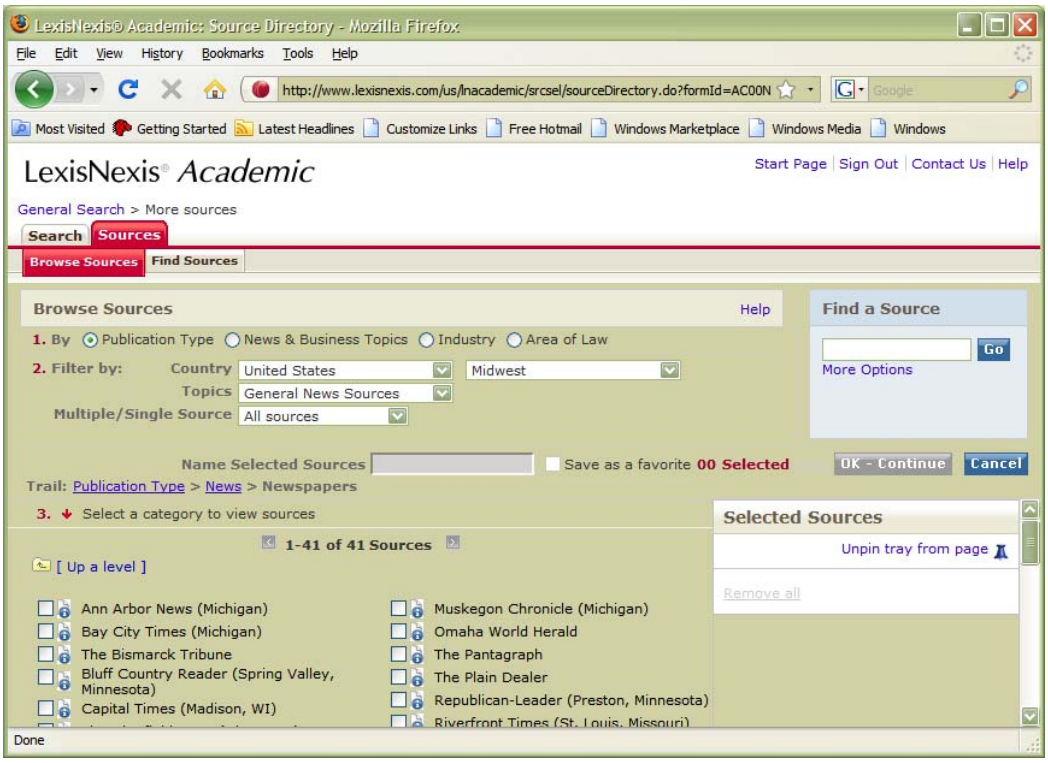


Figure 3.9: More sources form

Date selection

For the next step in the search process, a date or date range has to be specified. The Specify Date window gives the user these options through a drop-down menu: All available dates, Today, Date is…, Date is before…, Date is after…, Date is between…, Previous week, Previous month, Previous 3 months, Previous 6 months, Previous year, Previous 2 years, Previous 5 years, Previous 10 years, and Previous… . As more information is needed to complete the date specification, boxes to right will manifest. For example, in the case of Date

is…, a month will need to be selected by drop-down menu, and a day of the month and year will need to be entered in two separate boxes to the right. Also, selection of Previous… will cause two drop-down boxes to appear to the right. One contains selections of 1 to 100 and the next one contains these selections: Days, Weeks, Months, Years. This study used the Date is between… selection most frequently in connection with BVC searches. The broad variety of selections in terms of date range exemplify one attribute of the powerful flexibility of the *LexisNexis Academic* search interface. The BVC can be searched by a single day, week, month, year, or the complete span of ten years. Also, the BVC can be searched in terms of a custom range, such as a certain number of days, weeks or months, or years.

Construction of searches

The Power Search interface gives the user the choice of two search approaches: Terms and Connectors and Natural Language. Terms and Connectors is selected by default when the Power Search form is opened. *LexisNexis Academic* describes the function of its Natural Language feature: "When you search using the natural language feature, you can enter a search in plain English, without having to use any special terms or connectors" ("Developing a Search"). *LexisNexis Academic* notes the "Need to research general or conceptual issues, rather than very specific topics" as a motivation for the use of the Natural Language feature ("Developing a Search"). Because this study needed specific word level results, the use of Natural Language was not appropriate—especially because such results require the use of specific word-focused searches. Consequently, this study used the Terms and Connectors feature in *LexisNexis Academic* which proved to be very powerful and useful for searches of the BVC.

Terms and connectors

Before one can use Terms and Connectors successfully, the concepts behind *term* and *connector* need to be well understood. *LexisNexis Academic* explains that "Terms are the basic units of a search. A term is a single character or group of characters, alphabetic or numeric, with a space on either side" ("Developing a Search"). Still, even with this straightforward explanation, some character strings may not be immediately discerned as a single or multiple terms without some background information. For example, *F.B.I.* is one term, but *F. B. I.* is three terms ("Developing a Search"). In this case spaces are the determinant of whether a string of characters is a single term: "A period is treated like a space except when: the period is preceded by only one alphabetic character and followed (with no spaces in the sequence) by any number of single letters each of which is followed by a period" ("Developing a Search"). In these cases, *99.9* and *.999*, the period is not treated like a space because in the first case "The period is preceded and followed by a number;" in the second case, "The period is preceded by a space and followed by a number" ("Developing a Search"). Also, "Hyphens, slashes, and parentheses are treated as a space, so a hyphenated word or terms containing slashes or parentheses are seen as multiple words. When searching for terms or phrases that contain these characters, replace the hyphen, slash, or parenthesis with a space" ("Hyphens, Slashes, and Parentheses"). *LexisNexis Academic* supplies these examples: *co-operative* is read as two words (*co operative*); *401(k)* is read as two words (*401 k*); and *20/20* is read as *20 20* ("Hyphens, Slashes, and Parentheses").

The search for the form *go missing* presented some problems to the study. For example, the study encountered instances in which the terms, *go* and *missing*, were located properly in *LexisNexis Academic*, but they were interrupted by a period that is interpreted as a space by

LexisNexis Academic. So, results, such as ". . . that's the way things *go*. *Missing* still is this week's . . . ." had to be visually detected and removed from search results.

LexisNexis Academic recognizes ten connectors: AND, OR, W/n, AND NOT, PRE/n, W/p, W/seg, W/s, NOT W/n, NOT w/seg, NOT W/s, NOT w/p ("Hyphens, Slashes, Parentheses"). Note that some LexisNexis support pages on the Web will include W/para and W/sent and others include W/p and W/s for the same functions. This study used the latter forms of the connectors. A user's own search terms can be combined with connectors to create search strings that can deliver very focused results. Terms and Connectors in *LexisNexis Academic* are so flexible that searches can be broadened and narrowed to obtain desired results. Still, Terms and Connectors need to be used intelligently because their use alone does not ensure proper search results; however, their appropriate use should deliver proper search results. Table 3.10 shows the connectors and their functions; this information is taken from the LexisNexis help page, "Connector Order and Priority."

Table 3.10: Information provided by *LexisNexis Academic* about Connectors

Connector	Function
AND	Finds search words that may appear anywhere in a document.
OR	Can be used to connect words. One or both of the words may appear anywhere in the document.
W/n	Finds connected terms within a proximity of n number of words. n may not exceed 255. May be a better choice than the AND connector if proximity is important.
AND NOT	Excludes a word or phrase from the search.
PRE/n	Finds first connected term preceding the second term by no more than n words.
W/p	Finds connected terms within the same paragraph.
W/seg	Finds connected terms within the same document segment.
W/s	Finds connected terms within the same sentence.
NOT W/n	Finds first term but excludes documents in which the second term occurs within n words of the first term.
NOT W/seg	Finds first term but excludes documents in which the first and second term appear in the same segment.
NOT W/s	Finds the first term but excludes documents in which the second term occurs in the same sentence.
NOT W/p	Finds the first search term but excludes documents in which the second term occurs in the same paragraph as the first term.

In "Connector Order and Priority," *LexisNexis Academic* notes that connectors operate in the order of priority explained in Table 3.11.

Table 3.11: Connectors and priorities

Priority	Connector
1	OR
2	W/n, PRE/n, NOT W/n
3	W/s
4	W/p
5	W/seg
6	NOT w/seg
7	AND
8	AND NOT

Commands

In addition to connectors, these search commands exist in *LexisNexis Academic*: ATLEAST, ALLCAPS, CAPS, NOCAPS, PLURAL, SINGULAR ("Search Connectors and Commands"). ATLEAST is used in searches to specify a minimum number of occurrences for a term to occur in a result ("Search Connectors and Commands"). ALLCAPS, CAPS, and NOCAPS are used to search for specific forms of terms with regard to capitalization. ALLCAPS limits results to forms that are in all capitals; CAPS limits results to one or more capital letter and NOCAPS limits results to no capital letters ("Search Connectors and Commands"). Further, *LexisNexis Academic* allows the use of truncated and wildcard searches: "The truncation (!) and

wildcard (*) characters let you easily combine or eliminate search terms, making your search simpler" ("Developing a Search"). Terms and Connectors and a variety of commands provide the user an enormous amount of search possibilities. In order to support so many search features, *LexisNexis Academic* has established a system of priorities that govern search strings.

LexisNexis Academic shares the following example of a search string and follows with instruction about how the search string operates:

> bankrupt! W/25 discharg! AND student OR college OR education W/5 loan
>
> [The string above] is operated on in the following manner:
>
> - Because OR has the highest priority, it operates first and creates a unit of *student OR college OR education!* .
> - W/5, the smaller of the W/n connectors ties together the term *loan* and the previously formed unit of *student OR college OR education!*
> - W/25 operates next and creates a unit of *bankrupt! W/25 discharg!* .
> - AND, with the lowest priority, operates last and links the units formed in the second and third bullets above. ("Developing a Search")

The powerful nature of the search options within *LexisNexis Academic* enables the creation of extremely brief search strings that can be more powerful than much longer strings that may, at first, appear to be more exhaustive. For example, this study used the following string to create a corpus for a pilot study on *carb* forms:

> carb! w/25
>
> high or low or diet or atkins or (south pre/1 beach) and not carbohydrate
>
> and not carbon! and not carbide and not carburetor
>
> and not carberry and not carballo and not carbajal and not carbine

The truncation of *carbon* prevents the user from having to establish a long stop-list of *carbon* forms strung together with the AND connector. Heavy dependence on the AND connector can be the source of frustrating and overwhelming search results from *LexisNexis Academic*. If search results exceed 3,000, no results will be displayed. The truncation option (!) and W/n connector are easy to implement changes that can help users get such out-of-control searches under control. An understanding of how to modify search string strings makes *LexisNexis Academic* very approachable. Often around campus this phrase is heard: "With LexisNexis I either get nothing or too many results." The proper search approach will help the user take advantage of *LexisNexis Academic*, so the user can obtain useful results.

Another command that is available in *LexisNexis Academic* is the length command. A search string that employs the length command would use the symbol <, or the symbol >, in parentheses with an integer ("Commands"). In order to obtain search results in articles below 41 words, which a user might select in order to focus on captions, for *The Boston Globe* on January 28 1998, this string:

BODY (atlantic) AND LENGTH (<41)

resulted in the single result shown in Figure 3.10. The length command can be used more broadly as well. For example, for the same date and title, this command:

LENGTH (<100)

produced all the articles with word counts below 100 words, which in this case was a total of 18 articles.

Document section search

By default searches from the Power Search form will be conducted on all of the text of documents in *LexisNexis Academic*; however, the option to search specific sections of a document is available. First, note that more than the complete text of a news article constitutes a document in *LexisNexis Academic*. *LexisNexis Academic* lists 18 possible sections, but all possible document sections (sometimes referred to as segments) may or may not be present in every article based on a host of variables ("Document Section"). Some of the Standard News Sections of a news document include: Body (the text of the article); Headline; and Byline ("Document Section").

In Figure 3.10 a full document from *LexisNexis Academic* is shown, and note these sections in the document: Section, Length, Person, Organization, Country, State, Company, Subject, Load-date, Language, and Graphic. Neither the headline nor the text of the article are tagged with an all-capitals label, but they can be searched as a section also. In order to initiate a section-specific search, from the Power Search form, the plus sign to the left of "Show options to search specific document sections" must be clicked.

A document section needs to be selected from the drop-down box that appears when the down-arrow in the Select box is clicked. Next, a search term for that particular section needs to be entered into the box to the right of Terms. The process can be repeated, so searches in multiple sections can be performed at once. In the case of multiple section selections, *LexisNexis Academic* automatically installs the AND connector between each desired section search to maintain the proper syntax in the search string. Therefore, a possible search string could have this appearance: HEADLINE (speaking) and BODY (atlantic) and State (Massachusetts).

This search string resulted in the article in Figure 3.10, and the search terms, *speaking*, *Atlantic*, and *Massachusetts* were identified in red. January is also highlighted—perhaps because of the search specification for January 28, 1998.

Finally, the section search string can be entered manually without the use of the find and click method explained above. For example, State is not one of the choices offered from the section selection list, but after its presence was noted in the article, the search within it was successful. Here is one more example of both the flexibility of *LexisNexis Academic* and the need for users to become acquainted with how it operates. Note also that while three sections were searched, and one result was highlighted in each section, the results interface regards this article as one result. Still, the search terms are usefully highlighted in red, and through the Expanded List, which is explained below, the multiple terms for a search, such as this one may be previewed in context.

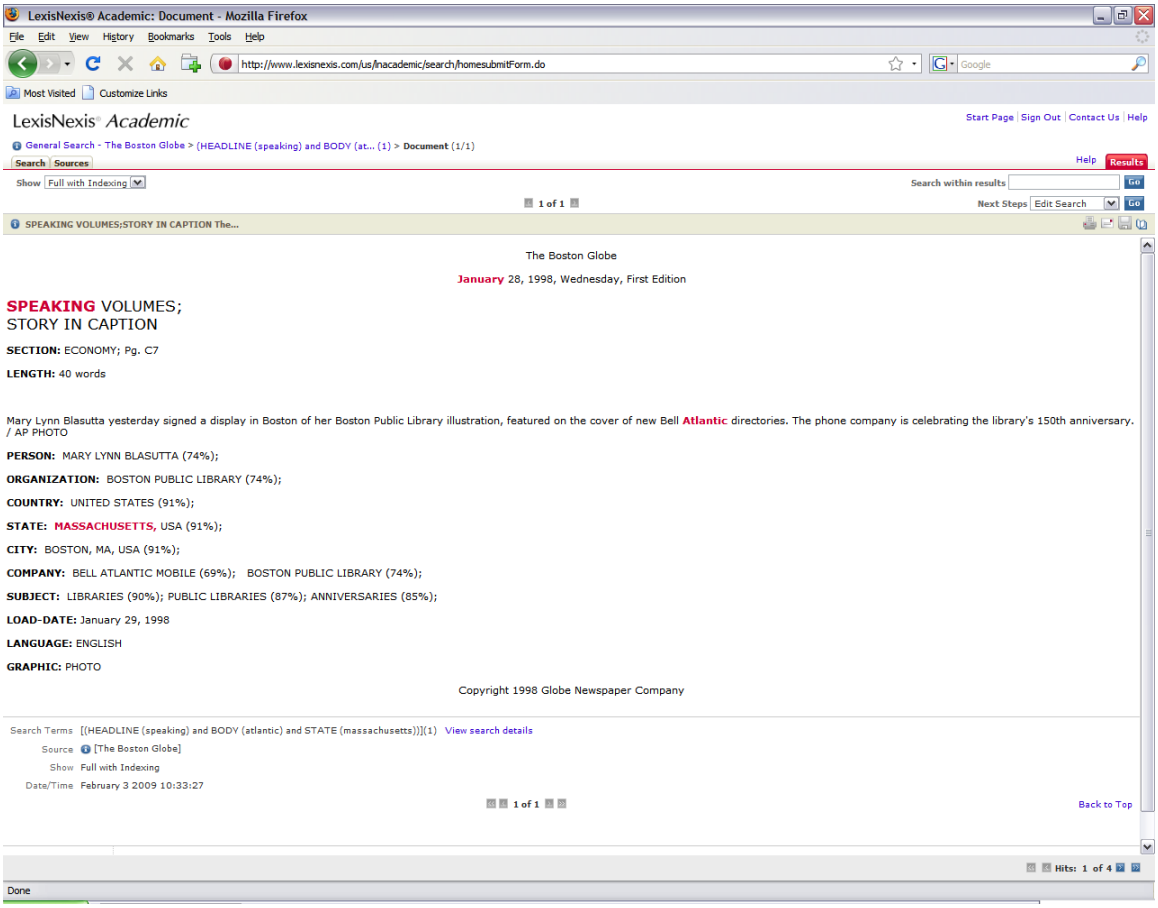


Figure 3.10: A full document with highlighted search terms in sections

Another way a document section search can be employed is a search for text within one section, such as sports. Such a string could be constructed in this way:

football pre/1 practice AND SECTION (sports)

This string would find the collocation *football practice* only in article sections, which are tagged as sports in the specified title(s) in the specified date range. In terms of the BVC, some forms that were pursued for study, such as *nickel back*, *nickel package*, and *nickel guy*, have a higher frequency in sports sections of newspapers; the ability to isolate one section for searching was useful for finding those forms in sports sections.

A comparison of sections (that is, frequency in one section compared to frequency in another section) was important to show empirically to what degree a specialized form has entered use in other sections of newspapers. The study used sections as domains in the BVC's newspapers to show the rise of use of the form *low carb diet*. For example, 2004 (the year with the highest number of occurrences for *low carb diet*) was the only year in which that form had occurrences in ten different newspaper sections. The year with the next highest number of sections was 2003 with seven sections. The ability to identify how forms manifest in sections allows a user of the BVC to go beyond frequency (which in itself is extremely significant) and make conclusions about a form's use across domains in addition to frequency.

Parentheses

Finally, the importance of parentheses in relation to search commands in *LexisNexis Academic* should be evident from the search string for the *carb* pilot corpus. The study used parentheses in this string to create priority over nearby connectors ("Developing a Search"):

high OR low OR diet OR atkins OR (south pre/1 beach)

The connection of *south* to the position immediately before *beach* is worked out before the OR connector; this priority ensured by the parentheses, in this example, maintains the connection of the *south beach* collocation in the search string.

Noise words

Because of the behemoth scale of the characters, words, and texts that *LexisNexis Academic* includes, the database maintains a list of words that cannot be searched. *LexisNexis Academic* explains, "Noise words, also called stop words, are commonly used English words like

any, *of*, and *all*. While we cannot provide a list of all noise words, here are the most common ones: any, at, as, The, my, are, and*, when, so, of, there, or, his, is, it" ("Noise Words"). The asterisk beside *and* indicates that it is "a reserved word," which is a word that is used as a connector in searches ("Noise Words"). *LexisNexis Academic* continues, "In most sources, you cannot search for noise words because the LexisNexis services disregard them. But in select sources and under certain subscriptions, you can search for noise words" ("Noise Words"). This study could not search for, or more importantly, exclude noise words and consequently had to create strategies, in some cases, to deal with their intrusion into search results. The study had to visually verify search returns because of noise words. The search string, *going missing*, was much more problematic than *go missing*, and it attracted many false returns in *LexisNexis Academic*, such as *going to be missing*, which includes the noise words *to* and *be*, that, again, had to be visually detected and removed from the results count.

Search display

One powerful feature in *LexisNexis Academic* is a search with no search terms. After a newspaper title has been selected, if the search window is left blank, and the "Date is…" has been selected, the search results will contain all the possible results within *LexisNexis Academic* for that newspaper title on that particular date.

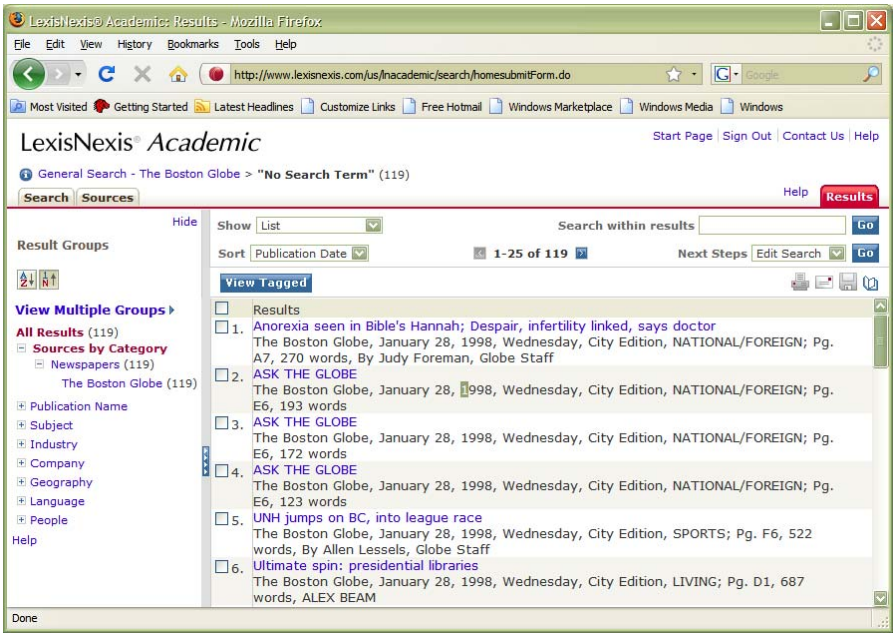


Figure 3.11: List of results

For *The Boston Globe* on 28 January 1998, the total number of results is 119. This is the display that makes an estimated word count possible for the BVC. Note, for example, in Figure 3.11 the bibliographic information contains a word count for each result, such as *270 words* for the first result. The collocation, *270 words*, included in the metadata for each result, is the critical element of the metadata for each result in the word count estimation process.

Users can select from the following display choices for search results: List (Figure 3.11), Expanded List (Figure 3.12), Full Document (Figure 3.10), KWIC, and Custom. The List display includes the title and other metadata for the articles in the results in increments of 25 results. *LexisNexis Academic* explains, "The Expanded List helps users to quickly determine if the document is relevant to the search and link to specific occurrences of their search terms within each document" ("Features"). Note in Figure 3.11 that the search term, *street*, is highlighted in blue and useful surrounding text is included. The search term, which is

highlighted in blue, can be clicked to deliver a bundle of all of the documents in that particular search that contain that term. After clicking the search term, the user would be directed to a Full Document that has the search term highlighted in red, and the user can click an arrow at the top of the document that will open the next Full Document with the search term, again, highlighted in red. Although, another display selection is referred to as KWIC, the Expanded List has more of the attributes of what a linguist would typically expect in a KWIC display because the search term is highlighted and surrounding text is displayed as strips of text in a preview format.

The Full Document displays the full text of the article that contains the search term, and the Full Document is displayed singly. The user can select the next Full Document by clicking an arrow above the document text.

The KWIC display allows the user to view ". . . each occurrence of the search terms surrounded by approximately 20 words of adjacent text" ("Features"). This display is isolated to one document per display, similar to the Full Document display, so the ability to view many previews at once, as the Expanded List allows, is impossible with the KWIC display. So, the isolated presentation of results information from the KWIC display is what, for this study, made the Expanded List much more powerful and consistent with the way keyword in context displays function in other applications.

The Custom display allows the user to change the results display according to document section. This application is a reverse application of the document section search that is available from the Power search form, but is not completely functional. For example, after results are displayed, and the Custom display is selected, the options for a document section search are displayed, but a search cannot be executed from this application. When Section is selected, no other information, such as a specific section (e.g. sports), can be entered, so for 119 articles, the

search apparently collapsed when that selection, section, was made because the Custom display only included one article. Again, this is likely a powerful tool, but a user will have to determine how it works and then decide how to make it work for a specific need. For example, in the case of multiple results, the Custom display could allow the user to display all the BODY sections of the search results, so the user could view the body text of an article singly without having to view an assortment of metadata with the body.

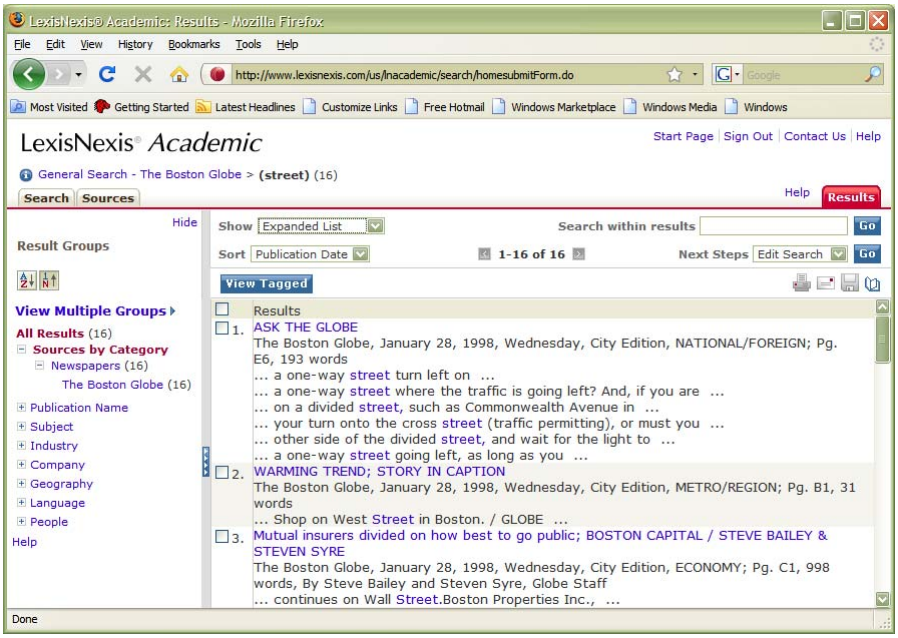


Figure 3.12: Expanded List of results

Downloading files from LexisNexis Academic

LexisNexis Academic will only display search results in increments of 25 results; also, *LexisNexis Academic* will not return any results for a search that produces over 3,000 results. This process could be disappointing to users who wish to obtain huge amounts of text quickly. The corpus linguist instinctively focuses on how to obtain large amounts of useful texts when

facing an interface, such as *LexisNexis Academic*, that contains access to a variety of useful texts. Such texts could be used for many different applications, and the texts need to be downloadable in a useful format. This study wished to do just that, download texts from *LexisNexis Academic*, and was able to through a special feature in the database that converts up to 500 search results to a single file. Above the results pane in *LexisNexis Academic* is a set of icons. One of these icons resembles a floppy diskette; this icon opens a downloading interface that is connected to the search results. The user can choose from several file types for the output, such as Word, HTML, Generic, Text, and PDF. The Word and Generic files both output files with an .rtf extension. This study used only the Text (.txt) output.

The .txt output is what the study needed for use in WordSmith Tools, which is a locally stored concordancing software application. No other file type was needed or used by the study. Still, the other file types can be very useful; the PDF format could be useful for preserving the original file in a stable format. The .rtf format (Word and Generic) could be useful for preserving special features, such as bold and color text, and the HTML format could be useful for placing articles online for student to view in a password-protected online interface, such as WebCT.

This conversion and delivery process can provide up to 500 results of any of the Document Views (List, Expanded List, Full Document, Full with Indexing, KWIC, and Custom) in any of the file format selections that *LexisNexis Academic* supports (Word, HTML, Generic, Text, and PDF). In the event of the need to convert over 500 results to a single file, the Select Items feature allows a user to conveniently organize large results into manageable increments. For example, 909 search results could be saved into 2 files; the first file could include 500 results (Select Items 1-500), and the second file could include 409 results (Select Items 501-909). This

process would yield two files that could be saved in one folder for later use, or the two files could be merged into a single file according to the user's preference. Figure 3.13 shows the All Documents selection that the user will be directed to if they select the download icon when multiple results have been generated, but non have been selected for individual viewing. Figure 3.14 shows the Current Document selection in the Download Documents form that a user would be face after selecting the download icon while a single result is displayed. Note that from either of these Download Documents forms, the user can select either one document or some combination of documents—up to a total of 500 documents for downloading.

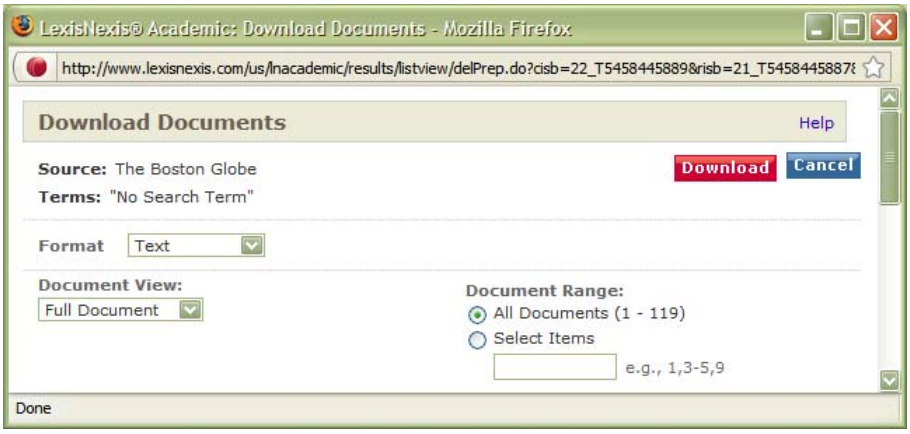


Figure 3.13: All Documents selection in Download Documents

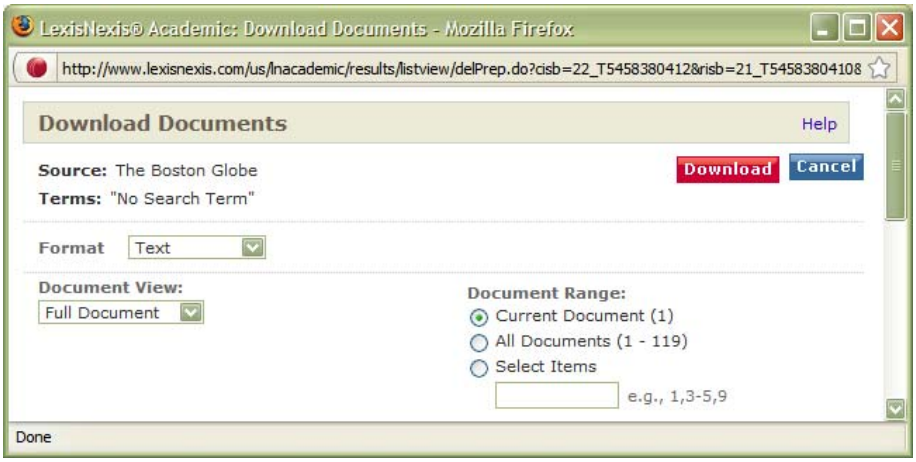


Figure 3.14: Current Document selection in Download Documents

Specialized terms related to use of LexisNexis Academic

This study refers to newspaper text as *speech*. Even though newspaper text is edited, this study regards newspaper text as the speech of a community. As mentioned above, *LexisNexis Academic* can generate results in the form of strips. This study refers to a single unit of such results as *a strip of text*. In the results that this study worked with, two types of complicated results were frequently encountered. First, *erroneously repeated* text refers to text that is unnecessarily repeated (possibly through some sort of computer malfunction); second, *recycled* text is text that is identical, but through no malfunction remanifests itself in the same or other newspapers in the BVC. A recurring book review could exemplify a recycled text in a single newspaper; a wire service article that is adopted by multiple newspapers in the BVC could exemplify recycled text between two or more newspaper titles.

The study downloaded texts from *LexisNexis Academic* for offline calculations with WordSmith Tools and refers to such a corpus as a *derived* corpus. Some of these derived

corpora were constructed from strips of text, and others were constructed from full texts downloaded from *LexisNexis Academic*. The study refers to former as a *strip corpus* and to the latter as a *full text corpus*.

Conclusion

LexisNexis Academic gives this study the tools to create very large virtual corpora of newspaper text. The large number of newspaper titles in *LexisNexis Academic* also created for the study the opportunity to balance the virtual corpora in terms of circulation and the location of the newspaper's publication. These elements of planning and balance follow in the tradition established by the classic corpora. Further, when a user finds useful search results in *LexisNexis Academic*, those results can always be traced to the original text that includes evidence the text's original date of use and location of publication.

LexisNexis Academic does much more than just provide text to the user; it also provides sophisticated search, document delivery, and previewing interfaces. These features cater well to the type of word-level research that this study embraces.

One interesting attribute of *LexisNexis Academic* is the fact that it not only supports research, but it also leads research. The powerful options related to search, results display, and document delivery have influenced the momentum of this study because they have caused the study to pursue new techniques for research as well as new research questions in themselves.

Chapter 4

A Case Study of a British Construction in American English: *go missing*

The BVC is a useful tool for the study of the recent career of *go missing* in American English because the form's career can be observed annually from 1998 to 2007 in both the American and British BVCs, and the form's use can also be observed regionally in the American BVC.

An approach to the question of *go missing*'s career in American English without the BVC would likely be anecdotal and based on electronic resources without an actual or estimated word count, such as the Web via Google or annual searches of single newspaper titles. These two approaches could likely produce evidence, but that evidence could not be placed in a meaningful context without the possession of critical information about the source texts—most importantly a word count. For example, the word count of the web is unknown, so X number of results cannot be described in a meaningful way because we can't calculate a rate of occurrence. Also, the selection of one or even several newspaper titles places a user at a similar disadvantage because the results themselves are not meaningful without a total corpus word count.

The BVCs have estimated word counts that enable a calculation of a rate of occurrence (per ten million words). This frequency makes a principled comparison of American results with British results possible because the results from each corpus are expressed in rates per ten million words. This ability to compare is especially important with *go missing* because its career as a current form in American English is still quite young, but the form has a well established and documented career of many decades in British English.

This chapter provides results of *go missing*'s use by year, title, region, and all years/regions from the American BVC. This chapter also provides results for the form in the British BVC by year, title, all years/titles. This fine-grained detail is useful because no principled quantitative study that compares this form's presence in British and American English has ever been conducted.

This chapter shows evidence of periods in American English in which the form's frequency is extremely low as well as periods in which its frequency increased significantly. Also, this chapter does not assume that the presence of the form in British English is high frequency; the form's presence in British BVC was pursued just as it was in the American BVC. This attention detail is useful both for the determination of frequency and identification of the form's variation in each BVC.

Innovation and American English

American English's heritage could be regarded as rebellious because American English has welcomed lexical and other linguistic innovations that highlight its separation from British English. For example, in the nineteenth century, American English coined a large selection of unusual words that Pyles refers to as "tall talk" (129). These words include *bodacious*, *absquatulate, cattywampus,* and *grandiferous* (Pyles 129), and they are examples of American English's ability and, sometimes, enthusiasm for moving away from British English.

Still, possibly surprising to some of its speakers, American English also currently welcomes innovations directly from British English. For example, the British English construction, *go missing* 'to disappear' has not only a presence in American English, but its

frequency in American English has shown evidence of an increase over the last decade. The presence of *go missing* constructions in American English has been noted by linguists including Algeo (1988), and Slotkin (1990), who described *go missing* as a British construction that has migrated into American English. Algeo notes, "A currently [c. 1988] fashionable collocation, which has also been the subject of popular comment, is *go missing*" (Algeo 30). As the frequency of *go missing* increased in American English, the reaction of speakers of American English to the form's use increased as well.

American newspapers chronicle reactions, sometimes emotional, to *go missing*, such as this 2007 headline: "Brace for Cheers if 'Went Missing' Ever Disappears" and an accompanying quotation: "Dozens of readers wrote in to tell me about the phrases that set their own teeth on edge, the all-time champion being: 'Went missing' (*LexisNexis Academic*). And from 2006, the more supportive, "Why have Americans, then and now, shunned gone missing? It's neither unclear nor ungrammatical (if you can go bankrupt, why not go missing?), and action-verb boosters should be applauding it as stronger than is missing" (*LexisNexis Academic*). The recent presence and use of *go missing* in American English has attracted attention both from the scholarly and popular communities. This attention is rooted in the form's newness in American English.

Historical Background of *go missing* in American English

Many of the earliest uses of *go missing* in American English that this study located share an interesting connection—maritime disaster. In January 1935, a *Christian Science Monitor* article includes: "One by one ships go missing at sea, or lose a suit of sails in a gale and go to the shipbreakers because it is inexpedient to provide new canvas" (*ProQuest Historical*

Newspapers). Later, in October 1935, The *Christian Science Monitor* included this *go missing* construction: "In the end, Acme went 'missing,' somewhere along the way to Cape Horn" (*ProQuest Historical Newspapers*). The quotation marks around *missing* appear to indicate that the author, or perhaps an editor, sensed their audience might find *went missing* to be a bit unusual. In *The New York Times*, a 1939 article explains, "When a vessel goes missing and the news becomes generally known, all kinds of theories are advanced, and in this respect the British freighter's disappearance brought forth an avalanche of guesses" (*ProQuest Historical Newspapers*). These articles do not show evidence of foreign authorship; for example, neither London nor any other foreign city is mentioned in the datelines.

This study also located some older *go missing* forms. Again, these forms also relate to maritime disaster, but they feature *a*-prefixing. For example, note this 1895 quotation from a *New York Times* article that explained the situation of a ship from Florida: "Nowadays when vessels go a-missing there are many theories advanced to account for their non-appearance" (*ProQuest Historical Newspapers*). *The New York Times* also included this construction in 1908: "But it is not every derelict that is so considerate as to sink, as is evidenced by the ships that go a-missing every once in a while for unaccountable reasons (*ProQuest Historical Newspapers*). Other noteworthy a-prefixing usages include *gone a-missing* in *New York Times* articles, also maritime in nature, from 1893, 1894, and 1896.

In addition to the early maritime uses of *go missing*, historical evidence also exists for British authorship of some *go missing* constructions in American newspapers. A 1913 *New York Times* article, whose dateline is London, features this use of *go missing*: "The frequency with which suitcases and other articles of portable property go missing has caused a rule to be enacted in some of these caravanseries that the club servants are not to be held responsible for anything

members may lose on the club premises" (*ProQuest Historical Newspapers*). Also, in a 1920 *NYT* article, with a London dateline, "To some women weeping a little in the crowd after an all-night vigil, he was their boy who went missing one day and was never found till now, though their souls went searching for him through dreadful places in the night" (*ProQuest Historical Newspapers*). *Go missing* has had some presence in American English historically, but its frequency, before recent years, seems limited to these scattered occurrences.

Analysis of evidence in the American BVC

The presence of *go missing* in recent American English can be demonstrated by combining the total number of all *go missing* forms that appear annually in the American BVC from 1998 to 2007. These totals include *go, goes, going, gone,* and *went missing* forms. These results were visually inspected and counted. This process was very time consuming, but such an approach was embraced because of a few reasons. First, the results in *LexisNexis Academic* sometimes repeat themselves—sometimes in pairs—possibly even in groups of four, so the presence of duplicate articles had to be acknowledged, so they could be removed from the count. All instances of *go missing* forms, including repeats of the same form in a single article were included, but commercial uses, such as racehorse names or titles of theatrical productions were not included.

LexisNexis Academic has a list of stop words (generally ultra high frequency forms) and characters that it will not recognize in searches, so a search for *go missing* can give an unwanted result, such as: "*go* from *missing* the playoffs." Also the mere presence of a *go missing* construction does not ensure that the construction is functioning grammatically in the sense of

'to disappear'—which was usually the result of intrusive search elements that *LexisNexis Academic* does not filter out, such as stop words and punctuation. Such forms were not included in the count. The following table shows the total number of all *go missing* forms by year in the American BVC. The italicized numbers are the occurrences expressed in rate per ten million words.

Table 4.1: Occurrences of all *go missing* forms by year in the American BVC

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	ten year total
98	173	230	338	481	599	765	888	942	1052	5566
1.755	*4.173*	*4.553*	*3.447*	*7.226*	*12.813*	*19.117*	*17.398*	*21.345*	*25.688*	*10.398*

Table 4.1 shows that from 1998 to 2007, the total number of *go missing* forms generally increased annually from 1998 to 2007. The rate of occurrence for each of the years of 2003 through 2007 exceeds the average rate of 10.398 per ten million words. Table 4.1 is a snapshot of the career of all *go missing* forms based on evidence from the BVC. The BVC makes both snapshots and closer analyses possible; for example, the bibliographic metadata that accompanies downloads from *LexisNexis Academic*, as well as the text itself, allows the researcher to determine trends within the text and the influence of cultural events on the occurrence of *go missing* forms.

Regional comparison of go missing in the American BVC

The study organized the occurrences of *go*, *goes*, *going*, *gone*, and *went missing* forms by form, year, and region. The following tables (4.2 through 4.6) include, for each form, the number of occurrences for each form by region and year and include annual and regional totals.

As in Table 4.1, the number in plain font is the number of results; the italicized number in the same cell is the frequency of that number of occurrences in a rate per ten million words. The per ten million word rate makes standardized comparisons possible.

The tables that show the results from the American BVC reflect rates per ten million words based on the whole corpus (5.3 billion words) and these subcorpora: year/region (e.g. 1998 NE); year/all regions (e.g. 1998 NE, SE, MW, W, CW); and ten years/one region (e.g. 1998-2007 NE). The tables include fifty interior cells that reflect calculations for each year/region subcorpus, and the right margin of the tables includes cells for the year/all regions subcorpora calculations. The bottom margin of the tables includes calculations for the ten years/one region subcorpora. Finally, the discussion that follows each table includes illumination of relevant, sometimes microscopic, factors that are not discernible from the tables alone.

Table 4.2: Results for *go missing* in the American BVC

Year	NE	SE	MW	W	CW	Totals
1998	6 *.376*	1 *.080*	1 *.137*	6 *.667*	1 *.089*	15 *.269*
1999	9 *.885*	8 *.813*	4 *.525*	1 *.160*	3 *.394*	25 *.603*
2000	4 *.353*	5 *.429*	4 *.556*	3 *.389*	2 *.157*	18 *.356*
2001	12 *.356*	6 *.266*	3 *.235*	8 *.774*	12 *.637*	41 *.418*
2002	25 *1.773*	15 *1.161*	11 *1.209*	9 *.944*	11 *.525*	71 *1.066*
2003	23 *1.846*	18 *1.767*	10 *1.397*	12 *1.734*	13 *1.291*	76 *1.625*
2004	20 *1.841*	13 *1.491*	13 *1.942*	9 *1.463*	8 *1.053*	63 *1.574*
2005	22 *1.637*	27 *2.178*	19 *2.829*	14 *1.503*	17 *1.851*	99 *1.939*
2006	26 *2.655*	23 *2.437*	13 *1.593*	11 *1.083*	22 *3.316*	95 *2.152*
2007	17 *1.590*	23 *1.965*	18 *3.833*	21 *2.724*	12 *1.948*	91 *2.222*
Totals	164 *1.152*	139 *1.141*	96 *1.241*	94 *1.132*	101 *.911*	594 *1.109*

Table 4.2 includes both the infinitive form as in this strip of words from a 2004 *Chicago Sun Times* article, “alone, bringing the number to go missing in November to more than” (*LexisNexis Academic*) as well as the present tense form as shown by this 2007 *Chicago Sun Times* article: “number of people who go missing in Chicago and around the” (*LexisNexis Academic*).

Table 4.2 displays the rise of *go missing*’s frequency. For example, in 1998 three of the regions have only one result for *go missing*. In 2007, the regions have a minimum of 12 results and a maximum of 23 results. Still, speech, which is so often categorized as being orderly and rule-based, follows its own tendencies toward disorder and variation rather than rules. For example, in the Northeast region, the number of results for 2002 is greater than the number of results for that region in 2007. The Northeast region does progress from 6 results in 1998 to 17 in 2008, but in addition to the 2002 results (25), results for 2003 (23), 2004 (20), 2005 (22), and 2006 (26) are also greater than the results for 2007 (17). Also, the 25 results for the 2002 Northeast region has only one skewing factoring involved; 3 of the results are within a book review that was published in the *New York Times* in three different weeks in March of that year. Even after consideration is given to this skewing factor, the 2002 Northeast results of 25 remains significant because the occurrences of *go missing* generally increased in the Amercan BVC between 1998 (15) and 2007 (91). Still in the Northeast region, the years with the highest results are 2002 (25) and 2006 (26).

Evidence from the BVC reflects the rise of *go missing*’s use in American English. Further, the BVC’s flexibility to isolate *go missing* results to a point as specific as a single title in a single year shows not only the granular changes that construct the overall increase, but also the BVC’s ability to show variation, such as the case in which the Northeast region’s

second-largest results are in 2002. Such conclusions could not have been found without the BVC's methodology that considers both geography and duration of time. Finally, the file delivery from *LexisNexis Academic*, which places search results in chronological order, gives the researcher a prefabricated file organization to assist linguistic analysis.

Table 4.3: Results for *goes missing* in the American BVC

Year	NE	SE	MW	W	CW	Totals
1998	6 *.376*	3 *.240*	1 *.137*	4 *.445*	2 *.179*	16 *.286*
1999	8 *.787*	6 *.609*	2 *.262*	2 *.320*	5 *.657*	23 *.554*
2000	5 *.442*	7 *.601*	4 *.556*	2 *.259*	2 *.157*	20 *.395*
2001	17 *.505*	5 *.222*	6 *.470*	7 *.677*	6 *.318*	41 *.418*
2002	18 *1.277*	33 *2.555*	4 *.439*	6 *.629*	16 *.764*	77 *1.156*
2003	23 *1.846*	20 *1.963*	9 *1.257*	7 *1.017*	13 *1.291*	72 *1.540*
2004	44 *4.051*	9 *1.032*	9 *1.344*	15 *2.439*	20 *2.632*	97 *2.424*
2005	33 *2.455*	24 *1.936*	12 *1.787*	24 *2.577*	11 *1.198*	104 *2.037*
2006	28 *2.870*	26 *2.755*	29 *3.553*	23 *2.266*	23 *3.467*	129 *2.923*
2007	28 *2.618*	34 *2.905*	28 *5.963*	24 *3.114*	21 *3.409*	135 *3.296*
Totals	210 *1.475*	167 *1.371*	104 *1.345*	114 *1.373*	119 *1.073*	714 *1.333*

Goes missing also shows an overall increase in use from 1998 (16) to 2007 (135). For 2002 the Southeast region has a significant increase in results (33). Of those 33 results, 15 are a single film review that was repeated on different days in the *Times-Picayune*. The BVC does attract multiple results for single events or more specifically in this case, an identical film review that is repeated on successive dates. That phenomenon appears to be a problem, but the file delivery from *LexisNexis Academic* (for a single title search, the results are ordered in the file by year from newest to oldest) and convenient bibliographic tagging that accompanies each result help a user to visually isolate single events as well as the number of recycled results (identical text, such as a film review, that appears on different dates). In this case, the headline for the strip of text frequently would be "Movies," but other times might be a broader title, such as "Straight-talking actor has a clear perspective" (*LexisNexis Academic*). The strip of text, "a nuclear bomb goes missing. Phil Alden Robinson's" can be easily identified, visually, in the results for 2002 for the *Times-Picayune*. Also, concordancing software, such as WordSmith Tools can be used to identify repeated text, but as stated above, this study visually counted all *go missing* results in a text file—rather than automating the process because an application, such as WordSmith tools cannot discern the presence of erroneously repeated results. WordSmith Tools can identify identical lines, but human intervention is needed to determine error-based repetitions.

The Midwest region has a small number of results in 1998 (1), and the results increase for 2006 (29) and 2007 (28). No skewing factor can be found for 2007, and for 2006, the *St. Paul Pioneer Press* has 15 of the region's 29 results. Of those 15 results, 7 are an announcement of what appears to be a single theatrical event, and this line includes: "good they had it when their teacher goes missing" (*LexisNexis Academic*). This single event skews the results slightly for the Midwest region in 2006.

For the Southeast region in 2006, the *Washington Post* has 16 of the region's total of 26 results, but no skewing factors can be found.

Of the 34 results for the Southeast region in 2007, 25 come from the *Washington Post.* In this case, one possible skewing factor is possibly the work of a single author. This possibility cannot be proven with the evidence that is available via *LexisNexis Academic*, but the database's bibliographic information that accompanies each set of results shows that 8 instances of *goes missing* appear in a recurring newspaper feature titled, "Highlights." This feature overviews recent events within television programs; perhaps one author with an affinity for *goes missing*, wrote all of these 8 articles in 2007. These 8 results do not appear to be connected to one event, such as a single television program.

In the Northeast region, the *New York Times* has a significant number of the region's total results for 2003 (12 of 23); 2004 (26 of 44); 2005 (15 of 33); 2006 (16 of 28). For 2006 only 2 of the 12 articles are connected to a single event; they are occurrences of a book review on separate dates. That same book review is repeated through 2005 and accounts for 10 of that year's 15 total results. In 2004, five events cause repetitions and account for 22 of the 44 results for *goes missing* in the *New York Times*: a repeated book review (3); a review of a performance (5); an article title (2); a book review (8); another book review (4). In 2003 the *NewYork Times* has 6 results that are a single recycled book review; that event is the only skewing factor for the *New York Times* for 2003.

Table 4.4: Results for *going missing* in the American BVC

Year	NE	SE	MW	W	CW	Totals
1998	1 .062	0 *0*	0 *0*	0 *0*	0 *0*	1 *.017*
1999	0 *0*	0 *0*	0 *0*	0 *0*	1 *.131*	1 *.024*
2000	2 *.176*	1 *.085*	0 *0*	0 *0*	0 *0*	3 *.059*
2001	0 *0*	2 *.088*	1 *.078*	1 *.096*	2 *.106*	6 *.061*
2002	0 *0*	1 *.077*	1 *.109*	0 *0*	2 *.095*	4 *.060*
2003	5 *.401*	1 *.098*	1 *.139*	0 *0*	1 *.099*	8 *.171*
2004	4 *.368*	3 *.344*	6 *.896*	0 *0*	1 *.131*	14 *.349*
2005	1 *.074*	1 *.080*	1 *.148*	3 *.322*	6 *.653*	12 *.235*
2006	5 *.512*	1 *.105*	6 *.735*	2 *.197*	1 *.150*	15 *.339*
2007	5 *.467*	3 *.256*	6 *1.277*	2 *.259*	1 *.162*	17 *.415*
Totals	23 *.161*	13 *.106*	22 *.284*	8 *.096*	15 *.135*	81 *.151*

Going missing is the lowest frequency *go missing* form in the American BVC. Multiple regions have zero results for 1998, 1999, and 2000. The West region has zero results for 2002, 2003, and 2004. No skewing factors could be identified in the BVC for the *going missing* results. The Coastal west region has zero results for 1998 and only one result for 2007; however, the Coastal west region has six results for 2005.

One interesting attribute of *going missing* is that it does not seem to occur with regard to lost pets as much as other forms of *go missing*. In a search of the *going missing* strip corpus, the forms *pet* and *dog* do not appear; however, the form *cat* does appear once. One factor that could account for this difference is the probability that, based on evidence from the American BVC, *going missing* is used much less frequently than the other forms of *go missing*. Also, the connection of two *–ing* forms to create a single verb phrase could likely be awkward for many speakers and writers of American English—which might explain the comparatively lower number of occurrences for *going missing*.

Table 4.5: Results for *gone missing* in the American BVC

Year	NE	SE	MW	W	CW	Totals
1998	7 *.438*	4 *.320*	4 *.551*	4 *.445*	4 *.358*	23 *.412*
1999	11 *1.082*	12 *1.219*	4 *.525*	5 *.801*	7 *.920*	39 *.940*
2000	17 *1.504*	17 *1.461*	6 *.834*	6 *.779*	17 *1.339*	63 *1.247*
2001	33 *.980*	21 *.932*	9 *.706*	18 *1.742*	14 *.744*	95 *.968*
2002	38 *2.696*	49 *3.794*	20 *2.198*	8 *.839*	17 *.812*	132 *1.983*
2003	41 *3.292*	34 *3.338*	18 *2.515*	19 *2.760*	24 *2.383*	136 *2.909*
2004	68 *6.260*	33 *3.785*	24 *3.586*	17 *2.764*	28 *3.686*	170 *4.248*
2005	63 *4.688*	33 *2.662*	30 *4.468*	30 *3.221*	26 *2.831*	182 *3.565*
2006	77 *7.894*	33 *3.497*	32 *3.921*	30 *2.955*	39 *5.879*	211 *4.781*
2007	59 *5.518*	53 *4.529*	41 *8.732*	25 *3.243*	30 *4.871*	208 *5.079*
Totals	414 *2.908*	289 *2.372*	188 *2.431*	162 *1.952*	206 *1.858*	1259 *2.352*

For the results of *gone missing*, several factors that are not apparent from an observation of the table alone influence the final results. For example, in 2002 the results are 49 for the Southeast region, and 43 of those results are from the *Washington Post*. That distinction alone is worthy of note; however, 28 of the 43 come from a single film review that is repeated across several weeks. The repeated strip of text, "at Bletchley -- and that Claire has gone missing. First-rate in" can be simply located both through a visual inspection of the file delivery from *LexisNexis Academic* and as well as with WordSmith Tools. Also, the results from *LexisNexis Academic* are numbered; this metadata attribute assists researchers in the observation of the burstiness (Kilgarriff (2001) 107) of a form, such as *go missing*. A researcher could create a spreadsheet, for example, and tabulate the article numbers in which each form of *go missing* occurs in a certain text. After tabulation, the researcher could determine what article number has the highest count of all forms of *go missing* or which article number has the highest count of a single form of *go missing*. Such built-in conveniences make the *LexisNexis Academic* interface a convenient and powerful tool for linguistic research.

For 2006 the Northeast region has 77 results; within those results, the *New York Times* has 22 results, and the *Boston Globe* has 32 results. A single article in the *Boston Globe* has 12 of the *Boston Globe*'s 32 total results; the article, "Missing in Action" is a meta-article, or a linguistic discussion of the use of the form *gone missing*. Again, *LexisNexis Academic* allows both the employment of analysis of the full-text of the articles in the BVC and the quick isolation of a single article that has multiple uses or burstiness (Kilgarriff (2001) 107). While this article may seem to be freakish as an exhibit of the use of *gone missing*, its validity is easy to defend. This article is a part of the situation and experience of the BVC's Northeast community for 2006; therefore, this article is evidence both of use within the community and the exposure of

community members to use of *gone missing*. Similarly, in 2004 the *New York Times* has one meta-article with 6 uses (of 27 total for the title) of *gone missing*.

In 2001, the *New York Times* has a single film review that occurs twice, and 8 uses are from a book review that was repeated 8 times. Also, for 2002 the *New York Times* has 14 results for the region's 38 total results, and 6 of those results are a single repeated book review.

The *New York Times* has one performance announcement repeated 6 times which accounts for 6 of the title's 28 occurrences of *gone missing* in 2005; one *New York Times* article in 2005 is a review of a performance titled, "Gone Missing." Those 4 uses in one article are not counted because of the study's plan to exclude commercial uses of *go missing* forms, such as theatrical titles and racehorse names.

Table 4.6: Results for *went missing* in the American BVC

Year	NE	SE	MW	W	CW	Totals
1998	14 *.877*	5 *.400*	8 *1.103*	6 *.667*	10 *.896*	43 *.770*
1999	27 *2.657*	26 *2.642*	9 *1.182*	13 *2.083*	10 *1.315*	85 *2.050*
2000	31 *2.742*	36 *3.095*	18 *2.503*	13 *1.689*	28 *2.206*	126 *2.494*
2001	39 *1.159*	38 *1.687*	18 *1.412*	25 *2.419*	35 *1.860*	155 *1.580*
2002	41 *2.909*	53 *4.104*	33 *3.627*	26 *2.728*	44 *2.102*	197 *2.959*
2003	53 *4.256*	80 *7.854*	57 *7.964*	43 *6.247*	74 *7.349*	307 *6.567*
2004	90 *8.286*	82 *9.405*	92 *13.747*	71 *11.547*	86 *11.321*	421 *10.520*
2005	102 *7.590*	93 *7.504*	130 *19.361*	95 *10.200*	71 *7.733*	491 *9.620*
2006	98 *10.047*	90 *9.538*	111 *13.602*	111 *10.936*	82 *12.362*	492 *11.148*
2007	120 *11.223*	131 *11.195*	150 *31.949*	94 *12.197*	106 *17.211*	601 *14.675*
Totals	615 *4.320*	634 *5.204*	626 *8.097*	497 *5.989*	546 *4.926*	2918 *5.451*

Above much discussion has focused on how a single factor or event can cause an unusually large number of results to occur in a single title. On the other hand, an unusually large number of results can occur from a single source without a skewing factor; the total number of results for *went missing* for the Southeast region in 2003 is 80; 43 of those results come from a single title, *The Washington Post*. Because *went missing* is the most frequent form of *go missing* in this study, the combination of that form with the title that has the largest circulation in the region could combine to create the possibility for an unusually large number of results. Also, as stated in earlier discussions, the activity of a speech community does not conform to pre-established rules, so naturally some results may be higher or lower than others—without an obvious explanation.

The results for the 2006 *Salt Lake Tribune* illustrate that use of *went missing* can associate with children, "young girl who went missing from her Ogden home Sunday was;" adults, "since Steve and Catheryn went missing Nov. 8.The couple were;" pets, "family's dog, Jake, went missing on July 4th;" a young man involved in an avalanche, "danger is high where the teen went missing, Gilbert said. A;" and also, money, "$27,000 of those funds went missing.And now the former executive;" (*LexisNexis Academic*). Finally, these results also show that the construction can be used in connection with the theft of an object: "a pottery fragment went missing in summer 2004.Therein" (*LexisNexis Academic*). The large scale of the BVC's inclusion allows observation of possibly unusual uses, such as the stolen piece of pottery mentioned above.

The Laci Peterson tragedy represents a multifaceted news event that allows the *went missing* strip corpus to be used in reverse. Instead of counting the presence of *went missing* forms and possibly connecting them to skewing factors, the *went missing* strip corpus (or any

other strip corpus in this study) can be searched by keywords that relate to a specific event in order to determine articles and occurrences that relate to that specific event.

Laci Peterson was last seen on 24 December 2002 (*LexisNexis Academic*). Her disappearance and the subsequent investigations and legal proceedings against her husband for her murder attracted much national media coverage. Scott Peterson was arrested on 18 April 2003, and he was sentenced on 16 March 2005 (*LexisNexis Academic*). These events only affected the Coastal west results for *went missing* with significance (*LexisNexis Academic*).

The term *Laci Peterson* occurs in the *went missing* strip corpus from the *San Francisco Chronicle* 4 times in 2004 and 3 times in 2003. In 2004 the *went missing* corpus from the *San Francisco Chronicle* includes 16 results that relate to Scott Peterson's trial; in the entire *went missing* corpus for all regions, except for the Coastal west, *Laci Peterson* occurs only two times (2 November 2004 *Washington Post*; 25 March 2005 *Denver Post*) and *Scott Peterson* occurs zero times. In the Coastal west region alone, *Laci Peterson* occurs 8 times and *Scott Peterson* occurs 5 times. Interestingly, in 2004, references by name to the trial were frequently made by first name (*Laci*) or last name (*Peterson*) only.

The American BVC presents evidence in support of the newness and also for the rise in frequency of *go missing* in the American BVC. The estimated word count in the BVC allows for the instances of *go missing* to expressed in rates per ten million words. These rates enable observation of the rise in frequency of *go missing* in the American BVC.

The *go missing* strip corpora can be analyzed offline as long as the files are saved electronically. This study did so, and the file saving process (made possible through *LexisNexis Academic*'s file delivery system) was very simple. Those saved files enabled analysis of significant repetitions and multiple events as discussed in the text that follows the relevant tables

in this chapter. This flexible attribute of the BVC, the ability to plan for future calculation needs, allows the researcher to analyze relevant data offline and even employ an additional application, such as WordSmith Tools, for further calculations, such as the generation of collocations.

Analysis of Evidence in the British BVC

Because this study is underpinned by the notion that *go missing* is a British construction that recently gained currency in American English, The British BVC allows a comparison between use of the form in the American BVC and use of the form in the British BVC. Additionally, because the size of the British BVC is similar to that of one region in the American BVC, comparisons between an American region and the British BVC are particularly useful because the corpus word counts are similar.

Table 4.7: Overview of totals for *go missing* forms in the British BVC

Construction	1998 total		2007 total		ten year total	
go missing	68	*7.228*	103	*11.004*	803	7.720
goes missing	49	*5.208*	84	*8.974*	574	*5.518*
going missing	51	*5.421*	90	*9.615*	578	*5.557*
gone missing	184	*19.558*	253	*27.030*	2006	*19.287*
went missing	501	*53.253*	921	*98.400*	5464	*52.535*

Table 4.7 shows that the results for all of the *go missing* forms show an increase from the 1998 results to the 2007 results in the British BVC. This brief overview table conceals many factors that can cooperate to influence the annual results for *go missing* forms in the British

BVC. The following tables (4.8 through 4.12) provide a closer analysis of the results by form, newspaper title, and year. The discussion that follows each of these tables addresses factors that cannot be seen in simple numbers of results. Several attributes of the BVC allow such factors to be identified and illuminated.

Table 4.8: Results for *go missing* in the British BVC

Year	*Daily Mail*	*Eve. Chron.*	*Herald*	*Independent*	*N. Echo*	Totals
1998	19 *10.349*	16 *7.965*	11 *4.442*	21 *8.705*	1 *1.482*	68 *7.228*
1999	25 *10.799*	14 *12.805*	11 *4.655*	18 *6.601*	13 *24.196*	81 *8.965*
2000	39 *4.779*	8 *9.343*	12 *2.803*	14 *2.826*	6 *10.382*	79 *4.196*
2001	21 *3.636*	5 *6.942*	4 *2.349*	16 *5.877*	9 *13.400*	55 *4.745*
2002	27 *9.687*	23 *47.457*	10 *6.942*	37 *12.740*	12 *47.995*	109 *13.856*
2003	31 *11.471*	6 *8.249*	7 *4.758*	21 *8.480*	4 *7.749*	69 *8.741*
2004	28 *10.794*	9 *11.925*	13 *11.570*	15 *4.591*	7 *9.922*	72 *8.526*
2005	29 *8.550*	15 *12.247*	13 *11.910*	14 *4.331*	4 *2.211*	75 *6.977*
2006	42 *9.001*	12 *10.269*	12 *8.937*	25 *7.833*	1 *2.156*	92 *8.493*
2007	50 *11.107*	12 *9.131*	10 *8.708*	19 *10.732*	12 *19.186*	103 *11.004*
Totals	311 *8.030*	120 *11.591*	103 *5.586*	200 *6.744*	69 *10.101*	803 *7.720*

For 2002, 16 of the 37 results for the *Independent* are from a single recycled film review. No other factor could be identified to account for the increase from a total of 55 results in 2001 to a total of 109 results in 2002.

A major news event in 2007, the disappearance of young Madeleine McCann on 3 May 2007 (*LexisNexis Academic*), did not significantly affect the results for *go missing*. Only one result (on 6 December 2007) is connected to that event; the article includes this use of *go missing* "torment of having a child go missing, like Vicky did all those," and the article explains the McCann family's gesture of support to another family affected by tragedy (*LexisNexis Academic*). For the purposes of the British *go missing* strip corpus, the evidence that is revealed in the above text strip is connected to the McCann event is the title, "McCanns in moving tribute to Vicky." Note that the strip corpus includes both the title and the text strip that includes the *go missing* form; also, the title could include the *go missing* form.

Because the McCann event was so intensely covered by news media in Great Britain, this study searched the *go*, *goes*, *going*, *gone*, and *went missing* British strip corpora for these terms in search of evidence for the McCann event as an influence on other search results for this chapter: *Madeleine*, *Maddie*, *Maddy*, and *McCann*,. The child's name is Madeleine McCann, but some media outlets refer to her by a shortened form of her first name; note the two spellings above for the shortened form.

Finally at the entire British BVC level, the average rate of 7.720 per ten million words for *go missing* is exceeded significantly at the annual level only in 2002 (13.856) and in 2007 (11.004).

Table 4.9: Results for *goes missing* in the British BVC

Year	*Daily Mail*	*Eve.Chron.*	*Herald*	*Independent*	*N. Echo*	Totals
1998	15 *8.170*	13 *6.471*	3 *1.211*	16 *6.632*	2 *2.964*	49 *5.208*
1999	11 *4.751*	15 *13.720*	5 *2.116*	14 *5.134*	5 *9.306*	50 *5.534*
2000	17 *2.083*	10 *11.679*	10 *2.336*	9 *1.817*	1 *1.730*	47 *2.496*
2001	16 *2.770*	6 *8.331*	6 *3.524*	18 *6.611*	4 *5.955*	50 *4.313*
2002	18 *6.458*	7 *14.443*	4 *2.776*	23 *7.920*	3 *11.998*	55 *6.991*
2003	22 *8.141*	0 *0*	10 *6.798*	27 *10.903*	4 *7.749*	63 *7.981*
2004	23 *8.866*	2 *2.650*	7 *6.230*	12 *3.673*	4 *5.669*	48 *5.684*
2005	20 *5.897*	7 *5.715*	6 *5.497*	11 *3.403*	12 6.633	56 *5.209*
2006	29 *6.215*	9 *7.702*	9 *6.702*	19 *5.953*	6 *12.940*	72 *6.646*
2007	43 *9.552*	11 *8.370*	14 *12.192*	9 *5.083*	7 *11.192*	84 *8.974*
Totals	214 *5.525*	80 *7.727*	74 *4.013*	158 *5.327*	48 *7.026*	574 *5.518*

The results for the 2003 *Evening Chronicle* point out that just because a construction is current in British English, a British newspaper may not have any results for the construction for an entire year. A researcher might have easily assumed that zero annual results would not have been possible for any form of *go missing* in any of the British newspaper titles in this study. The evidence that the British BVC provides in this instance, zero results, could likely challenge the preconceived assumptions of a researcher (Sinclair 100). The presence of a title, in this case the *Evening Chronicle*, which has zero results for one year while having an average of 7.727 results per ten million words over the ten year period exemplifies the variety of sources included in the BVC. The other titles have these results for 2003: *Daily Mail* (22); *Herald* (10); *Independent* (27); *Northern Echo* (4). Still, perhaps surprisingly, for the ten year period, the *Evening Chronicle* has the highest average rate of occurrence for *goes missing*.

The *Independent* has single event factors that affect the results for 2002 and 2003. In 2002, 11 of the 23 results are a single, recycled film review, and in 2003, 9 of the 27 results are also a single, recycled film review.

Table 4.10: Results for *going missing* in the British BVC

Year	*Daily Mail*	*Eve. Chron.*	*Herald*	*Independent*	*N. Echo*	Totals
1998	17 *9.260*	9 *4.480*	7 *2.826*	14 *5.803*	4 *5.929*	51 *5.421*
1999	16 *6.911*	18 *16.464*	10 *4.232*	6 *2.200*	4 *7.445*	54 *5.976*
2000	10 *1.225*	15 *17.519*	16 *1.211*	6 *10.382*	6 *10.382*	53 *2.815*
2001	13 *2.251*	8 *11.108*	2 *1.174*	20 *7.346*	8 *11.911*	51 *4.400*
2002	17 *6.099*	12 *24.760*	16 *11.170*	6 *2.066*	3 *11.998*	54 *6.864*
2003	21 *7.711*	11 *15.124*	9 *6.118*	14 *5.653*	2 *3.874*	57 *7.221*
2004	22 *8.481*	11 *14.575*	11 *9.790*	12 *3.673*	5 *7.087*	61 *7.223*
2005	19 *5.602*	13 *10.614*	7 *6.413*	16 *4.950*	3 *1.658*	58 *5.395*
2006	24 *5.143*	14 *11.980*	3 *2.234*	3 *.940*	5 *10.784*	49 *4.523*
2007	46 *10.218*	12 *9.131*	12 *10.450*	3 *1.694*	17 *27.181*	90 *9.615*
Totals	205 *5.293*	123 *11.881*	93 *5.043*	100 *3.372*	57 *8.344*	578 5.557

The British BVC results for *going missing* stand as a useful contrast to the *going missing* results in the American BVC. Of the fifty cells in Table 4.4 (each cell represents the results for *going missing* in one American region for one year), sixteen cells have zero results. No cells in the results for *going missing* in the British BVC have zero results.

In the 2007 results for the *Daily Mail and Mail on Sunday*, 4 articles relate to the McCann event. The articles refer to the event through the names, *Madeleine*, *Maddie*, and *McCann*. This headline, for example, includes the child's shortened first name and family surname: "The Maddie Files; Five of Britain's top crime experts were sent to Portugal to give their verdicts on the McCann case" (*LexisNexis Academic*). The text that accompanies this headline refers more to the general concept of a missing child rather than to the specific McCann event: "following a child going missing is known as the 'Golden Hour'" (*LexisNexis Academic*). Another quotation refers to the child by her complete first name: "us the way Madeleine going missing did. We have our own heartache and grief" (*LexisNexis Academic*). The remaining 2 of the 4 total articles refer to the child as *Madeleine*. The use of *Madeleine* or *Maddie* or *Maddy* without a surname could be an expression endearment to the missing child and her family.

Table 4.11: Results for *gone missing* in the British BVC

Year	*Daily Mail*	*Eve. Chron.*	*Herald*	*Independent*	*N. Echo*	Totals
1998	56	32	43	37	16	184
	30.503	*15.930*	*17.364*	*15.337*	*23.719*	*19.558*
1999	53	22	36	46	17	174
	22.894	*20.122*	*15.237*	*16.869*	*31.641*	*19.258*
2000	58	41	37	41	20	197
	7.107	*47.885*	*8.644*	*8.278*	*34.608*	*10.463*
2001	64	24	21	48	14	171
	11.083	*33.326*	*12.335*	*17.631*	*20.845*	*14.753*
2002	77	28	31	44	27	207
	27.626	*57.773*	*21.520*	*15.151*	*107.989*	*26.314*
2003	75	34	22	58	17	206
	27.753	*46.749*	*14.956*	*23.423*	*32.934*	*26.099*
2004	70	31	27	45	16	189
	26.985	*41.075*	*24.031*	*13.775*	*22.679*	*22.381*
2005	91	34	21	49	8	203
	26.832	*27.761*	*19.239*	*15.159*	*4.422*	*18.885*
2006	90	25	22	57	28	222
	19.289	*21.394*	*16.384*	*17.806*	*60.390*	*20.494*
2007	110	32	35	35	41	253
	24.435	*24.351*	*30.480*	*19.770*	*66.554*	*27.030*
Totals	744	303	295	460	204	2006
	19.211	*29.268*	*15.998*	*15.511*	*29.864*	*19.287*

The raw number of occurrences for *gone missing* in the *Daily Mail* nearly doubled from 1998 (56) to 2007 (110); however, the rate per ten million words actually reduced because the *Daily Mail*'s estimated word count for 2007 is 45 million as opposed to 18 million for 1998. The largest change in terms of rate of occurrence is for the *Northern Echo* which has 16 results or 23.719 per ten million words for 1998 and 41 results or 66.554 per ten million words in 2007.

In the results for the 2007 *Daily Mail*, in terms of the McCann event, 3 articles refer to *Madeleine*; 3 refer to *Maddie*; one refers to *Maddy*; 2 refer to *McCann*. These references occur in a total of 7 articles. As mentioned previously, many references use the child's first name only in reference to her, such as this line from the *Daily Mail* in September 2007: "Wales? Imagine if Maddie had gone missing from a seaside boarding" (*LexisNexis Academic*). Still this line from the Independent in September 2007 refers to the child by first and last name: "reports that Madeleine McCann had gone missing from the holiday apartment where she and her family were" (*LexisNexis Academic*). This kind of variation in reference to the child is why this study elected to search all of the British *go missing* strip corpora by multiple relevant keywords in an attempt to illuminate to what degree the McCann event affects the results of this chapter's *go missing* searches in the British BVC.

Table 4.12: Results for *went missing* in the British BVC

Year	*Daily Mail*	*Eve. Chron.*	*Herald*	*Independent*	*N. Echo*	Totals
1998	128 *69.722*	159 *79.152*	87 *35.132*	71 *29.431*	56 *83.018*	501 *53.253*
1999	123 *53.132*	122 *11.591*	87 *36.823*	81 *29.704*	53 *98.647*	466 *51.577*
2000	136 *16.666*	99 *115.625*	89 *20.794*	141 *28.469*	44 *76.138*	509 *27.035*
2001	165 *28.575*	58 *80.538*	104 *61.091*	150 *55.098*	33 *49.136*	510 *44.000*
2002	168 *60.275*	73 *150.624*	110 *76.362*	121 *41.666*	50 *199.980*	522 *66.358*
2003	200 *74.009*	83 *114.122*	114 *77.499*	142 *57.346*	73 *141.426*	612 *77.537*
2004	165 *63.608*	81 *107.326*	77 *68.535*	100 *30.611*	29 *41.107*	452 *53.526*
2005	178 *52.484*	105 *85.732*	67 *61.383*	65 *20.109*	30 *16.584*	445 *41.399*
2006	247 *52.938*	79 *67.606*	76 *56.602*	90 *28.200*	34 *73.331*	526 *48.559*
2007	425 *94.409*	124 *94.363*	153 *133.243*	126 *71.174*	93 *148.696*	921 *98.400*
Totals	1935 *49.964*	983 *94.954*	964 *52.280*	1087 *36.654*	495 *72.464*	5464 *52.535*

The frequency of *went missing* in the British BVC allows both the analysis of *went missing* in connection with the McCann event as well as other, less famous events. For example, one event is the disappearance of a little boy, Wesley who vanished around June 1998 (*LexisNexis Academic*). The *Evening Chronicle* mentioned the Wesley event in 8 articles in 1998; in 6 articles in1999; in 3 articles in 2000; and in 2 articles in both 2002 and 2003. Only one other title mentioned the Wesley event; the *Northern Echo* referred to the event two times in 1998. This event is not significant as a single factor that could affect the overall results of *went missing* in the British BVC; however, this event testifies to the BVC's ability to track lesser known events, and, in doing so, present variety in terms of British speech communities.

Table 4.13 refers to results from the British *went missing* strip corpus for 2007. These results are article counts; frequency counts would likely be higher as, for example, a single name could be mentioned in a single article multiple times. Also, the article counts in Table 4.13 do not reflect any control for overlap, so one (or more) articles could possibly be counted for a title in all four name categories.

Table 4.13: McCann event keywords

Title	*Madeleine*	*McCann*	*Maddie*	*Maddy*
Daily Mail	54	34	25	8
Evening Chronicle	11	5	1	8
Herald	27	14	0	0
Independent	29	12	0	0
Northern Echo	18	0	0	0

The *went missing* results for 2007 that relate to the McCann event show not only evidence for the McCann event's impact on the total number of results in the British BVC, but they also stand as another example of the variety within the British BVC. In the *went missing* strip corpus (for 2007), the *Northern Echo* refers to *Madeleine* in 18 articles; however, that title does not refer to *Maddie*, *Maddy*, or *McCann*. In contrast, the *Daily Mail* refers to *Madeleine* (in 54 articles); *Maddie* (in 25 articles); *Maddy* (in 8 articles); and *McCann* (in 34 articles). As an example of the overlap that may exist in the article counts in Table 4.13, in one *Daily Mail* article from December of 2007, *Madeleine*, *Maddie*, and *McCann* all occur in the article titled, "The McCann detectives." (*LexisNexis Academic*). The article's text includes, "days after Madeleine went missing on May 3.Then;" and "year-old Maddie went missing she and Robert were setting up" (*LexisNexis Academic*).

Also, the *Herald* and the *Independent* both refer to *Madeleine* and *McCann*, but neither refers to *Maddie* or *Maddy*. The variety that these 5 titles show in terms of name selection in reference to the McCann child is another testimony to the BVC's ability to capture variation microscopically within the larger British speech community.

Only two of the five titles in the British BVC use the form *Maddy* in the *went missing* strip corpus in 2007. These titles are the *Evening Chronicle* (8 articles) and the *Daily Mail* (8 articles). The articles that use *Maddy* occur from 5 May to 10 October in the *Evening Chronicle* and from 6 May to 28 July in the *Daily Mail*. Perhaps, as this study previously hypothesized in relation to a television column in an American newspaper, authorial preference could have influenced the use of the less frequent spelling of the child's first name.

The frequency of all forms of *go missing* increased over the ten year span of the British BVC. The frequency of all *go missing* forms stayed fairly consistent in the British BVC between 1998 and 2006; however, in 2007 the frequency of *go missing* forms increased sharply in the British BVC. Because of this increase, this study pursued possible skewing factors—especially single events. The raw number of results for *went missing* in the British BVC for 2007 is 863. Within those 863 results, the study searched for the term *mccann*. That new search yielded 208 results. The McCann event shows that as the BVC reflects culture, life, and society, single events can become significant skewing factors. Subsequently, the study specifically pursued the McCann event as an influence of the *go missing* search results from the British BVC.

The rate of all *go missing* forms in the British BVC is lower for the year 2000. Even though the raw number of results for 2000 is higher than that of the previous two years, the estimated word count of the subcorpus for 2000 is much higher. The randomly selected sampling date of 4 November 2000 is a Saturday; this particular Saturday yields an estimated annual word count that is twice as large as the estimated annual word count for 1999. In terms of math, because the total number of results for 2000 (885) is similar to that of 1999 and 2001 (825 and 837), the frequency per ten million words is much lower because the estimated word count for 2000 is much higher.

Discourse Prosody

Stubbs notes that the discourse prosody of a text is a reflection of the attitude of the speaker of the text or a reflection of the function of the text (88). In the case of *go missing*, some analysis of the form's discourse prosody is useful in the comparison of the form's use in the American and British BVCs. That is, evidence from the BVCs, should show that *go missing* forms occur in similar or contrastive (such as positive, negative, neutral) discourse situations in the American and British BVCs.

For the calculation of the presence of instances of *go*, *goes*, *going*, *gone*, and *went missing* in the American and British BVCs (presented earlier in this chapter), the study used strip corpora; here is an example of one strip of text from the American *went missing* strip corpus:

... college student who went missing June 23.More than ...

Those strips combine to create a focused corpus that allows the node (a form of *go missing*) and its nearby collocates of a few words to the left and right to be closely observed. These collocating forms will inform the study as to the discourse prosody of the form in the American and British BVCs.

The *went missing* strip corpus derived from the American BVC consists of 130,266 words; the *went missing* corpus derived from the British BVC consists of 226,370 words. The study selected *went missing* as the focal point for analyzing discourse prosody because that form is the most frequent form of *go missing* in both the American and British BVCs.

In order to view the collocates of *went missing* that are content words, instead of function words, the study created a simple stop list that consists of the 100 most frequent forms from the Brown Corpus. The use of this stop list in connection with the generation of collocates of

went missing from the American and British BVCs caused collocates, such as *the* to be removed, so lower frequency content words, such as *action* could be observed in relation to *went missing*. Table 4.14 shows the collocates of *went missing* in the American BVC that result after the stop list mentioned above is in place; the correlating number is the number of raw results for each form in its respective position. The form *went missing* is the node or central form for Table 4.14 and Table 4.15; L1 indicates the position that is one word to the left—L2 and L3 continue similarly to the left. R1 indicates the position that is one word to the right—R2 and R3 continue similarly to the right.

Table 4.14: Top ten collocates of *went missing* by location in the American BVC

L3	L2	L1	R1	R2	R3
old 70	*since* 50	*plane* 30	*during* 104	*action* 49	*ago* 52
days 50	*day* 47	*wife* 29	*last* 67	*july* 22	*days* 31
since 40	*old* 33	*daughter* 26	*while* 54	*week* 22	*months* 15
year 39	*woman* 27	*cat* 19	*saturday* 23	*night* 20	*while* 15
weeks 26	*girl* 26	*0* 18	*tuesday* 22	*month* 19	*home* 14
day 22	*night* 24	*son* 17	*april* 19	*afternoon* 18	*found* 13
months 22	*boy* 20	*woman* 16	*friday* 14	*year* 18	*vietnam* 13
hours 19	*died* 18	*girl* 14	*monday* 14	*several* 17	*during* 12
containing 15	*0* 17	*sjodin* 14	*nov* 13	*30* 16	*hours* 11
night 15	*soldiers* 17	*money* 13	*shortly* 13	*june* 16	*p* 11

In Table 4.14 the forms in the L1 position, tend to represent the item or individual that vanished; one of the forms is an individual's name, *sjodin,* and the name *peterson* has the same frequency as *money* but is not included in Table 4.14 because only ten forms per collocate position are included, and *money* preceded *peterson* alphabetically. Names themselves are neutral, but they can cooperate with other collocates (or reflect certain cultural events) to create a negative prosody. The highest frequency L1 collocate of *went missing* in the American BVC is *plane*. The sense of *plane* in the case of this collocate is 'aircraft.' All of the collocations of *plane* in the L1 position report or describe an accident or disaster that involves an aircraft. The other nouns in the L1 position are items or people who have disappeared; the one exception is *0*.

This zero actually represents *000* as in the case of a numerical description of thousands of items that disappeared.

In the R1 position, all of the ten most frequent forms indicate some sort of temporal expression; these forms in themselves are neutral in terms of prosody, but they can cooperate to construct a negative prosody, such as in this case: “man who went missing last fall.” Table 4.15 shows the collocates of *went missing* from the British BVC; these results reflect the application of the same stop list that was used for Table 4.14.

Table 4.15: Top ten collocates of *went missing* by location in the British BVC

L3	L2	L1	R1	R2	R3
old 124	*since* 116	*madeleine* 122	*while* 125	*days* 75	*home* 197
days 102	*day* 103	*girls* 63	*during* 122	*home* 58	*ago* 190
year 77	*old* 68	*old* 44	*last* 110	*weeks* 41	*days* 54
day 64	*night* 58	*daughter* 41	*three* 22	*July* 40	*night* 44
hours 60	*year* 53	*0* 30	*near* 21	*week* 37	*last* 42
months 50	*woman* 44	*sarah* 25	*just* 19	*august* 34	*family* 30
weeks 43	*mr* 41	*money* 23	*shortly* 19	*september* 32	*hours* 29
night 34	*girl* 38	*milly* 22	*eight* 18	*sunday* 32	*found* 27
since 33	*boy* 29	*jessica* 21	*between* 17	*leaving* 29	*while* 24
british 28	*miss* 23	*girl* 20	*six* 16	*action* 28	*months* 21

Table 4.15 shows that the most frequent form in the L1 position is *madeleine* (e.g. Madeleine McCann), and that form's frequency shows how significant the McCann event is to the study of *go missing* in the British BVC. *While* is the most frequent form in the R1 position of *went missing* in the British BVC. The forms in the R1 position tend to be a temporal description, and the forms in the R3 position tend to identify a place or duration of time.

During is the most frequent R1 form for *went missing* in the American *went missing* strip corpus, and Table 4.16 shows the ten most frequent collocations for *went missing during* (*during* plus two words to the right), and Table 4.17 shows the ten most frequent collocations for *went missing while* in the British *went missing* strip corpus (*while* plus two words to the right in the British BVC). In each table the collocations are listed in order of their frequency; the number of occurrences for the collocation and the number of connected events, and the context are also presented.

Table 4.16: Ten most frequent collocations for *went missing during* in the American BVC

Collocation	Occurrences	Number of events	Context
during the Vietnam	8	7	various people who vanished
during a storm	7	2	vanished aircraft (6 events) vanished person (1 event)
during world war	5	4	vanished people (3 events) vanished aircraft (1 event)
during the war	4	4	soldiers who vanished solitarily or in groups
during the fighting	3	1	people who vanished during a war
during the junta	3	1	people who vanished
during her tenure	2	1	Political legislation
during a snowstorm	2	1	people who vanished
during racial unrest	2	1	a statue that vanished
during one of	2	1	a prisoner of war

Table 4.17: Ten most frequent collocations for *went missing while* in the British BVC

Collocation	Occurrences	Number of events	Context
while on a	5	4 (plus 1 unclear)	vanished person or people
while he was	6	4 (plus 1 unclear)	person of interest, vanished child, vanished person, vanished clothing
while walking home	6	2 (plus 1 unclear)	vanished child
while they were	5	5	a possession or a pet
while on their	4	2	vanished children (1 event) vanished people (1 event)
while on her	4	2 (plus 1 unclear)	vanished child (2 events)
while playing on	4	2 (plus 1 unclear)	vanished person
while exploring a	4	Unclear	vanished diver or divers
while she was	3	3	vanished person or people
while being walked	3	3	vanished dog or dogs

The collocations in Table 4.16 and Table 4.17 have been carefully examined, so duplicated occurrences in a single text were removed. In the case of two (or more) articles that contain the same collocation and were identical (such as a wire article or another article that concerned the same topic), the collocations were counted as long as the articles were from different titles or different dates within the same title. The typical case for such an included repetition would be a wire article that manifested itself in newspapers across the country.

In the case of single events that attracted multiple collocations, the British BVC tended to present collocations from multiple, different articles that refer to the same event; however, the American BVC has a greater tendency to present the collocations in identical wire articles across regions.

The collocates of *went missing during* in the American BVC tend to describe situations and conflicts that are clearly negative and point to the national and world scene of events. The discourse prosody constructed by these collocations is consistently negative. The collocates of *went missing while* in the British BVC also construct a negative discourse prosody; however, the significant contrast between the collocations in Table 4.16 and Table 4.17 rests in the association of the British collocations with daily life pursuits, such as a lost pet. Note for example that in Table 4.17, the most frequent R3 form of *went missing* in the British strip corpus is *home*. In Table 4.16 the most frequent R3 form for the American BVC is *ago*, and *home* occurs as the fifth most frequent form in that position. Perhaps, the newness of *went missing* to American English causes its use to focus away from daily life pursuits; similarly, evidence from the British BVC that shows *went missing*'s use in association with family and home could be the result of the form's established presence and use in British English.

Evidence from Additional Sources

The question of what career could *go missing* have had in American English between those dates is worthy of consideration. In pursuit of that question the study selected three famous historical events, each related to the disappearance of a person, to be examined both in historical texts and contemporary texts in the pursuit of forms of *go missing* in connection with these events.

The child, Charles Augustus Lindberg, Jr., frequently referred to in the media as "the Lindberg baby" was kidnapped from his residence on 1 March 1932 (*ProQuest Historical Newspapers*). The study conducted a search in the following historical newspaper titles: *Atlanta Constitution, Christian Science Monitor, The New York Times, The Wall Street Journal,* and

The Washington Post for any *go missing* forms between 1 March 1932 and 31 December 1932. The study found no uses of any *go missing* forms. A search in the above titles for "Lindbergh baby" produced 800 results, and Table 4.18 includes a sample of some phrases from articles published on 2 March 1932. These phrases contextualize the stealing or vanishing of the child.

Table 4.18: Historical newspaper phrases related to the disappearance of the Lindbergh baby

Relevant verb phrase in italics
Child *stolen*
Boy, 20 months old, *gone*, in night robe
Was kidnapped between 8:30 and 10 o'clock
The baby *had disappeared*
The baby actually *was gone*
The crib *was empty*
The baby *had been kidnapped*
Lindbergh baby, *kidnapped* at night
Taken from his crib
Most famous baby in U.S. *stolen*
Famous baby who *was kidnapped*
Lindbergh's baby *kidnapped* in Jersey
Child *is spirited away* from home
That her grandson *had been kidnapped*

Note that while no form of *go missing* appears in the lines in Table 4.18, the semantic field that the phrases point to are very similar to many in which this chapter has shown recent uses of *go missing*. Also, note that two of the verb phrases in this historical selection build on *gone*—but they are different grammatical constructions from *go missing* constructions as these forms (*gone*, and *was gone*) build directly on the verb *go* without the inclusion of a participle form.

The form *Lindbergh* does not appear in the American BVC *go missing* strip corpora, but the study also sought texts on the Web via Google that might connect the Lindbergh event with *went missing*. The search string "Lindbergh baby" "went missing" produced 16 results with these advanced search features selected: Language, English; Country, United States; Date, past year.[1] Of these results, only 2 directly connected the Lindbergh event with *went missing*. Still, these small numbers present evidence for a change within the American English community in terms of use of *go missing*.

Amelia Earhart's disappearance was another famous event that attracted much U.S. media attention. Amelia Earhart disappeared on 2 July 1937 (*ProQuest Historical Newspapers*). The five historical newspapers listed above were searched in the frame of 2 July 1937 to 31 December 1937, and no form of *go missing* was found. Table 4.19 includes a selection of phrases taken from newspaper articles published on 3 July and 4 July 1937.

[1] This Google search was conducted on 5 April 2010.

Table 4.19: Historical newspaper phrases related to the disappearance of Amelia Earhart

Relevant verb phrase in italics
Wide search for *missing* plane is ordered
Fliers are believed *to have been forced down*
Miss Earhart *forced down* at sea
If she *were forced down*
Amelia Earhart feared *forced down*
When she *was forced down*
Fliers *are feared down* in Pacific
Amelia *lost* in mid-Pacific
Landing field where fliers *are long overdue*
Earhart was believed *to have fallen* into the Pacific shortly after 4 p.m.
Lost aviatrix
Aviatrix *is missing*
Aviatrix *was still missing* late yesterday

The phrases in Table 4.19 taken from historical news texts, again, do not use *go missing* constructions, but three of these phrases employ *missing*. Still, all of theses uses of *missing* (*missing* plane, is *missing*, was still *missing*) differ grammatically and semantically from *go missing* constructions.

The American BVC *go missing* strip corpora contain one strip of text that relates to Amelia Earhart. Initially the study could not determine if this strip authentically connected

Miss Earhart and *gone missing*: "say, Amelia Earhart -- has "gone missing." Guillaume, who teaches" (*LexisNexis Academic*). The full text explains the context: "But it seems Guillaume is a native of Britain, where it's fairly common to talk about how someone -- say, Amelia Earhart -- has "gone missing." But it seems Guillaume is a native of Britain, where it's fairly common to talk about how someone -- say, Amelia Earhart -- has "gone missing" (*LexisNexis Academic*). The study regards this use as meta-discussion; still this one article testifies to an American author's noticing the apparently unusual form and in discussion inventing an illustration that features a reference to Miss Earhart.

The study pursued results from the Web with this string "Amelia Earhart went missing" and the same advanced search selections described above for the Lindbergh event search. The search produced 18 results, and all of the results connect with the Earhart event.[2] Of the 18 results, one is a meta-discussion of the use of *went missing* that employs *Amelia Earhart* in an invented illustration that features *went missing*.

Later in the twentieth century, on 5 February 1974 Patricia Hearst was kidnapped, and the event was a major media focal point (*LexisNexis Academic*). Four of the historical titles mentioned above (all but *The Atlanta Constitution*) covered the span of 5 February 1974 to 31 December 1974, and that span was searched for *go missing* forms, but none were found. No reference to Ms. Hearst could be found in the BVC *go missing* strip corpora.

[2] This Google search was conducted on 5 April 2010.

Next, because of the availability of full-text downloadable articles from the *New York Times* (available via *LexisNexis Academic*), the study searched the 5 February 1974 to 31 December 1974 span for *go missing* forms and found none. The study also searched the same time span within the *New York Times* with the following string:

(Patricia pre/1 Hearst) or (Patty pre/1 Hearst)

and 154 search results were produced. Those 154 articles were downloaded in full-text format from *LexisNexis Academic*.

A concordance of these 154 articles with WordSmith tools, with *Hearst* as the node form, produced the following (unfiltered) results for the R1 position: *and* (27), *was* (9), *corp* (9), *in* (9), *is* (9). The form *corp* refers to the Hearst Corporation. The study made another concordance with *Hearst was* as the node form, and Table 4.20 shows the 9 results—in all of which *Hearst* refers to Patricia Hearst. The table includes *Hearst was* plus the verb phrases that build to the right.

Table 4.20: *Hearst was* collocates from *The New York Times*

Date of article	*Hearst was* plus verb phrase
5 April	Hearst was forced to record
6 April	Hearst was coerced into making
18 April	Hearst was not a reluctant participant
18 April	Hearst was willing participant
24 April	Hearst was kidnapped
18 May	Hearst was in house
20 May	Hearst was introduced to him
28 May	Hearst was turned into terrorist
25 October	Hearst was said to be living

Table 4.20 Shows no evidence for any construction similar, gramtically or semantically, to a *go missing* form. Perhaps the past participle *kidnapped* is the closest semantically to *went missing*, for example, but *kidnapped*'s meaning rests on the action of one party in violation of the will of another party. That concept of acting is quite different from *went missing* which carries with it a sense of 'vanishing.'

In addition to the downloadable full-text *New York Times* articles, the study pursued the Web via Google for possible evidence of recent use of *went missing* in connection with the Patricia Hearst event. The study used Google advanced search with the selections described above and obtained zero results with these two search strings: "Patty Hearst went missing" and "Patricia Hearst went missing." Next the study obtained 11 results with this search string:

"Patricia Hearst" "went missing."[3] Only one of these 11 results connects *went missing* with the Hearst event. The results generally overviewed a variety of crimes or events, and the Hearst event appears as just one within discussions of several, usually famous, events. The studied also pursued the string "Patty Hearst" "went missing" and the only result that connected the Hearst event with *went missing* is the same result from the "Patricia Hearst" "went missing" search explained above.

The investigation of the Lindbergh, Earhart, and Hearst events in terms of *go missing* is useful because of the absence of *go missing* forms in the available historical texts published in proximity to the respective events. Further, the Google searches show little evidence for use of went missing in connection to the Lindbergh and Hearst events; however, the 18 search results that connect the Earhart event with *went missing* persuasively testify, in concert with this chapter's search results from the American BVC, that American English has begun to use *went missing* with some degree of frequency.

This chapter has pursued a study of a low frequency form; however, the large size of the BVC has presented a critical, useful linguistic mass for analysis. With other corpus approaches, a researcher would likely face less evidence for the use and collocational tendencies of *went missing*. For example, in Davies's COCA *went missing* occurs 272 times.[4] The most frequent R1 collocate of *went missing* in that corpus (after a stop list of the 100 most frequent words from the Brown Corpus is applied) is *last*. That form appears in the R1 position only 3 times, which is not a large enough sample to observe collocating habits of the R1 position in this case. Even though COCA has 272 results for *went missing*, if the most frequent form in R1 position only has 3 occurrences, a researcher will have little linguistic mass to analyze. In the American BVC the

[3] The Google search for "Patricia Hearst" "went missing" was conducted on 5 April 2010, and the Google search for "Patty Hearst" "went missing" was conducted on 11 April 2010.
[4] This search was conducted on 12 February 2010.

most frequent R1 form of *went missing* is *during*. The form *during* has 104 results in that position in the American BVC, but that form only generated 4 separate collocational constructions (Table 4.15) that are connected to multiple events—rather than single events.

In the BNC, which this study accessed via Davies's online, public interface, *went missing* occurs 120 times.[5] After the stop list mentioned above is applied to search results from it for *went missing*, only two words remain in the R1 position of *went missing* (*four* and *last*), and they each occur only once. Again, in order to determine the collocating habits of the R1 position (as well as the node form), a larger corpus, such as the BVC, is needed.

Analysis of *go missing* through the BVCs has brought two complications to the surface. First, single events can distort search results, and the BVC may include erroneously repeated search results. These complications are a result of the BVCs massive nature and virtual construction. More importantly, these complications can be addressed.

Also, because the *go missing* strip corpora downloads from *LexisNexis Academic* are labeled with the title and date of their occurrence, accidental duplication of search results can be determined visually. In the case of the analysis of the discourse prosody of *went missing*, erroneously repeated text strips from the *LexisNexis Academic* database can be easily identified because the overall number of American and British collocations is so small.

Conclusion

This chapter presents evidence not only for *go missing*'s presence in American English but also evidence for its career. That is, the American BVC shows that over the span of ten years, the form has moved from little use to a position of currency in American English. Note,

[5] This search was conducted on 13 February 2010.

for example, the all-region 1998 total for *went missing* is .770 per ten million words, but the 2007 all-region total is 14.675 per ten million words.

The methodology of this chapter has allowed the form's use to be compared between American English and British English; such a capability allows the American BVC's 2007 all-region total for *went missing* of 14.675 per ten million words to be compared with the 2007 all-title results for *went missing* from the British BVC of 98.400 per ten million words. Just as *went missing* can be compared in 1998 and 2007 in terms of the American BVC, results from the American BVC and British BVC can be compared, so the story that the BVCs tell about *go missing* crosses both years and cultures.

Another interesting occurrence in the British BVC is the zero result for *goes missing* in the *Evening Chronicle* for the year 2003. Also, for the same form, the *Northern Echo* has only one result for 2000. The BVC allows microscopic glimpses into the form's presence in the BVC, and these observations might not be surmised from a glance at either the ten-year totals for the particular title or the annual total for all the titles. The point of analysis of the BVC can shift from microscopic (such as results in one article) to macroscopic (such as results for one year of text for one newspaper title). Each point of analysis tells an important story about the form's career; further, because BVC results from the less microscopic views do not always show important variation events, such as a zero result, the microscopic and macroscopic views work together in the quantitative description of the form's career.

The American BVC has shown the gradual increase in use of *go missing* forms in American English. The study begins with texts that show use of *go missing* forms in connection with American maritime disasters in the late nineteenth and early twentieth centuries; later texts

between 1998 and 2007 (the BVC) show the rise of *go missing* in American English. Many years intervene between the anecdotal maritime examples and the earliest scope of the BVC.

Finally, this chapter presents more than just a quantitative observation of the young career of *go missing* in American English or evidence of the well established career of *go missing* in British English. This chapter shows that American English, which we so often think of as breaking from British English in the seventeenth or eighteenth century never to return again, has reconnected with British English in the late twentieth and early twenty-first centuries.

Chapter 5

A Case Study of Neology in American English: *carb*

The popular low carb diet, sometimes referred to as the *Atkins diet*, influenced the dietary habits of many Americans. Many media forms noted the momentum of the diet by building on the form *carb* in some way. The belief that the so-called *low-carb revolution* has ended might not spark controversy, but research through the BVC can show evidence for *carb* as an extremely low frequency form that was catapulted to greater frequency and later returned to much lower frequency. One author notes, "When low-carb mania peaked in February 2004, about 27 million Americans (or 9.1 percent) were following Atkins" (*LexisNexis Academic*). So, the year 2004 could well be the high water mark for this particular diet (both as a corporate commodity and a cultural practice) in the United States. The BVC allows a structured analysis of the use of forms related to the low-carb diet.[1]

Pilot Study

In order to determine high frequency forms (within the population of *low carb* related forms), this study constructed a pilot corpus based on the American BVC. The pilot corpus was constructed by searching the American BVC with this search string:

---

[1] In January, 2010, access to *The Boston Globe* was not available to the study through *LexisNexis Academic*. In the absence of that title, the study added *The Boston Herald* to the BVC's Northeast region. *The Boston Globe* had a circulation of 323,000 and a ten-year estimated word count of 312 million. *The Boston Herald* has a circulation of 138,000 and a ten-year estimated word count of 137 million. It was noted in the Methodology that circulation information was originally unavailable for the *Herald*. In January, 2010, circulation information was available for the *Herald* through the SRDS database. With the addition of the *Herald*, the estimated word count for the BVC is 5.1 billion words; formerly, the estimated word count for the BVC was 5.3 billion words.

carb! w/25

high or low or diet or atkins or (south pre/1 beach) and not carbohydrate

and not carbon! and not carbide and not carburetor

and not carberry and not carballo and not carbajal and not carbine

Next, the study downloaded the full text of all the articles that resulted from this search. The pilot corpus created by these full-text articles consists of 2.6 million words. Within the pilot corpus, these results were not arranged by title or region; rather, all of the articles were placed in a single directory without regard to title, geography, or year of publication because the purpose of this pilot corpus was to identify high frequency forms related to *carb* and their collocates.

WordSmith Tools was used to generate a word list from the pilot corpus. These results are raw; they have not been searched or analyzed in terms of repeated articles, single events, or any other skewing factor. Still, the raw frequencies of the forms in the list tell an important story about high frequency forms in the pilot corpus. The search term for the word list is carb*, and the following are the most frequent forms that build on the root *carb* and conceptually relate to the *low carb* diet: *carb* (3,366 occurrences), *carbs* (939 occurrences), and *carbo* (233 occurrences). All other forms in the corpus, which are morphologically related to *carb*, and could be determined to have a direct connection to the *low carb* movement, had fewer than 10 results.

In the search for collocating forms for *carb*, the study used WordSmith Tools to locate the most frequent forms that occupy the R1 position of the node *carb*. The study employed the 100 most frequent forms from the Brown Corpus as a filter to remove ultra-high frequency forms. Table 5.1 includes the ten most frequent forms that occupy the R1 position of *carb* in the pilot corpus.

Table 5.1: Most frequent R1 forms for *carb*

R1 forms and number of occurrences
diet 375
diets 244
craze 142
high 75
low 62
dieters 44
foods 43
kitchen 40
beer 39
atkins 36

The most frequent L1 form for *carb* was *low* which had 2742 occurrences; the next most frequent form was *high* with 123 occurrences. The study concluded from these preliminary investigations that *low carb diet* was the most frequent collocation (L1 to R1) of the node *carb* in the pilot corpus.

The pilot study shows that *low carb diet* is the most frequent (L1 to R1) *carb* collocation and should be the central focus of the study's analysis of the recent career of *carb* in the American BVC. *Low carb diet* is not the only form the study will pursue for analysis; however, the study will use *low carb diet* (actually the search string low pre/1 carb pre/1 diet) as the search term for the collection of full-text articles from *LexisNexis Academic* database to further analyze *carb*'s career. This search string should provide articles that include both *low carb diet* and

low carb diets. An article could include the singular and plural forms, but in the event that an article contains only the plural form, that article should be included in the results because as *LexisNexis Academic* explains, "Using the singular word form will retrieve the singular, plural, and possessive forms of most words. For example, city would find city, cities, city's, and cities'" ("Developing a Search"). Therefore, the search string of low pre/1 carb pre/1 diet should provide the articles that include the two highest *low carb diet* forms from the pilot study—*low carb diet* and *low carb diets* which by extension should provide the BVC's articles that are the most relevant to the low carb diet movement.

Low Carb Diet Corpora

The study searched the American and British BVCs with the search string low pre/1 carb pre/1 diet and downloaded the full text of the articles that resulted from this search. The study organized the results in directories by year in anticipation of the need to show the trajectory of the form's use across the ten year coverage of the American BVC.

It is useful to note that the articles in these search results contain the form *low carb diet*, but because the form *low carb diet* is a conceptualizing factor in the development of the articles, other *carb* forms may be present in the articles as well. Also, factors other than dietary practices could influence the presence and development of the articles.

The American low carb diet corpus

The American low carb diet corpus consists of 1089 articles and a total of 1 million words. The British low carb diet corpus consists of 68 articles and 69,000 words. The pilot corpus for this chapter that was created from the American BVC contains 2.6 million words; the

American low carb corpus is less that half the size of that corpus; however, the focus of the articles is more deliberately directed toward discourse that is related to the low carb diet. Also worthy of note is that both the American and British low carb corpora have zero results for 1998, and because 2004 is regarded by some as the high water mark of the low carb diet, the BVCs should be able to provide evidence for both the rise and fall of the form *carb*.

The study next endeavored to discover extremely fine grained variations of the form *low carb diet* in the American low carb diet corpus. That is, just as this study held curiosity about the possible presence of a variety of *carb* forms, (i.e. *carb, carbs, carbo*) in the pilot corpus, the study investigated how many variants of *low carb diet* exist in the low carb diet corpus. The study achieved this determination through the use of the search string low carb diet* on the low carb diet corpus with WordSmith Tools. Table 5.2 shows the variations of the form *low carb diet* in the American low carb diet corpus.

Table 5.2: *Low carb diet* and variants in the American low carb diet corpus

low-carb diet	864
low-carb diets	818
low-carb dieters	102
low-carb dieting	33
low carb diet	13
low carb diets	11
low-carb dieter	7

The search term low carb diet* was used in WordSmith Tools to generate all results that would include both *low carb diet* and other forms that build on the form. The highest frequency variant of *low carb diet* is *low carb diets*, so the singular and the plural forms account for 1706 or 92 percent of the 1848 total variant forms. The 1706 figure includes 24 non-hyphenated results: 13 singular and 11 plural.

Table 5.2 shows one interesting aspect of offline searches with WordSmith Tools. The search string used on the BVCs produced both hyphenated (the majority of the results) and unhyphenated results. WordSmith Tools discriminates the hyphen (which is regarded as a space to searches within the *LexisNexis Academic* database), so a search for only the hyphenated forms, for example, is possible with WordSmith Tools. This particular matter of hyphen non-recognition is not regarded by the study as a weakness of the *LexisNexis Academic* interface; rather, the fact that the database provides robust downloading processes that can be connected to another interface (which in this case can discern hyphens) makes the *LexisNexis Academic* database highly flexible.

The study next sought to determine the collocational habits for *low-carb diet* (note hyphen) in the low carb diet corpus. As stated above, the study again used the previously mentioned stop-list to filter out ultra-high frequency forms and cause content forms to rise to the top of the collocate ranks. Table 5.3 includes the L3 to R3 collocates for *low-carb diet* in the American low carb diet corpus.

Table 5.3: L3 to R3 collocates for *low-carb diet* in the American low carb diet corpus

L3	L2	L1	R1	R2	R3
you're 13	*high* 44	*rigorous* 36	*craze* 95	*snacks* 36	*included* 36
part 11	*tried* 28	*protein* 28	*fad* 14	*2004* 7	*atkins* 9
those 10	*following* 20	*fat* 17	*phenomenon* 12	*might* 7	*peaked* 9
adults 9	*atkins* 9	*atkins* 12	*plans* 12	*says* 7	*affect* 6
atkins 9	*eat* 9	*controversial* 7	*trend* 11	*posthumously* 6	*weight* 6
ever 9	*defended* 8	*edition* 7	*according* 6	*given* 5	*followed* 5
popular 9	*follow* 6	*popular* 7	*because* 6	*william* 5	*tuesday* 5
pounds 9	*declaring* 5	*strict* 7	*byline* 6	*exercise* 4	*yancy* 5
going 8	*ate* 4	*whose* 5	*doctors 6*	*because* 3	*february* 4
people 8	*blaming* 4	*current* 4	*flared* 6	*byline* 3	*9* 3

Table 5.3 shows several important elements of the story of *carb* in American English. First, the last name of Dr. Robert Atkins, one of the most famous promoters of the diet, appears as one of the ten most frequent collocates of *low-carb diet* in four (L3, L2, L1, R1) of the six collocate positions. Another surname within the group of collocates is *yancy* which appears as the number eight collocate in the R3 position. Yancy is a physician who conducted a six month study of a group of individuals who were practicing low carb dieting principles (*LexisNexis Academic*).

In the L2 position, *high* has 44 occurrences. In a separate concordance of the American low carb diet corpus, the most frequent collocate in the R1 position of *high* is *protein* with 327 occurrences. The next most frequent form in that position is *fat* with 150 occurrences, and the

next is *carb* that has 97 occurrences. *Craze* is the most frequent form in the R1 position of *low carb diet* and apparently reflects authorial motivation to place the low carb diet in the category of a passing fad.

Craze stands as an ominous predictor of the diet's rise and fall. In the low carb diet corpus, *craze* appears only twice from 1998 to 2002; the form appears 27 times in 2003, and in 2004 the form appears 241 times. In the years from 2005 to 2007 the form sharply reduces in frequency for a total of 56 occurrences in the three-year period. The BVC is a witness to how the career of *craze* in the American low carb diet corpus mirrors the explosive career of *carb* in American English. The American BVC provides an overview the rise and fall of *carb* in American English through annual article counts in the low carb corpus as shown in Table 5.4.

Table 5.4: American low carb diet corpus article count by year

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Articles	0	24	30	13	40	158	597	142	53	32

Table 5.4 shows the zero point from which the low carb corpus begins in 1998. Later, from 1999 to 2002 the article count increases slowly from 24 to 40. As mentioned earlier, 2004 has been noted in some news media as the peak of the low carb diet (*LexisNexis Academic*). Still, the 158 results for 2003 represent a large increase. One contributing factor for the increase in 2003 is the death of Dr. Robert Atkins on 17 April 2003 (*LexisNexis Academic*). In the American low carb diet corpus, the form *died* occurs 14 times in reference to the passing of Dr. Atkins. In terms of Dr. Atkins's posture within the low carb movement, one newspaper referred

to Dr. Atkins (after his death) as "the father of low-carb eating" (*LexisNexis Academic*). Other articles in 2003 related to low carb matters may have appeared because his death may have made low carb dieting a newsworthy topic.

2004 represents the apex of coverage of the low carb movement in the low carb corpus with 597 articles, and in 2005 the count decreases to 142. The waning presence of the low carb movement is keenly reflected by the subsequent decreases in article counts to 53 in 2006 and 32 in 2007. These article counts show that the ten year scope of the BVC is able to tell the story of the beginning, rise, and fall of *low carb diet* through the BVC's reflection of American culture.

Because of the large mass of evidence contained in 2004, the study analyzed the 2004 texts in more specific detail. In order to do so, the study determined the top ten named newspaper sections, by article occurrence, in the low carb corpus for 2004. A section identified by a single alphabet character was regarded as unnamed; also, sections labeled by the abbreviation for page, PG, were regarded as unnamed. The study used WordSmith Tools to determined the article occurrences by section by concordancing the 2004 texts in the American low carb diet corpus with this simple search string: Section: . The R1 form of *section:* will be the article's section name within the article's metadata. Many complexities are involved in such a determination; for example, a section may be called *lifestyle* in one newspaper, and another newspaper may have a similar section labeled *lifestyle and living*. In order to bypass a number of variables that would likely inspire a fruitless, time intensive search, the study simply used the single form in the R1 position of *Section:* as the section label. Here the study uses one attribute of the BVC's flexibility (analysis of article metadata) to analyze the career of *carb* in greater

specificity. Table 5.5 shows article counts by section for each of the ten years of the American low carb diet corpus. The ten sections for 2004 with the highest article counts are the ten sections that are used for all ten years in the table.

Table 5.5: Annual article count by section

Year	Book	Business	Editorial	Financial	Food	Health	Life	Lifestyle	Living	News
1998	0	0	0	0	0	0	0	0	0	0
1999	0	0	0	0	0	4	0	0	0	0
2000	0	0	0	0	5	0	0	0	0	1
2001	0	0	0	0	0	0	0	0	0	0
2002	0	0	0	0	4	5	5	0	0	3
2003	16	15	0	0	25	7	8	0	6	19
2004	19	75	12	31	60	24	21	14	17	61
2005	2	18	0	0	12	2	9	7	4	16
2006	0	6	0	0	0	7	5	0	0	8
2007	0	0	0	0	0	0	0	0	0	7
Totals	37	114	12	31	101	49	48	21	27	115

The single word labels that the study used for the top ten sections for 2004 definitely have some possible overlap; for example, note the similarity between these section labels: Life, Lifestyle, and Living. More importantly, note the sharp contrast between the section labels, *health* and *financial*. Only 2004, the year with the highest number of articles in the American

low carb diet corpus, includes articles from the sections, Editorial and Financial. Only in 2003, 2004, and 2005 have articles in the Business category, but 2004 has 75 of the three-year total of 114 articles.

Financial difficulty of the corporation that markets the Krispy Kreme doughnut was a significant event in 2004 that influenced the presence of *low carb diet* in the BVC. The company complained of poor sales and attributed such losses to the low carb movement (*LexisNexis Academic*). This event increased the form's use in 2004, and the form *krispy* appears in six of the sections in Table 5.5. In order to determine the presence of the Krispy Kreme event as a factor within the sections in 2004, the study searched each section only for the form *krispy* because the longer lexical units that build on *krispy* have frequent variations, such as plural and possessive inflection that were easily avoided by searching exclusively for *krispy*. *Krispy* appears in the following sections (number of occurrences in parentheses): Business (184); Financial (58); News (38); Editorial (4); Lifestyle (2).

Krispy is an excellent example of both the influence of a single event and the BVC's ability to determine what domain of the newspapers are influenced by the event. In this case, the influence of *krispy* is important to the study because of the presence of the form affects multiple newspaper sections; still, 242 of the form's 291 occurrences are in the Business and Financial sections. The title of a 2004 editorial is telling in terms of the doughnut seller's situation and the low carb movement: "Diet fad won't fry Krispy Kreme" (*LexisNexis Academic*). The 184 occurrences of *krispy* make the form the sixth most frequent form in the Business section for 2004.

In order to determine any regional variation, the study used the 2004 American low carb diet corpus directory that had been saved by region to determine occurrences of *low carb diet* by

region. The study searched the 2004 directory by region with WordSmith Tools using the search term, low-carb diet*. Also, in order to obtain the most accurate frequencies possible, the study visually searched each region's corpus for erroneously repeated texts. Two erroneous articles were removed from the Northeast region; no articles were removed from the Southeast region; one article was removed from the Midwest region; three articles were removed from the West region; three articles were removed from the Coastal west region.

Table 5.6 shows the total occurrence of forms by region and the total at the BVC level for 2004; the italicized number represents the rate of occurrence (per ten million words) for the forms.

Table 5.6: *Low-carb diet* occurrences by region for 2004

Region	Northeast	Southeast	Midwest	West	Coastal west	Total
Occurrences	208	298	267	172	157	1102
	1.665	*2.446*	*3.453*	*2.072*	*1.416*	*2.128*

The results for the low-carb diet* search in WordSmith Tools includes four forms: *low-carb diet*, *low-carb diets, low-carb dieters,* and *low-carb dieting*. For each region, *low-carb diet* and *low-carb diets* are the two most frequent forms. Only the Northeast and Southeast have results for *low-carb dieting*; the Northeast has 9 results, and the Southeast has 5 results. Each of the five regions have results for *low-carb dieters*. In each region, the sum of the *low-carb diet* and *low-carb diets* forms account for at least 90% of each region's results.

In terms of skewing factors, the Northeast region contains no recycled articles; the Southeast region contains one article of book announcements that appears five times between October and April in the *Washington Post* and includes *rigorous low-carb diet*; another article

appears 3 times in the *Washington Post* and includes the same *rigorous low-carb diet* strip; the Southeast region also includes 2 identical articles from the *Charleston Gazette* that include commentary on dieting and food choices.

The rates for the Northeast (1.665) and the Coastal west (1.416) regions are well below the average of 2.128. The rate for the Southeast (2.446) is slightly above the average, and the West is slightly below average at 2.072. Finally, the Midwest region has the highest rate of occurrence for low carb diet forms at 3.453 per ten million words.

The British low carb diet corpus

The study built a British low carb diet corpus that was derived from the British BVC to mirror the American low carb diet corpus. The British low carb corpus contains 68 articles and 69,000 words. Table 5.7 shows the annual article count for the corpus from 1998 to 2007.

Table 5.7: British low carb diet corpus article count by year

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Articles	0	0	1	1	5	11	19	20	6	5

The article count for the British low carb corpus clearly shows that the zero period of use for *carb* in the British BVC is longer than that of the American BVC and that even with consideration for the size of the smaller British BVC, the use of the form *carb* appears culturally keyed to American English. The British corpus has zero results for 1998 and 1999 and peaks at a count of 20 articles in 2005 before ending at a count of 5 articles in 2007.

In the British low carb diet corpus, *low-carb diet* occurs 47 times; *low-carb diets* occurs 30 times; and *low carb diet* occurs 6 times. After the stop list filter was applied, the only collocate that remained for *low-carb diet* was *atkins* with a total of 5 occurrences. Therefore, beyond the cultural separation that *low-carb diet* illustrates between British and American English, evidence for the form *low carb diet* in the British corpus does not have enough linguistic mass to justify further analysis.

Additional evidence for the low carb diet movement as a British practice is present in the British low carb diet corpus. The form, *Atkins diet* has 77 occurrences in the British corpus, so *Atkins diet* may be the preferred British form of reference for the practice of the low carb diet. In contrast, the form *Atkins diet* has 525 occurrences in the American low carb diet corpus; *low-carb diet* has 864 occurrences and (*low-carb diets* has 818 occurrences), so based on evidence from the American low carb diet corpus, *low-carb diet* is the preferred form of reference for this dieting practice in American English.

Conclusion

This chapter has shown that *carb* has an identifiable career in recent American English. Evidence from the American BVC identifies a zero point, a peak, and a concluding point with a sharply reduced frequency. The full-text articles in the American low carb diet corpus provide fine-grained details of the career of *carb* in American English. For example, the financial problems of the Krispy Kreme doughnut company increased the presence of *low carb diet* in six newspaper sections in 2004—most notably the Financial section whose only results for the ten-year period occurred in 2004. The *LexisNexis Academic* database allows convenient organization of the low carb diet corpus articles by section; Table 5.5 shows the representation of

articles by year and section in the low carb diet corpus. The progression of this table shows more than a start, a peak, and a drop-off. The table shows that in the years in which *low carb diet* was most frequent (2003, 2004, and 2005), the articles in those years reflected the form's use in a greater variety of sections than in other years. Consequently, the BVC reveals more than just frequency. The BVC, through newspaper articles that reflect a cross-section of American culture, tells a story of *low-carb diet* as a cultural phenomenon.

Finally, based on evidence from the BVCs, the cultural presence of the form *low carb diet* is greater in the United States than in the United Kingdom. Still, the BVC is able to show the rise and fall of *carb* in American speech, and the BVC also shows that in British English, *Atkins diet* challenges *low-carb diet* as the culturally preferred reference to this dietary practice. In the American low carb diet corpus, the most frequent R1 form for *low-carb diet* is *craze*. The collocation, *low-carb diet craze* well describes the trajectory of *carb* in American English because *low-carb diet* appears in zero articles in the BVC in 1998; in 597 in 2004; and in only 32 in 2007. The structure and organization of the BVC texts, along with their being full-text, highlights the varied discourses that construct the career of *carb* in American English.

The BVC allows panoramic analysis of the discourse that surrounded the concept of dieting from 1998 to 2007. From a broader linguistic view of dieting, the BVC can show what forms preceded *low carb* in terms of frequency in the years of 1998 to 2003 as well as what forms followed *low carb* in the years of 2005 to 2007. Just as the American low carb diet corpus allows a focused glimpse into the career of *low carb diet* in American English, another corpus, derived from the American BVC, can show what other forms were used in high frequency before, after, and alongside *low carb diet* from 1998 to 2007 in the American BVC. The study

constructed a derived corpus, the American diet corpus, to provide such evidence. The significance of this derived corpus is its ability to frame and highlight the career of *low carb diet* within a larger discourse context.

The American Diet Corpus

The American diet corpus was constructed by searching all 25 titles in the American BVC annually in *LexisNexis Academic* with the term diet! . This search term should locate all articles in each year that contain a form that builds on the root *diet*. These forms include, *diet, diets, dieter, dieters, dieting, dietitian,* and *dietitians*. The study is not necessarily interested in the specific forms *dietitian* or *dietitians*, for example, but it is interested in the context, and most importantly, the lexical items that surround the use of such forms.

Construction of the American diet corpus was tedious and time consuming because of the magnitude of texts involved. Also, because related forms in the text that congregate around the search form were significant to the study, the search results were downloaded in full-text format. The downloading of so many full-text articles (approximately 1100 articles per month over the ten-year span of the American BVC) was time consuming because the delivery of such a large amount of text requires more time than customary, smaller downloads. Still, the downloading was completed, and the final derived corpus consists of 139,675,264 words. As mentioned above, the corpus is organized in directories by year, so the texts for one year consist of approximately 13 million words.

This derived corpus was larger than most that the study had worked with before; however, the only inconvenience the larger word count presented was a wait of a few minutes when calculations were performed on the corpus with WordSmith Tools at the ten-year level.

Such calculations required a few minutes; calculations with WordSmith Tools on each annual subcorpus were not unusually slow. Tables 5.8 through 5.17 include results for the node *diet* for 1998 through 2007.

Table 5.8: Results for the node *diet* for 1998

L3	L2	L1	R1	R2	R3
2	*0*	*calorie*	*coke*	*exercise*	*robert*
624	621	648	368	282	60
fat	*low*	*fat*	*drug*	*redux*	*fruits*
117	331	269	209	59	48
eat	*high*	*steady*	*drugs*	*help*	*exercise*
64	175	255	184	56	47
part	*atkins*	*healthy*	*pills*	*says*	*fat*
58	81	216	116	42	43
Go	*plant*	*vegetarian*	*pepsi*	*nutrition*	*high*
43	55	206	108	39	35
Eating	*change*	*balanced*	*exercise*	*byline*	*health*
42	51	186	96	37	33
breast	*cancer*	*american*	*revolution*	*combination*	*losing*
35	44	128	93	37	30
protein	*exercise*	*healthful*	*plan*	*fitness*	*weeks*
35	41	95	89	35	26
1998	*fat*	*poor*	*pill*	*lifestyle*	*byline*
33	39	89	80	35	25
weight	*health*	*daily*	*soda*	*lower*	*calcium*
32	39	85	77	35	24

Table 5.8 shows the results for the node *diet* for the 1998 subcorpus. The node itself occurs 11,944 times. The forms, *atkins* in the L2 position and *revolution* in the R1 position relate to the low carb diet; Atkins wrote the popular book, *Dr. Atkins' New Diet Revolution*; references to that title create the high frequency results for those two forms in proximity to *diet*. The form *new diet revolution* occurs 89 times in the 1998 subcorpus. In the L2 position, the form *atkins* is the fourth most frequent form with 81 results.

The most frequent form in the R3 position of *diet* is *robert*. The 60 results for *robert* in that position all are references to Dr. Atkins; a typical line from these 60 results is: "'Dr. Atkins' New Diet Revolution' by Robert C. Atkins (Avon)" (*LexisNexis Academic*). This line, as well as many of the other concordance lines, contains the posting of the title on a best-seller list. The form *carb* appears 8 times in this subcorpus; two of those eight occurrences are the form *low-carb* that appears in the title of one cookbook. The other 6 occurrences include these lines or collocations: *my carb level*, *high-carb*, *your carb intake*, *super-carb drink*, and *high-carb diet* and *carb up for energy*. The complete sentence that includes the last form clarifies its interesting use: "As for the carb up for energy advice for the sports-minded, Young points to the potatoes, salad and soft drink during dinner and the sports drink consumed during the game, in Jordan's case" (*LexisNexis Academic*).

The most frequent L1 collocate for *diet* is *calorie*. The most frequent collocation of *diet* is *2,000 calorie diet* with 618 occurrences. The majority of these occurrences come from the *Chicago Sun-Times* which includes percentage values for recipe items. At the end of a recipe, this, or a similar sentence, is frequently included with nutritional information that relates to the recipe: "Percentages of daily value based on 2,000-calorie diet" (*LexisNexis Academic*).

Other *diet* collocates include *low fat diet* with 197 occurrences; and *low calorie diet* has 14 occurrences. The proprietary name, *diet coke* has 364 occurrences, and *coke* is the most frequent form in the R1 position of *diet*. The second most frequent form in the R1 position is *drug*, and the next is *drugs*. The drug name, *redux* is the second most frequent form in the R2 position of diet. Also, in the 1998 subcorpus, but not shown as a result in Table 5.8 is the drug name, *fen-phen* or *fen/phen* which has 311 occurrences.

Some of the major building blocks of low carb diet discourse are included as collocates of *diet* in Table 5.8; however, these forms (*robert*, *atkins*, and *revolution*) are overshadowed in terms of frequency by the more dominant collocations that involve dietary standards: *2,000 calorie diet* (618 occurrences); *steady diet* (255 occurrences); *healthy diet* (212 occurrences); and *low-fat diet* (197 occurrences). Also, the combined occurrences for the collocates *diet drug* and *diet drugs* have 376 occurrences. The presence of Dr. Atkins's book title on best-seller lists places the forms, *robert*, *atkins*, and *revolution* in close proximity to the form *diet*, but the form *carb* has only 8 occurrences in the subcorpus. The 1998 subcorpus shows that Dr. Atkins's name and book title received coverage, but the 1998 subcorpus does not show evidence for a *low-carb craze*.

Table 5.9: Results for the node *diet* for 1999

L3	L2	L1	R1	R2	R3
2	*0*	*calorie*	*drug*	*exercise*	*fen*
648	644	763	330	293	58
fat	*low*	*fat*	*coke*	*combination*	*fruits*
133	420	339	290	144	52
part	*high*	*steady*	*revolution*	*lifestyle*	*robert*
81	301	264	184	55	49
eat	*atkins*	*balanced*	*pills*	*robert*	*exercise*
61	149	197	154	37	47
protein	*weight*	*healthy*	*drugs*	*help*	*losing*
57	99	197	146	33	41
fed	*fen*	*protein*	*exercise*	*health*	*help*
56	71	132	102	32	38
eating	*change*	*vegetarian*	*pill*	*nutrition*	*diet*
43	52	106	99	29	36
go	*plant*	*mediterranean*	*rich*	*byline*	*percent*
43	52	100	77	28	36
diet	*exercise*	*daily*	*plan*	*breast*	*fat*
36	49	92	73	27	35
resolutions	*reduced*	*atkins*	*books*	*business*	*weight*
34	44	89	68	24	34

Table 5.9 includes collocates for *diet* for the 1999 subcorpus. The node *diet* has 12,197 occurrences in the 1999 subcorpus. Lexical items in Table 5.8 that relate to the low carb diet include *robert* (R2 and R3 positions), *atkins* (L1 and L2 positions), and *revolution*. As mentioned above, *revolution* is a part of the title of Dr. Atkins's popular book about the low carb diet. As was the case for the 1998 subcorpus, *2,000 calorie diet* is the most frequent collocation of *diet* in the 1999 subcorpus, and, again, the majority of those occurrences come from nutritional data connected to recipes in the *Chicago Sun-Times*.

Other high frequency collocations of *diet* include: *diet drug* and *diet drugs* (476 total occurrences) *diet coke* (290 occurrences); *steady diet* (263 occurrences); *low-fat diet* (220 occurrences); *balanced diet* (194 occurrences) and *healthy diet* (196 occurrences).

The collocation *diet pill*, along with *diet pills* and *diet-pill* occur a total of 250 times in the 1999 subcorpus. The drug name, *fen-phen* occurs 859 times—a large increase from 311 results in the 1998 subcorpus. The most frequent form in the R1 position of *diet*, *drug*, occurs 330 times. Of the ten years of coverage in the American diet corpus, only in the 1999 subcorpus is *coke* not found as the most frequent form in the R1 position of *diet*. The two most frequent R1 collocates of *diet drug* in the 1999 subcorpus are *combination* and *cocktail*. These labels are used in reference to a variety of drugs—especially *fen-phen*; however, *redux* is also appears in some of the related concordance lines.

The 1999 subcorpus shows that the use of *carb* forms increased from 1998 to 1999. The form *carb* appears 126 times in the 1999 subcorpus which is a significant increase from 8 occurrences in 1998. The most frequent collocate of *carb* in this subcorpus is *low* which is also

the most frequent form in the L1 position of *carb*. The form that was the focus of many of the earlier sections of this chapter, *low carb diet*, does not appear in the 1998 subcorpus, but that form has 17 results in the 1999 subcorpus.

Another attribute of the collocating habits of *diet* is the variety of words that gravitate to *diet* because of the variation that American society and business can create; for example, *mediterranean* does not appear in Table 5.8, but it does in Table 5.9 that is based on the 1999 subcorpus. That form has 100 occurrences in the L1 position of *diet* in the 1999 subcorpus. Apparently, the *mediterranean diet* was a trend that American dieters embraced to achieve the healthful attributes of certain European dietary practices. Also, the popularity of Dr. Atkins's book likely motivated occurrences of *low-carb diet* in the 1999 subcorpus.

Table 5.10: Results for the node *diet* for 2000

L3	L2	L1	R1	R2	R3
2	*0*	*calorie*	*coke*	*exercise*	*robert*
556	549	588	288	307	63
fat	*low*	*steady*	*pills*	*combination*	*exercise*
125	322	273	274	57	60
eat	*high*	*healthy*	*drug*	*nutrition*	*fruits*
64	262	259	212	57	59
protein	*atkins*	*fat*	*revolution*	*lifestyle*	*fat*
63	102	236	159	51	39
eating	*change*	*balanced*	*exercise*	*help*	*low*
61	57	182	112	40	37
part	*carbohydrate*	*special*	*pill*	*health*	*weight*
49	48	127	97	39	36
go	*exercise*	*vegetarian*	*drugs*	*byline*	*people*
45	48	126	96	33	32
fed	*concerning*	*protein*	*rich*	*prescribed*	*vegetables*
39	40	115	89	31	32
resolutions	*weight*	*mediterranean*	*nutrition*	*robert*	*c*
39	40	102	67	30	31
staple	*plant*	*fiber*	*books*	*includes*	*day*
32	39	97	60	28	28

Table 5.10 shows the collocates of the node *diet* in the subcorpus for 2000. The node *diet* has 11,983 occurrences in the subcorpus. As in the subcorpora for 1998 and 1999, the forms *robert, atkins,* and *revolution* appear in Table 5.10. Also, Dr. Atkins's middle initial, *c,* appears as the next to last collocate in the R3 column in Table 5.10. The most frequent form in the R3 position, *robert,* appears to be exclusively from best seller lists that feature Dr. Atkins's book title and his name. As was the case with the 1999 collocates, in the 2000 results, *robert* appears in both the R2 and R3 columns in Table 5.10.

In Table 5.10, *2,000 calorie diet* is the most frequent collocation of *diet* and has 544 occurrences. High frequency collocates in the 2000 subcorpus include *steady diet* with 273 occurrences; *healthy diet* with 257 occurrences; *balanced diet* with 180 occurrences; and *low-fat diet* with 146 occurrences. *Diet pill* and *diet pills* have 371 total occurrences; *diet drug* and *diet drugs* have 308 total occurrences; and *diet coke* has 288 occurrences.

One significant change in the 2000 subcorpus is that the form *carbohydrate* is present; that form is not present as a collocate of *diet* in the collocational results for 1998 or 1999, and its presence could be a signal of an increase in the social momentum of the low carb diet. The form *carb* appears in the 2000 subcorpus 121 times—slightly fewer occurrences than 1999's 126 occurrences. *Low-carb diet* appears 22 times in the 2000 subcorpus.

Also, in terms of change, the forms *pill, pills, drug,* and *drugs* appear in Table 5.10 in the R1 position of *diet,* but no names of drugs appear in Table 5.10. Drug names appear in Tables 5.8 for 1998 (*redux*) and 5.9 for 1999 (the *fen* portion of *fen-phen*).

Table 5.11: Results for the node *diet* for 2001

L3	L2	L1	R1	R2	R3
2	*low*	*steady*	*coke*	*exercise*	*exercise*
107	236	298	259	363	61
fat	*high*	*healthy*	*pills*	*modification*	*heart*
86	128	219	153	50	52
eat	*0*	*balanced*	*exercise*	*blood*	*fruits*
62	106	174	123	43	41
eating	*program*	*fat*	*drug*	*nutrition*	*food*
61	60	173	91	38	38
seniors	*change*	*calorie*	*pill*	*fitness*	*weight*
50	50	145	78	35	38
registered	*dietitians*	*special*	*soda*	*help*	*glucose*
49	49	133	62	31	35
part	*exercise*	*vegetarian*	*rich*	*lifestyle*	*fat*
43	48	105	61	31	34
week	*atkins*	*daily*	*byline*	*combination*	*foods*
37	47	83	58	28	32
classes	*changes*	*american*	*pepsi*	*low*	*speak*
35	45	77	58	27	31
fed	*plant*	*healthful*	*high*	*says*	*health*
33	30	63	53	26	29

For the 2001 subcorpus, *diet* has 9,959 occurrences, and Table 5.11 shows a significant change in the way *calorie* collocates with *diet* in the 2001 subcorpus. The form *2,000-calorie diet* has 104 occurrences in the 2001 subcorpus, and the same form had 544 occurrences in the 2000 subcorpus. Within the confines of the American diet corpus, a definitive explanation likely cannot be presented to answer why that form's use dropped so sharply from 2000 to 2001. The *Chicago Sun-Times* is the nearly exclusive source for the occurrences in both 2000 and 2001, and, unfortunately, an explanation is not feasible. Whatever caused the *2,000 calorie diet* occurrences to drop in 2001 likely is also the source for the drop in the number of occurrences for the node, *diet* in 2001 as well. The average node count (for *diet*) per year in the American diet subcorpora is 11,350, and the count for 2000 is 11,983 which is above the yearly average average. The node count for 2001, 9959, is 2024 less than the previous year and 1391 less than the average node count per year in the subcorpora.

The most frequent collocation of *diet* in the 2001 subcorpus is *steady diet* with 296 occurrences; *healthy diet* has 217 occurrences; *balanced diet* has 173 occurrences; and *low-fat diet* has 117 occurrences. *Diet coke* has 258 occurrences, and *diet pill*, *diet pills* and *diet-pill* have a total of 230 occurrences in the 2001 subcorpus, but no drug names appear in Table 5.11.

Carb has 111 occurrences in the 2001 subcorpus; *low carb* has 42 occurrences; *low carb diet* has 9 occurrences in 6 articles. Those 6 articles come from all of the BVC's regions except for the Midwest region.

In the 2001 subcorpus the most frequent collocation of *diet, steady diet,* replaces *2,000 calorie diet* which had been the most frequent collocation of *diet* in the 1998, 1999, and 2,000 subcorpora. Only one cell (47 results for *atkins* in the L2 position) in Table 5.11 has a form that relates significantly to the low carb diet. Almost all of these 47 occurrences result from the title

of Dr. Atkins's book. Some part of Dr. Atkins's name or the book's title was reflected in 5 cells in Table 5.10 for the 2000 subcorpus; evidently, in 2001 Dr. Atkins's book was featured on fewer best seller lists in the BVC's newspapers than in 2000 as only one cell in Table 5.11 features an element of his name or his book's title.

Table 5.12: Results for the node *diet* for 2002

L3	L2	L1	R1	R2	R3
fat	*low*	*healthy*	*coke*	*exercise*	*robert*
95	327	277	295	390	81
part	*high*	*steady*	*pills*	*nutrition*	*heart*
55	135	272	169	43	72
seniors	*atkins*	*atkins*	*exercise*	*fitness*	*speak*
51	106	238	116	42	50
registered	*exercise*	*fat*	*revolution*	*lifestyle*	*exercise*
49	58	181	115	34	47
eat	*change*	*balanced*	*pepsi*	*help*	*food*
44	51	151	98	31	44
eating	*dietitians*	*vegetarian*	*rich*	*health*	*fruits*
44	49	141	95	29	40
go	*raw*	*special*	*drug*	*includes*	*small*
44	40	120	86	29	38
low	*changed*	*strict*	*soda*	*byline*	*byline*
35	36	97	58	25	31
fed	*watch*	*american*	*needs*	*says*	*foods*
27	31	81	52	25	27
foods	*weight*	*protein*	*pill*	*eating*	*lemon*
27	27	72	50	21	27

In the 2002 subcorpus, the node *diet* occurs 10,727 times. The collocate *2,000* which has been a prominent collocate of *diet* in previous tables above, does not appear in Table 5.12. The collocation, *2,000 calorie diet*, only appears 17 times in the 2002 subcorpus.

The most frequent collocation of *diet* in the 2002 subcorpus is *diet coke* with 286 occurrences. The most frequent collocation of *diet* created by a form from the L1 position is *healthy diet* with 273 results. Other high frequency collocations include *steady diet* with 272 occurrences; *atkins diet* with 231 occurrences; *diet pill and diet pills* with 213 total occurrences; and *low-fat diet* with 123 occurrences.

The form *carb* appears 195 times in the 2002 subcorpus, and *low-carb diet* appears 23 times. Table 5.12 shows some significant collocational habits on the part of forms related to the low carb diet. *Atkins* and *revolution* appear in the top half of the occurrences for the L2 and R1 positions of *diet* respectively in Table 5.12, and these forms connect directly to the title of Dr. Atkins's book. More significant is the presence of *atkins* in the L1 position; the collocation *atkins diet* refers to the cultural practice of low carb dieting, and the form *atkins* is the third most frequent form in the L1 position in the 2002 subcorpus. In the 2002 subcorpus *atkins diet* occurs 231 times. Before 2002 the only year in which *atkins* appears in the L1 position in a collocate table for *diet* is 1999. Table 5.9 shows *atkins* with 89 results in the L1 position in the 1999 subcorpus. In 1999 and in 2002 the collocation *atkins diet* is employed in articles as a label for the practice of the low carb diet; these results (separate from Dr. Atkins's book title which involve *atkins* in the L2 position) in the 1999 and 2002 subcorpora do not show evidence of being caused by a single factor, such as a recurring article title, a book title, or a recurring news announcement.

This study has noted 2004 as the peak for use of the form *low-carb diet*, and the presence of *atkins* as the third most frequent form in the L1 position of *diet* in the 2002 subcorpus is a major step toward achievement of that peak.

Table 5.13: Results for the node *diet* for 2003

L3	L2	L1	R1	R2	R3
fat	*low*	*atkins*	*coke*	*exercise*	*robert*
94	480	755	315	356	121
eating	*south*	*steady*	*revolution*	*beer*	*agatston*
57	229	298	240	89	84
exercise	*atkins*	*healthy*	*pills*	*arthur*	*heart*
57	201	232	146	87	52
low	*high*	*beach*	*supplement*	*lifestyle*	*exercise*
55	177	229	106	44	45
part	*exercise*	*balanced*	*pill*	*weight*	*weight*
54	59	181	105	41	45
eat	*ephedra*	*fat*	*pepsi*	*says*	*fat*
53	52	176	100	39	42
go	*change*	*vegetarian*	*drug*	*dr*	*speak*
53	43	164	91	38	40
protein	*changes*	*carb*	*root*	*byline*	*diet*
53	42	158	88	36	38
fed	*changed*	*special*	*exercise*	*help*	*saturated*
46	41	123	87	34	38
following	*dietitians*	*carbohydrate*	*rich*	*nutrition*	*byline*
42	40	121	77	34	37

In the 2003 subcorpus, the node *diet* appears 12,936 times. As was the case with 2002, the collocation *2,000 calorie diet* does not appear as a collocate of *diet* in Table 5.13. Still the node quantity increases sharply in 2003.

The most frequent collocation of the node *diet* in the 2003 subcorpus is *atkins diet* with 686 occurrences. The other high frequency collocations of *diet* include *steady diet* with 298 occurrences; *healthy diet* with 232 occurrences; *balanced diet* with 179 occurrences; *low-carb diet* with 144 occurrences; *low-fat diet* with 142 occurrences; and *low-carbohydrate diet* with 88 occurrences. The collocation *south beach diet* occurs 228 times and is a portion of a book title written by Arthur Agatston; the collocation *arthur agatston* spans the R2 and R3 positions with 87 and 84 results respectively. In the entire 2003 subcorpus, *arthur agatston* has 133 occurrences, and the high frequency of his name and *south beach diet* both are connected to book lists and other book-related discourse that include the book's title and author in the same fashion that cause Dr. Atkins's name and elements of his book's title (such as *revolution*) to appear, sometimes in high frequency, in previous tables. *Diet coke* continues to be a high frequency collocation with 315 occurrences.

Table 5.13 shows some important changes in comparison to the *diet* collocation tables for preceding years. Neither *2,000* nor *calorie* is present in Table 5.13; *2,000 calorie diet* has only 9 occurrences in the 2003 subcorpus, but unlike 2002, the node count for 2003 increases. The 2002 subcorpus has 10,727 occurences for *diet*, and the 2003 subcorpus has 12,936 occurrences. Occurrences of *low carb diet* likely causes an increase of occurrences for the node form, *diet*, from 2002 to 2003; note, for example, *atkins* has 755 occurrences in the L1 position of *diet* in Table 5.13. In contrast, in Table 5.12 for 2002, the most frequent form in the L1 position of *diet* is *healthy* with 277 occurrences.

Table 5.14: Results for the node *diet* for 2004

L3	L2	L1	R1	R2	R3
eat 78	*low* 840	*atkins* 783	*coke* 309	*exercise* 381	*agatston* 133
fat 69	*south* 687	*beach* 695	*craze* 139	*arthur* 133	*good* 84
part 68	*high* 168	*carb* 548	*exercise* 122	*physical* 73	*weight* 70
low 60	*atkins* 84	*healthy* 284	*pills* 103	*fats* 66	*arthur* 61
fed 58	*changes* 51	*steady* 245	*books* 100	*lifestyle* 61	*exercise* 61
go 50	*change* 50	*balanced* 185	*pepsi* 93	*health* 52	*low* 57
day 49	*exercise* 42	*fat* 176	*revolution* 89	*nutrition* 48	*inactivity* 49
protein 44	*food* 41	*poor* 157	*book* 84	*byline* 45	*diet* 40
diet 40	*weight* 40	*carbohydrate* 130	*plan* 80	*regular* 42	*food* 38
eating 39	*gluten* 38	*vegetarian* 98	*cookbook* 79	*says* 41	*heart* 37

In the 2004 subcorpus, the node *diet* has 14,162 occurrences, which is the highest annual node count (for *diet*) in the American diet corpus; that number also represents an increase from the node count for 2003 which is 12,936. *2,000 calorie diet* is not present in Table 5.14, so in two respects the 2003 and 2004 subcorpora are similar. First, for each subcorpus the occurrences of *diet* increase, but what had been an extremely high frequency form for previous years (*2,000 calorie diet*) does not appear as a collocate in Table 5.13 or Table 5.14.

The most frequent form in the L1 position of *diet* is *atkins* for the 2003 and 2004 subcorpora. *Atkins diet* has 758 occurrences in the 2004 subcorpus. In the 2004 subcorpus, the form *south beach* joins *atkins* as a major collocate of *diet*. *South beach diet* occurs 684 times and is the second most frequent collocate of *diet* in the 2004 subcorpus, and *arthur agatston* occurs 326 times. *Low-carb diet* has 495 occurrences; *low-carbohydrate diet* has 103 occurrences; and *low-fat diet* has 116 occurrences.

The following forms occur in the 2004 subcorpus: *healthy diet* (283 occurrences), *steady diet* (245 occurrences), and *balanced diet* (182 occurrences). *Coke* is the most frequent collocate of *diet* in the R1 position in Table 5.14., and *diet coke* has 309 occurrences in the 2004 subcorpus.

The presence of the low carb diet is evident in Table 5.14. *Carb* moves up to the third most frequent form in the L1 position, and the ominous form *craze* appears in Table 5.14 as the second most frequent form in the R1 position. Interestingly, as the low carb diet reaches its high water mark in 2004, the South Beach diet is a significant source of collocates, as mentioned above, for the node *diet* in the 2004 subcorpus as well. Low carb diet discourse maintains

possession of the collocate *craze*; *low-carb diet craze* has 68 results, and no collocations of forms specifically relative to the South Beach Diet collocate with *craze* in the 2004 subcorpus. The low carb movement's strong association with the form *craze* reflects the media frenzy that surrounded the diet and foreshadows *carb*'s fall from high frequency use.

Table 5.15: Results for the node *diet* for 2005

L3	L2	L1	R1	R2	R3
fat 64	*low* 322	*healthy* 263	*coke* 353	*exercise* 345	*agatston* 105
eat 55	*south* 232	*steady* 251	*pepsi* 131	*arthur* 105	*exercise* 60
go 49	*high* 120	*beach* 235	*exercise* 113	*lifestyle* 54	*heart* 55
part 46	*change* 62	*atkins* 181	*soda* 91	*fitness* 52	*foods* 38
fed 38	*exercise* 38	*balanced* 160	*pills* 77	*health* 36	*diet* 31
2 32	*plant* 35	*fat* 141	*plan* 70	*nutrition* 36	*regimen* 31
diet 31	*heart* 33	*carb* 101	*books* 67	*drinks* 34	*food* 29
eating 30	*0* 31	*calorie* 91	*craze* 52	*help* 30	*health* 29
fiber 29	*changed* 31	*vegetarian* 76	*don't* 52	*regular* 28	*help* 29
low 29	*raw* 29	*american* 75	*rich* 51	*cancer* 27	*fruits* 27

Occurrences for the node *diet* dropped from 14,162 in 2004 to 10,487 in the 2005 subcorpus. For 2005, *atkins* is the fourth most frequent form in the L1 position of diet with 181 occurrences, and Table 5.15 shows a significant reduction of occurrences for forms related to the South Beach Diet and the low carb diet. For 2003 and 2004 *atkins* was the highest frequency form in the L1 position of *diet*. Table 5.15 shows that *healthy* (with 263 occurrences) is the highest frequency form in the L1 position for the 2005 subcorpus. As evidence for the influence of the waning momentum of the low carb diet movement on occurrences of the node *diet*, for 2003 in the L1 position of *diet*, *atkins* has 755 occurrences and for 2004 in the L1 form of *diet*, *atkins* has 783 occurrences. Table 5.15 shows that *healthy* is the highest frequency form in the L1 position and has 263 occurrences. The comparatively lower number of occurrences for low carb diet forms and South Beach Diet forms in 2005 likely causes the number of occurrences of the node *diet* in the 2005 subcorpus to be significantly lower than in 2004.

Healthy diet has 260 occurrences in the 2005 subcorpus; *steady diet* has 250 occurrences, and *balanced diet* has 160 occurrences. In terms of low carb diet discourse, *atkins diet* has 176 occurrences; *low-carb diet* has 88 occurrences; and *low-carbohydrate diet* has 25 occurrences. *south beach diet* has 230 occurrences and *arthur agatston* has 138 occurrences. *Low-fat diet* has 117 occurrences and *diet coke* is the most frequent *diet* collocation in the 2005 subcorpus with 353 occurrences.

Tables 5.13 (2003) and 5.14 (2004) show the significant influence of the low carb movement and the South Beach Diet on the collocates of *diet*. The BVC presents evidence, beyond many obvious media pronouncements, in support of the notion that the low carb movement and South Beach Diet were fads. First the node (*diet*) counts are above the average annual amount in 2003 and 2004 when forms related to these diets reached their highest

frequencies. Second, as the frequencies for these diet-related forms drop, the node counts in 2002 (10,727) and 2005 (10,487) drop to below average. Finally, 2002 and 2005 dovetail with each other in that for both years *healthy* is the highest frequency L1 form. Forms related to the low carb movement and the South Beach Diet appear in both the 2002 and 2005 subcorpora, but the peak for these forms' frequencies is 2004. The similarities between 2002 and 2005 suggest that forms related to these two diets significantly increased in frequency in the discourse of American dieting in 2003 and reached their peak in 2004. The *diet* node count of 10,487 for the 2005 subcorpus is below the annual corpus average of 11,350; also, in 2005 as *healthy* resumes its position as the highest frequency L1 form of diet in the 2005 subcorpus, American diet discourse undergoes a kind of return to mainstream lexical use.

Table 5.16: Results for the node *diet* for 2006

L3	L2	L1	R1	R2	R3
fat 75	*low* 323	*fat* 272	*coke* 329	*exercise* 377	*heart* 65
go 74	*south* 127	*steady* 228	*pepsi* 144	*nutrition* 45	*exercise* 57
eat 59	*high* 114	*healthy* 214	*soda* 134	*arthur* 38	*speak* 45
registered 45	*fat* 57	*balanced* 136	*pills* 128	*help* 33	*agatston* 38
eating 41	*dietitians* 45	*beach* 126	*exercise* 94	*includes* 31	*food* 34
fed 39	*change* 44	*american* 106	*sodas* 69	*lifestyle* 31	*diet* 33
part 39	*changed* 44	*special* 94	*plan* 68	*weight* 31	*chicken* 31
diet 33	*exercise* 39	*atkins* 77	*needs* 59	*byline* 30	*fat* 27
exercise 29	*weight* 32	*calorie* 73	*drug* 52	*health* 28	*regimen* 25
percent 28	*health* 31	*poor* 68	*books* 47	*drinks* 27	*fruits* 24

The frequency of the node *diet* drops from 10,487 in 2005 to 9883 in the 2006 subcorpus. Table 5.16 shows that *coke* is the most frequent form in the R1 position of *diet*, and *diet coke* is the most frequent collocation of *diet* in the 2006 subcorpus with 328 occurrences.

Healthy is the most frequent form in the L1 position of *diet*. In the 2006 subcorpus, *steady diet* has 228 occurrences; *healthy diet* has 214 occurrences; *low-fat diet* has 204 occurrences; and *balanced diet* has 135 occurrences.

Forms in the 2006 subcorpus that relate to the low carb diet or South Beach Diet include *south beach diet* with 126 occurrences; *arthur agatston* with 51 occurrences, *atkins diet* with 75 occurrences; *low-carb diet* with 40 occurrences; and *low-carbohydrate diet* with 14 occurrences.

The form *craze* does not appear in Table 5.16; however, *craze* occurs in the R1 position of *diet* in Table 5.14 (2004) and Table 5.15 (2005). Some forms that relate to the low carb diet and the South Beach Diet are evident in Table 5.16, but the most frequent collocation of *diet* is *diet coke* which, as stated above, is a beverage and not a cultural practice as the low carb diet and South Beach Diet are.

Table 5.17: Results for the node *diet* for 2007

L3	L2	L1	R1	R2	R3
eat	*low*	*healthy*	*coke*	*exercise*	*exercise*
49	208	246	412	345	42
go	*best*	*steady*	*soda*	*nutrition*	*greene*
47	93	175	124	49	38
eating	*high*	*fat*	*pepsi*	*bob*	*food*
46	75	118	105	38	34
fat	*south*	*balanced*	*exercise*	*lifestyle*	*manual*
37	71	101	88	37	34
fed	*change*	*life*	*pill*	*owner's*	*fat*
31	68	84	74	35	31
exercise	*health*	*american*	*pills*	*says*	*heart*
28	42	82	73	32	31
calorie	*exercise*	*free*	*drug*	*health*	*f*
25	41	77	71	31	27
high	*never*	*vegetarian*	*book*	*weight*	*feline*
24	36	73	70	31	27
low	*raw*	*beach*	*plan*	*d*	*weight*
24	35	72	53	30	23
part	*fat*	*atkins*	*byline*	*michael*	*diet*
24	33	71	52	28	22

The 2007 subcorpus has the lowest number of occurrences for the node *diet* with 9229 occurrences. Table 5.17 shows that *healthy* is the most frequent form in the L1 position, and *healthy diet* is the second most frequent collocation of *diet* with 243 occurrences. *Steady diet* has 175 occurrences, *balanced diet* has 100 occurrences, and *low-fat diet* has 85 occurrences. *Diet coke* is the most frequent collocation of *diet* with 411 occurrences.

Collocations in the 2007 subcorpus that relate to the low carb diet and the South Beach Diet include *atkins diet* (68 occurrences), *low-carb diet* (27 occurrences), *south beach diet* (71 occurrences), and *arthur agatston* (21 occurrences).

Two factors contribute to the comparatively low node count for *diet* in 2007. First, the 2007 subcorpus has only eight occurrences for *2,000 calorie diet*; second, *diet* collocates related to the low carb diet and the South Beach Diet dropped significantly from 2004 to 2007. For example, Table 5.17 shows that for the L1 position of *diet*, *beach* (72 occurrences) and *atkins* (71 occurrences) are the ninth and tenth most frequent forms in the 2007 subcorpus.

Table 5.17 also shows that *healthy* and *steady* are the first and second most frequent forms in the L1 position for *diet* respectively. *Healthy* and *steady* are consistently high frequency forms in the L1 position of *diet* in all of the subcorpora. It is significant to note, for example, that in the 2004 subcorpus, the most frequent forms in the L1 position are (in order from high to low) *atkins*, *beach*, *carb*, *healthy*, and *steady*. After the popularity of the low carb movement and South Beach Diet had waned, *healthy* and *steady* become the most frequent collocates of *diet* in the L1 position in the 2007 subcorpus.

Tables 5.8 through 5.17 display collocates of *diet* by year, and *healthy* and *steady* function as markers in the L1 position. These two forms never leave the R1 position, and as fad forms, such as *atkins* and *carb*, increase in popularity, their occurrences surpass those of *healthy*

and *steady*. Similarly, as the fad-related terms decrease in frequency, *healthy* and *steady* return closer to the top of the L1 column of *diet* as noted in 2007 in which *healthy* is the most frequent form in the L1 position and *steady* is the second most frequent form in that position.

Discussion of Select *diet* Collocations

The American diet corpus, derived from the American BVC, provides evidence that allows an understanding of the habits of some collocations that build on the form *diet* and occurred before, during, and after *low carb diet*'s peak in 2004. *Low carb diet* evolves alongside these other *diet* collocations, and Table 5.18 shows the occurrences of 6 diet collocations in terms of three eras of the ten-year span of the American diet corpus.

The eras are delineated by the rise and fall of the form *carb*. In the years 2003, 2004, and 2005, *carb* is present in the L1 position of *diet*, so those years form the carb or middle era. The first era is formed by the years 1998 through 2002, and the last era is formed by the years 2006 through 2007. The forms listed in Table 5.18 were selected in terms of high frequency at both the ten-year level and at the level of the three eras to reflect activity of a variety of high frequency collocations that build on *diet* in all three eras. In the case of this table, high frequency is established by a form's presence as one of the three most frequent L1 forms of *diet* at either the ten-year level or one of the three era levels.

The three most frequent forms in the L1 position of *diet* at the ten-year level are *steady*, *calorie*, and *healthy*. *2,000* is the most frequent L1 form of *calorie*, so the collocations *steady diet*, *healthy diet* and *2,000 calorie diet* are included in Table 5.18. In the first era, *calorie*, *steady* and *fat* are the most frequent forms in the L1 position of *diet*, and as at the ten-year level,

2,000 calorie diet is the most frequent collocation that builds on *calorie diet*. Also, in each of the eras, *low* is the most frequent form in the L1 position of the form *fat diet*. Consequently, *low fat diet* is included in Table 5.18.

The three most frequent forms in the L1 position of *diet* in the carb era are *atkins*, *beach*, and *carb*. *Low* is the most frequent form in the L1 position of *carb diet* in the carb era, and because *low carb diet* was the central focus of much of this chapter, *low carb diet* was also selected for Table 5.18. Finally, because *low calorie diet* is the most frequent collocation of *calorie diet* in the carb and last eras, that form is also included in Table 5.18.

Table 5.18: Occurrences of select *diet* collocations by corpus region

diet collocation	1998 to 2002 first era	2003 to 2005 carb era	2006 to 2007 last era	Ten year total
steady diet	1359 *195.626*	793 *178.381*	403 *156.501*	2555 *182.924*
healthy diet	1155 *166.260*	775 *174.332*	457 *177.471*	2387 *170.896*
2,000 calorie diet	1935 *278.540*	46 *10.347*	16 *6.213*	1997 *142.974*
low fat diet	819 *117.893*	375 *84.354*	289 *112.230*	1483 *106.174*
low carb diet	71 *10.220*	727 *163.534*	67 *26.018*	865 *61.929*
low calorie diet	139 *20.008*	97 *21.819*	34 *13.203*	270 *19.330*

In the case of Table 5.18, rates of occurrence are shown in a rate per ten million words in italics immediately below the raw number of occurrences. The table shows that for the first era, *2,000 calorie diet* is the most frequent of the forms in the table, and *steady diet* is the second most frequent. For the *carb* era, *steady diet* is the most frequent and *healthy diet* is the second most frequent form. For the last era, *healthy diet* is the most frequent form and *steady diet* is the second most frequent form.

The most frequent collocation of diet in Table 5.18 is *steady diet*. In each of the three eras, the most frequent form in the R1 position of *steady diet* is *of*. At the ten-year level, all of the forms with 10 or more occurrences in the R1 position of *steady diet of* are included in Table 5.19.

Table 5.19: Forms in the R1 position of *steady diet of* at the ten-year level

Forms in the R1 position of *steady diet of* at the ten year level	Occurrences
fastballs	24
breaking	23
news	21
running	19
anti	16
junk	15
big	14
fast	14
curveballs *double* *television* *violence*	13 13 13 13
bad	12

All instances of the two most frequent forms in the R1 position, *fastballs* and *breaking* (47 total occurrences), relate to baseball. This connection of use and context is significant to the study because in each of the three eras, *steady diet* is either the most or second most frequent collocation in Table 5.18. A connection to athletics occurs with other forms in the R1 position of *steady diet of* at the ten-year level. For example, all of the 19 occurrences of the fourth most frequent form, *running*, are connected to athletics; 14 of the occurrences are connected to football; 2 are connected to basketball; 2 are connected to baseball; and one is connected to marathon training. *Curveballs* and *double* each have 13 occurrences in the R1 position of *steady diet of*. All of the occurrences of *curveballs* are connected to baseball, and in the case of the 13

occurrences of *double*, 10 are connected to football; and 2 are connected to basketball. Within the 15 occurrences of *junk* in the R1 position, 3 of the occurrences are connected to athletics; similarly, 5 of the 14 occurrences of *big* are connected to athletics; and one of the 12 occurrences of *bad* in the R1 position is connected to baseball.

Of the 210 total R1 occurrences in Table 5.19, 100 of them or 47% refer to athletics. Of the R1 forms in Table 5.19, the only form whose occurrences exclusively refer to food intake is *fast* with 14 occurrences. These 14 collocations build on *food* to the right and include *steady diet of fast food* (12) and *steady diet of fast food burgers* (2).

Steady diet's presence as a high frequency collocation of *diet* is significant to the understanding of how words collocate with *diet*. Many collocates of *steady diet* are connected to athletic contexts; however, the form *steady diet of* also attracts many single instance forms. Still, it is interesting to note that so many high frequency collocations of *steady diet* do not refer to dieting or eating. The trajectory of *steady diet* is flat across the three regions; the average rate of occurrence at the ten-year level is 182.924 per ten million words. That rate is exceeded in the first era (195.626) and the carb era and last era have lower rates of 178.381 and 156.501 respectively.

The R1 output for *steady diet of* exemplifies what is referred to as an A-curve (Kretzschmar 198). That is, *steady diet of* has a few forms in the R1 position that each have a large number of results, and in the R1 position, many forms have few or single occurrences. The total output for the R1 position (after a filter of the 100 most frequent forms from the Brown Corpus was applied) includes 276 forms. Before these results are adjusted for erroneously repeated occurrences, only 16 of those 276 forms have 10 or more occurrences; at the other end of the A-curve, 86 forms have only one occurrence. The single occurrence forms bear

significance to this discussion in two ways. First, the BVC has the mass to produce significant evidence in the R1 position of *steady diet of* (276 forms). Second, the mass of the BVC that makes the 86 single occurrences possible, shares forms that are important to the story of *steady diet of*. These single occurrence forms show the broad range of possibilities that can occur in the R1 position of this collocation. Kretzschmar explains that such low frequency forms can contain significant meaning even though they may have only single occurrences (201). In other words, the higher frequency forms, such as *fastballs* work together with the lower frequency forms, such as *wine* to create an illustration of the career and use of *steady diet of*. Appendix B contains the complete R1 output for *steady diet of*. The numbers in Appendix B may vary slightly from those in Table 5.19 because the occurrences in the Table have been adjusted for erroneously repeated texts. Appendix B contains the raw output.

Possibly some of the low carb discourse of the carb era could have encroached on *steady diet*'s share of diet discourse; even so, with a reduced rate in the carb era, *steady diet* is still the most frequent collocation in Table 5.18. In the last era, *healthy diet* has a rate of 177.471 and overtakes *steady diet* that has a rate of 156.501.

Healthy diet is similar to *steady diet* in that this form also ranks as one of the three most frequent forms in each era in Table 5.18. In all three eras, the most frequent phrase that builds on *healthy diet* is *a healthy diet and exercise.*

At the ten-year level, the most frequent collocates in the R1 position of *a healthy diet and* are *exercise, regular, lifestyle, getting,* and *hygiene*. The most frequent form in the R1 position of *a healthy diet and regular* is *exercise*. Similarly, the most frequent form in the R1 position of *a healthy diet and getting* is *exercise.* The collocational habits of *healthy diet* point clearly to the practices of exercise and healthful living. The study did not identify any other high frequency

habits of *healthy diet* that related to any other practices or concepts. *Healthy diet* has a flat trajectory; its average rate of occurrence at the ten-year level is 170.896 per ten million words. The rates for the three eras are 166.260 for the first era; 174.332 for the carb era; and 177.471 for the last era, so the frequency of *healthy diet* remains steady across the three eras.

From 1998 to 2001 the collocation *2,000 calorie diet* is a high frequency form in the American diet corpus. The source, as has been mentioned previously, for the majority of these occurrences is nutritional data that accompanies recipes in the *Chicago Sun Times*. It is interesting to note that while *2,000 calorie* only appears in this chapter's *diet* collocation tables for the years 1998 through 2001 (Tables 5.9 through 5.11). The mass of use of *2,000 calorie diet* in those few years was significant enough to cause *calorie* to be the second most frequent collocate in the L1 position of *diet* at the ten-year level.

Also, the flexibility of the BVC enables discernment of what texts caused *calorie* to rise to prominent frequency, and this methodology also highlights the presence of these texts in 1998 through 2001 and the sharp drop in frequency for the form in the following 6 years. The trajectory of *2,000 calorie diet* includes a high frequency of occurrences in the first era that is followed by a severe and consistent drop in the carb and last eras. In the first era *2,000 calorie diet* occurs at a rate of 278.540 per ten million words; in the carb era, *2,000 calorie diet* has a rate of 10.347; and in the last era, the form has a rate of 6.213. As has been stated earlier, this study cannot provide an answer for why *2,000 calorie diet* fell from high frequency use, but the study can point to specific texts, in which the form was used with high frequency between 1998 and 2001.

Low-fat diet is the fourth most frequent collocation in the first era of Table 5.18; it occurs in the same position in the carb era; and it rises to the third most frequent form in the last era. Its

ten-year average rate of occurrence is 106.174 per ten million words, and that rate is slightly exceeded in the first and last years (117.893 and 112.230 respectively); however, in the carb era, occurrences of *low fat diet* fall short of the average with a rate of 84.354. In the carb era, *low carb diet* is the third most frequent form; *low fat diet* is in the fourth position, but its rate is almost half that of *low carb diet*; also, in the carb era, *2,000 calorie diet* drops to the lowest frequency position in Table 5.18. The drop of *2,000 calorie diet* causes *steady diet* to move to the most frequent position in the carb era; *low fat diet* does not ascend in rank of frequency because *low carb diet* takes the place immediately above it. The likely reason why the rate of occurrence for *low fat diet* drops severely in the carb era is the significant share of discourse that was devoted to the low carb diet discourse.

Low carb diet has the lowest rate of occurrence of all the collocations in Table 5.18 for the first era. In the carb era, *low carb diet* moves to the third most frequent position—after *steady diet* and *healthy diet*. In the last era, *low carb diet* is the fourth most frequent form with a rate of 26.018, and its rate exceeds only *low calorie diet* at 13.203 and *2,000 calorie diet* at 6.213 per ten million words. Also, the rate of the third most frequent form, *low fat diet*, is a significant increase at 112.230. At the ten-year level, *low carb diet* is the next to last in terms of frequency, but its trajectory is clear: low in the first era; high in the carb era; and low in the last era.

Low carb diet's frequency in the last era suggests a very sharp drop in frequency. Table 5.4 shows that in 2005 (the last year of the carb era) low carb diet appears in 142 articles in the BVC; *low carb diet* appears in 53 articles in 2006 and in 32 in 2007. *Low carb diet*'s trajectory is a slow increase from 1998 to 2002 and an explosive increase to 2004 and a steady drop after 2004. One hallmark of *low carb diet*'s career is the way that *steady diet* and *healthy diet* remain

the most frequent collocations for the carb era. In the carb era, *low carb diet* has a frequency of 163.534 per ten million words and approaches *healthy diet*, the second most frequent form, which has a rate of 174.332.

Low calorie diet is the least frequent form in Table 5.18 at the ten-year level. In the first era, the only form that is less frequent is *low carb diet*; in the carb era *2,000 calorie diet* is the only form that is less frequent than *low calorie diet*. In the last era, again, *low calorie diet* is the fifth most frequent form, and *2,000 calorie diet* is the least frequent form. *Low calorie diet* does not experience the kind of frequency that *low carb diet* or *2,000 calorie diet* does; however, *low calorie diet* shows an important attribute of collocates of *diet*.

2,000 calorie and *low carb* are, at times, high frequency collocates of *diet*, and they have careers that are trademarked by peaks and valleys. *Low calorie diet* is a lower frequency form that preserves a mostly flat trajectory across the three eras in Table 5.18. The form's average rate of occurrence at the ten-year level is 19.330 per ten million words; that rate is exceeded slightly in the first era (20.008) and in the carb era (21.819). In the last era, the frequency of *low calorie diet* drops to 13.203. The BVC is able to show that a lower frequency form can have a career that includes variation in terms of frequency .

Conclusion

The American BVC provides evidence that illuminates the career of *carb* from 1998 to 2007. Evidence from the BVC also allows identification of collocations that build on the form *diet*. Such collocations, along with their frequencies, provide a reflection of the diet-related discourse in which *carb* is a low frequency form, a high frequency form, and again a low frequency form.

The collocations show that *steady diet* and *healthy diet* are a block of forms that remain high frequency in all three ears; only in the first era is a form more frequent than either of these—that form is *2,000 calorie diet*. That form's frequency is established by the inclusion of *2,000 calorie diet* at the end of recipes—especially in the *Chicago SunTimes*. *Low carb diet* approaches the frequency of *steady diet* and *healthy diet* in the *carb* era as shown Table 5.18. Later, as *low carb diet* falls from high frequency, it is replaced by *low fat diet*, which is the third most frequent form in the last era. The BVC allows the development of *low carb diet* to be analyzed, but the BVC also makes observation of surrounding forms, such as *low fat diet*, possible in order to provide an explanation for the linguistic climate that *low carb diet* endured from 1998 to 2007. Further, because of the significant amount of text in the BVC, analysis of the collocational habits of forms, such as *steady diet of* is possible. Analysis of that particular form reveals that 47% of the texts that produce the 11 most frequent forms in the R1 position connect directly to discussion of athletics. At a glance, *steady diet of* might not appear to be a sports-related collocation; however, all the top two R1 forms (27 concordance lines) come from discussions of baseball. A methodology that could not access such a large amount of text might identify mostly single-occurrence forms in the R1 position of *steady diet of*. In the case of

steady diet of, the BVC identifies 2439 occurrences of the node, *steady diet of*. Along with that node, a total of 275 forms are identified in the R1 position of the node. As Table 5.19 shows, 13 of the 275 forms occur more than ten times; additionally, 84 of 275 forms are single occurrences. Some of these forms relate to athletics; some relate to broad discussions (*steady diet of news*), and some relate to food (*steady diet of fast food* and *steady diet of fast food burgers*). The BVC is able to identify the fine-grained elements, such as the two most frequent collocates in the R1 position work to create collocations which relate exclusively to baseball commentary. The low frequency collocates, which the BVC provides, are a part of the story of *steady diet of* as well because they show uses that isolate themselves to single occurrences, such as *steady diet of commercials* and *steady diet of entertainment*. This picture of the use of *steady diet of* in American English shares both high and low frequency collocates and shows the sometimes surprising contexts, such as baseball, in which they may be used.

In the matter of *carb*, *low carb diet* has been identified as the most frequent collocation of *carb*, and in 1998 that form has no occurrences in the BVC. Later, in 2004, that form appears in 597 articles in the BVC, and in 2007 the form appears in only 27 articles. Additionally, the trajectories of other collocations that surround *low carb diet* can be studied, so the diet-related discourse in which *carb* appears may be understood better. The low carb diet may be regarded as a fad, but the BVC provides evidence which linguistically highlights the form's unique career in American English.

Chapter 6

Practical lexicography: A Pilot Study toward Updating

A Dictionary of Americanisms on Historical Principles

The process of identifying and defining an Americanism is a challenge. Clearly, American English branches off from British English; however, individual forms can only be identified as Americanisms after relevant historical and recent evidence are obtained and considered. Algeo (1992) explains that certain challenges congregate around the pursuit of identifying what an Americanism is and what a Briticism is. Still, Algeo clarifies that such pursuits can be engaged in a highly principled fashion:

> Because of their parallelism, a comparison of Briticisms with Americanisms is inevitable, but Americanisms are easier to define. A diachronic Americanism is an expression that originated in America, whatever its current use may be. A synchronic Americanism is an expression with characteristic form or use in America, whatever its origin may have been. Thus *hamburger* is historically an Americanism, but the name along with the thing has spread internationally, so the term is no longer synchronically an Americanism. On the other hand, *fall* for the season of the year was in British use at one time and may still be in regional dialects, but it is now rare in England. It is however typical in America and so has become a synchronic Americanism, although not one in origin. *Hamburger* and *fall* are both Americanisms, though of radically different kinds. (287-288)

Algeo notes the importance of understanding a form's career across decades in the determination of whether it is an Americanism. Algeo continues: "We must remind ourselves that when two 'branches' of a language grow apart, they are not categorically distinct like the branches of a real tree but continue to exchange influences and may grow back together, as British and American seem now to be doing, and eventually remerge" (289). Algeo's explanation of synchronic and

diachronic forms underscores a researcher's need to provide both historical and recent evidence in both telling the story of an Americanism within American English and comparing its American career to its career in British English.

Foundational Works

Several American scholars have contributed significant word-level research to the scholarly community that contributes to the identification and study of Americanisms. The study of American English has produced several works that have contributed important research toward identifying and describing Americanisms. The perspective of these works varies and could embrace a focus that is historical, current, or combined.

A Dictionary of American English on Historical Principles (1938-44) (*DAE*) followed rigorous principles of support through historical quotations which had been established by the *OED 1*. This work is significant because it is an American project that highlights Americanisms and embraces a close connection between entry forms and evidence in the form of historical quotations. The Preface to this work explains, "The end of the nineteenth century has been selected as a fitting point at which to terminate the admission of new words, however common some of these may have become in recent use. The illustration of those already current before that date, however, is frequently carried into the first quarter of the present century" (v). This statement shows that the focus of the *DAE* was generally historical; however, the Preface notes that some illustrations could come from as recent as 1925—which is to a certain extent a focus on recent use as the years of the *DAE*'s publication were between 1938 and 1944. The focus of the *DAE* is, therefore, both historical and current.

Allen Walker Read systematically explained the origin of the famous Americanism, *O.K.*, through manual research of newsprint on paper. Read's work in a variety of areas has been a major contribution to the study of American English. Still, with regard to the current study, Read's work on *O.K.* is especially relevant because he gravitated to newspapers for his research. Read's work that explained the etymology of *O.K.* was published in 1941 (Read (2002) 123). He explained Boston newspapers' fondness for the use of abbreviations beginning in 1838 (123). Read followed a strict research process, and he relied on real use through the evidence he found in newsprint. An editorial comment in *Milestones in the History of American English* notes about Read's work on *O.K.* twenty years after he had published his first work on the topic: "Even then, he did not realize that he was embarking on one of the longest and most sustained inquiries into the history of an American word" (123). Much of Read's work focused on words that relate to the American experience and culture and the description of American English. In the matter of *O.K.*, Read's focus was historical as he was seeking the origin of a form in current use.

A Dictionary of Americanisms on Historical Principles (1951) endeavored to record Americanisms and connect them with illustrative quotations in the tradition of the *OED 1* and the *DAE*. The *Dictionary*'s Preface explains the following about what an Americanism is: "As used in the title of this work, 'Americanism' means a word or expression that originated in the United States" (v). The *Dictionary*'s Preface further explains "The purpose of this dictionary is to treat historically as many as possible of those words and meanings of words which have been added to the English language in the United States" (v). The Preface continues: "In trying to identify those words and word meanings which came first into the English language in the United States, one of the procedures followed has been to examine carefully the evidence in the *OED*—both the main work and its *Supplement*—and in the *EDD*" (v). Evidence was clearly important to Mitford

Mathews who was the editor of the *Dictionary*. The *Dictionary*, which has not been updated since its release in 1951, stands as an important dictionary that was compiled with a historical perspective of words that originated in the United States.

Thomas Pyles's *Words and Ways of American English* (1952) focuses on how American English, which is by origin connected to British, but through American cultural events and human experience has become a separate variety of English. Pyles connects concepts, such as the new American frontier, the growth of the nation, and commercialization with layers of words and linguistic mores that contribute to the creation of American English and linguistic distinction from British English. For example, Pyles notes many lexical examples of frontier tall talk including, *absquatulate*, *cattywampus*, *grandiferous*, and *monstracious* (129). Pyles also gives these examples of clipped forms that reflect the American experience: *gas, photo, pep, bike, ad, bunk, auto, prof, taxi*, *tux* (185). Pyles's work reflects much about the careers of words and their building blocks. He also highlights word elements, such as *-orium*; *-mobile*; *-buster*; and *-conscious* in the reflection of the American experience through words (188-89). Pyles begins with a historical focus on American English, and he concludes with a more recent (to 1952) perspective of American English in a chapter named, "Later American Speech: Adoptions from Foreign Tongues." Also, in the chapter "American and British Word Usages" Pyles addresses recent and historical contrasts between American and British English in terms of borrowings and variation of uses of words that are shared in both forms of English.

The *Dictionary of American Regional English* (1985-) (*DARE*) uses personal interviews to obtain paticipants' responses and word use (Cassidy). *DARE* also reinforces information obtained from interviews with evidence from a variety of print sources that can be historical and recent. Maps accompany some entries to indicate the region of the country in which the

informants reside. *DARE* also has a set of usage labels that indicate, beyond the primary concern of region, *DARE* is also concerned with a variety of aspects of word use. The categories of the *DARE* usage labels are: amount of use, currency, type of user, and manner of use (Cassidy 1987). Within the category of currency, these labels exist: obsolete, archaic, old fashioned, historical (Cassidy 1987). Still, *DARE* has some focus on recent use because, for example, Volume IV (2002) includes a quotation from a 2001 print source in support of *raft duck* (438), so *DARE* has approach that combines historical and recent illustrations of regional speech.

The *DAE*, Allen Walker Read, *A Dictionary of Americanisms on Historical Principles,* Thomas Pyles, and *DARE* have established a foundation for scholars of American English to build on through new electronic corpus-based processes. These foundational scholars and works have shared evidence that has highlighted the distinctness of American culture and experience through American English, especially at the word level, and the focus of these works has been historical as well as a mixture of historical and recent. Tools are now available that allow current researchers to abandon the tedious work of reading paper newspapers one at a time, as Read did, and embrace a corpus linguistics approach that gives researchers access to enormous amounts of recent text that is organized by principles of balance. Current researchers can also use online historical databases in which multiple titles can be searched across decades in a single search operation.

Pilot Study Background

This chapter documents a pilot study that is focused on an update of *A Dictionary of Americanisms on Historical Principles*. It reviews a limited range of entries and analyzes them with current resources and strategies and proposes additions based on the same resources. Such

a pilot study is important because an update gives a 21st century perspective to the 1951 work. Much has changed in the American experience since 1951. Technologies, entertainment forms, and inventions that did not exist in 1951 surround us each day. A lexicographic record of the changes in American speech since 1952 would be a valuable linguistic and historical record for the scholarly and broader American communities.

The resources that give such a modern perspective to the *Dictionary* include recent dictionaries as well as historical evidence that is available through *ProQuest Historical Newspapers*, *APS Online*, and *NewspaperARCHIVE.com*, which are electronic databases. These databases highlight the early careers of Americanisms and provide antedatings for original entry forms in the *Dictionary*. The BVC can provide evidence both in American and British texts for the recent careers of forms, so the BVC, through recent frequency, can show whether a prospective Americanism has currency in British English as well as American English. Thus, the BVC can provide recent evidence of the careers of original entry forms in the *Dictionary* and supply new forms that should be included in an update. The organizing principles of the BVC reflect speech use from five regions of the United States. The BVC can identify forms, which are likely low frequency, that are generally used only in one geographic area. Such forms could illuminate the differences of one region of the U.S. from the others—just as Pyles highlights words whose use has made American English move further away from British English.

Today, through electronic technology, dictionary-making projects can do more. The BVC works with a word mass that is unusually large compared to current projects in the field; however, even though the BVC's greatest contribution to the study of American English is likely the collection of an estimated 5 billion words for linguistic study, the flexibility of the BVC should not be overlooked. The BVC enables immediate access to the original full-text source

documents; the BVC also has powerful downloading capabilities, so files can be organized for analysis by an offline interface.

In order to update the *Dictionary*, a researcher should work with a certain range of entry forms at a single time. The researcher should maintain a spreadsheet or similar application that includes the current range of forms that are being researched; the earliest illustrative quotation date for each form in the *Dictionary*; the earliest date of illustration that the *OED* provides if applicable; and the dates of any independently obtained antedatings. For the active range of entry forms, a researcher should attempt to acquire antedatings from the *OED* initially. The *OED* employs teams of professional researchers who have access to many rich resources, so its value to the antedating process should be prized. After consultation with the *OED*, a researcher should refer to historical databases, such as *ProQuest Historical Newspapers*, *APS Online*, and *NewspaperARCHIVE.com* to determine whether such original research could produce antedatings.

Next, the original entry forms should be analyzed for recent presence in American and British English with the BVC, which reflects both principles of balance in its construction and a recent snapshot of a form's career in both American and British English.

While still functioning under the process of determining recent frequencies, a researcher should mine (in the case of this study) the BVC for new forms that may not be present in the *Dictionary* because of, for example, changes in American speech over the last 50 years. Such a mining process may include slightly complex steps, such as the downloading of a derived corpus that can be analyzed offline with a flexible concordancing application, such as WordSmith Tools. As a parallel to the mining process, the *OED*, and multiple, recent American dictionaries should be consulted for the presence of possible new forms. As new forms in the active range of

forms are identified as candidates for inclusion (from either the BVC or recent American dictionaries), those forms should be evaluated, just as the original forms in the active range have been, for their frequency in the American and British BVCs.

After all forms (original entry forms and new candidates for inclusion) have been evaluated for recent frequency in the American and British BVCs, a standard should be adhered to for inclusion. In the case of the pilot study, the standard for answering the question, "What is an Americanism?" in terms of recent frequency is at least 100 occurrences in the American BVC, and that number of occurrences must be greater than the number of occurrences for the form in the British BVC by at least one order of magnitude.

The new candidates for inclusion that meet the inclusion standard should be saved in a file that reserves them specifically for inclusion. This judgment, as explained above, is based on recent frequency in American and British English, but the original entry forms need to be judged as well. Labels for the extremely low frequency forms should be developed; perhaps only three would be enough: obsolete, rare, and historical. The placement of these labels would assist in the creation of a 21st century view of the *Dictionary*.

Finally, the completed, updated, *Dictionary* must be placed in a package that is accessible to both the researcher and the end-user. One possible format that would yield accessibility to both the researcher and the end-user would be an online format. First, the online format would allow the researcher to publish results periodically; for example, as an active range of forms is updated, the results could be published online. The original text of the *Dictionary* could be featured as a scan; in a parallel frame (beside or below), newly included forms as well as antedatings and frequency-generated labels for the original forms can be shared as well. The

copyright date of the *Dictionary* is 1951, so the text should be in the public domain which would allow free publication of the scanned (original) pages online.

Among currently available electronic resources, the BVC has the largest body of evidence. Also, the BVC is constructed solely of full-text newspapers. Newspapers reflect daily life events, such as those at work, home, school, as well as recreation. The BVC contains a vast array of text types from obituaries to elementary school lunch menus. Such variety and span of 10 years of coverage make the BVC the ideal tool to identify recent frequency of forms in American English. Also, the presence of those forms in the British BVC would be useful evidence in the determination of convergence or divergence of American English from British English. The pilot study toward updating the *Dictionary* employs a 21st century perspective and tools. This study includes decision-making that is based on real use as reflected by the American BVC. The pilot study avoids the subjectivity of personal opinion and depends on the specially designed BVC to support decisions related to inclusion and exclusion. Also, the British BVC can show if a prospective Americanism has reached currency in the British BVC. Further, the the pilot study also uses *ProQuest Historical Newspapers*, *APS Online*, and *NewspaperARCHIVE.com* to antedate historical quotations. Thus, the pilot study uses the BVC for a study of recent frequency, and the pilot study uses the online databases mentioned above to discover antedatings of extant historical quotations. These resources, the BVC and these electronic historical databases, constitute a methodology that addresses the recent and historical careers of words.

Pilot Study

The present study selected a range of entries in the *Dictionary* that begin with the letter N. The pilot range begins with the original entry form *Niagara* and concludes with the original entry form *nickelodeon*. The complete range of N entries has been recently updated in the *OED Online* project, so the presence of highly useful evidence from the *OED* is just as relevant today as it was when the *Dictionary* was being compiled. A scanned image of the pilot pages as well as scans of some other relevant pages from *A Dictionary of Americanisms on Historical Principles* can be found in Appendix C.

All of the forms in the entry range of the pilot study (*Niagara* - *nickelodeon*) were tested in the American BVC to determine their frequency between 1998 and 2007. Table 6.1 shows an overview of the pilot range within the *Dictionary*.

Table 6.1: Overview of original entries in the pilot study

Original entry form	Number of senses and/or Collocations
Niagara n.	9
nibbler n.	1
Nicholas n.	1
Nicholite n.	1
Nicholson n.	1
nick n.	1
nick n.	1
nickel n.	13
nickelodeon n.	5

Antedating of Illustrative Quotations

The study initially endeavored to antedate the evidence (illustrative quotations) that accompany the *Dictionary*'s entries in the pilot study range. In Tables 6.2 through 6.6, original entry forms from the pilot study range are featured, and the three columns to the right include the *Dictionary*'s earliest illustrative quotation; the earliest illustrative quotation for the same form from the *OED*; and the date of an independently obtained antedating in the event that the study could obtain such. The study used two electronic databases within the University of Georgia Libraries, *Proquest Historical Newspapers*, and *APS Online* as tools to search for antedatings. The study also used the public, commercial Web-based database, *NewspaperARCHIVE.com*, to search for antedatings as well. If the form in question could not be located in the *OED*, then Not included appears in the appropriate cell; if an antedating for the earliest *Dictionary* or *OED* form could not be independently located, then None appears in the appropriate cell. Also, in a few cases, the *Dictionary* presents entry forms but does not include any relevant illustrative quotations; in this case, None appears in the appropriate cell. Finally, the year of the earliest evidence for each form is indicated in bold font. Table 6.2 includes antedatings for original *Niagara* forms from the *Dictionary*.

Table 6.2: Antedatings of original *Niagara* forms

Original entry form	*DA*	*OED*	Other sources
Niagara of X n.	1843	**1841**	None
Niagaras n. 'curls'	1865	**1864**	None
Niagara n. 'grape'	1884	Not included	**1882**
Niagara cane n.	**1891**	Not included	None
Niagara green n.	1901	Not included	**1888**
Niagara gudgeon n.	**1842**	Not included	None
Niagara limestone n.	**1862**	Not included	None
Niagara shale n.	**1878**	Not included	None
Niagara thyme n.	**1843**	Not included	None

In the matter of the nine *Niagara* forms in Table 6.2, the study antedated two forms independently (*Niagara* 'grape' and *Niagara green*) and two through evidence in the *OED*. The *Dictionary* includes this 1884 quotation as its earliest evidence for *Niagara* as a variety of grape: "Concord is still far in the lead, though Worden has many friends as the 'coming' black, and Niagara is being most largely tested among the whites" (1129). The study located this 1882 occurrence for *Niagara* from the *Iowa State Reporter*: "We have heard so much about the two New White Grapes of the North, Niagara and Pocklington, that our curiosity was greatly aroused" (*NewspaperARCHIVE.com*). Both of these quotations seem to place the *Niagara* grape in a context of newness.

The *Dictionary* includes this 1901 quotation for *Niagara green*: "Running through the whole plan from the deeper barbaric primary colors to the delicate blue on the propylaea there greets you everywhere at intervals the Niagara green" (1130). The definition that the *Dictionary*

presents is: ‘bluish-green’ (1130). The 1901 quotation, which is the only illustrative quotation for *Niagara green* that the *Dictionary* includes, refers to specifically to a natural context in the vicinity of Niagara Falls; the definition refers generically to a color. The study located this 1888 quotation for *Niagara green* from *The Boston Sunday Globe*: “She is dressed in a rich robe of Niagara green crepe; lined with silk to match, interlined with down, and heavily embroidered in vines of pink rosebuds” (*NewspaperARCHIVE.com*). The study regards this quotation as an antedating even though the context is that of clothing and not Niagara Falls itself. Table 6.3 includes antedatings for original *Nibbler*, *Nicholite*, *and Nicholson* forms.

Table 6.3: Antedatings of original *Nibbler*, *Nicholite*, *and Nicholson* forms

Original entry form	*DA*	*OED*	Other sources
Nibbler n. ‘cunner;’ ‘bergall’	**1842**	1842	None
Nicholite n.	**1786**	1786	None
Nicholson n.	1870	Not included	**1864**

Of the three forms, *Nibbler*, *Nicholite*, and *Nicholson* in Table 6.3, the study was only able to antedate one form, *Nicholson*. The *Dictionary* includes this 1870 quotation for its earliest illustrative quotation for *Nicholson*: “In St. Louis I observed some streets floored with iron gratings, others macadamized, and others paved with wooden bricks laid on a floor of sanded planks, and cemented with asphalt. This is called the Nicholson Pavement, and is found in New York, Chicago, and other cities as well as St. Louis” (1130). The study found this illustrative quotation in an 1864 *Dawson’s Fort Wayne Daily Times* (Indiana) article that includes *Nicholson pavement*: “FIRE!—On yesterday evening, a kettle of pitch, which the workmen were using in laying the Nicholson pavement on Columbia street, caught fire” (*NewspaperARCHIVE.com*).

Interestingly, the 1870 illustration is self-defining—an indication of the form's newness; however, the 1864 occurrence does not self-define. Perhaps the form *Nicholson pavement* was already in use conversationally in Fort Wayne at the time of publication in 1864, or the variable of authorial discretion could have prevented the definition of *Nicholson pavement* from being shared. Table 6.4 includes antedatings of original *nick* forms.

Table 6.4: Antedatings of original *nick* forms

Original entry form	*DA*	*OED*	Other sources
nick n. 'breeding process'	1889	**1824**	None
nick n. 'short for nickel'	**1857**	Not included	None

Table 6.4 shows that for the two senses of *nick* included in the *Dictionary*, the *OED* provides an antedating for 'breeding process;' however, *nick* 'short for nickel' is not included in the *OED* and could not be antedated independently. Table 6.5 includes antedatings of original *nickel* forms.

Table 6.5: Antedatings of original *nickel* forms

Original entry form	*DA*	*OED*	Other sources
nickel bank n.	**1888**	Not included	None
nickel cent n.	1863	Not included	**1860**
nickel-in-the-slot machine n.	1893	Not included	**1889**
nickel novel n.	1896	1930	**1880**
nickel nurser n.	1924	**1916**	None
nickel show n.	1914	1954	**1870**
nickel theater n.	1912	Not included	**1908**
five cent nickel n.	None	Not included	**1874**
liberty head nickel n.	None	Not included	**1927**
plugged nickel n.	None	**1883**	None
nickel-in-the-slot scheme n.	1889	1889	**1888**
nickel slot-machine n.	1947	Not included	**1892**

Of the twelve entry forms in Table 6.5, the study antedated nine. These nine forms are *nickel cent*, *nickel-in-the-slot machine*, *nickel novel*, *nickel show*, *nickel theater*, *five cent nickel*, *liberty head nickel*, *nickel-in-the-slot scheme*, and *nickel slot-machine*.

The *Dictionary* includes this quotation from 1863 for *nickel cent*: "I shall by and by throw you a paltry nickel cent for your tropical dreams" (1130). The following antedating refers to the practice of using coinage as a good luck charm for fishing, and the form in the quotation is labeled as "new." This 1860 quotation comes from the *Christian Inquirer*: "Among them was one clipped so as to form a cross, and there were also some of the new coinage—the nickel cent—showing that the practice is still continued" (*APS Online*).

The *Dictionary* includes this 1893 illustration for *nickel-in-the-slot machine*: "(In Jacksonville) there were the same . . . nickel-in-the-slot machines (as in Asbury Park)" (1130). The study located this 1889 antedating from the *San Antonio Daily Light* (Texas): "A nickel-in-the-slot machine is bedevised [sic] to take instantaneous photographs" (*NewspaperARCHIVE.com*). One interesting attribute of this antedating is that this illustration identifies a specific application for the machine.

This 1896 quotation appears in the *Dictionary* to illustrate *nickel novel*: "Pistols and bandits abound in a nickel-novel atmosphere" (1130). This illustration uses the form as an expanding compound noun as *nickel novel* expands in this quotation to include *atmosphere*. The 1880 antedating illustration from the *Janesville Daily Gazette* (Wisconsin) includes both the entry form, *nickel novel* as well as *5-cent novel*: "A 5-cent novel, entitled 'Mark the Fearless; or, Hoeing His Own Row.' Another nickel novel, entitled 'Brooding Thrust; or, the Hermit of the Hills.'" (*NewspaperARCHIVE.com*).

The *Dictionary* uses this 1914 quotation to illustrate *nickel show*: "Ragged and dirty children attending 'nickel shows' and buying quantities of cheap candy" (1130). The study antedated this illustration of *nickel show* with an 1895 illustration from the *Wellsboro Agitator* (Pennsylvania): "It's the best nickel show you'll; ever get; and don't you forget it when he passes the hat" (*NewspaperARCHIVE.com*). These two quotations work together in both reflecting moments almost a decade apart from each other. Also, the 1914 illustration reports the habits of children, and the 1895 antedating is an admonition to an apparently adult audience. The study frequently encountered false results in the antedating process and initially selected the next quotation as a worthy antedating for *nickel show*. From 1870, this quotation in *The Democratic Pharos* (Indiana) apparently includes the entry form and follows the form with a definition that

shows it is not worthy for use as an antedating: "'But, as a general thing, the city fellows grab everything, and don't leave an outsider even a nickel show.' 'And what do you mean by a 'nickel show?' 'I mean that they don't leave a fellow even a nickel to show for his share of the damages.'" (*NewspaperARCHIVE.com*). This quotation is an interesting coinage, but this use shows the need for a researcher to be cautious and precise in gathering antedatings.

The *Dictionary* shares this 1912 quotation for *nickel theater*: "We have had . . . the excitement provided by the 'nickel theater'" (1130). The study found this antedating for *nickel theater* from 1907 in the *Fort Wayne Sentinel* (Indiana) under a small headline of "Moving Pictures": "The nickel theater has a legitimate place in the life of the people" (*NewspaperARCHIVE.com*). Both of these quotations are useful in showing the connection between *nickel theater* and the general public.

The *Dictionary* includes theses two forms in reduced font size without definitions or illustrative quotations: *five cent nickel*, *liberty head nickel*. This 1874 quotation for *five cent nickel* comes from the *Galveston News* (Texas): "The reason appears to have been that it would, on the resumption of specie payment, be likely to expel from circulation and drive to the treasury for redemption the five cent nickel copper coins" (*NewspaperARCHIVE.com*). The *OED* does not provide this form as an entry, so no possible antedating is available from that source. Similarly, neither the *Dictionary* nor the *OED* provide illustrations for *liberty head nickel*. The earliest illustration the study could locate for *liberty head nickel* comes from a 1927 advertisement from a collector in the *San Antonio Light* (Texas): "Offer $100 for 1894 dime S. mint; $ 50 for 1913 Liberty Head Nickel (not Buffalo) and hundreds of other amazing prices for coins" (*NewspaperARCHIVE.com*). This quotation reflects an historical context and not the

current contextualization that is a hallmark of evidence included in the *Dictionary*. Still, this quotation reflects the earliest use of the form that the study could find.

The *Dictionary* uses this 1889 quotation to illustrate *nickel in the slot scheme*: "The latest nickel-in-the-slot scheme is really a stroke of genius and is destined to revolutionize cheap literature in this country" (1130). The study located an 1888 illustration from the *Salt Lake Daily Tribune* (Utah) for *nickel-in-the-slot scheme*: "Its aim is to provide theater-goers with opera glasses upon the 'drop a nickel in the slot' scheme (*NewspaperARCHIVE.com*). Each of these illustrations provides a different context for the application of nickel-vending in the late 1800's in the United States. From a linguistic viewpoint, the uses of the entry form in these two illustrations is different. In the antedating, the form functions as part of a verb phrase that builds on the verb *drop* while in the original illustration, the form functions independently as a noun phrase. Still, the study feels that the phrase in the antedating illustrates the form and meaning of the original entry form.

The *Dictionary* features this 1947 illustration for *nickel slot machine*: "In fact, my $1.10 is sitting in a nickel slot-machine there" (1130). The study located this antedating from 1892 in *The Daily Northwestern* (Wisconsin): "I will say to them in conclusion that the people of this great country are going to have a happy and a prosperous new year in spite of the democratic party, the gentleman from Indiana (Mr. Dolman) and his nickel slot machine" (*NewspaperARCHIVE.com*). The original quotation illustrates a moment from a daily life experience, and the antedating is connected to a political dialogue; the antedating is followed immediately by "(Applause on the republican side)" (*NewspaperARCHIVE.com*). Table 6.6 includes antedatings of original *nickelodeon* forms.

Table 6.6: Antedatings of original *nickelodeon* forms

Original entry form	*DA*	*OED*	Other sources
nickelodeon n. 'theater'	**1888**	1888	None
nickelodeon n. 'place of amusement'	1913	Not included	**1911**
nickelodeon n. 'juke box'	**1938**	1938	None
nickelodeon machine n. 'video machine'	**1944**	Not included	None

For *nickelodeon* 'place of amusement' the *Dictionary* includes this 1913 illustration: " . . . a place of amusement generally charging no admission fee, containing various automatic machines, such as cinematographs, graphophones, etc., which may be used by patrons for a small charge" (1130). The study located an antedating in the headline of a 1911 *Washington Post* article that explains how sideshow entertainers were losing their ability to preserve a high salary because of the inexpensive film viewing that *nickelodeons* provided: "The Fat Woman, the Human Skeleton, the Dog Face Boy, and the Circassian Beauty Meet a Waterloo in the numerous Nickelodeons" (*ProQuest Historical Newspapers*). As mentioned above, these two illustrations work together; the original quotation defines this sense of *nickelodeon*, and the antedating illustrates one highly specific effect the technology had on one group within American culture.

The sources for the antedatings in the pilot study include the *OED* and these online databases: *APS Online*, *Proquest Historical Newspapers*, and *NewspaperARCHIVE.com*. For the 30 entry forms that the study sought antedatings for, the *OED* provided antedatings for 5 entry forms. These antedatings are a significant contribution to the pilot study, but the *OED* must be augmented by additional electronic resources, such as *APS Online*, *Proquest Historical Newspapers*, and *NewspaperARCHIVE.com* in order to obtain relevant antedatings.

The additional electronic databases as a group provided 13 antedatings for the study. The importance of these resources cannot be emphasized enough. They have provided original research that the *OED* could not. Further, these databases can be found quickly online, and they are simple to use.

The *OED* is a powerful research tool, but for the research and antedatings of Americanisms, the *OED* is not a complete tool. The fine-grained nature of the research of Americanisms requires consultation of original, American texts. The electronic databases mentioned above provide those texts and, consequently, the support necessary for a project that has a uniquely American perspective. Table 6.7 shows how the *OED* functioned in support of the pilot study.

Table 6.7: Support from the *OED* for the pilot study

Original entry form	Included in the *OED*	The Pilot Study antedated the *OED*'s illustration
Niagara n. 'grape'	No	
Niagara cane n.	No	
Niagara green n.	No	
Niagara gudgeon n.	No	
Niagara limestone n.	No	
Niagara shale n.	No	
Niagara thyme n.	No	
Nicholson n.	No	
nick n. 'short for nickel'	No	
nickel bank n.	No	
nickel cent n.	No	
nickel-in-the-slot machine n.	No	
nickel novel n.		Yes
nickel show n.		Yes
nickel theater n.	No	
five cent nickel n.	No	
liberty head nickel n.	No	
nickel-in-the-slot scheme n.		Yes
nickel slot-machine n.	No	
nickelodeon n. 'place of amusement'	No	
nickelodeon machine n. 'video machine'	No	

Conclusion

The study pursued antedatings for 30 entry forms from the *Dictionary*. The *OED* provided 5 antedatings, and through the online databases mentioned above, the study located 13 antedatings. The antedatings are important part of the pilot study because their evidence increases the illustrative evidence for those 18 entries. The antedatings frequently broaden the historical picture of the entry form's use as the antedatings have frequently provided situations and contexts that the original, earliest illustrations do not. Still, the original earliest quotations and the antedatings work together in illustrating the entry form's career.

BVC as Source of Evidence for Original Entry Forms

Through a direct search of *LexisNexis Academic*, no evidence in the American BVC could be found for these forms: *Nibbler*, *Nicholite*, *Nicholson*, *nick* 'breeding process', and *nick* 'short for nickel.' In order to strategically identify the presence and careers of original entry forms in the American and British BVCs, the study constructed three strip corpora derived from each BVC. The first corpus consists of search results for the search term *Niagara*. The second corpus consists of search results for the term *nickel*. The third corpus consists of search results for the term *nickelodeon*. The corpora files that were derived from the American BVC, were stored in directories according to region, so geography could be considered as a factor if needed. Table 6.8 shows which original entry forms in the *Dictionary* are supported by BVC evidence.

Table 6.8: Evidence from the American BVC for original entries

Original entry form	Evidence for an original entry form or collocation in the American BVC
Niagara n.	Yes
Nibbler n.	No
Nicholas n.	Yes
Nicholite n.	No
Nicholson n.	No
nick n. 'breeding process'	No
nick n. 'short for nickel'	No
nickel n.	Yes
nickelodeon n.	Yes

Table 6.8 shows that the American BVC provides evidence for four of the nine entry forms, *Niagara*, *Nicholas*, *nickel*, and *nickelodeon*. In the matter of *Nicholas,* the study did not pursue further evidence because the *Dictionary* notes *Nicholas* as the second part of the form *Saint Nicholas* which the *Dictionary* notes is not of American origin (1130). Tables 6.9 through 6.12 share BVC occurrences for *Niagara*, *nickel*, and *nickelodeon*.

Niagara

The study did not pursue evidence for the presence of *Niagara* as a place or landmark name in the BVC; however, such uses were clearly evident as was the presence of other institutional uses for *Niagara*, such as *Niagara University*. Table 6.9 shows occurrences of original *Niagara* entry forms in the BVC. Italicized numbers beside the number of occurrences in this and subsequent tables is the rate per ten million words for the number of occurrences in its respective BVC.

Table 6.9: Occurrences of original *Niagara* forms

Original entry form	Am BVC	Br BVC
Niagara of X n.	112 *.216*	38 *.365*
Niagaras n. 'curls'	0	0
Niagara n. 'grape'	3 *.005*	0
Niagara cane n.	0	0
Niagara green n.	1 *.001*	0
Niagara gudgeon n.	0	0
Niagara limestone n.	0	0
Niagara shale n.	0	0
Niagara thyme n.	0	0

The study pursued evidence for the presence of what the *Dictionary* refers to as "transferred uses" of *Niagara* (1129). The *Dictionary* provides evidence for the construction, *Niagara of* X, in which X is an item that appears or flows in large quantities. The *Dictionary* provides these examples: "Niagara of red silk;" "Niagaras of lager beer;" and "Niagaras of hot water" (1129). The Niagara corpus shows that, in the R1 position of *Niagara of* (combined with the plural form) the two most frequent forms (*information* and *words*) have 5 occurrences apiece; the rest of the forms in the R1 position have fewer than 5 occurrences apiece.

Niagara of, has 98 occurrences, and the following are five examples of low frequency R1 forms (number of occurrences in parentheses): *ideas* (3), *cash* (2), *hype* (1), *laundered drug money* (1), and *misfortune* (1). The plural *Niagaras of* is a much lower frequency form with 14 occurrences and includes these five examples in the R1 position: *sound* (2), *cascading sound* (1), *octaves* (1), *private tears* (1), and *velvety aromatic chocolate* (1). The low frequency R1 collocates for both the singular and plural form cover a wide range of semantic fields. *Ideas* with 3 occurrences is semantically similar to *words* and *information* that have 5 occurrences apiece. Still, note the semantic variation that exists among the R1 forms that represent music, personal emotion, fine dining, and money. Within the BVC, *Niagara of* is a low frequency form, but its presence is an important testimony that an Americanism from the *Dictionary* is still in use in American English to some extent.

The next use of *Niagara* that the *Dictionary* presents is *Niagaras* in reference to a woman's hair—especially curls. The *Dictionary* labels this form with *obs.* or obsolete, so this label indicates that the editor did not feel that the form was current around the time of the *Dictionary*'s publication in 1951. In the Niagara corpus, no results occurred for the search terms *Niagara curls* or just *curls*; however, one result occurred for *hair*. This quotation from 2000 in

The New York Times preserves the meaning of *Niagaras*, in this context, but the lexical use is less direct as the construction depends on *like*: "hair cascading like Niagara Falls down her shoulder" (*LexisNexis Academic*). This example from 2000 is useful because it connects the metaphor of *Niagara Falls* with a woman's hair, but this example uses *Niagara Falls* in a simile, and the *Dictionary* provides this illustrative quotation from 1865: "An elastic pipe must have passed through one of the 'Niagaras' or 'cataract curls'—the name given to the shower of true or false ringlets the ladies are in the habit of wearing at the back of their heads" (1129). In the case of the use of *Niagaras* in reference to a woman's hairstyle, the BVC provides no evidence; however, the one example of a related use of a *Niagara Falls* simile shows that the metaphor still exists—albeit at an extremely low frequency in the BVC.

The *Dictionary* presents evidence for *Niagara* as the name of a grape. The study found the low frequency form, *Niagara grape* with 36 occurrences (singular and plural combined) in the Niagara corpus. Again this example is a low frequency form, but its presence allows the BVC to demonstrate its ability to provide fine-grained detail about the use of a low frequency form. For both the singular and plural forms, the Northeast region is the source of most of the forms' occurrences with 8 of 11 occurrences for the singular form and 16 of 25 for the plural form. The singular form has three more occurrences—2 in the Coastal west and 1 in the West. The plural form has 7 occurrences in the Southeast and 2 in the Coastal west. Also, the form *Niagara green grapes* occurs in one article in the *Charleston Gazette* in 2005, and that title repeated the same line in another article one week later. So, for the name of a grape, the BVC presents evidence for *Niagara grape*, *Niagara grapes*, and *Niagara green grapes*; however, each

of the three illustrative quotations in the *Dictionary* employs the lone form *Niagara* as a term in reference to a kind of grape. For example, the *Dictionary* shares this quotation from 1945: "Grapes: Concord, Fredonia, Wordon, Golden Muscat, Niagara, Portland are all hardy" (1130).

In response to the *Dictionary*'s use of *Niagara* in the illustrative quotations, the study attempted to locate occurrences of *Niagara* as a single, non-collocated lexical form that represents a kind of grape. The study created a concordance of the form *Niagaras* (which has 34 occurrences in the BVC) and three of the occurrences refer to a type of grape. One of the references from 2001 in the *Charleston Gazette*: "Ohio and, soon, green Niagaras from the Amish country too" employs the collocate *green* that was mentioned above (*LexisNexis Academic*). Another line includes, "small bunches of fresh Niagaras" (*LexisNexis Academic*); and finally, "Their Delawares and Niagaras are in a separate" (*LexisNexis Academic*). The brief text strip in the last quotation is slightly vague, but the study readily included it as a reference to a grape because *Pittsburgh Post-Gazette*'s article's title (from 2003) is: "A Grape Escape; Finger Lakes Wineries Offer An Intoxicating Getaway" (*LexisNexis Academic*).

In the Niagara corpus, the form *niagara* appears 15,945 times. This mass of occurrences for the singular form makes isolation of uses of *Niagara* that represent a grape extremely difficult. The study created a concordance for the form *grape*, and with the 32 results, the form *Niagara* in reference to a grape occurs 7 times. Three of the occurrences are very similar to the *Dictionary*'s illustrative quotations: "grape varieties such as Concord and Niagara, or French Hybrid varieties" from the *Dayton Daily News* in 1998 (*LexisNexis Academic*); again in the same title in 2002: "grape varieties such as Catawba, Niagara and Concord" (*LexisNexis Academic*); and from 2005 in the *Pittsburgh Post-Gazette*: three (grape) varieties -- Niagara, Fredonia and Diamond" (*LexisNexis Academic*).

The remaining occurrences include the following lines that appeared on successive days in the *Oregonian*: “grape crushing with a Niagara variety” (*LexisNexis Academic*); “the green gold Niagara” (*LexisNexis Academic*); and “one sip of Niagara” (*LexisNexis Academic*). The “sip of Niagara” evidently refers to the wine made from the *Niagara grape*; the title of the article is, “A Berry Good Start to Wine” (*LexisNexis Academic*).

The final *Niagara* forms that the *Dictionary* presents are the following collocations: *Niagara cane*, *Niagara green*, *Niagara gudgeon*, *Niagara limestone*, and *Niagara shale* (1130). *Niagara green* is the only collocation from the five *Niagara* collocations above that the study found evidence for in the Niagara corpus, and the single occurrence comes from the *Chicago Sun-Times* in 2001: “‘Niagara Green’ is the color water from the Niagara River turns as it rushes over Niagara Falls” (*LexisNexis Academic*). This term is perhaps largely confined to conversations around the immediate mileu of Niagara Falls; note, for example, that the title of the quotation’s speaker is Niagara Falls historian (*LexisNexis Academic*). In the case of *Niagara green*, the BVC provides little evidence; however, this single occurrence could be significant because that form may only rarely be used in print. Also, this quotation connects directly with the 1901 illustration in the *Dictionary* that refers directly to Niagara Falls.

nickel

Table 6.10 shows BVC evidence for occurrences for *nickel* collocations from the *Dictionary*. The BVC provides many occurrences for *nickel*, ‘five U.S. cents’ and ‘five cent U.S. coin.’ These occurrences range from practical discussion of tax increases to a new U.S.

minting in 2003. *Nickel* also has occurrences in the BVC as a surname and appears in a variety of institutional names, such as the band *Nickel Creek* and a jazz album whose title is *Live at the Plugged Nickel*.

Table 6.10: Occurrences of original *nickel* forms

Original entry form	Am BVC	Br BVC
nickel bank n.	0	0
nickel cent n.	0	0
nickel-in-the-slot machine n.	3 *.005*	0
nickel novel n.	0	0
nickel nurser n.	0	0
nickel show n.	0	0
nickel theater n.	0	0
five cent nickel n.	3 *.005*	0
liberty head nickel n.	60 *.115*	2 *.019*
plugged nickel n.	47 *.090*	4 *.038*
nickel-in-the-slot scheme n.	0	0
nickel slot-machine n.	0	0

Table 6.11 shows that *nickel-in-the-slot machine* has 3 occurrences. This line from 2006 appears to be a historical reference: “coinage.The jukebox debuted as the Nickel-in-the-Slot machine” (*LexisNexis Academic*). The next occurrence from the *Washington Post* in 2003 also appears to be historical: “The nickel-in-the-slot machine has been” (*LexisNexis Academic*); that line is followed by: “more or less filled with nickels of the unfortunates who had called”

(*LexisNexis Academic*). The headline for the article is: “PAST POST: 1894; Big Payoff From a Slot Machine” (*LexisNexis Academic*). The last occurrence for this form also appears to be historical: “so many. ‘You forget the nickel-in-the-slot machine.’” (*LexisNexis Academic*). That line from a 1999 *Telegraph Herald* article is preceded by this metadata: “Telegraph Herald, circa 1899” (*LexisNexis Academic*). All of the uses of this low frequency form appear to be historical. One matter of note is that the 2006 occurrence refers to a jukebox, the 2003 occurrence clearly refers to a gambling device, and the final occurrence’s reference is ambiguous.

The BVC actually has one occurrence for *nickel show*; this occurrence does not seem to be historical, but its use is not clear: “deflect cigarette smoke. Nickel show: Here's a good” (*LexisNexis Academic*). The article’s metadata cites the author’s “Las Vegas Advisor monthly Newsletter” (*LexisNexis Academic*). This use of *nickel show* appears to be associated with a gambling environment, but whether the author is referring to an admission fee is unclear.

The form *five cent nickel* has 3 occurrences in the BVC and they appear to be historical. Two appear to be an almanac quotation; the other includes historical minting information for the earliest *five cent nickel* (*LexisNexis Academic*).

The form *liberty head nickel* has 60 occurrences. These occurrences are generally in articles that refer to collecting. This headline overviews one event that motivated the publication of several of the articles which include *liberty head nickel*: “$1 million offered for elusive nickel” (*LexisNexis Academic*). One line from that 2003 article notes, “The Liberty Head Nickel was minted from 1883 to 1912, when” (*LexisNexis Academic*). These articles generally use this coin’s name in terms of collecting and appraisal, so these uses are contemporary, but they are framed in a historical sense because the coin itself is a historical artifact. The British BVC has 2

occurrences for *liberty head nickel*, and these occurrences report the million dollar offer mentioned above.

Plugged nickel has 47 occurrences. Additionally, 19 separate occurrences are for Jazz music related references—mostly one album, *Live at the Plugged Nickel* (*LexisNexis Academic*). The 47 occurrences tend to highlight lack of worth, such as this 2003 headline: "'Joe Millionaire': Not Worth a Plugged Nickel" (*LexisNexis Academic*). These occurrences do not appear to have the historical character of, for example, *liberty head nickel*. These occurrences seem to reflect daily American life situations from local utility disputes to sports gambling (*LexisNexis Academic*). The British BVC has 4 occurrences for *plugged nickel*, and these uses highlight a *plugged nickel* as a worthless item, such as this line from a 2005 article titled, "New Labour needn't lose more sight over a scandal that's old hat": "scandals go, this was plugged nickel and wooden dime. Even the" (*LexisNexis Academic*).

The illustrative quotation for *nickel slot machine* in the *Dictionary* does not clearly highlight the form as a gambling device; *nickel slot machine* has 35 occurrences in the American BVC. Of those 35 occurrences, 26 clearly refer to a gambling device (those occurrences are treated in Table 6.10), and 9 refer to a machine but the application or function of the machine is ambiguous—just as the *Dictionary*'s illustration is ambiguous: "In fact, my $1.10 is sitting in a nickel slot-machine there" (1130). Because the *Dictionary*'s sense for *nickel slot-machine* is unclear, and the 9 occurrences mentioned above are also unclear, the study recorded zero results for *nickel slot machine* in Table 6.10.

nickelodeon

The American BVC has 183 lines of text that include the lower case form, *nickelodeon.* The American Nickelodeon corpus has 10,976 total occurrences for *nickelodeon*, but the majority are uses in the sense of the television network. The study filtered out the majority of these uses by searching for lower case forms only. Many of the *nickelodeon* forms in the 183 lines mentioned above are framed in nostalgia, history, and recollection. Still, the problem remains of whether these uses refer to 'theater;' 'place of amusement;' 'juke box;' or 'video machine'. The references are often unclear, for example: "live bands replacing the nickelodeon, Messer said. We held" (*LexisNexis Academic*). This line does not clarify between a juke box or a player piano or possibly even another item. Table 6.11 shows BVC evidence for original *nickelodeon* forms.

Table 6.11: Occurrences of original *nickelodeon* forms

Original entry form	Am BVC	Br BVC
nickelodeon n. 'theater'	25 *.048*	2 *.019*
nickelodeon n. 'place of amusement'	0	0
nickelodeon n. 'juke box'	0	0
nickelodeon machine n. 'video machine'	0	0

Within the 183 lines mentioned above, the study located 25 occurrences of *nickelodeon* within 5 forms to the left or right of *theater*, *film*, or *movie*. These results were inspected, and they appeared to fit semantically into the use of *nickelodeon* as 'theater.'

The study also found two occurrences that appeared to be in the sense of *nickelodeon* as 'juke box,' but the lines are not completely clear, so the study recorded zero occurrences for this

sense; for example note this quotation from 1998: “downtown Tulsa with a nickelodeon providing the music;” and from 2000, “got music from a nickelodeon. Remember now? Remember” (*LexisNexis Academic)*. These unclear occurrences highlight the ambiguity of many of the 183 occurrences of *nickelodeon* that the study encountered; nevertheless, even with more clarification *nickelodeon*, beyond the name of the television network, is a low frequency form whose current career is largely rooted in historical reference. Table 6.12 shows the original *Dictionary* forms that have zero results in the BVC.

Table 6.12: Original *Dictionary* forms with zero BVC occurrences

Niagaras n. ‘curls’	*nickel bank* n.
Niagara cane n.	*nickel cent* n.
Niagara gudgeon n.	*nickel novel* n.
Niagara limestone n.	*nickel nurser* n.
Niagara shale n.	*nickel show* n.
Niagara thyme n.	*nickel theater* n.
Nibbler n. ‘cunner;’ ‘bergall’	*nickel-in-the-slot scheme* n.
Nicholite n.	*nickel slot-machine* n.
Nicholson n.	*nickelodeon* n. ‘place of amusement’
nick n. ‘breeding process’	*nickelodeon* n. ‘juke box’
nick n. ‘short for nickel’	*nickelodeon machine* n. ‘video machine’

Table 6.12 shows that 22 of the 30 original entry forms have zero results in the American BVC. Because of the contemporary nature of the evidence in the BVC, this large number of entry forms with zero occurrences indicates that the entry forms within the pilot study generally reflect historical use. From a 21st century perspective, this collection of zero evidence gives

credence to the likelihood that the *Dictionary* is largely a historical work. The BVC is a witness to the lack of use for these 22 entry forms, but the BVC can also provide new forms and senses for the pilot study.

BVC as Source of Evidence for New Senses and Collocations

The study pursued new forms and collocations by observing results of searches for the original entry forms, and the study also created two additional derived corpora for the purpose of locating new forms. Table 6.13 shows the entry forms for which new senses or new collocations were located in the American BVC.

Table 6.13: Evidence from the American BVC for new senses or collocations

Original entry form	Evidence from the BVC for new senses or collocations
Niagara n.	Yes
nibbler n.	Yes
Nicholas n.	No
Nicholite n.	No
Nicholson n.	No
nick n. 'breeding process'	No
nick n. 'short for nickel'	No
nickel n.	Yes
nickelodeon n.	No

The first corpus of the two additional derived corpora is based on the form *nia* and was constructed with the motivation for finding words that begin with *nia*, excluding *Niagara* forms

in the BVC. This corpus did not produce any candidates for inclusion as new words, but the corpus did produce a variety of acronymical organization names (that likely incorporate *National* in their titles) as well as place and personal names.

The second corpus is based on the root *nick*, but the search string for *LexisNexis Academic* is extremely long (16 lines) and excludes many surnames. Further, the corpus also excludes the root form, so the personal name, *Nick*, would be excluded. The Nick corpus brings up the sometimes challenging nature of the BVC. While the BVC's massive size is a distinct plus, that size can create baffling complications. For example, the *nick* search string for *LexisNexis Academic* begins with: "nick! and not nick and not nickel and not nickelodeon and not nicklaus." As mentioned above, the majority of the "and not" elements are surnames, and some of them are high frequency forms. The downloading of files, had all of the results for the form, *nick*, been included, would have been exhausting and wasteful in terms of time and return on effort. The only high frequency form that resulted from this corpus was *nicked*, which is the past participle of the well-documented verb in American English, *nick*.

The exclusion of the form *nick* from this corpus is both an achievement (*LexisNexis Academic* executed a useful search on the root form without including the root form) and a handicap. The study initially needed to explore the use of *nick* in the two senses of 'crossbreeding' and 'a nickel.' The study pursued evidence for these senses of *nick* directly from the BVC as an antidote for *nick*'s being excluded from the derived corpus. The process for investigating the presence of these senses was systematic but simple. Also, even though the notion of creating a derived corpus to be used with WordSmith Tools seems disconnected from the use of the *LexisNexis Academic* search interface, the two processes can be highly connected because complex search strings are sometimes necessary for both the BVC searches for a single

form in the *LexisNexis Academic* interface as well as for the construction of corpora derived from the BVC. Tables 6.14 through 6.17 show the new senses or collocations that the study located through analysis of the BVC and corpora derived from the BVC.

Niagara

In order to locate a possible verb sense for *Niagara*, the study searched the Niagara corpus for *Niagaras* and then arranged the results alphabetically in terms of the R1 form, so possible context for use of *Niagara* as a verb could be identified. The R1 form, from was located and this quotation appeared in consecutive weeks in 2003 in *The New York Times*: "tangerine-colored chiffon that Niagaras from ceiling to floor to form" (*LexisNexis Academic*). The same strategy of monitoring the R1 form of the singular form was employed, but no verb senses for the singular form could be located.

nibbler

After the study noted uses of *nibbler* in reference to people, a Nibbler derived corpus was built from the American BVC. This corpus has 143 occurrences of *nibbler*; many have to do with eating habits and generalizations about taking a small amount of something, but the study felt the lone, significant discovery in the Nibbler corpus is the use of *nibbler* as a label for a baseball pitcher. Of the 143 occurrences for *nibbler*, 23 refer to baseball pitchers. These two quotations from 2003 and 2007 respectively cooperate in making this sense of *nibbler* clear: "After that, he became a different type of pitcher, a nibbler"; and "Boston's lineup will turn the boldest of flamethrowers into an unabashed nibbler" (*LexisNexis Academic*). *Nibbler* is not a high frequency form, but evidence from the BVC helps makes this one aspect of the American experience more understandable for those who are not highly familiar with the game of baseball.

nickel bank

The entry *nickel bank* 'a kind of game in which nickels are used' has an obsolete label, but the study pursued evidence for it as well as any other form labeled obsolete in the study (1130). The study found one occurrence of the form: "50 cents and a nickel bank may never be" (*LexisNexis Academic*). This reference seems to be addressing ways to save money as a form of advice; the headline of the 1999 *Florida Times-Union* article is: "Newspaper urged prudence while giving its money away" (*LexisNexis Academic*). Again here is a form, which is a very low frequency form in the BVC, but the BVC provides evidence for its use. Table 6.14 shows BVC evidence for new senses of *Niagara* and *nibbler* and the new *nickel* form, *nickel bank.*

Table 6.14: Occurrences of general new forms

Form	Definition	Am BVC	Br BVC
Niagara v.	To flow with great volume or force.	2 *.003*	0
nibbler n.	A pitcher who characteristically places pitches in the corner of the strike zone.	23 *.044*	0
nickel bank n.	A savings device.	1 *.001*	0

Nickel forms related to football

As the study pursued research on the form *nickel* in the BVC, references to football quickly became evident. Table 6.15 shows occurrences for 10 forms that build on the form *nickel*, and 7 forms that build on the form *nickel and dime*.

Table 6.15: Occurrences of *nickel* forms related to football

Form	Definition	Am BVC	Br BVC
nickel back n.	A fifth defensive back.	2462 *4.754*	0
nickel package n.	A defensive formation that includes a fifth defensive back.	892 *1.722*	0
nickel defense n.	A defensive formation that includes a fifth defensive back.	774 *1.490*	0
nickel situation n.	Situation that motivates the nickel package.	222 *.428*	0
nickel corner n.	A fifth defensive back.	111 *.214*	0
nickel coverage n.	A defensive formation that includes a fifth defensive back.	79 *.152*	0
nickel defensive back n.	A fifth defensive back.	36 *.069*	0
nickel pass rusher n.	A fifth defensive back.	25 *.048*	0
nickel defensive package n.	A defensive formation that includes a fifth defensive back.	21 *.040*	0
nickel pass defense n.	A defensive formation that includes a fifth defensive back.	14 *.027*	0
nickel and dime package n.	Defensive formations that include a fifth and sixth defensive back.	249 *.480*	0
nickel and dime defense n.	Defensive formations that include a fifth and sixth defensive back.	134 *.258*	0
nickel and dime situation n.	Situation that motivates the nickel And dime packages.	56 *.108*	0
nickel and dime defensive package n.	Defensive formations that include a fifth and sixth defensive back.	45 *.086*	0
nickel and dime coverage n.	Defensive formations that include a fifth and sixth defensive back.	40 *.077*	0
nickel and dime pass coverage n.	Defensive formations that include a fifth and sixth defensive back.	13 *.025*	0
nickel and dime formations n.	Defensive formations that include a fifth and sixth defensive back.	10 *.019*	0

These *nickel* and *nickel and dime* forms are arranged by number of occurrences, but some comments on their semantic associations is appropriate. First, from the first seven forms, four of the forms refer to a player who is a fifth defensive back; these forms with their number of occurrences in parentheses are: *nickel back* (2462); *nickel corner* (111); *nickel defensive back* (36); *nickel pass rusher* (25). The second semantic grouping is five forms that refer to the plan

or strategy that includes the change of the defense to include a fifth back: *nickel package* (892); *nickel defense* (774); *nickel defensive package* (21); *nickel pass defense* (14). Finally, the last form of the first ten refers to the situation that motivates the change to include a fifth defensive back: *nickel situation* (222). The BVC's mass allows for the identification of multiple forms within a single semantic field.

Six of the seven forms that build on *nickel and dime* share the same semantic value; these forms are: *nickel and dime package* (249); *nickel and dime defense* (134); *nickel and dime defensive package* (45); *nickel and dime coverage* (40); *nickel and dime pass coverage* (13); *nickel and dime formation* (10). The last form of these seven is analogous to *nickel situation*; the form is *nickel and dime situation* (56). One contrast between the *nickel* forms and the *nickel and dime* forms is that the *nickel and dime* forms do not include a term for the player as in the *nickel back* in the *nickel* forms.

The *nickel* and *nickel and dime* forms work together because their meanings have an intersection in that the *nickel package* is a part of the *nickel* (package) *and dime package*. Still, note the number of occurrences for *nickel back* (2462) in comparison to those for *nickel and dime packages* (249) which has the most occurrences for the *nickel and dime* forms. The *nickel* football forms are clearly more frequent than the *nickel and dime* forms. Also, it is interesting to note that *nickel and dime* in this football sense does not mean '5 and 10;' rather, the sense is '5 and 6.' *Nickel and dime* will be revisited from another angle in the discussion of Table 6.18.

Nickel forms related to gambling and illegal drugs

Table 6.16 includes eight *nickel* forms, and, as was the case with *nickel* forms that relate to football, these forms have a tendency to collect themselves as groups in a single semantic range. The semantic ranges for these eight forms are 'game,' 'game machine,' and 'game player.'

Table 6.16: Occurrences of *nickel* forms related to gambling

Form	Definition	Am BVC	Br BVC
nickel slots n.	A gaming machine that accepts nickels.	205 *.395*	0
nickel machine n.	A gaming machine that accepts nickels.	123 *.237*	0
nickel slot machine n.	A gaming machine that accepts nickels.	109 *.210*	0
nickel game n.	A game that accepts money in five cent increments.	107 *.206*	0
nickel video slot n.	A gaming machine that accepts nickels.	33 *.063*	0
nickel video game n.	A gaming machine that accepts nickels.	18 *.034*	0
nickel slot player n.	A person who plays nickel slots.	13 *.025*	0
nickel video poker n.	A gaming machine that accepts nickels.	12 *.023*	0

Five of the forms refer to the concept of 'game;' these forms are *nickel slots* (205); *nickel game* (107); *nickel video slot* (33); *nickel video game* (18); and *nickel video poker* (12). Two of the forms in Table 6.10 represent the narrower concept of 'gaming machine' that accepts nickels; these forms and their number of occurrences are: *nickel machine* (123) and *nickel slot machine* (109). The remaining form is solitary in its field 'game player,' *nickel slot player* (13). Note that the majority of the evidence for *nickel* forms related to gambling refer to the concepts of a device or a particular game played on such a device, and only *nickel slot player* (13) refers to a

person who gambles with these devices and games. Table 6.17 includes occurrences of a *nickel* form related to illegal drug use.

Table 6.17: Occurrences of a *nickel* form related to illegal drug use

Form	Definition	Am BVC	Br BVC
nickel bag n.	Five dollar's worth of an illegal drug.	80 *.154*	1 *.009*

One matter of note with regard to Table 6.17 is that *nickel bag* in the search results can contain a variety of illegal substances, such as heroine, marijuana, or cocaine (*LexisNexis Academic*). Also in the case of the lone British occurrence, that form occurred in an article that was a 1999 review of a book that was set in a U.S. city.

Evidence from other Dictionaries

After using the BVC to independently locate additional forms in the pilot range of *Niagara – nickelodeon*, the study consulted the *OED Online* (*OED*), *The American Heritage Dictionary* (Fourth edition) (*AHD*), and *The New Oxford American Dictionary* (*NOAD*) to find additional, prospective Americanisms to test in the BVC for possible inclusion in the update of the *Dictionary*. Tables 6.17 through 6.19 include forms that were obtained from these dictionaries and their occurrences in the American and British BVCs.

Table 6.18: Occurrences of *nickel and dime* forms

Form	Definition	Am BVC	Br BVC
nickel and dime adj.	Insignificant or substandard.	103 *.198*	12 *.115*
nickel and dime v.	To exact money steadily in small increments.	570 *1.100*	4 *.038*
nickel and dime v.	To gain yardage slowly in a football game.	14 *.027*	0

All three dictionaries offer some entries for *nickel and dime*, but the *OED* provides the most variety, including the football sense for *nickel and dime*. The study used the American Nickel corpus and WordSmith Tools to identify these forms. Some functions of this software that were especially valuable include case sensitive searching and isolation of the node's (form of *nickel and dime*'s) R1 form as a strategy to identify the node's part of speech. Even though WordSmith Tools assisted this process enormously, a large amount of time and concentration were still necessary to connect the nodes with their meanings as shown in Table 6.18.

The case sensitive search option allowed commercial forms to be isolated; for example, a book title that included *nickel and dimed* had 831 occurrences in this derived corpus. Also, the study sought occurrences of the verb, nickel and dime in the football sense, but had trouble. Finally, a proximity search in WordSmith Tools located the majority of the forms; with the search term *nickel and dime*, the study added "within 5 to the left or right of down." The key phrase in this use is *down the field*. The study also used a variety of other search terms, including ball which produced this lone result from 2005: "halfbacks, but they like to nickel and dime you and keep the ball" (*LexisNexis Academic*).

Table 6.19: Occurrences of additional *Niagara* forms

Form	Definition	Am BVC	Br BVC
Niagaran adj.	Of or related to Niagara Falls.	0	1 *.009*
Niagarean adj.	Of or related to Niagara Falls.	0	0
Niagarian adj.	Of or related to Niagara Falls.	0	0
Niagara-like adj.	Of or related to Niagara Falls.	7 *.013*	6 *.057*

Table 6.19 includes four *Niagara*-related forms that were obtained from the *OED*. As Table 6.19 indicates, these forms are extremely low frequency in both the American and British BVCs. Table 6.20 includes eight new forms; all of these forms are extremely low frequency in the American and British BVCs.

Table 6.20: Occurrences of other forms

Form	Definition	Am BVC	Br BVC
nice nelly n. and adj.	A prude; prudish.	1 *.001*	0
nice nellie n. and adj.	A prude; prudish.	2 *.003*	0
nice nellyism n.	The practice of prudish ways.	2 *.003*	0
nick n.	Five cents.	0	0
nick n.	Five year prison term.	0	0
nickel n.	Five year prison term.	1 *.001*	0
nickel note n.	Five dollar bill.	1 *.001*	0
nickelodeon n.	A player piano.	7 *.013*	0

The one occurrence for *nickel note* is an example of meta-use as it was included in a glossary of "Swing Talk" (*LexisNexis Academic*). The *nick* forms were extremely challenging to search for in *LexisNexis Academic* because of its popularity as a man's first name. In some cases, the study gathered several high frequency last names (judges, defendants, etc.) and used the AND NOT connector in the *LexisNexis Academic* search interface to focus the search results.

Inclusion

In order to determine what forms should be included, a candidate form (a prospective Americanism) must first have at least 100 occurrences in the American BVC, and second, the number of occurrences in the American BVC must be greater than the occurrences for the same form in the British BVC by at least one order of magnitude.

Forms that have fewer than 100 occurrences in the American BVC will not be recognized by the study for inclusion in the pilot study's update of the *Dictionary*; however, these forms are worthy of recognition as a group because their occurrences represent a class of forms that may be well known many speakers of American English. Still, evidence from the American BVC shows that the group of forms in Table 6.21 is less frequent than the inclusion class, so these forms must not be included in the pilot study's update of the *Dictionary*.

Table 6.21: Forms with 30 to 99 occurrences in the American BVC

Form	Occurrences
nickel coverage n.	79 *.152*
nickel defensive back n.	36 *.069*
nickel and dime situation n.	56 *.108*
nickel and dime defensive package n.	45 *.086*
nickel and dime coverage n.	40 *.077*
nickel video slot n.	33 *.063*
nickel bag n.	80 *.154*

In Table 6.21, 5 of the 7 forms relate to football; *nickel video slot* relates to gambling, and *nickel bag* relates to illegal drug use. Each form in Table 6.21 has zero occurrences in the British BVC. Table 6.22 includes forms with 100 or more occurrences in the American BVC.

Table 6.22: Forms with at least 100 occurrences in the American BVC

Form	Occurrences in the American BVC
nickel back n.	2462 *4.754*
nickel package n.	892 *1.722*
nickel defense n.	774 *1.490*
nickel situation n.	222 *.428*
nickel corner n.	111 *.214*
nickel and dime package n.	249 *.480*
nickel and dime defense n.	134 *.258*
nickel slots n.	205 *.395*
nickel machine n.	123 *.237*
nickel slot machine n.	109 *.210*
nickel game n.	107 *.206*
nickel and dime adj.	103 *.198*
nickel and dime v.	570 *1.100*

Table 6.22 includes 13 forms with at least 100 occurrences in the American BVC. These forms have met the first requirement for inclusion in the pilot study's update of the *Dictionary*. Of these 13 forms, only 2 have occurrences in the British BVC. *Nickel and dime* (adj.) has 12 occurrences in the British BVC, and *nickel and dime* v. ('to exact money steadily in small increments') has 4 occurrences in the British BVC. The American BVC results (103) for the form *nickel and dime* adj. does not meet the requirement of being larger than the British results (12) by one degree of magnitude, so that form cannot be included in the pilot study's update of the *Dictionary*. In the matter of *nickel and dime* v., the American BVC results (570) are greater

than the British BVC results (4) by one order of magnitude. Therefore all of the forms in Table 6.22 can be included in the pilot study's update of the *Dictionary* except for *nickel and dime* (adj.). Table 6.23 includes forms that are recognized for inclusion in the pilot study's update of the *Dictionary*.

Table 6.23: Forms for inclusion

Form	Definition	Am BVC	Br BVC
nickel back n.	A fifth defensive back.	2462 *4.754*	0
nickel package n.	A defensive formation that includes a fifth defensive back.	892 *1.722*	0
nickel defense n.	A defensive formation that includes a fifth defensive back.	774 *1.494*	0
nickel situation n.	Situation that motivates the nickel package.	222 *.428*	0
nickel corner n.	A fifth defensive back.	111 *.214*	0
nickel and dime package n.	Defensive formations that include a fifth and sixth defensive back.	249 *.480*	0
nickel and dime defense n.	Defensive formations that include a fifth and sixth defensive back.	134 *.258*	0
nickel slots n.	A gaming machine that accepts nickels.	205 *.395*	0
nickel machine n.	A gaming machine that accepts nickels.	123 *.237*	0
nickel slot machine n.	A gaming machine that accepts nickels.	109 *.210*	0
nickel game n.	A game that accepts money in five cent increments.	107 *.206*	0
nickel and dime v.	To exact money steadily in small increments.	570 *1.100*	4 *.038*

Table 6.23 shows that 12 forms are included in the pilot study's update of the *Dictionary*. 7 of these forms relate to football; 4 relate to gambling; and one form, *nickel and dime* (v.) resides in a more generalized semantic field. Of the 12 included forms, only *nickel and dime* (v.) has occurrences in the British BVC.

Conclusion

The study pursued two pathways of research to achieve a 21st century perspective of *A Dictionary of Americanisms on Historical Principles* through a pilot study. First, antedatings were obtained through *ProQuest Historical Newspapers*, *APS Online*, and *NewspaperARCHIVE.com*. These illustrations of original entry forms that antedate illustrations in the *Dictionary* are useful in achieving a reflection of a form's historical career. These electronic databases are large, powerful, and approachable to users. Still, the research that these tools provide is historical and does not inform us of recent activity of forms in question.

The BVC informs a researcher of recent careers of forms that are being researched. The BVC can provide evidence in the form of occurrences per ten million words for a particular form in both American and British texts. Such evidence reflects use in a recent decade (1998 to 2007) and could represent uses and forms that vary from historical contexts. Also, with the BVC, zero results can occur; these results are based on one decade only, and a researcher cannot responsibly estimate or suppose what the results for a previous decade would be without a BVC or similar tool that reflects linguistic use for a previous decade. Still, given the size of the American BVC, an estimated 5 billion words, such zero evidence is a significant result because the breadth of the evidence within the BVC is expansive.

The American BVC assists identification of Americanisms that are more frequent today in American English than British English. Evidence from each BVC can be compared with evidence from the other, so both British and American evidence are available through the BVCs. The BVC cannot answer questions about use outside the decade of 1998 to 2007; however, its ability to provide evidence for speech use within that decade is incredibly powerful and especially relevant to the lexicographic needs of the pilot study to update the *Dictionary*.

Chapter 7

Conclusion

The construction of BVCs functions under the hypothesis that these specially designed corpora will provide the basis for a consistent methodology that can evaluate lexical frequency over the ten year span of 1998 to 2007. The BVCs also allow lexical frequency to be determined by year and region of the U.S. because the texts in the American BVC consist of twenty-five newspaper titles that are organized into five regions with five newspaper titles in each of these regions: Northeast, Southeast, Midwest, West, and Coastal west.

The smaller British BVC allows frequency to be determined by year and by newspaper title. The British BVC consists of five titles that were selected as much as possible to reflect geographical variety and a broad range of circulation. Within each of the regional groups of American newspaper titles, each title was selected with regard for variety in terms of circulation.

Comparable Projects

In terms of comparable projects, the classic and next generation corpora must be recognized not only as pioneers of the field that the BVC resides in but also for their utility in comparison to the BVC. In addition to the classic and next generation corpora, additional projects and applications must also be identified, so the utility of the BVC within the current scholarly field can be identified with regard to these existing applications.

Classic and next generation corpora

The classic and next generation corpora established principles of balance, and this study has maintained principled, balanced construction within the BVC. The full-text newspapers themselves have a variety of sections that create variety among the newspaper texts. The five titles in the British BVC reflect a variety of locations and circulation sizes. Each American BVC region contains five titles that as a unit represent variety in terms of geographic representation and newspaper circulation size. The regions of the American BVC allow results to be displayed and compared in terms of geographic divisions. The classic and next generation corpora do not have principles within their construction that allow for geographic representation. The combination of geographic reflection, broad textual variety, and massive quantity of text (much larger than the classic and next generation corpora) cause the BVC to deliver, in essence, a slice of American culture in which recent habits of American English can be conveniently observed.

Additional projects and applications

Kilgarriff and Grefenstette comment on a situation that the current study faces—low frequency forms (336). The current study is interested in several forms that are low frequency in the 5 billion word American BVC; those forms would likely have no or very little presence in a one, or even 100, million word corpus. The BVC takes advantage of the opportunity to access extremely large numbers of texts and words to deliver to the user a large scale virtual corpus that has a larger word count than typical next generation corpora.

The present study avoids some of the problems outlined in "Googleology is Bad Science" in relation to use of the Web for linguistic purposes. First, the *LexisNexis Academic* interface gives its user complex search options that can create highly useful linguistic searches. The

LexisNexis Academic interface is especially useful in terms of proximity searches. Some of the basic syntax of the *LexisNexis Academic* search interface is overviewed in Chapter 3. In further contrast to web-based (via Google) virtual corpus research, the texts in the BVC approach have delineation in terms of text-type (newsprint), newspaper title (and therefore a connection to geography), and time frame (1998-2007). Also, BVC search results can be quickly connected with the title, date and other bibliographic information quickly as search results in *LexisNexis Academic* are actually generated as numbers (of articles) beside newspaper titles—not exclusively as words in context. Finally, in this article, Kilgarriff notes the strength of the Web, large numbers of words, brings with it difficulties. The present study focuses on making very large amounts of words, much larger than most well known constructed corpora, available without some of the complications of Google-based approaches; therefore, because of size, manageability, and utility, the present study stands in between the next generation corpora and a Google approach to corpus linguistics.

The BVC approach, in contrast to Glossanet, yields search results immediately. Even though the BVC is not downloaded, the search process functions as though it is. The delay that is necessary with Glossanet is avoided with the BVC. Also, the boundaries of the BVC are formulated so a total word count can be estimated. Further, the BVC is planned so judgments about regional lexical use can be made. Also, in contrast to the BVC approach, "GlossaNet uses offline linguistic processing tools and therefore permits more complex linguistic queries" (Fairon and Singler 333). The BVC approach functions only online, which means the user must use the search interface connected to LexisNexis Academic. Still, this study shows that the flexible syntax in the *LexisNexis Academic* interface makes robust searches and downloads possible.

In contrast to the NNC and HF corpora, the BVC approach does not require an ongoing time frame to access large amounts of newspaper texts as the texts for the BVC are currently available within an online database; also, the BVC method, while virtual, has some flexibility that these downloadable newspaper corpora likely do not have. For example, the BVC can produce search results for one region or title at a time for precise discernment about lexical evidence. So, the BVC has the advantage of very large total word count, but the BVC can also pursue linguistic questions with the simplicity and focus of searching a single newspaper title.

The CNN corpus differs from the present study's BVCs in that these texts are all transcriptions of spoken texts; however, the newspaper articles in the BVCs do contain some spoken English that passes through the respective newspaper's editorial process. Still, for example, evidence of recent trends in American and British English can be easily found in the American BVC, such as quotative *like* and a variety of neologisms, such as this study's analysis of the trajectory of the recent career of the neologism *carb*. Some advantages of the BVC in comparison to the CNN corpus include larger size in terms of word count; geographic representation of use; and the ability to compare corpora of both British and American English.

Davies's TIME and COCA projects have organized large amounts of text that lend themselves to diachronic investigations as well as questions about more recent matters of American English use. He uses advanced methods to tag the corpora, so users can search for a particular part of speech. These corpora are valuable contributions to the linguistics community, but they do present concern in terms of the study of American English. These corpora do not allow the user to search for texts that were produced in a particular region of the country. The BVC approach does allow such a comparison, so the BVC does not just reflect American English, it reflects regional use of American English.

In terms of the TIME corpus, the devotion of so much text to a single title can create a complication; in house style sheets and other directives could cause some constructions to fail to appear because of editing processes. The balanced nature of Davies's COCA is the antidote to the single title complications in terms of editing. Also, the scale of word count in comparison of the BVC and Corpus of Contemporary American English is worthy of note. The Corpus of Contemporary American English contains 385 million words, and the American BVC contains an estimated 5 billion words. The union of massive word count and a variety of regional texts makes the BVC more powerful in the production of evidence of words and their careers over the span of 1998 to 2007.

In the present study, one factor stands above all in preventing many of Kilgarriff's corpora comparison analyses, which are mentioned in his study, "Comparing Corpora" to be replicated. The present study's corpora are not constructed and cannot be linguistically analyzed as a single downloaded file or directory. For example, the present study is unable to generate a word list that features every form in the American and British BVCs; the *LexisNexis Academic* interface is not that flexible. Still, searches can be performed in *LexisNexis Academic*, and 500 documents can be downloaded at a time which can be used to construct derived corpora (such as the *go missing* strip corpora) that can be studied in a variety of ways with an offline interface, such as WordSmith tools. While the present study cannot respond to the question of how statistically similar the BVCs are, the BVCs themselves provide evidence, through search results of American and British newsprint, of similarities and differences between American and British English. This study's work with *go missing* highlights how contrastive the ten-year careers of *go missing* in British and American English are.

Conclusion

The BVC gives the researcher access to an extremely large corpus that maintains principles of balance that were established by the classic corpora. The estimated word count for the BVC (5 billion words for the American BVC and 1 billion words for the British BVC) is significantly greater than the 1 million word counts for the classic corpora. This significantly larger corpus makes access to fine-grained evidence possible that helps to tell a more detailed story of a form's career—simply due to the fact that more evidence resides in the BVC than in a classic corpus or next generation corpus.

In terms of construction, the BVC is bounded in contrast to Google, which is another virtual corpus. The BVC is organized under principles of sampling; for example, the circulation of each of the 25 newspaper titles was considered, and each the selection of each was justified in terms of circulation. Also, each newspaper title in the BVC resides in a region in which newspaper titles with a variety of circulation sizes also reside. Further, search results from a single title within one region could represent a region within a region—one more example of the BVC's ability to provide fine-grained evidence.

With regard to the next generation corpora and the additional projects and applications mentioned above, the BVC stands above them through its larger word count and ability to reflect geographic use of speech. Also, many of the additional projects and applications required considerable time and expense to construct. A researcher who has access to the *LexisNexis Academic* database, can quickly replicate the construction of the BVC or even create a customized BVC under the same principles of balance that this study's BVCs adhere to; that researcher's institution will have had to pay for access to *LexisNexis Academic*, but the researcher will not have to pay for the access and opportunity to replicate BVCs.

Attributes of the BVC

The BVC establishes itself as an innovative contribution to the field through providing evidence that reflects both the geography of the United States and an extremely large estimated word count. Still, the BVC must be used cautiously in order to preserve its ability to function as a principled research tool. In order for a researcher to use the BVC successfully, the following attributes of the BVC must be understood by a researcher: management of BVC searches and results; the fleeting nature of the BVC; corpus and file management relative to the BVC; and flexibility and replication of the BVC.

Management of BVC searches and results

Searches within the BVCs have to be performed judiciously; for example, *LexisNexis Academic* has a list of common words that are not recognized in searches; this search feature can cause unwanted search results. Also, because of the connection between newspapers and culture, a single cultural event can cause occurrences of a particular form to increase sharply. For example, some event might be described as an X, and that form could have previously had a low frequency of use for several years, but a single event could cause the use of that form to increase significantly. In such an instance (and in others that this study encountered), the user can form strategies to address complications. These strategies include a visual scan of the context of the search results as well as a modified search to create more selective results.

In the *LexisNexis Academic* database search results are indicated numerically to the right of the title in parentheses. The results number actually represents articles instead of a total number of results. Therefore, a title with a two beside it could actually represent six occurrences

of the search form because the search form could possibly occur three times in each of the two articles. In other words, a result of 2 beside a title indicates that in two articles, the search form occurs at least once.

The article count search display in *LexisNexis Academic* could actually be regarded by linguists as an internal search refinement. Kilgarriff (2001) explains, "Where a word occurs once in a text, you are substantially more likely to see it again than if it had not occurred once" (107). Kilgarriff continues, "Some words are 'clumpier' or 'burstier' than others; typically content words are clumpier than grammatical words" (107). The process in which *LexisNexis Academic* displays an article count, rather than a word count, provides the user with an effective overview of a form's use. Also, the article count is helpful with regional searches, so, at a glance, a user can discern possible contrasts in a form's use between different regions.

The use of an additional (offline) interface, such as WordSmith tools, can determine the total number of search forms an article (or selection of articles) contains. *LexisNexis Academic* delivers files (up to 500 per download), and one of the file formats is plain text or .txt. This format is compatible with the WordSmith Tools interface that provides output, such as concordances and word lists. WordSmith Tools can amplify the specificity of searches performed in *LexisNexis Academic*. For example, WordSmith Tools does not have blocked words or symbols—including punctuation. Therefore, WordSmith Tools allows offline, highly specific searches to be performed on search results from the BVC that cannot be performed directly from the *LexisNexis Academic* interface.

The study faced a challenge with search results from *LexisNexis Academic* that were erroneously duplicated. While it is not good or reasonable scholarship to decide, based on intuition, what a satisfactory number of search results should be, judgment still needs to

accompany search results obtained in *LexisNexis Academic*. One way to exercise such judgment is to visually scan the Expanded List of results to detect duplications, such as an identical string of words that repeats throughout the results.

With the conversion of the Expanded List of results into a single text file, which interfaces with WordSmith Tools when downloaded in the .txt format, one important point must be understood. An integer or numerical assignment precedes each result. If results 4 and 5 are identical, WordSmith Tools will not discriminate the duplication because the integers that precede each result, 4 and 5 in this example, cause WordSmith Tools to regard each result as unique—because the integer changes as the results advance. In order for WordSmith Tools to recognize duplication of results, the integer would have to be removed. This procedure should be able to be handled with an advanced string command in the WordSmith Tools application; this study did not pursue this option because the duplication problems the study faced were handled visually. In the end, this visual approach was valuable to a degree because of the education gained about idiosyncrasies of *LexisNexis Academic* in terms of search results.

In the American and British BVCs in *LexisNexis Academic*, noise words could not be searched or removed from searches. Actually, this study wished to exclude some of them. For example, a search for variants of *go* in a study of *go missing* in the American BVC, *go*, *goes*, *gone*, and *went* generally produced accurate results when the search string consisted of the *go* form connected to PRE/1 missing.

On the contrary, the search string going PRE/1 missing was not as successful as the other *go* form searches were because of two major problems that relate to the way *LexisNexis Academic* recognizes characters. First, the *LexisNexis Academic* search interface generally recognizes periods as spaces; also, *LexisNexis Academic* appeared to recognize commas as

spaces as well. The noise word situation, as well as comma and period recognition matters, resulted in some unusual, unwanted results in the search for *going missing*.

The recognition of the period as a space in *LexisNexis Academic* searches caused results that spread across sentence boundaries, such as: "demand to keep the momentum going.' What was missing in each instance was" (*LexisNexis Academic*). Also, this example shows noise words that are unwelcome intruders, *what was*, cannot be strategically removed because either each word, or the construction, is a noise word in *LexisNexis Academic*. Another search malfunction involved the comma and a construction, such as: "never got his driver going, missing eight fairways" (*LexisNexis Academic*). *LexisNexis Academic* produces this result for one of this study's *going missing* searches because the comma is invisible to the search interface. The result in this case is an unwanted form that is grammatically and semantically different from the desired search result.

The most frequent unwanted result of noise words in *LexisNexis Academic* that this study encountered is the intrusive *to be*, such as the example above that extends across sentence boundaries or within a single sentence, such as: "whole list of solutions, he's going to be missing an opportunity" (*LexisNexis Academic*). After much experimentation, the results were refined through a visual scan of the Expanded List of results. This activity was inconvenient, but not so time consuming as to be considered impossible.

Single cultural events should be rationalized and considered as possible skewing factors, and results can be visually inspected to detect recurring events. In this study events that surround certain people and events sometimes created distorted results, so the study was able to refine searches (through the Search Within Results feature in *LexisNexis Academic*) to show that many articles, in some searches, were focused on concepts connected to a single person or event.

The study also used WordSmith Tools for finer grained analysis of single events, such as the determination of collocates of *went missing*; that analysis revealed that in the British BVC, *Madeleine* (the child who vanished) is the most frequent form in the position immediately to the left of *went missing*.

In the BVCs, because of the full-text nature, many unwanted forms can complicate searches. A search form may yield a large number of results, but after closer investigation, an unwanted form, such as an e-mail address, an institution name, a place name, or an author's name could cause a majority of the results. Again visual inspection, which is most convenient in the Expanded List of search results display in *LexisNexis Academic*, can assist the detection of unwanted forms. After the forms are detected, new strategies to remove unwanted forms can be performed.

Another complicating form in BVC searches is sense of the search form. For example, the form *Niagara* could manifest itself as: 'the river in New York and Canada' or in a transferred sense as 'a deluge' which only relates metaphorically to the actual river. The search strategies highlighted above, with regard to refinement of search strings and visual scanning of results, assisted the study with refining results in search of a specific sense of forms.

Fleeting nature of the BVC

Since the BVCs are stored online, their presence is not guaranteed for any length of time. UGA librarians confirmed that none of this project's titles had moving walls of coverage that would shift forward as the years progress; however, the titles could be removed by the vendor without notice. Many variables stand between the researcher and accessibility of the BVC. Business agreements, subscription changes, and matters relating to copyright could all cause the

texts of the BVC to become inaccessible; therefore, research calculations on a BVC should be done quickly in preparation for the possibility that availability of the texts could terminate.

In order to perform necessary searches quickly, a researcher must have a research plan that includes specific goals. One such goal could be a derived corpus, such as this study's *went missing* strip corpus. Once that derived corpus is downloaded, it can be returned to repeatedly without influence from vicissitudes that could compromise online access to the BVC via *LexisNexis Academic*. Also, as this study pursued research on the BVCs, through the months, spot checking of the titles within *LexisNexis Academic* was practiced to ensure that the coverage of newspaper titles remained available.

Corpus and file mangaement

The size of the BVCs, if completely downloaded and saved as a single file or directory, would constitute a very large file; therefore, convenient access, via the Web, to an estimated 5 billion word virtual corpus that is organized in terms of several factors of representativeness contradicts any notion of inconvenience relative to the virtual nature of the BVCs. Still, the study estimated the file size of the American BVC, based on one full day of articles (155) from the *New York Times* downloaded from *LexisNexis Academic*. The 155 articles totaled 104,432 words and were copied and pasted until a new file that consisted of 10 copies of the original 155 articles, or a total of 1.04 million words resulted; that count includes whatever metadata, such as title and author that *LexisNexis Academic* might append to the full text of the articles. The file size of that 1.04 million word corpus was 6710 kilobytes. The product of that, approximately 1 million word, file size, 6710 kilobytes, was multiplied by 5,000 to create an estimation of the file size of a 5 billion word corpus. The estimated file size is 31.99 gigabytes. The estimated file

size for the British BVC is 6.38 gigabytes, so the total file size of these BVCs, downloaded locally, would be about 38 gigabytes.

Files of this size would be burdensome to store and difficult to move from place to place. Beyond the inconvenient size of the locally downloaded files for an undertaking of this magnitude, two other factors must be considered. First, such a downloading project would be burdensome and time consuming. Second, the legality of such a full-text locally downloaded corpus is questionable—especially if the corpus were moved from the domain of the subscribing library. Finally, libraries and database vendors monitor repetitive downloading activity from a single IP address, so even a researcher who has subscription access to a database, such as *LexisNexis Academic*, would likely be required to use it within limits established by their library or the database vendor. Such limits would likely prevent such a large scale downloading project.

Flexibility of the BVC

One of the BVCs greatest strengths as a tool for linguistic study is its ability to produce useful evidence for both high frequency and low frequency forms. For example, *go missing* and its variations are low frequency forms both in American and British English. Still, the BVC has been able to do more than simply justify the label, low frequency. The BVC has provided evidence that tells the story of the form's increase in frequency in American English as well as the form's well established presence in British English. Further, evidence from the BVC shows years in which the form's use increased in the American and British BVCs as well as some factors, such as single events that influenced the form's use.

The BVC can with similar convenience report evidence of higher frequency forms, such as *nickel back* in American English. In the event that a derived corpus is needed for offline

study, *LexisNexis Academic*'s 500 unit (full-text or expanded list display) per download makes the construction of such a corpus similarly convenient for both low and high frequency forms because of the downloading process can be promptly repeated in the event that a large number of articles needs to be downloaded.

The downloading process of *LexisNexis Academic* points out one of the most important attributes of the BVC: all searches can be immediately connected with the original full-text document. The full-text document can be extremely useful because, for example, a study of a strip corpus might reveal a form whose sense is unclear. Access to the original full-text source document can clarify such problems. Also, because *LexisNexis Academic* is an online database, users will be able to use it with the convenience and accessibility of other online applications—both on-site and off-site according to the arrangements of their home library.

The convenient file delivery interface of the *LexisNexis Academic* database allows its user to connect the downloaded files to an offline application, such as WordSmith Tools. Such an offline application can take a corpus derived from a *LexisNexis Academic* search and identify patterns, especially in terms of collocations, that would likely be inconvenient or even impossible within the *LexisNexis Academic* interface. For example, an introductory search for *nibbler* in the *LexisNexis Academic* interface revealed some results that were vague; others that referred to a fish; and finally, some that were connected to baseball. The study created a derived Nibbler corpus, and in WordSmith Tools, *the* was included as a search term. The results of that offline search revealed that 15 of 143 results for *nibbler* actually were a restaurant name. Searches for *the* are not possible in *LexisNexis Academic*.

The use of an offline interface also speeds up the analysis of words in terms of their collocates; for example, the study was quickly able to identify, offline, that the most frequent

form in the R1 position of *steady diet* is *of*. After that determination, the study was able to determine the collocating habits of *steady diet of*. The collocation feature in WordSmith Tools displays a node's collocates from the L5 position to the R5 position, and quickly gives total numbers of occurrences for any of these positions. These types of calculations cannot be made through a visual search of results in the *LexisNexis Academic* database.

Still, the *LexisNexis Academic* interface can assist offline searches in derived corpora. Collocates of the node, *nibbler*, do not transparently reveal the form's use; that is, some of the occurrences clearly relate to baseball, but those occurrences are not readily identifiable through their collocates. A search in *LexisNexis Academic* for *nibbler* only in newspaper sections labeled Sports provided 50 results. Those 50 results were fed back into the offline interface and a determination could be made in terms of primarily baseball, and some other uses of *nibbler*.

A very large number of newspaper titles are available within *LexisNexis Academic*. Some of these titles have decades of full-text coverage, and others cover only a few years. Because the total number of newspaper titles in *LexisNexis Academic* is so large, a variety of types of newspaper texts is available. This variety includes representation of a variety of regions, circulation sizes, and in-house nuances toward news reporting. Such a variety to choose from creates the possibility for rich research opportunities for the user.

This study regards all Web texts as copyrighted and upholds Kilgarriff's admonition that permission be granted from all copyright holders to keep corpus linguists safe from legal repercussions ("Legal Aspects"). Because access to the BVCs rests on institutional subscription, the user pays no immediate fees for use of the BVC. Still, users need to refer to rules of use associated with their library's subscription to *LexisNexis Academic* for their particular research projects. Also, no copyright clearances must be maintained for academic use of the texts in the

BVC. These matters of legality, lack of immediate cost to user, and freedom of use for academic purposes make the BVC both affordable and easy to use; these positive attributes would likely not be associated with the offline construction of a corpus that approaches the size of the BVC.

Replication

The American and British BVCs in this study can be replicated successfully by another user. In the event that a user encounters problems with the BVC through replication, solutions do exist. These solutions result from the flexibility of the *LexisNexis Academic* database and the user's own experience with attributes of the database.

The first malfunction that a user might encounter could be the absence of one of the BVC titles from *LexisNexis Academic*. This absence could seem mysterious, based on the titles and research in this study, but what might not be well understood by library patrons is the fact that *LexisNexis Academic* is a subscription, and one library's subscription could differ from that of another library's subscription. The difference of subscriptions could result in a user's not having access to the same titles in *LexisNexis Academic* that this study had.

In the event of a user's being unable to congruently replicate the study, then with a list of their subscription's newspaper titles and research of circulation and, finally, an understanding of the geographical region in which they need to find a replacement, the user would likely be able to find a replacement title within their *LexisNexis Academic* subscription for their replicated study. The appendix of this study contains instructions for the process of estimating word counts for titles that would allow the user to subtract the estimated word count of the title they cannot access, so they could add an estimated word count (that they calculate) to complete a new estimated word count for their replicated study.

A researcher can also select titles from *LexisNexis Academic* to build their own BVC. The selections could parallel the 25 titles in this study, or according to the needs of a particular researcher, the titles could reflect a region that is delineated according to the needs of a specific research application.

The process of calculating estimated word counts for the newspaper titles is a tedious activity because of the multiple interfaces (file delivery from *LexisNexis Academic* and offline calculations with WordSmith Tools and Microsoft Excel) that are involved. No antidote exists for the time consuming nature of the estimated word count process, but one problem needs to be carefully anticipated. Just as preceding chapters have overviewed the possibility of erroneously duplicated search results, the same problem can also present itself in the estimated word count process because the estimated word count process begins with an article search in *LexisNexis Academic*. The estimated word count involves all of the available articles for one day of one title, so this process causes the results to equal all of the available articles for that title on that date.

When the study encountered erroneously duplicated articles on a large scale, the situation made itself clear. In some instances, 50% percent of the results included word counts that were shared by two or more articles; such a result is immediate evidence of erroneous duplication of results. Fortunately, such a problem was quickly evident; unfortunately, the solution was tedious and time consuming. The *LexisNexis Academic* search results were visually inspected, and the duplications were eliminated. Often, the duplicated results followed a rhythmic pattern of every two or three results; at other times the duplications were distributed without predictable explanation. The theory behind cautiously removing these duplications is obvious; an accurate estimated word count was the study's goal.

Summary

The BVC tells a story that a glance at a newspaper, a Google search, or classic or next generation corpus cannot. The American and British BVCs testify to the cultural and linguistic separation of two communities in the case of *nickel back* and its related forms; however, the BVCs provide evidence of convergence through the rise of *go missing* in American English. Corpus linguistics needs a tool, such as the BVC approach, to provide evidence of such changes, connections, and separations between American and British English because our introspection alone will not be enough to discern what is really happening linguistically with these two communities (Sinclair 100). The BVC approach allows a new degree of specificity to be applied to comparisons of British and American English because of the mass of texts within each corpus and the flexibility of the *LexisNexis Academic* interface.

Works Cited

"About the Collins Corpus and the Bank of English." Web. 14 Oct. 2008.

Algeo, John. "British and American Grammatical Differences." *International Journal of Lexicography* 1.1 (1988): 1-31. Print.

Algeo, John. "What Is a Briticism?." *Old English and New*: *Studies in Language and Linguistics in Honor of Frederic G. Cassidy*. 287-304. New York, NY: Garland, 1992. Print.

"American National Corpus." Web. 16 Oct. 2008.

APS Online. University of Georgia Libraries. Web.

Baker, Paul, Andrew Hardie, and Tony McEnery. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh UP, 2006. Print.

"The Bank of English." Collins. Web. 14 Oct. 2008.

Burchfield, Robert. "The Oxford English Dictionary." *Lexicography, an Emerging International Profession*. Fulbright papers, v. 1. Ed. Robert Ilson. 17-27. Manchester, U.K.: Manchester University Press in association with the Fulbright Commission, London, 1986. Print.

"Cambridge International Corpus." Web. 14 Oct. 2008.

Cassidy, Frederic G., ed. *Dictionary of American Regional English*. Cambridge, MA: Belknap, 1985– . Print.

Cassidy, Frederic G. "The Dictionary of American Regional English as a Resource for Language Study." *Studies in Lexicography*. Ed. Robert Burchfield. 117-135. Oxford: Clarendon, 1987. Print.

"Commands." *LexisNexis Academic*. Web. 29 December 2009.

"Composition and Structure". *Ask Oxford*. Web. 14 Oct. 2008.

"Connector Order and Priority." *Research Help*. *LexisNexis Academic*. Web. 13 April 2008.

Craigie, William A. and James R Hulbert, eds. *A Dictionary of American English on Historical Principles*. 4 vols. Chicago, IL: U of Chicago P, 1938-44. Print.

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present. Available online at http://www.americancorpus.org.

Davies, Mark. (2007-) TIME Magazine Corpus (100 million words, 1920s-2000s). Available online at http://corpus.byu.edu/time.

Davis, Lawrence M. *Statistics in Dialectology*. Tuscaloosa: University of Alabama P, 1990. Print.

"Developing a Search (with Natural language)." *Research Help*. *LexisNexis Academic*. Web. 3 January 2009.

"Document Section: Description of Common Sections." *LexisNexis Academic*. Web. 28 November 2008.

The Europa World Year Book 2007 Volume II Kazakhstan-Zimbabwe. 4645-4648. London: Routledge, 2007. Print.

Fairon, Cedrick, and John V. Singler. "I'm Like, 'Hey, It Works!': Using GlossaNet to Find Attestations of the Quotative (Be) Like in English-Language Newspapers." *The Changing Face of Corpus Linguistics.* 325-336. Amsterdam, Netherlands: Rodopi, 2006. Print.

"Features." *LexisNexis Academic*. Web. 7 February 2009.

Firth, J. R. *Studies in Linguistic Analysis*: Special *Volume of the Philological Society*. Oxford: Blackwell, 1957. Print.

Furiassi, Cristiano, and Knut Hofland. "The Retrieval of False Anglicisms in Newspaper Texts." *Corpus Linguistics 25 Years on*. Ed. Roberta Facchinetti. Amsterdam, NE; Netherlands: Editions Rodopi, 2007. 347-363.

Hockey, Susan M. *Electronic Texts in the Humanities*: *Principles and Practice*. Oxford: OUP, 2000. Print.

Hoffmann, Sebastian. "From Webpage to Mega-corpus: the CNN transcripts." *Corpus Linguistics and the Web*. Language and computers, no. 59. Ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. 69-85. Amsterdam: Rodopi, 2007. Print.

Hofland, Knut, and Stig Johansson. *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities, 1982. Print.

"Hyphens, Slashes, and Parentheses." *LexisNexis Academic*. Web. 19 April 2008.

"International Corpus of English." Web. 16 Oct. 2008.

Jackson, Howard. *Lexicography*: *An Introduction*. London: Routledge, 2002. Print.

Johnson, Samuel. *A Dictionary of the English Language*. 1755.

Kilgarriff, Adam. "Comparing Corpora." *International Journal of Corpus Linguistics* 6.1 (2001): 97-133. Print.

---. "Googleology is Bad Science." *Computational Linguistics* 33.1 (2007): 147-51. CSA *Linguistics and Language Behavior Abstracts*. Web. 20 Oct. 2008.

---. "Legal aspects of corpora compiling." Web. 1 Oct. 2002. 29 Dec. 2009.

Kilgarriff, Adam, and Gregory Grefenstette. “Introduction to the Special Issue on the Web as Corpus.” *Computational Linguistics* 29.3 (2003): 333-47. Print.

Kretzschmar, William A. *The Linguistics of Speech*. Cambridge: Cambridge UP, 2009. Print.

Kucera, Henry, and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Providence: Brown University Press, 1967. Print.

Landau, Sidney. *Dictionaries*: *The Art and Craft of Lexicography* Second Edition. Cambridge, Cambridge UP, 2001. Print.

Leech, Geoffrey, and Roger Fallon. “Computer Corpora—What do they tell us about culture?” *ICAME Journal* 16.Apr (1992): 29-50. Print.

LexisNexis Academic. University of Georgia Libraries. Web.

Lighter, J. E., ed. *Random House Historical Dictionary of American Slang.* New York, NY: Random House, 1994–. Print.

MacQueen, Donald S. “Developing Methods for very-Large-Scale Searches in Proquest Historical Newspapers Collection and Infotrac the Times Digital Archive: The Case of Two Million Versus Two Millions.” *Journal of English Linguistics* 32.2 (2004): 124-43. CSA *Linguistics and Language Behavior Abstracts*. Web. 4 Nov. 2008.

Mathews, Mitford, ed. *A Dictionary of Americanisms on Historical Principles*. Chicago, IL: U of Chicago P, 1951. Print.

McEnery, Tony, and Andrew Wilson. *Corpus Linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh: Edinburgh UP, 2001. Print.

McEnery, Tony, Richard Xiao, and Yukio Tono. *Corpus-Based Language Studies*: *An Advanced Resource Book*. Routledge applied linguistics. London: Routledge, 2006. Print.

McKean, Erin, ed. *The New Oxford American Dictionary* Second Edition. New York, NY: OUP, 2005. Print.

Meyer, Charles F. *English Corpus Linguistics*. Cambridge: CUP, 2002. Print.

Moon, Rosamund. "Sinclair, Lexicography, and the Cobuild Project: The Application of Theory." *International Journal of Corpus Linguistics* 12.2 (2007): 159-181. *MLA International Bibliography*. EBSCO. Web. 14 Oct. 2008.

Mugglestone, Lynda. *Lexicography and the OED*: *Pioneers in the Untrodden Forest*. Oxford: OUP, 2000. Print.

Murray, James A. H, et al. *The Oxford English Dictionary*: *Being a Corrected Re-Issue with an Introduction, Supplement, and Bibliography of a New English Dictionary on Historical Principles*. Oxford: Clarendon Press, 1933. Print.

NewspaperARCHIVE.com. Heritage Microfilm. Web.

"Noise Words." *Support Center*. *LexisNexis Academic*. Web. 5 September 2008.

"The Norwegian Newspaper Corpus." Web. 23 Oct. 2008.

Oakes, Michael P. "Contrasts Between US and British English of the 1990s." *Research and Scholarship in Integration Processes*: *Poland, USA, EU*. Ed. Elzbieta Oleksy and Barbara Lewandowska-Tomaszczyk. Lodz: Lodz UP, 2003. 213-222 Print.

Oakes, Michael P. *Statistics for Corpus Linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh: Edinburgh University Press, 1998. Print.

OED Online. OUP. Web.

Pickett, Joseph P., ed. *The American Heritage Dictionary of the English Language* Fourth Edition. Boston, MA: Houghton Mifflin, 2000.

Popik, Barry. "Digital Historical Newspapers: A Review of the Powerful New Research Tools." *Journal of English Linguistics* 32.2 (2004): 114-123. *MLA International Bibliography*. EBSCO. Web. 8 Mar. 2008.

ProQuest Historical Newspapers. University of Georgia Libraries. Web.

Pyles, Thomas. *Words and Ways of American English*. New York, NY: Random House, 1952. Print.

Read, Allen Walker. "The History of Lexicography." *Lexicography, an Emerging International Profession*. Fulbright papers, v. 1. Ed. Robert Ilson. Manchester, U.K: Manchester University Press in association with the Fulbright Commission, London, 1986. 28-50. Print.

---. *Milestones in the History of English in America*. PADS, no. 86 edited by Richard W. Bailey. Durham, NC: Duke UP, 2002. Print.

"Search Connectors and Commands." *Research Help*. *LexisNexis Academic*. Web. 5 January 2009.

William, Safire. "Gone Missing." *New York Times Magazine* (2004): 21. *LexisNexis Academic*. Web. 16 Aug. 2008.

Scott, Mike. WordSmith Tools version 4, Oxford: OUP, 2004.

Simpson, John, ed. *The Oxford English Dictionary* Second Edition. Oxford: Clarendon, 1989. Print.

Sinclair, John. *Corpus, Concordance, and Colllocation*. Oxford: OUP, 1991. Print.

Slotkin, Alan. “To Go Missing ‘to Disappear’: Another British Syntactic Intrusion?” *American Speech* 65.2 (1990): 196. *MLA International Bibliography*. EBSCO. Web. 11 Mar. 2008.

SRDS Media Solutions. University of Georgia Libraries. Web.

Stubbs, Michael. *Words and Phrases*. Oxford: Blackwell, 2002. Print.

Teubert and Cermakova. “Directions in Corpus Linguistics.” Ed. M. A. K. Halliday. *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum, 2004. 113-165. Print.

Winchester, Simon. *The Meaning of Everything*: *The Story of the OED*. Oxford: OUP, 2003. Print.

Zelinsky, Wilbur. *The Cultural Geography of the United States*. Englewood Cliffs, N.J.: Prentice-Hall, 1973. Print.

Appendix A

Steps for Calculating an Estimated Word Count

1) Pick one month (selected at random) and one day (selected at random) for each year.

2) For that day/title in *LexisNexis Academic*, click search with search box empty. This process will retrieve all article results for that day/title.

3) Through the file download interface in *LexisNexis Academic*, download all of the search results. This interface will allow up to 500 units of text to be downloaded, so in the event of very large amounts of text to be downloaded, the sum of texts can be downloaded in increments of 500. Still, for the word count estimation process, all needed articles for one day/title should be fewer than 500 units. Save this file in .txt format in a folder.

4) In WordSmith Tools Select File→ \: to locate the folder that contains the necessary file. This study used 5 folders; each one corresponded to one of the 5 geographic regions of the BVC.

5) Highlight the folder and in WordSmith Tools, highlight the necessary file and move it to the right side of the screen. Be sure to check the green check mark to finalize the selection.

6) Click Concord, File→ New OK

Click Compute

Search word = *words*

Select clusters—clusters need to be set-up to search 1L (to search for the integer and zero forms to the right).

7) Scan the KWIC output for word environments in which *words* is preceded by a non-integer such as *good*. These results should be removed with the Zap feature in WordSmith tools or manually.

8) Copy the clusters (integer and the form *words*) and their frequency—which is the first integer to the right, for example: 124 WORDS 3

9) Paste the complete file of clusters and frequencies into Microsoft Excel. The first step in Excel is to remove *WORDS* with Find and Replace.

10) Determine if one or more of the word count results has a frequency of 2 or more. In this case the word count values need to be multiplied by their respective frequencies to determine extended word counts. The extended column needs to be totaled to determine the total word count. If no frequencies are above 1, then the word counts can be immediately totaled in Excel after WORDS has been removed with the Find and Replace application.

11) To verify precision, total the frequency column—this total should equal the total number of articles that were downloaded from *LexisNexis Academic* for the particular file.

Appendix B

Complete R1 Output for *steady diet of*

fastballs	**26**	*american*	**6**	*tough*	**5**	*5*	**3**
breaking	**25**	*changeups*	**6**	*tranquilizers*	**5**	*baseball*	**3**
news	**22**	*corn*	**6**	*western*	**5**	*books*	**3**
running	**20**	*hard*	**6**	*white*	**5**	*both*	**3**
anti	**17**	*hot*	**6**	*2*	**4**	*bush*	**3**
big	**16**	*leafy*	**6**	*city*	**4**	*change*	**3**
junk	**16**	*losing*	**6**	*classic*	**4**	*cheap*	**3**
fast	**15**	*losses*	**6**	*close*	**4**	*cheeseburgers*	**3**
bad	**14**	*meat*	**6**	*comic*	**4**	*classical*	**3**
double	**14**	*movies*	**6**	*convenience*	**4**	*commercial*	**3**
curveballs	**13**	*right*	**6**	*curves*	**4**	*country*	**3**
television	**13**	*sliders*	**6**	*daily*	**4**	*films*	**3**
violence	**13**	*violent*	**6**	*dirt*	**4**	*four*	**3**
ballpark	**10**	*ahman*	**5**	*fish*	**4**	*full*	**3**
blitzes	**10**	*antibiotics*	**5**	*food*	**4**	*golf*	**3**
high	**10**	*caffeine*	**5**	*fried*	**4**	*israel*	**3**
carries	**9**	*cold*	**5**	*harrison*	**4**	*john*	**3**
good	**9**	*cuts*	**5**	*hollywood*	**4**	*less*	**3**
low	**9**	*eight*	**5**	*human*	**4**	*lobs*	**3**
toxins	**9**	*female*	**5**	*ice*	**4**	*love*	**3**
zone	**9**	*government*	**5**	*imported*	**4**	*milk*	**3**
fairways	**8**	*heavy*	**5**	*just*	**4**	*music*	**3**
free	**8**	*mice*	**5**	*live*	**4**	*old*	**3**
inside	**8**	*negative*	**5**	*milkshakes*	**4**	*playing*	**3**
off	**8**	*nothing*	**5**	*pizza*	**4**	*power*	**3**
anything	**7**	*obesity*	**5**	*rate*	**4**	*program*	**3**
burgers	**7**	*pop*	**5**	*reading*	**4**	*propaganda*	**3**
chicken	**7**	*radio*	**5**	*rice*	**4**	*protein*	**3**
left	**7**	*ramen*	**5**	*sensational*	**4**	*seahawks*	**3**
outside	**7**	*reruns*	**5**	*short*	**4**	*soul*	**3**
political	**7**	*second*	**5**	*small*	**4**	*success*	**3**
sex	**7**	*self*	**5**	*soccer*	**4**	*sugar*	**3**
tv	**7**	*stories*	**5**	*3*	**3**	*sugary*	**3**

take	**3**	*popular*	**2**	*class*	**1**	*programming*	**1**
talk	**3**	*pork*	**2**	*coming*	**1**	*programs*	**1**
tennis	**3**	*practice*	**2**	*commercials*	**1**	*public*	**1**
three	**3**	*pressure*	**2**	*conference*	**1**	*rats*	**1**
weight	**3**	*quality*	**2**	*cover*	**1**	*ready*	**1**
action	**2**	*reality*	**2**	*cream*	**1**	*recent*	**1**
being	**2**	*road*	**2**	*day*	**1**	*runs*	**1**
black	**2**	*rock*	**2**	*deep*	**1**	*same*	**1**
cartoons	**2**	*run*	**2**	*defense*	**1**	*saturday*	**1**
cheese	**2**	*senior*	**2**	*end*	**1**	*screens*	**1**
chinese	**2**	*seven*	**2**	*entertainment*	**1**	*show*	**1**
chips	**2**	*shots*	**2**	*feeding*	**1**	*sitcoms*	**1**
cigarettes	**2**	*shows*	**2**	*foods*	**1**	*soap*	**1**
comedies	**2**	*soda*	**2**	*fries*	**1**	*speed*	**1**
competition	**2**	*soft*	**2**	*game*	**1**	*sports*	**1**
computer	**2**	*strikes*	**2**	*green*	**1**	*state*	**1**
defenses	**2**	*teams*	**2**	*greens*	**1**	*steaks*	**1**
dillon	**2**	*those*	**2**	*half*	**1**	*strong*	**1**
drugs	**2**	*tomlinson*	**2**	*hit*	**1**	*summer*	**1**
e	**2**	*video*	**2**	*home*	**1**	*sunday*	**1**
eating	**2**	*water*	**2**	*how*	**1**	*sure*	**1**
family	**2**	*young*	**2**	*images*	**1**	*thomas*	**1**
fat	**2**	*arab*	**1**	*joe*	**1**	*throwing*	**1**
fiction	**2**	*attack*	**1**	*little*	**1**	*tigers*	**1**
games	**2**	*attacks*	**1**	*mayhem*	**1**	*too*	**1**
hamburgers	**2**	*ballads*	**1**	*mostly*	**1**	*training*	**1**
hatred	**2**	*better*	**1**	*nba*	**1**	*turnovers*	**1**
health	**2**	*blood*	**1**	*network*	**1**	*vegetables*	**1**
hip	**2**	*boys*	**1**	*night*	**1**	*watching*	**1**
information	**2**	*bread*	**1**	*people*	**1**	*weeks*	**1**
invective	**2**	*british*	**1**	*performances*	**1**	*whole*	**1**
long	**2**	*brown*	**1**	*pirates*	**1**	*why*	**1**
media	**2**	*business*	**1**	*players*	**1**	*wine*	**1**
murder	**2**	*cable*	**1**	*plays*	**1**	*words*	**1**
national	**2**	*cases*	**1**	*poor*	**1**	*work*	**1**
passes	**2**	*cats*	**1**	*products*	**1**	*world*	**1**

Appendix C

Relevant Pages from *A Dictionary of Americanisms on Historical Principles*

This appendix includes scans of 5 pages from *A Dictionary of Americanisms on Historical Principles* (1951). Figure C.1 is a scan of the title page from Volume II of the *Dictionary*; the range of the pilot study (*Niagara - nickelodeon*) in Chapter 6 comes from Volume II and is displayed below in figures C.4 and C.5. Figures C.2 and C.3 come from the front matter contained in Volume I of the *Dictionary*. Figure C.2 is an explanation of special lettering and symbols, and Figure C.3 is a list of abbreviations for the *Dictionary*.

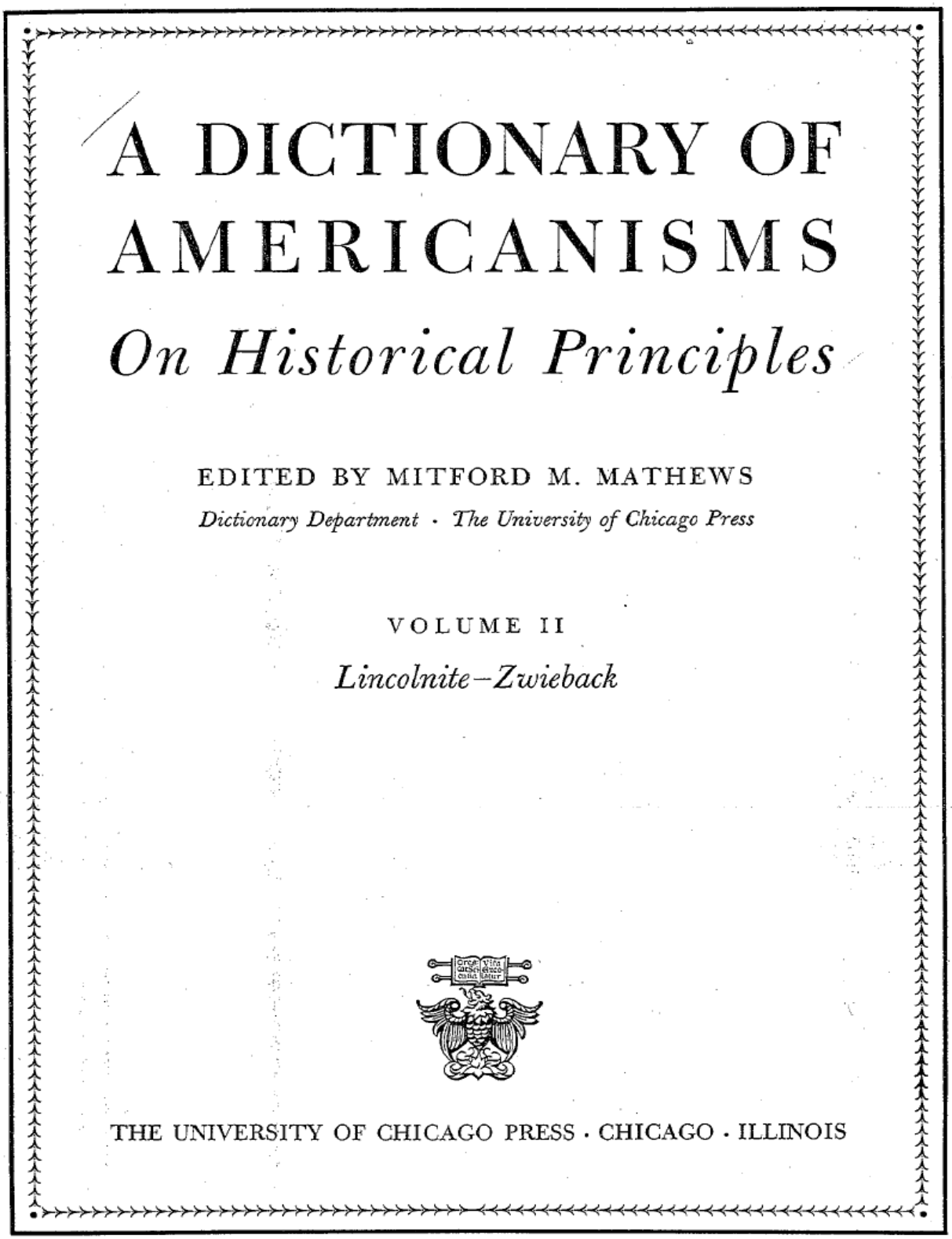


Figure C.1: Title page of Volume II

EXPLANATION OF SPECIAL LETTERING AND SYMBOLS

All entry words are given in boldface. When reference is made to them in the definitions or etymologies, they are in lightface if followed by *q.v.*

* indicates that the word or expression before which it appears did not come first or independently into English in the United States. In the lists of combinations the star on the second element denotes that the combination is not of United States origin.

( ) are used in entry words sometimes to inclose a letter which may or may not be found in the spelling. In phrases parentheses are used about a word to indicate that it may or may not occur in the expression. Regularly, in the definitions of transitive verbs, parentheses are used for the direct object or to indicate the nature of the object.

[ ] regularly contain the etymologies. Brackets are also used to inclose matter supplied by the editor. Quotations are inclosed by brackets when they have some bearing upon, but do not contain, just the term being illustrated or when the quotation can hardly be cited as a legitimate or valid occurrence of the word in question.

† is used before spellings that are obsolete.

> is used for "whence," i.e., from which is derived.

< is used for "from, derived from."

+ is used for "and" in etymologies.

* is used before a hypothetical form.

The quotations are given in the form usual in historical dictionaries. Large capitals indicate volumes or other large divisions, small capitals parts or sections, and lower-case letters chapters or prefatory pages, e.g., III, II, xxi. When a subordinate sense marked **b** follows a sense which is not marked **a**, this signifies that the subordinate sense is regarded as being derived from the antecedent one.

Figure C.2: Explanation of special lettering and symbols

LIST OF ABBREVIATIONS

a (before a date) . . . = *ante*, before
a. . . . . = adjective
absol. . . . . = absolute, -ly
adv. . . . . = adverb
advt. . . . . = advertisement
Amer. . . . . = America(n)
Amer. Sp. . . . . = American Spanish
app. . . . . = apparently
attrib. . . . . = attributive, attributively
B. '48, etc. . . . . = J. R. Bartlett, *Americanisms* (1848, 1859, 1877)
B. and L. . . . . = Albert Barrère and C. G. Leland, *Dictionary of Slang* (1888–90)
Br. Wtb. . . . . . = *Versuch eines Bremischniedersächsischen Wörterbuchs*
c (before a date) . . . = *circa*, about
Camb. . . . . = Cambridge
Can. F. . . . . = Canadian French
cap. . . . . = capitalized
Cent. . . . . . = *Century Dictionary* (1889–91)
cf. . . . . = confer, compare
collect. . . . . = collective, -ly
colloq. . . . . = colloquial -ly
combs. . . . . = combinations
DAB . . . . . = *Dictionary of American Biography*
DAE . . . . . = *Dictionary of American English*
dial. . . . . = dialect
dim. . . . . = diminutive
Doc. . . . . = Document, -ary
Du. . . . . = Dutch
Econ. . . . . = Economics
ed. . . . . = edition
EDD . . . . . = *English Dialect Dictionary*
Educ. . . . . = Education, -al
ellipt. . . . . = elliptical
Eng. . . . . = English
esp. . . . . = especially
F. . . . . = French; J. S. Farmer, *Americanisms* (1889)
f. . . . . = from
F. and H. . . . . = J. S. Farmer and W. E. Henley, *Slang* . . . (1890–1904)
f(f). . . . . = following
fig. . . . . = figurative
fl. . . . . . = *floruit*, flourished (followed by date)
freq. . . . . = frequently
G., Ger. . . . . = German
Geol. . . . . = Geology
Gk., Gr. . . . . = Greek
hist. . . . . = historical
i.e. . . . . . = *id est*, that is
imper. . . . . = imperative
interj. . . . . = interjection
intr. . . . . = intransitive
irreg. . . . . = irregular, -ly
L. . . . . = Latin
LG . . . . . = Low German
masc. . . . . = masculine
MF . . . . . = Middle French (14th–16th cent.)
misc. . . . . = miscellaneous
n. . . . . = noun
N. Amer. . . . . = North America(n)
naut. . . . . = nautical
N. Eng. . . . . = New England
no. Eng. . . . . = northern England (English)
n.s. . . . . = new series
obs. . . . . = obsolete
OED . . . . . = *Oxford English Dictionary*
OHG . . . . . = Old High German
opp. . . . . = opposite
orig. . . . . = originally
p. . . . . = page
Pa. G. . . . . = Pennsylvania German
pass. . . . . = passive
Pg. . . . . = Portuguese
pl. . . . . = plural
Polit. . . . . = Political
poss. . . . . = possibly
prec. . . . . = preceding
prep. . . . . = preposition
prob. . . . . = probably
pron. . . . . = pronoun
pt. . . . . = part
quot(s) . . . . . = quotation(s)
q.v., *qq.v.* . . . . . = *quod vide*, which see
R. . . . . = Robert L. Ramsay, *Mark Twain Lexicon* (1938)
R., r. . . . . = river
Russ. . . . . = Russian
S. . . . . = Southern, South
sc. . . . . . = *scilicet*, understand or supply
Ser. . . . . = Series
Sess. . . . . = Session
sing. . . . . = singular
specif. . . . . = specifically
Stand. . . . . . = *Standard Dictionary* (Funk & Wagnalls, 1893–95)
Supp. . . . . . = *Supplement*
S.W. . . . . = Southwest
Sp. . . . . = Spanish
s.v. . . . . . = *sub verbum*, under the word
Th. . . . . = R. H. Thornton, *Glossary* (1912)
theat. . . . . = theatrical
Th. Supp. . . . . = R. H. Thornton, Supplement in *Dialect Notes*, Vol. VI
tr. . . . . = transitive, translation
transf. . . . . = transferred (sense)
usu. . . . . = usually
v. . . . . = verb
var. . . . . = variant
Ver. . . . . = Verwijs en Verdam, *Middelnederlandsch Woordenboek*
W. . . . . = West, Western, Webster
We. . . . . = Joseph A. Weingarten, *Supplementary Notes to the Dictionary of American English* (1948)
WNT . . . . . = de Vries en te Winkel, *Woordenboek der Nederlandsche Taal*

Figure C.3: List of abbreviations

of this species is considered more musical than that of the Louisiana Water-Thrush. — (13) 1844 *Nat. Hist. N.Y., Zoology* II. 78 The New York Water Thrush. *Seiurus Noveboracensis*. . . . This musical little bird . . . is partial to the neighborhood of brooks, in search of insects.

c. In the names of plants and fruits, as (1) **New York fern,** (2) **Gloria Mundi,** (3) **shield fern,** (see quots.).

(1) 1943 SHIMER *Plant Names* 47 New York fern, Dryopteris noveboracensis. — (2) 1817 W. COXE *Fruit Trees* 117 Monstrous Pippin, or New-York Gloria Mundi. This apple originated on Long Island, state of New-York; it is of an uncommonly large size. — (3) 1843 TORREY *Flora N.Y.* II. 497 *Aspidium Noveboracense*. . . . New-York Shield-fern. Moist woods and thickets. 1901 MOHR *Plant Life Ala.* 316 *Dryopteris noveboracensis*. . . . New York Shield Fern. . . . Alleghenian and Carolinian areas.

d. Designating a **biscuit, cracker, cupcake,** such as were formerly popular in New York. *Obs.*

1714 SAMUEL SEWALL *Diary* II. 440, I had my New York Biscuit to eat, and a Bottle of Wine. 1846 W. G. STEWART *Attowan* I. 14 Their contents consisted of . . . the biscuit root, tasting exactly like a New York cracker newly baked. 1853 WEBSTER *Improved Housewife* 111 New York Cup Cakes.

2. In derivative expressions: (1) **New Yorkeress,** a female New Yorker; (2) **Yorkese,** a variety of English regarded as characteristic of the inhabitants of New York City, cf. **Bostonese 2;** (3) **Yorkish,** *a.* characteristic of New York; (4) **Yorkism,** a term in New Yorkese; (5) **Yorky, =New Yorkish.**

(1) 1871 HOWELLS *Wedding Journey* i. 10 The New-Yorkeress was stylish, undeniably effective. — (2) 1894 *Harper's Mag.* Oct. 695/1 'Cafe' . . . is New Yorkese for dram-shop. 1948 *Time* 13 Sep. 83/1 Wanted for radio series: one girl who speaks New Yorkese, has bad diction and careless enunciation. — (3) 1894 HOWELLS in *Harper's Mag.* May 822/2 The Nation was always more Bostonian than New-Yorkish by nature. — (4) 1832 *N.Y. Mirror* 12 May 359/2 The fashion of *moving* (is it not a wretched New-Yorkism?) has all my life given me a great annual disturbance.

(5) 1908 E. WHARTON *Hermit* 150 To be compared to her next! to be accused of being 'New Yorky'!

b. Esp. **New Yorker,** a native or inhabitant of New York.

1756 WASHINGTON *Writings* I. 315 The Jerseys and New Yorkers, I do not remember what it is they give [to their soldiers]. 1884 MATTHEWS & BUNNER *In Partnership* 127 'Are you a New Yorker, sir?' 'From the north of the State.' 1948 *N.Y. Star* 30 June 14/3 The Board of Transportation is appealing to New Yorkers to put up patiently with the confusion.

New York Indians. Indians of various tribes that formerly lived in the state of New York.

1827 *Spirit of Seventy-Six* (Frankfort, Ky.) 4 Oct. 2/4 The Menominie, Chippewa, Winnebago, and New York Indians, and a few of the Ottawas, were parties to it. 1894 ROBLEY *Bourbon Co., Kans.* 7 These various tribes of New York Indians, consisting of the remnants of the Senecas, Onondagas, Cayugas [etc.] . . . were called the 'Six Nations.' 1946 FOREMAN *Last Trek* 335 A number of these so-called New York Indians were living in Canada.

New-Yorkize n(j)uˈjȯrkaɪz, *v. tr.* To give (something or someone) the character or appearance of the institutions or people of New York City. *Colloq.*

1867 *Atlantic Mo.* March 342/2 What a reproach to Tammany, that a politician in far-off Chicago should have been the first to see the mode of New-Yorkizing the politics of the South! 1871 HOWELLS *Wedding Journey* i. 33 Broadway had filled her length with . . . that easily distinguishable class of lately New-Yorkized people from other places. 1942 LILLARD *Desert Challenge* 94 Will it, like much of Florida and some of southern California, be New Yorkized?

New York shilling. A shilling forming part of the New York currency *q.v. Obs.*

"At the time when the decimal system was adopted by the United States, the shilling or twentieth part of the pound in the currency of New England and Virginia was equal to one sixth of a dollar; in that of New York and North Carolina, to one eighth of a dollar" (*Cent. s.v. shilling*).

1834 *Knickerb.* III. 349 A levy was a coin; corresponding . . . to a New York shilling. 1836 CROCKETT *Exploits* 19 [The barkeeper] knew that a coon was as good a legal tender for a quart in the west, as a New York shilling, any day in the year. 1879 WILLIAMS *Pacific Tourist* 277/1 Carriages to any part of the city may be had for 'four bits'; the 'bit' being equivalent to the old New York shilling.

New York Stock Exchange. An organization of brokers in New York for buying and selling securities according to established rules; also the place where the trading is done. Also attrib.

The New York stock exchange is the oldest of its kind in the United States. On May 17, 1792, twenty-four brokers met on a spot across from 60 Wall St., and drew up a working agreement. Formal organization was effected in 1817.

[1842 (*title*), Report of the Committee of the New York Stock and Exchange Board.] 1862 *Amer. Ann. Cyclo. 1861* 307/2 The highest, lowest, and average quotations for 1859, 1860, and 1861, at the New York Stock Exchange for the stocks most largely dealt in. 1900 NELSON *A B C Wall St.* 141 New York Stock Exchange seats command . . . $40,000. 1949 *Sat. Ev. Post* 29 Oct. 144/3, I was doing my yelling just then to a certain quotation clerk on the floor of the New York Stock Exchange.

***next,** *a.* and *adv.*

1. Aware of things, "wise," esp. in phrases. *Slang.* Cf. ***get,** *v.* **7. i.**

1896 G. ADE *Artie* xvi. 146 I've been next, I'd tell you those. 1910 W. M. RAINE *B. O'Connor* 225 Mrs. Mackenzie will put you next to the etiquette wrinkles when you are shy.

2. next man, anyone taken at random, the next comer. Usu. in comparative phrases introduced by *as.*

1857 *Lawrence* (Kans.) *Republican* 18 June 2 The Judge . . . will probably talk as long to a crowd without tiring them as the next man. 1902 S. G. FISHER *True Hist. Amer. Revol.* 146 We do not surrender our property to the next man who is an abler business manager. 1908 *N.Y. Ev. Post* 29 June 4 Mr. Bryan knows this as well as the next man.

Nez Percé. [F. "pierced nose," though there is no proof that the Nez Percé Indians practiced nose-piercing.] An Indian of the principal tribe of the Shahaptian family, discovered by Lewis and Clark in 1805 in what is now western Idaho. Also *pl.*, the tribe. Cf. **Pierced Nose.**

1832 in *Overland to Pacific* IV. (1934) 120 Here [hunters' rendezvous near Lewis River] we found about 120 Lodges of the Nez Perces and about 80 of the Flatheads. 1837 IRVING *Bonneville* I. 169 In another part of the field of action, a Nez Perce had crouched behind the trunk of a fallen tree and kept up a galling fire from covert. 1884 BARROWS *Oregon* 121 The Rev. Mr. Parker joined himself to the Nez Percés, and under their . . . protection, threaded his way to Walla Walla. 1947 DEVOTO *Across Wide Missouri* 97 He might have added Kanakas, Irish, Bannocks, Nez Perces, and Flatheads to the melting pot.

attrib. 1812 in *S. Dak. Hist. Coll.* IV. (1908) 157 The . . . Nez Perce nation have a tradition that the human race spring from this dog [prairie dog] and the beaver. 1832 *Ev. & Morning Star* (Independence, Mo.) Oct. 7/1 There were, of Capt. S's Fur company, Capt. Wythe's Oregon company, &c. about 250; of the Nepersee Indians, making a force of 300 against from 80 to 100 of the Black feet, Indians. 1884 BARROWS *Oregon* 121 [Whitman and Parker] met the Nez Percé Flat-Heads. 1949 *Pacific Discovery* May–June 16/1 According to some it is derived from the Nez Percé word meaning 'muddy water.'

N.G., *n.* and *a.* Abbreviation for "no go" or "no good." *Colloq.*

1839 *N.O. Picayune* 21 April 2/4 Though his grey-headed rival tried to win, it was n.g. (no go!) 1840 *St. Louis D. Pennant* 20 June (Th.), The bells, boys, and engines tried to get up a fire last night, but it was N.G. 1888 *Cin. W. Gazette* 22 Feb. (F.), Hill claims . . . that he will make the farmer sweat who have been asserting that his claim was N.G. 1928 *Amer. Mercury* Aug. 477/2 It's N.G.! Let's go!

Niagara naɪˈægərə, *n.* [Iroquoian. Name of a river and a famous waterfall between Lakes Erie and Ontario.]

1. In transferred uses (see quots.).

1843 STEPHENS *High Life N.Y.* II. 243 The winders were . . . kivered from top tu bottom with a hull Niagara of red silk. 1872 MARK TWAIN *Roughing It* 530 The flaming torrent . . . remains there [in the Hawaiian Islands] to-day all seamed, and frothed, and rippled, a petrified Niagara. 1912 COBB *Back Home* 321 Rivers of red pop had already flowed, Niagaras of lager beer and stick gin had been swallowed up. 1947 BEEBE *Mixed Train Dly.* 99 For two dollars we mounted to a bedroom whose Irish linen sheets, shaded bed lamps and Niagaras of hot water could have spelled luxury in New York.

b. (See quot.) *Obs.*

1865 SALA *Diary* II. 180 An elastic pipe must have passed through one of the 'Niagaras' or 'cataract curls'—the name given to the shower of true or false ringlets the ladies are in the habit of wearing at the back of their heads.

2. A kind of green grape descended from the Concord.

1884 *N.Y. Wkly. Tribune* 6 Feb. 10/2 Concord is still far in the lead, though Worden has many friends as the 'coming' black, and Niagara is being most largely tested among the whites. 1916 ALWOOD *Chemical Comp. Amer. Grapes* 6 [Grapes tested by government analysts include] Niagara, North Bass, Ohio. 1945 *Greeley* (Colo.) *D. Tribune* 15

Figure C.4: Page 1129

March 10/8 Grapes: Concord, Fredonia, Wordon, Golden Muscat, Niagara, Portland are all hardy.

3. In special combs.: (1) **Niagara cane,** ?a walking cane obtained as a souvenir at Niagara Falls, *rare;* (2) **green,** bluish-green; (3) **gudgeon,** (see quot.); (4) **limestone,** a limestone formation occurring in the Upper Silurian in New York; (5) **shale,** (see quot. 1891); (6) **thyme,** (see quot.).

(1) **1891** WELCH *Recoll. 1830–50* 147 He wore a short, drab brown, sack coat, the prevailing fashion of those days, his hands stuck in his low coat pocket, his right hand holding a crooked neck Niagara cane, wrong end up, handle in the pocket. — (2) **1901** *World's Work* Aug. 1035/1 Running through the whole plan from the deeper barbaric primary colors to the delicate blue on the propylæa there greets you everywhere at intervals the Niagara green. — (3) **1842** *Nat. Hist. N.Y., Zoology* IV. 394 The Niagara Gudgeon. *Gobio cataractae.* . . . Body elongated and rounded. — (4) **1862** *Amer. Jrnl. Sci. & Arts* 2 Ser. XXXIII. 47 The regularly bedded layers present the usual lithological characteristics of the Niagara Limestone as it appears in Northern Illinois and Iowa. **1899** U.S. Geol. Surv. *Water Supp. Paper* No. 21, 14 At about 110 feet Lockport (often called Niagara) limestone is entered, which furnishes water containing sulphureted hydrogen.

(5) **1878** *Geol. Survey Ohio Rep.* III. 386 The Niagara shales here overlie the Clinton limestones. **1891** *Cent.* 5545/3 Niagara shale, a division of the Niagara group, especially interesting from its relation to the recession of Niagara Falls. It is there a shaly rock, and it underlies a more compact limestone, each division being at the present Falls about 80 feet thick. The shale wears away more rapidly than the limestone, which is thus undermined and breaks off in large fragments, greatly aiding the work of the water in causing the recession of the Falls. — (6) **1843** TORREY *Flora N.Y.* II. 67 *Micromeria glabella.* . . . Niagara Thyme. . . . On calcareous rocks, about the Falls of Niagara; Goat Island, and on Table Rock.

∗**nibbler,** *n.* The cunner, noted for nibbling the bait off hooks. *Colloq.* Cf. ∗**nipper 2,** and ∗**scamp.** — **1842** *Nat. Hist. N.Y., Zoology* IV. 173 The Bergall has various popular names: Nibbler, from its vexatious nibbling at the bait thrown out for other fishes [etc.]. **1859** BARTLETT 58 *Burgall, (Ctenolabrus cerulous).* . . . Other names are Nibbler, . . . and in New England, those of Blue Perch and Conner.

∗**Nicholas,** *n.* As the last term in **Saint, son of Saint Nicholas.**

Nicholite ˈnɪkl̩aɪt, *n.* Also **Nicolite.** [Origin unknown.] A member of a religious sect in some respects resembling the Quakers. In full **Nicholite Quaker.** *Obs.*

1786 *Md. Journal* 21 Feb. (Th.), Not a Presbyterian, Baptist, Methodist, Dunker, Menonist, Nicolite nor even the peaceable old Quaker can now be prevailed with to contribute a single farthing towards the good design. **1804** DOW *Journal* (1814) 196 At night I lodged with one of the Nicholites, a kind of Quakers who do not feel free to wear coloured cloaths. *c***1870** CHIPMAN *Notes on Bartlett* 292 *Nicholites, Nicholite Quakers.* A sect in and about Delaware, about 1750–90.

∗**Nicholson,** *n.* [f. the name of the inventor.] (See quot. 1870.) Also **Nicholson-paved.** *Obs.*

1870 MACRAE *Americans* II. 172 In St. Louis I observed some streets floored with iron gratings, others macadamized, and others paved with wooden bricks laid on a floor of sanded planks, and cemented with asphalt. This is called the Nicholson pavement, and is found in New York, Chicago, and other cities as well as St. Louis. As long as it lasts, it is as safe, smooth, and noiseless a road as could be desired. **1881** *Harper's Mag.* April 711/1 Its broad, Nicholson-paved business streets are bounded . . . with warehouses. **1884** SWEET & KNOX *Through Texas* xxii. 301 'This,' said he, tapping with his cane one of the bowlders on the pavement, 'is none of your slippery asphalt or Nicholson.'

∗**nick,** *n.*[1] Crossbreeding in which a superior offspring results, or the animal secured in this way.

1889 WARFIELD *Cattle-Breed* 26 This thing of a 'nick,' or a successful cross, is as difficult as determining beforehand how much an animal will inherit from one or the other of its parents. **1897** *Outing* XXIX. 484/1 In time Star, a good one in the field, was bred to Druid, and Mr. Wells made a record with this nick. **1949** *Time* 1 Oct. 90/2 Actually, a breeder may run through hundreds of combinations before he hits a 'nick'—trade slang for a good hybrid.

nick nɪk, *n.*[2] Short for ∗**nickel 1.** *Colloq.* — **1857** *N.Y. Herald* 27 May 5/2 The bags containing the 'Nicks' were neat little canvass [*sic*] arrangements, each of which held five hundred [of the new coins]. **1865** SALA *Diary* II. 54 Two sticks of lollipops are to be had for two 'nicks.'

∗**nickel,** *n.*

1. Any one of various coins made in part of nickel, esp. a one-cent piece authorized in 1857 and discontinued in 1864. *Obs.* Cf. **three-cent,** *a.*, **nickel cent.**

1857 *N.Y. Herald* 27 May 4/6 'Nary red' will soon be an obsolete phrase among the 'boys,' and 'nary nickel' will take its place. **1863** *Chi. Tribune* 1 May 2/3 The heavy coinage of 'nickels' still continues, the number last week made at the mint in Philadelphia being 53,000. **1918** in *Pub. Col. Soc.* XX. 222 By 'nickels' are meant one-cent coins made of copper-nickel, first coined (of that material) . . . in 1858.

b. A five-cent piece made of three parts copper and one part nickel. Also fig.

1881 ROMSPERT *Western Echo* 233 My sales ran from seventy-five to one hundred and fifty dollars per day for several weeks. I shall leave the reader to *guess* at the *margins,* and only say that we did not *deal* in nickels. **1903** ALDRICH *Ponkapog P.* 120 Reaching out with some other man's hat for the stray nickel of your sympathy. **1949** *Nat. Geog. Mag.* Oct. 506/2 He had to import sacks of nickels and dimes.

2. In combs.: (1) **nickel bank,** a kind of gambling game in which nickels are used, *obs.;* (2) **cent,** = ∗**nickel 1,** *obs.;* (3) **-in-the-slot machine,** see as a main entry; (4) **novel,** a cheap, paper-covered, trashy novel, also attrib., cf. **dime novel;** (5) **nurser,** a tightwad, one who is stingy, *slang;* (6) **show,** a show to which a five-cent admission is charged, cf. next; (7) **theater,** =**nickelodeon 1.**

(1) **1888** *N.Y. World* May (Farmer), He started a nickel bank of his own, and won both fame and fortune as a gambler. — (2) **1863** G. HAMILTON *Gala-Days* 305, I shall by and by throw you a paltry nickel cent for your tropical dreams. **1872** *Harper's Mag.* Aug. 347/1 To make amends, he gave it a handful of nickel cents. — (4) **1896** SHINN *Story of Mine* 1 Pistols and bandits abound in a nickel-novel atmosphere. **1944** *Reader's Digest* July 71 A potbellied stove glowed in a corner and Harry, stretched on his back, read nickel novels all day and drank from an endless supply of beer bottles stored under the bed.

(5) **1924** *Cosmopolitan* Dec. 70/2 'A proper nickle-nurser, what I mean!' is Jerry Murphy's verdict. 'He is too stingy to harbor a doubt!' **1929** *Cent. Mag.* Autumn 64 'Bonehead,' 'nickel-nurser,' and flat-tire' are original tokens of his esteem for humanity. — (6) **1914** LOWRY *Himself* 165 Ragged and dirty children attending 'nickel shows' and buying quantities of cheap candy. — (7) **1912** BRECKINRIDGE & ABBOTT *Delinquent Child* 157 We have had . . . the excitement provided by the 'nickel theater.' **1914** *Collier's* 17 Jan. 4/3 There are too many nickel theatres around here.

As the last term in **five-cent, liberty head, plugged nickel.**

nickel-in-the-slot machine. A machine in which, by the insertion of a nickel in a slot, certain gears are, or may be, moved, thereby releasing gum, a bar of candy, etc. Also **nickel-in-the-slot scheme, nickel slot-machine.**

1889 *Tacoma* (Wash.) *News* 13 Dec. 3/5 The latest nickel-in-the-slot scheme is really a stroke of genius and is destined to revolutionize cheap literature in this country. **1893** *Harper's Mag.* March 494 [In Jacksonville] there were the same . . . nickel-in-the-slot machines [as in Asbury Park]. **1914** *Calif. Appellate Rep.* XXIII. 769 The statute under discussion relates only to gambling machines, and does not purport to make the business of manufacturing 'coin operating machines commonly known as nickel-in-the-slot machines' unlawful. **1947** *So. Sierran* Nov. 1/3 In fact, my $1.10 is sitting in a nickel slot-machine there.

nickelodeon ˌnɪkl̩ˈodrən, *n.* [**nickel** (sense **1b.**)+ ∗*odeon* (var. of ∗*odeum*), a music hall.]

1. In the early days of motion pictures, a theater where a short picture, with or without singing, dancing, etc., could be seen for five cents. Now *hist.* Cf. **nicolet.**

1888 *Boston Transcript* 26 Nov. 5/7 Austin's Nickelodeon. . . . Open Day and Evening. Shows Hourly. **1908** *World To-day* Oct. 1053/1 There is no town of any size in the United States which does not contain at least one nickelodeon [moving-picture show]. **1949** *Business Week* 1 Oct. 6/3 He has been a success in the movie business since 1907, when he opened his first nickelodeon.

b. nickelodeon machine, ?=**mutoscope.**

1944 HOLTON *Yankees* 63 Needless to say Mother was completely enamored of them from the minute she looked into the top of the first little nickelodeon machine.

2. (See quot.)

1913 *Stand.* 1673/2 *Nickelodeon.* . . . a place of amusement generally charging no admission fee, containing various automatic machines, such as cinematographs, graphophones, etc., which may be used by patrons for a small charge.

3. A juke box.

1938 *Fla. Review* Spring 25/1 The requisites of a place entitling it to the name *jook* are . . . presence of the nickelodeon, and . . . of the dance-floor [etc.]. **1949** *Sat. Ev. Post* 15 Jan. 88/3 A nickelodeon at the end of the street emits a tinny piano tinkle.

Figure C.5: Page 1130