Construction and Analysis of the University of Georgia Tobacco Documents Corpus

by

Clayton M. Darwin

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

Abstract

This dissertation provides a detailed description of the construction and analysis of the University of Georgia Tobacco Documents Corpus, a representative corpus of tobacco-industry documents designed to serve as a norm of written tobacco-industry discourse for the University of Georgia Tobacco-Documents Project (2001–2004). The Tobacco Documents Corpus was constructed as part of the National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (NIH-NCI) grant 1 RO1 CA87490-01, 'Linguistic Analyses of Tobacco Industry Documents.' This description is provided primarily as a means of demonstrating the viability of the given premise, that it is possible to manage and describe large document sets—apart from extensive review of individual texts—by using a combination of Corpus Linguistics, Humanities Computing, and Statistics methods. Secondarily, it provides the specifics of the project necessary to 1) properly implement the resultant corpus as a norm for comparison studies and interpret related data, and 2) use the Tobacco Documents Corpus as a model for similar projects. In particular, this work presents the underlying theory, implementation, and results of each step in the process of corpus creation and description, from the initial sampling and conversion of documents, through the statistical description and analysis of the resultant corpus, and

ultimately (although in a limited form) to the distribution of the corpus and associated analyses via Compact Disc and the Internet (http://www.tobaccodocs.uga.edu/TDC). Subtopics addressed include category theory (categorization and classification), statistical sampling, text markup using Extensible Markup Language (XML), text extraction using Extensible Stylesheet Language (XSL) and XSL transformations (XSLT), tokenizing, parsing, count methods, and proportions analysis. To a limited extent, this work addresses scripting using the Python programming language as a tool for corpus construction and analysis, and the Internet as a means for displaying corpus data and analyses. Based on the overall success of the Tobacco Documents Corpus, it is believed that this process description will be a contribution to the developing field of Corpus Linguistics, particularly in the area of large-scale document analysis and text-mining.

INDEX WORDS: Corpus Linguistics, Humanities Computing, Markup schema, Statistical sampling, Text mining, Tobacco documents, Dissertations (academic)

Construction and Analysis of the University of Georgia Tobacco Documents Corpus

by

Clayton M. Darwin

B.A., Central Washington University, 1995

M.A., Central Washington University, 1997

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

2008

Construction and Analysis of the University of Georgia Tobacco Documents Corpus

by

Clayton M. Darwin

Approved:

Major Professor: William A. Kretzschmar, Jr.

Committee: Donald L. Rubin
Michael A. Covington
Marlyse Baptista

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

Dedication

There are certainly a great number of candidates for this dedication, but given that this is primarily an academic work, it seems best that it be dedicated to the individual most influential in this realm of my existence. Undoubtedly, this is my professor and friend Dr. William A. Kretzschmar.

Although it is known to many that Dr. K has long served as my academic mentor and major professor, less known is that it was Dr. K who originally encouraged me to come to the University of Georgia, who recommended me for my initial university-funded Research Assistantship, who later helped me land a second Research Assistantship with the Tobacco Documents Project, who encouraged me to choose the TDC as a topic for my dissertation, who introduced me to my current employer, and who has continued to encourage me as I worked through this last step. Thus Dr. K's influence on my career extends beyond the classroom, and I believe, beyond what is expected of a professor. But not just my career. I have seen this same attention to student needs with all of my peers.

As much as the above items are appreciated, I have chosen to dedicate this work to Dr. K for a different reason. During my first semester as a doctoral student, as I was being delivered from my academic naiveté, constantly bombarded by opposing theories and conflicting results, I found myself in despair (academically speaking mind you) and voiced such in class. Dr. K. simply said this (paraphrasing), 'There is something useful in every work. Find out what it is, learn it, leave the rest, and keep going.' And now, these ten years later, I find myself graduating, already employed as a linguist and exploring the fringes of the field with my own work. But what I am as a linguist, what I believe, and how I approach my work, are based on that single principle of which Dr. K so timely reminded me. It allowed me to continue past that first semester, and it has allowed me to continue in the field. It's all

about looking for the good and useful, being willing to give up the unproductive, and moving forward.

Acknowledgments

The University of Georgia Tobacco Documents Project was generously supported by a grant entitled 'Linguistic Analyses of Tobacco Industry Documents' (RO1 CA 87490) from the National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services. The views expressed in this dissertation, however, are solely those of the author.

I begin these acknowledgments with the confession that what follows is incomplete and wholly inadequate to express the gratitude I have for all the individuals that have employed, supported, helped, encouraged, and prodded me through this ten-year adventure of completing my dissertation and degree. Time and space constraints prevent me from mentioning them all, and my rhetorical ineptness prevents me from expressing my true appreciation. However, there are several persons yet unmentioned who have had such an impact on the process that this dissertation would not have been completed without their help. These deserve formal recognition.

To begin, I would like to express my sincere appreciation to Dr. Donald Rubin for hiring me as a Graduate Assistant for the Tobacco Documents Project (and allowing me to hang out with him for three years). Not only did he give me a job, but he also gave me the freedom in that job to explore ideas. My belief is that this freedom is what allowed the project to reach its full potential, and ultimately what prepared me for the position I have now.

As well, I would like to acknowledge the work done by the project archivists. In particular, the project is deeply indebted to our lead archivist Anastasia (Stacy) Wright, but also Cati Brown (now Dr. Brown), Brooke Heller, Sean Matthews, Elizabeth Taxel, and Hollie White. Ultimately, the completion and success of the TDC Project (and therefore this dissertation)

was dependent on their diligence in performing a tedious and often monotonous task, and I thank them for the hard work.

Also, many thanks to Drs. Michael Covington and Marlyse Baptista. Although not directly involved in the Tobacco Documents Project, they readily agreed to support me in this endeavor by serving as committee members and reading this dissertation (every word, even the appendices).

I am particularly grateful for the support of my boss and friend, Wayne West, during the final months of writing this dissertation. Although the 60-hour work weeks over the last few years did take their toll on the progress of this work, the 5-hour work weeks at the end allowed it to be completed. Thanks.

And finally, I conclude by saying very simply that I would not have finished this work had it not been for the constant support and patient encouragement of my wife Kim. She is my truest friend.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the fall of 1998 the National Association of Attorneys General (NAAG) reached a settlement with the seven leading United States tobacco industry organizations. These organizations include the manufacturer American Tobacco, Brown and Williamson, Lorillard, Philip Morris, and R. J. Reynolds, and the research and publicity groups Council for Tobacco Research and the Tobacco Institute. The purpose of the settlement, which is known as the Master Settlement Agreement or MSA, was to impose regulatory measures on the tobacco industry, particularly in areas related to advertising and public disclosure. As part of the MSA, the tobacco industry (that is, those seven organizations listed above) was and is required to release to the public all industry documents which are not considered attorney-client privileged and do not contain trade secrets (NAAG 1998, Section 4). The first deadline imposed for this disclosure was June 1999, and at that time the document set known as the NAAG Snapshot (all available industry documents up to the deadline) was made available to the public. This was done electronically via company web sites and physically in depositories in Minneapolis (the site of the original trial) and Guildford, England. As well, large collections of these documents are now available in electronic form on a number of non-tobacco-industry websites such as the Legacy Tobacco Documents Library at the University of California San Francisco (http://legacy.library.ucsf.edu, cited as 'Legacy') and Tobacco Documents Online (http://tobaccodocuments.org). This initial release contained approximately 3.5 million documents totaling over 30 million material pages. Since the snapshot, the tobacco industry has continued releasing documents, to include most documents which

had previously been classified as attorney-client privileged. The available set of tobacco documents currently totals over seven million and will continue to grow until at least 2010 when the MSA disclosure requirement ends (Legacy).

From the standpoint of linguistics, the number of documents alone represents a unique opportunity for study. However, the true uniqueness of the tobacco documents (TDs) and their value to linguistics and other fields comes from the fact that the document disclosure was forced, meaning the tobacco industry had little say about which documents were released. The result is that the TDs are a vastly superior representation of true industry discourse compared to document sets which were previously available. The documents are primarily intended for industry-internal audiences, and they cover the full range of business document types, from airline tickets, dinner receipts and inventory reports, to memoranda and policy letters, to research reports and court transcripts. As well, the TDs vary widely by date, from the 1800s to the present; by length, a few words to hundreds of pages; by implication, from incriminating to benign; and by style, from hand-written notes to highly edited press releases. Thus the TDs are very much the opposite of what has typically represented business discourse to date, namely business documents intended for public release which have undergone numerous revisions. This is simply because companies are reluctant to expose methods and strategies by releasing internal documents, even if their operations are ethically sound.

Following the initial release of the TDs, the primary use has been and continues to be for litigation against the tobacco industry and to support tobacco-control policy by exposing the unethical practices of the tobacco industry. However, over the last few years researchers from a variety of other fields have begun to discover the value of the TDs, and studies less directly related to tobacco control are becoming more frequent. The TDs are now serving as a primary data source for research in business ethics and policy making (see the continuing works of Robbin Derry as in Derry 2008), language and deception (Rubin 2001, Brown 2007), business methods (Malone 2003), business litigation (Daynard 2003), and biochemistry

(Pankow 2003). Although each of these projects is associated with tobacco control (the National Cancer Institute is the major provider of funding for TD research), one would expect the trend of moving away from purely tobacco-control studies to continue. This is particularly the case for linguistics, which until the MSA has not had access to a representative corpus of internal business documents of the type or magnitude provided by the TDs. Even the recent release of email documents from the Department of Justice investigation of Enron Incorporated (Cohen 2004), although certainly valuable for analysis, does not compare to the snapshot documents in terms of scope or magnitude, being only 517,431 non-duplicate documents in a single format (email) from a limited number of individuals (150).

Unfortunately, the study of this new corpus is not without trouble. What makes the TDs so attractive, the huge number and range of documents, also makes them difficult to approach systematically. For an individual, the number of documents is unreadable (roughly 1,000 pages a day for 82 years), and the range of document types is unmanageable. This is compounded by the fact that in general the document text is not in an easily readable form (i.e. low-quality digital images rather than text). Because of these issues, the method of study which has become the norm for TD research is to search document indexes and archives by whatever means are available and examine potentially useful documents case by case (Brown 2004). While this type of study can answer (although unreliably in terms of negation) the initial question of existence, which may be all that is necessary for litigation, it fails to answer the question of extent, which is fundamental for scientific study. Yet, as researchers move from the study of TDs for tobacco-control purposes towards the study of TDs on their own merit, it becomes more and more necessary for a general and principled description of the TDs as a whole to be set forth, as well as normative values for key indices. With this available, researchers would be able to determine not only that a topic exists in the TDs, but also how much, at what time, and in what context, such that a clear relationship could be established between the topic and the entirety of TDs. Of course, this would have to be coupled with new methods of conveying these norms to others quickly, bypassing the

traditional method of long-term exposure to the documents (i.e. familiarity), which is neither reliable nor unbiased.

1.2 Proposal

The real question to ask is whether or not it is possible to approach the documents systematically. Can one come to terms with such a large document set and obtain an unbiased, general understanding of its content, topics, and distributions? And if so, is it possible to make this knowledge available in such a manner that others can quickly reach a similar level of understanding, both of the data and the method used in obtaining them? If this is possible, then the implications are much wider reaching than the study of tobacco-industry documents. The methods become useful for the study of all large document sets. This is the case in the academic environment, but much more so in the commercial setting where the management and understanding of large document sets has become a prime area of interest. More and more in the current electronic age companies are finding themselves faced with the daunting task of making sense of terabytes of language data, whether in litigation (document disclosure orders), federal compliance requirements, employee and customer email, or archive management. What has been learned from the release of corporate documents such as the TDs and the Enron Incorporated emails, and from document disclosures in recent litigation, is that very little of a company's history is not documented and preserved in an analyzable form. This in turn has prompted changes in both governmental and legal industry procedures. Notably, there have been recent changes to the Federal Rules of Civil Procedure[1] in relation to the archival, disclosure, and destruction of electronically stored information (see FRPC 2006), with parallel sets of guidelines being released by prominent groups in the legal

[1]On December 1, 2006, changes in the Federal Rules of Civil Procedure (FRCP), USCS took effect. The changes to Rules 16, 26, 33, 34, 37 and 45 offer guidelines for counsel and the bench as they make decisions about the relevance, discoverability, production and costs associated with email, word processing documents, spreadsheets, databases and other forms of electronically stored information.

profession, such as The Sedona Conference (2007). Thus the problem is not a lack of text data, but making sense of the overwhelming volume.

My proposal is that there is a straightforward and reliable manner for approaching large document sets. The solution comes in the adaptation of proven methods from Corpus Linguistics, Humanities Computing, and Statistics. By modifying and integrating these methods, researchers are able to quickly analyze very large document sets and reach a general understanding of the document types, content, events and structures they contain, apart from extensive manual review of the texts. As well, the resultant data can be made readily accessible to others interested in the document set, and thus provide a reliable and statistically-sound foundation (starting point) for more in-depth study.

My avenue into the study and development of methods for describing large documents sets, as one might suspect at this point, is the tobacco-industry documents. More specifically, it is my involvement with the Tobacco Documents Project at the University of Georgia. This was a three-year project (Rubin 2001) funded by the National Cancer Institute[2] and lead by Donald L. Rubin, PhD (Principal Investigator, University of Georgia), along with Norbert Hirschhorn, MD (Co-Investigator), and William Kretzschmar, PhD (Investigator, University of Georgia); with the assistance of Douglas Biber, PhD (Northern Arizona University), Roderick Hart, PhD (University of Texas), and Roger W. Shuy, PhD (Georgetown University). The focus of the Tobacco Documents Project was the rhetorical analysis of deception in the tobacco documents through the examination of successive document drafts. However, early on it was realized that there was no suitable reference corpus with which to compare findings in order to determine if they were the result of deceptive strategies, or simply the norm for this specific genre of text. There was simply no reliable way to judge what the norm of tobacco communication might be. To remedy this, I was given the task of assembling and describing a representative text corpus of tobacco-industry documents to serve as that

[2]National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (NIH-NCI) grant 1 RO1 CA87490-01, 'Linguistic Analyses of Tobacco Industry Documents.'

norm. In essence I was asked to make the tobacco document set approachable and usable by providing researchers with avenues for discovery and for testing data and hypotheses. This reference corpus, which became known as the University of Georgia Tobacco Documents Corpus or TDC, is now complete, and this dissertation will present the specifics of the corpus project, both its successes and shortcomings, from the initial sampling and conversion of documents, through the statistical description and analysis of the resultant corpus, and ultimately (although in a limited form) to the distribution of the corpus associated analyses via Compact Disc and the Internet (http://www.tobaccodocs.uga.edu/TDC). The intent is to provide a detailed description of the underlying theory, implementation, and results of each step, and in this way present the project as a package, from conception to product, in sufficient detail that it can be easily replicated by researchers faced with the same task. This will demonstrate my premise.This process, I believe, will be a contribution to the developing field of Corpus Linguistics, particularly in the area of large-scale document analysis and text-mining.

Overall, the description of the TDC Project will consist of six chapters. This first chapter provides a general introduction to the tobacco-industry documents, the issues associated with analysis, and the proposed solution for overcoming these issues. The second is a general discussion of the concept of category as it applies to linguistics and the analysis and description of the TDC. In particular, it outlines how the success of the project and the correct interpretation of analyses are based on the careful application of categories. The third and fourth chapters provide detailed description and discussion of the methods, procedures, and decisions made in construction of the TDC. The fifth chapter describes the analysis methods used for discovery and description of corpus content. It provides descriptive statistics comparing the TDC to other established corpora as well as a number of examples (results) from TDC analysis that illustrate the analysis methods. The final chapter is a brief summation of the above.

1.3 Limits of Study

As interesting as the content of the Tobacco Documents may be, there are time when lines need to be drawn, and this is one of them. By necessity, the focus of the following chapters will be on methods and procedures used for the discovery of content, not on the content itself. In particular, I will not attempt to address the topic of deception. This has been investigated by other Tobacco Documents Project members (see Rubin 2004, Shuy 2003, Hirschhorn 2004, and Brown 2005). The examples used in the chapters are intended only to be illustrations of the given points and methods. In other words, all tobacco documents and related data presented in this work, however interesting and/or informative they are in their own right, were selected for their value as methodological illustrations, not as smoking guns implicating the tobacco industry for the murder of modern society. These types of judgments are left to the reader. The intent here is to provide an understanding of the underlying methods and related data such that through the use of the TDC and Toolkit the reader can rapidly move to a well-informed position from which to make such judgments.

It is also in my nature to revisit and reconsider past work ad nauseam and to continue with investigation and improvements. However, as much as I would like to discuss the continuation of the methods which will be presented in the following chapters, what is presented will be limited primarily to the work done on the TDC from June of 2001 to the project's end in June of 2004. As well, this work does not include any detailed discussion of programming and data preparation specifically related to the UGA Tobacco Document Corpus and Toolkit. In cases where particular issues have proven themselves problematic for subsequent TDC work, I have reserved a section in the conclusions chapter titled 'Possibilities and Improvements.' Here, those items now seen as errors will be noted as such and in sufficient detail to allows other to avoid them. As well, suggestions for further refinement will be made. However, none of these will be fully developed.

Finally, it is not my intention to mislead the reader into concluding that I alone am responsible for the whole of the TDC and its analysis. As the project supervisor, I have been

heavily involved in each of the steps which I will describe in the following chapters, but I have not acted alone. In particular, corpus construction (Chapters 3 and 4) was done primarily under the guidance of William Kretzschmar, with considerable assistance with archival from Anastasia Wright. The data analysis and internet presentation (found in Chapter 5) were supervised by Donald Rubin.

Chapter 2

Categorization

2.1 Introduction

Before beginning a more in-depth description of the reference corpus (hereafter referred to as the Tobacco Documents Corpus or TDC), I will diverge in this chapter to discuss categorization. The reason for this is that the defining of categories (categorization) and the placement of items into those categories (classification) are fundamental to all that follows. Although not specifically the focus of the next chapters, the necessity of clearly defined categories is a common thread that runs throughout this text, and to a large extent the success of the TDC Project is based on the categories used. They form the foundation of how the TDC documents were sampled, how they were archived, and ultimately how they were analyzed statistically, given that at its base the type of analysis used for the general description of the Tobacco Documents Corpus is an examination of how one category compares to another.

2.2 General Category

If there is single cognitive trait common to all humanity, my suspicion is that it would be a predisposition to divide, that is, to *categorize*. As a means of data management, we regularly devise taxonomies and paradigms that help catalog the vast amounts of information we encounter daily, just as I have done in the above paragraph by pointing out the differences between *categorization* and *classification*, and just as I will continue to do throughout the remainder of this chapter. At times the taxonomies we create are formalized, such as

mammals being that sub-group of animals with the features [+warm blood, +mammary glands, +hair, −feathers], but often they are informal, culturally-biased perceptions, such as Southerners being [+sweet tea, +overalls, −shoes, −education]. While the former example may indeed hold true, the latter obviously does not. Being both Southern and moderately educated, I prove it false on at least one account. Whatever the case, to divide is a fundamental process, and questions concerning it are common. 'Whose side are you on?' 'Is it right or wrong?' 'Is it plant or animal?' 'Is it bigger than a breadbox?' 'Are you a Democrat or a Republican?' 'What's your sign?' 'What do you do for a living?' 'Where did you go to school?' In other words, every person and every item we encounter, be it tangible or not, we strive to place into the correct category box, even if that box is labeled *don't care*. In fact, being able to categorize quickly causes us to be labeled ourselves (i.e. classified) and placed in the *decisive* box, which has a very positive connotation in Western[1] culture, rather than being dropped in the *fickle* box, which is far from positive. This perspective of believing that the ability to classify is a desirable trait may of course stem from the Judeo-Christian heritage of most Westerners. It is at least part of it. In fact, we find in the first words of the first book of the Mosaic Law two strict division, 'In the beginning God created the heavens and the earth' (Gen. 1:1 ASV). They are that 1) the creator is not the created, and that 2) the heavens and whatever they entail are not the earth and what it entails. We also find in the last book of the Christian New Testament this message to the church in Laodicea:

> I know thy works, that thou art neither cold nor hot: I would thou wert cold or hot. So because thou art lukewarm, and neither hot nor cold, I will spew thee out of my mouth (Rev. 3:16–17 ASV).

There is little room for fickleness and gray areas here, lest we be spewn out. Even still, the history of the Americas is very much a history of one group in opposition to another, from

[1]The term *Western* is used loosely in this chapter and refers generally to any cultural type similar to that of the author, a middle-class American of Northern-European descent.

Cortez's conquest of the Aztecs, to the arrival of the Jamestown Congregationalists, to the current debate on border control.

Although we regularly and with great insistence categorize and classify nearly everything we encounter, we seldom consider the reasons for, or the results of, these events. The consequence is that we end up with a great number of poorly defined categories and misclassified items.[2] These undefined categories and misclassifications are certainly functional in everyday life (why else would we use them?), yet when we move into the realm of science it becomes essential that our categories be more empirically based and that the rules for classification be defined well enough that classification can be replicated without significant error. And I confess here that my bias is that the field of Linguistics should be Science rather than Arts. While I do support the study of language as art, I do not classify that study as Linguistics. Nor do I consider lists of word types and their relative frequencies to be poetry. This being said, in the remainder of this section I will provide and discuss a general definition for category which allows a more scientific approach to the process of categorization and classification, and in the sections that follow relate the discussion of general category to empirical and applied linguistic category.

2.2.1 Defining Category

From the beginning of this chapter we have continually returned to the ideas of categorization and classification. Categorization can be formally defined at this point as the process of creating categories, and classification as the processes of testing and assigning items to the categories previously created. In other words, you first must make some boxes, and once you have them you can place items into them. Of course, what both of these processes hinge on is the idea of *category*, which can be defined simply as *a concept of definable difference*. This definition, as simple as it is, actually addresses both the processes of categorization and classification.

[2]Not stereotypes and prejudices, which are instead assumed overlaps of categories. For example, that *blond*, *not intelligent*, and *fun* define the same group.

The first item to note about the above definition is the idea that when we talk about category, what we reference is concept. That is, if we create a category (categorize), or simply accept one that has been suggested by others, what we have is essentially an idea, not any tangible item. Categories are mental constructs created for intellectual exercises. In other words, they are imaginary. It may be that the *difference* to which a given category refers is tangible, that the category has a distinct purpose, or that it fits a particular intellectual or cultural model (paradigm/taxonomy). Still, none are fixed. As Berlin and Kay demonstrate in their 1967 study *Basic Color Terms*, there is nothing absolute about the most common, everyday categorizations of the physical world, even those which we learn as children. The effect Berlin and Kay were able to describe, regardless of whether the true cause has been determined, is that the number of basic color categories recognized by humans varies widely from group to group. In their studies, Berlin and Kay found that some people groups[3] have as many as eleven basic color categories, while there are others that have as few as three. Thus, for me, *orange* is a category of color because I accept it as such, but it could just as well be *red* or *yellow*, or even *white* had I grown up influenced by another culture. Even in modern Western culture where it is commonly known that sunlight contains the complete visible spectrum of electromagnetic radiation, and where even cheap computer monitors allow 65,536 colors, 'How many colors are in a rainbow?' will rarely retrieve a count that exceeds that of a 95-cent pack of crayons. Thus, the continuum of visible light gets divided not by a universally correct system, but by what is convenient for the user in any given situation.

Of course, the idea of convenience as a basis for category means that categories are essentially arbitrary, varying according to the needs of the user. Although it is commonly said that 'one man's pleasure is another man's pain,' the formal acknowledgment of this idea is counterintuitive to the Western mind. Traditionally we categorize even the process

[3]The label *people group* is used here to avoid entering the nature-versus-nurture debate (i.e. culture versus genetics). Saunders 1998 provides an opposing argument to Berlin and Kay in relation to cause, although effect is not challenged.

of categorization itself, finding it as either inductive or deductive depending on whether it is governed by observation or logic. Regardless of which is chosen, the expectation is that it is governed by one or the other, with the implication that there is some absolute (correct) form. Yet being categories themselves, the concepts of induction and deduction, although useful, are far from fixed and certainly not mutually exclusive as categories. Consideration of existent categories indicates that in common practice categories are not strictly governed by logic, although they may be, nor are they strictly governed by observation, although it certainly may play a part. For example, in another study of category, *Categories of Eating in Tzeltal and Navaho* (1967), Brent Berlin notes that for speakers of Tzeltal

> . . . mushrooms are (metaphorically) meat in that they occur obligatory as direct objects of the verb *-ti* 'to eat meat, flesh.' [For example,] *ya hti' chewchew* 'I am eating mushrooms' and *ya hti' ti'bal* 'I am eating meat (unspecified)'. Informants will state that *-ti'* is the appropriate verb (vs. *-lo'* 'eat soft foods', *-kux* 'eat crunchy foods', *-we'* 'eat bready food, e.g., tortillas') because mushrooms do have the texture of flesh. Mushrooms are also referred to as *lumilal ti'bal* 'flesh of the earth'.

The fact that we (Westerners) allow biological taxonomies, which separate not only plants from animals but also mammals from mammals, to impinge on our culinary taxonomies and produce beef, pork, and mutton, is due more to our cultural and intellectual heritage than pure, unbiased observation and/or logic. As the Tzeltal show, they could just as well be divided functionally into how well they chew if the structure better served our immediate purposes. Thus in the above case, at least in terms of eating, one finds the highest forms of Animalia, represented by meat or flesh, in the same category as one of the lowest forms of Plantae, a fungus. Of course, this is only the case if we still accept 19th-Century taxonomies which classify fungi as 'plants without chlorophyll' (*Webster's New Dictionary*). Now the categories have changed. For the modern Western taxonomist, the mushroom is currently in its own kingdom *Fungi*, which is parallel to the kingdoms *Animalia* and *Plantae*, all

within the domain *Eukaryota* (http://www.ucmp.berkeley.edu/exhibits/historyoflife.php). This again is an illustration of category as an adaptable, malleable concept rather than having a fixed form (i.e. being real).

Once we are able to view category as concept rather than fixed reality, the overarching reason for category differences across group boundaries is much more straight-forward. That is, if category is concept rather than reality, then the expectation that follows is that categorization be driven by the needs of the users. Thus, differences between groups should be the norm rather than the exception. Continuing with the above example, the fact that Tzeltal taxonomies are different from Western taxonomies should not come as a surprise. The two groups are culturally diverse, and therefore have very different needs in terms of categorization. For the Tzeltal, given that the mushroom is named '*lumilal ti'bal* 'flesh of the earth' ' it is reasonable to assume that food quality is of greater cultural value than Darwinian explanations of origin. For the Western scientific mind, biological relationships are clearly of prime cultural importance. This is evident in the fact that even the humble *white button* mushroom, commonly found in cans and on pizzas throughout the Western world, has been given the lofty name *Eukaryota Fungi Dikarya Basidiomycota Agaricomycotina Agaricomycetes Agaricomycetidae Agaricales Agaricaceae Agaricus bisporus* ('*A.* bisporus' for short) which establishes its relationship to all other categorized entities (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?lvl=0&id=5341).

The question that remains, of course, is one of diachronic change. How is it that the button mushroom moved from the kingdom *Plantae* into the newly-formed kingdom *Fungi*? This is answered by the second half of the category definition, namely *definable difference*. When a category is created or accepted, the two primary implications it carries are that 1) there is a difference between what is included in the category and what is not, and that 2) the difference can be described. Without these two conditions being met, accurate classification is impossible and the category becomes meaningless simply because there is no way to determine if an item is or is not a member. This lack of definition is a key concept

(which will be addressed in more detail in Section 2.2.2), but to answer the above question of category change we need to look at the converse. It is the ability to describe difference that allows one to create meaningful categories. Thus, as that ability changes, categories too should naturally evolve, just as with *Fungi* becoming its own kingdom, or with changes in how color is described (by wavelengths or RGB values rather than names). Of course, *ability* to describe is as much dependent on audience as it is presenter, so for my youngest daughter a mushroom is 'kind of like a plant' and rainbows have about four colors when put to paper.

Another way to conceptualize the idea of *definable difference* is to consider the definition as a series of tests which together describe the meaningful difference between what is a member of a given category and what is not. In other words, the sum of the tests (whether one or many) forms a boundary around the category. What passes all the tests is included as a member, and what fails any single test is excluded. Borrowing from mathematics, this can be visualized more clearly using a Venn diagram and equating *category* with *set*.

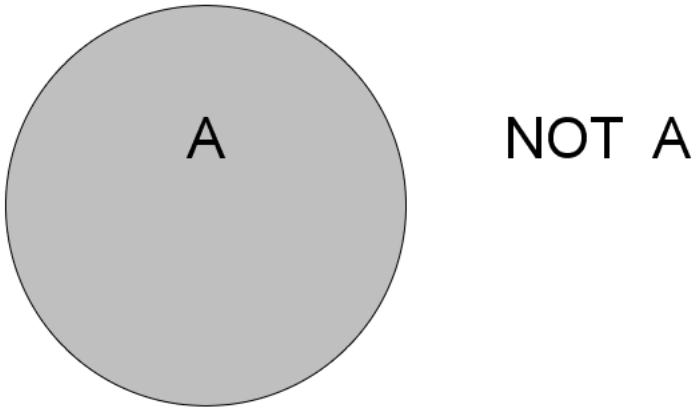


Figure 2.1: Venn Diagram 1: Single Category

In Figure 2.1, the category is A. The tests for being an element of A are represented by the border of the circle. Anything that passes the tests is a member of A, and what remains (everything else) is $NOT\ A$. As simple as this sounds, it illustrates a key characteristic of category, which is the fact that classification in its root form is a binary decision. That is,

the choice in classification is not actually whether an item belongs in category A or B, but more simply whether it is A or $NOT\ A$. Does it pass the tests which separate items that are A from those items that are not? This is a separate question from whether or not the item is B or $NOT\ B$, or C or $NOT\ C$.

Although tedious in conversation, these types of distinctions are common in programming where the explicit definitions required for variables are built from multiple layers of binary judgments. For example, given the above situation, in the Prolog programming language the categories A and B are represented by the following rules which define the arbitrary variables A and B:

```
'A'(X):-
    test1(X),
    test2(X),
    test3(X).
'B'(X):-
    test4(X),
    test5(X),
    test6(X).
```

The combination `:-` is read as a logical IF, and commas are logical ANDs. We also must assume that the test rules (which themselves are categories) already exist in the knowledge base. So, for each category A and B, the boundary between what is and is not a member is the sum of the tests the rule prescribes. Thus, X is A if X is $test1$, and X is $test2$, and X is $test3$. This is the definition of category A. Returning to the original issue, to determine if an item is A or B requires as many as two separate queries, one to determine if the item is a member of A, `'A'(item)`, and if that query fails, another to determine if it is in B, `'B'(item)`. In everyday life this is somewhat deceptive because the human mind makes decisions fast enough that they are considered simultaneous. So, if we are tasked with separating spoons from forks, on the surface the decision is 'spoon OR fork,' but on a lower level the logic is actually 'spoon OR (NOT spoon)' combined with 'fork OR (NOT fork).' This prevents the logical fallacy of either-or reasoning (which is actually either-or classification and is a much higher level function) and therefore prevents us from being confused by knives '(NOT spoon) AND (NOT fork)' and sporks 'spoon AND fork'. The end result of this view of category and

classification is that determining the *definable difference* (or boundary) is greatly simplified. This is because any single definition involves only one concept rather than multiple ones. In this way, 'spoon *OR* fork' becomes the higher level function such as '(spoon *AND* (*NOT* fork)) *OR* (fork *AND* (*NOT* spoon))' which by the way is an exclusive *OR*, meaning sporks are not allowed in this category. In Prolog, a more complex query could be used, or more rules could be added to the knowledge base (using A and B again):

```
'A or B'(X):-
    'A'(X),
    \+ 'B'(X).
'A or B'(X):-
    'B'(X),
    \+ 'A'(X).
```

The combination `\+` is read as a logical *NOT*, such that X is A or B if X is A and not B, or if X is B and not A. But still there are two separate decisions (rules to evaluate). If the lower-level definition actually determined 'spoon *OR* fork,' then there would also have to be additional definitions for 'spoon,' 'fork,' and 'spoon *AND* fork.' Theoretically this would amount to 2^n definitions, where n is the number of concepts included. With the binary view, there need only be n definitions, which can later be combined as needed for the desired result. This is critical because, as mentioned above, the creation of meaningful categories is based on one's ability to clearly define the tests which form the boundary. If rather than in or out a definition allows a maybe, meaning that it is impossible to tell if an item passed the inclusion tests or not, then the category itself has little or no value.

The last point that needs to be made concerning the definition of category is that *difference* implies two groups, which is just what should be expected for a binary decision. However, this expectation is more often not the case. Referring back to Figure 2.1, we see that the line drawn by the tests separates the field (often called the domain) into two parts, A and $NOT\ A$. Although the focus of attention is on A, the category also carries with it the complement $NOT\ A$, which is just as meaningful as A. The reason for this is that both the category and the complement are defined by the same set of tests, and consequently, one

cannot exist without the other. Even the domain of ALL things has $\emptyset$ (the empty or null set) for its complement, and a category definition that produces $\emptyset$ has ALL as a complement. Thus whenever a category is defined, the complement is also defined, and although usually hidden, it remains attached and should not be forgotten. This point will become more obvious and significant in Sections 2.2.2 and 2.4 below.

2.2.2 Category Statistics

The primary use we have for categories is in the building of taxonomies, and in this role they have the function of storage boxes. In other words, they hold similar items together. Once items are assembled together in a low-level-category box, this box can be grouped together with like categories into higher-level boxes. Beginning at the lowest levels, the boxes become the foundation of higher level structures, being grouped together into categories themselves, and so forth and so on until ultimately a model of a given domain is constructed. This is similar to the Prolog example above in which the tests, which themselves may be categories, were assembled to define a higher-level category, which then becomes a test for the next level of category. Because the domains being defined combine to form a model of the world in which we live (and I will not venture to say what that might be), there often comes a point in which some form of validation is desired. We want to know (or at least we should) that the categories which have been created are useful for describing relationships and making discoveries. This is the role of descriptive statistics, both in formal and informal categorization.

Single Category Statistics

There are two key ideas to have in mind when dealing with a single category. The first is that the decisions being made in classification, as discussed above, are fundamentally binary, determining if any given item is an element of the category or an element of the complement. The category elements are defined by the tests, as are the complement elements. Thus the entirety of the relationship between the category and the complement is described in the

definition. This relationship is simply that one is not the other. If this is not the case, the definition is illogical such that A is $NOT\ A$. The second key idea is that the domain is undefined. In other words, single categories select from the set of ALL items, which is infinite. If the domain were bounded, then there would no longer be a single category, but a category and subcategory. This is because any domain itself is a category, having a definition and perhaps some elements. Thus the count of category and complement elements for a single category is never a fixed number. There may be another item discovered and added at any time. In fact, any item discovered belongs to either the category or complement (for all categories).

Having these points in mind, it becomes easy to see that there can be no meaningful statistic produced from a single category. The reason for this is that statistics are for depicting relationships. Yet, for single categories the conceptual relationship between the category and complement is all that exists, and this is already fully described by the definition. As well, all counts are open-ended. Thus from a logical or mathematical standpoint, there is little to say. There must always be at least two categories for meaningful statistical measure (although those two will, of course, be intertwined with others).

Although the above may seem like an hypothetic discussion, it is worth the effort and ink because there is a subtle logical fallacy associated with single categories. Namely, making the complement into a category, which we will call *elevating* the complement. Because there is a logical necessity to have two categories when making comparisons, given a single category there is the tendency to elevate the complement to the status of a being a separate category, and consequently to assign importance to differences between it and the actual category. This is particularly the case when categories are poorly defined.

Returning to the flatware example used above, this type of error can be illustrated simply. If the category is *fork*, and the given definition is 'has tines,' then having classified some items as either *fork* or *NOT fork*, there is little value in statistics depicting the ratio of forks to non-forks. This is for two reasons. The first and most obvious is in relation to probabilities.

Given that the idea of *tines* is included in the definition of the category, the expectation should be that those elements in the *fork* category will have tines, and that those in the complement will not. Thus to present a statistic predicting this outcome shows no great discovery. It simply follows the definition. The second and less obvious reason relates to percentages. With a single category (which is unbounded), there is no way to know that the items currently classified will accurately represent the distribution of *fork* and *NOT fork* in the universe of all things. It could be that classification began in the fork slot of the silverware tray in the top, left kitchen drawer, and that the current statistic is that ten out of ten items (100 percent) have been classified as *fork*. Obviously, this should not be generalized, lest we all be considered forks. Conversely, the classification may have begun outside the kitchen and 100 percent of the classified items are *NOT fork*. Because the domain is unbounded, the single-category sample can never be stratified, and consequently cannot accurately represent the domain. To present a statistic describing this outcome is of little use because of its unreliability.[4]

The real error in the above is that in providing a statistic one makes the complement independent from the category. It becomes elevated to the status of being its own entity (category) and is then compared to the true category. While it is not expected that one would make this type of error in a case like the one above in which the category has a relatively clear definition, problems do arise as categories become more poorly defined. For example, if the definition of *fork* were given as 'looks like a fork,' then errors are more likely. Clearly, 'looks like a fork' is a poor definition. However, it is not an uncommon example of how categories are defined, both in informal and formal settings. In this case, we are much more tempted to produce a statistic describing or predicting the ratio of elements with tines in the *fork* and *NOT fork* pseudo categories. The reason is that the idea of *tines* is not

[4]To make the statement that a percentage statistic reliably describes a trend, one would have to limit the statistic to the domain of previously classified items. However, in this case, there are two domains (and thus two categories): the unbounded domain of all things, and the bounded domain of classified items. Thus, the statistic does not involve a single category but at least two: things *classified* (with the complement *unclassified*), and *fork* (with the complement *NOT fork*).

explicit in the definition of *fork*, so it is not clear if one has a category and complement, or two categories. However, because one of the primary characteristics of being a fork is having tines, then certainly the idea *tines* is included in determining if an item 'looks like a fork' and should be placed in the *fork* category. If this is the case, then we have the same situation as above, although less obvious. The category is being compared with the complement, which is not at all useful. In reality, however, the result of a poor definition is often that one simply does not know what the situation is. Just as with the fork example, poor definitions allow great variation in classification methods. Was the existence of *tines* included in the classification process? or was the classification based solely on non-*tine*[5] ideas? This cannot be determined from the definition. The result is that any subsequent use of the category or application of data produced by its analysis is suspect. We will see an example of this below in Section 2.4, Linguistic Category.

Because there has been so much emphasis on the shortcomings of poor category definitions, I would be remiss if before leaving this portion of the discussion I did not provide some indication of what constitutes a good or strong definition. I am getting a bit ahead of the discussion, however, given that determining the strength of a definition actually requires duel-category statistics, which is to be discussed in Section 2.2.2 below. The concern, when referring to definition strength, is with the internal relationship between the category concept and the tests associated with it (which form the definition itself). That is, strength is measured by how well the tests match with the concept. This relationship cannot be discovered by examining the category and complement alone. Instead, it is discovered by examining the difference between a theoretical category and complement and an observed category-complement pair, where *observed* implies being derived from an actual (real) classification process. Given the same domain (a large set of elements), the expectation for a strong definition is that both the theoretical and observed categories (or both complements) contain the same elements. In other words, the definition matches the concept well enough

[5]Of course, all of this depends on having a clear definition for *tine*, which itself is a category.

that no misclassification occurs (i.e. no errors). Although in theory well defined categories permit classification with no error, error rates of 0.0 are rarely the case in application. For this reason it is generally acceptable to have a small amount of error. This will be discussed in Section 2.2.2 below. It should also be noted that in this case misclassification refers only to definition, not process. The assumption is that any error is the result of a poor definition, not a mechanical problem. There is the very real possibility of having a very strong definition, but not having the ability to measure accurately. However, mechanics is beyond the scope of this discussion. Regardless of the disclaimers, we will say that the strongest category definitions allow classification with the least error.

Multiple Category Statistics

Having dispensed with single categories, we can now turn to the discussion of multiple-category statistics, which conceptually is much more approachable. The reason for this is that statistics always imply a relationship between at least two categories. Even a very simple measure such as 'ten percent of the forks malfunctioned,' which clearly describes a relationship between the given category *malfunctioning* and its complement *NOT malfunctioning*, only becomes meaningful in relationship to other categories. At a very minimum, the question of which forks must be answered. In other words, the larger domain must be bounded before we can begin to interpret the smaller. We need to know that it was ten percent of the forks from set A, whatever that entails, that malfunctioned. Yet even at this minimal level, relationships have been established between multiple categories. Already the categories *all*, *fork*, A, and *malfunctioning* are involved. However, for the statistic to become useful, it must be further compared[6] to the norm of fork malfunctioning. That is, if ten percent of forks malfunction in category A (the given fork set), is this the same rate found in category B (the normal fork set)? More concisely, how does A compare with B (or NOT compare)? It is this particular relationship that ultimately allows the statistic to be interpreted in a useful

[6]And/or contrasted. Following the above logic, differences will be noted using NOT, as in *NOT similar*, so additional terminology to denote such is not necessary.

manner. This of course is the desired statistic which prompted the formation of the Tobacco Documents Corpus. The goal was to establish a reference such that subsequent measurements could be made meaningful through comparison (which will become more evident in Chapter 5).

Of course, in order to understand the relationship between category A and category B, we must first understand not only the category hierarchy of A, as described above, but also that of category B, which is equally complex. Thus, the understanding of any single relationship between categories is influenced by one's understanding of the complex network of secondary relationships associated with each of the primary categories. However, we can simplify this discussion by focusing only on dual-category statistics. The reason for this is that the root question is always how one category relates to another. If this is known, then the results can be combined with others to form multi-category (multi-dimensional) studies. In other words, the logic used in duel-category statistics can be applied at any level, but it is more easily discussed at the lowest.

When given two categories and a statistic describing some relationship between them, what one has is a depiction of how the category elements coincide, i.e. what part of one set is also part of the other. As simple as this sounds, this is the extent of what can be done with categories themselves. In subsequent chapters, additional statistical measures will be discussed. However, these are secondary procedures which compare and contrast the type of descriptive statistic being discussed here.

There may also be the tendency to attempt a comparison in reverse by beginning with so called 'undefined' sets of elements. However, it must be remembered that according to the above definition of category, apart from a definable difference (a definition) there is no category. Sets proceed from definitions, meaning the original must already exist, even if the task is to discover additional tests which group the same set of elements together. This being the case, a category (with definition) must also exist, even if it is as simple as *items in my pocket*. Thus, the comparison still proceeds from category to elements.

Here we will stay strictly with the comparison of categories, which again is the root question in the TDC study to be presented. What we always want to know is whether a given category definition A groups together any of the elements from a given category definition B. Do definitions A and B describe any of the same elements? If we return to set theory as a means of illustration, the given statistic describes what elements from a given set A are also elements of the given set B. In other words, it describes how sets A and B overlap. This can be seen in Figure 2.2. The given statistic is represented by the darker portion denoting the overlap of A and B.

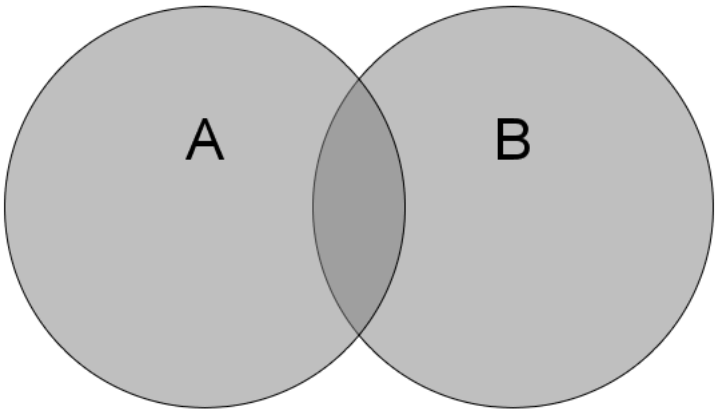


Figure 2.2: Venn Diagram 2: Overlapping Categories

In this case, *SOME* of the elements in A are also elements of B and *SOME* of the elements in B are also elements of A (the overlap), which indicates that there are similarities between the categories. Exactly how much remains undetermined in this illustration. However, this is an important discovery. We now know something about the relationship between category A and category B, and are provided a point of entry for further investigation.

Of course, Figure 2.2 does not illustrate the only option for a relationship between sets (categories) A and B. Logically there is also the possibility that ALL of the A elements are in B, and ALL of B is in A. This denotes, as illustrated in Figure 2.3, that A and B are identical or equal, meaning that category A has an alternate definition (or vice versa). This

would be a significant discovery. As mentioned above, and as we will see in Chapter 5, this is a return to the root question being asked in TDC study, 'Do definitions A and B describe the same set?'

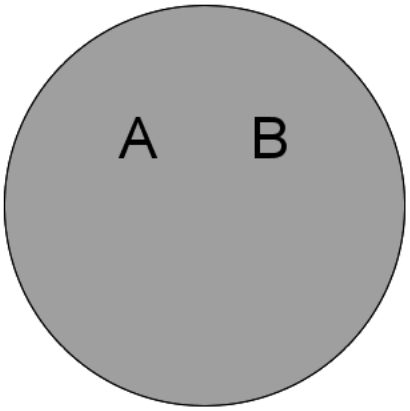


Figure 2.3: Venn Diagram 3: Equal Categories

There is also the possibility that ALL elements of B are elements of A, but only $SOME$ elements of A are elements of B (or vice versa). In other words, B is a part of the larger set A. In terms of category we could say that B is a sub-category of A. This is illustrated by Figure 2.4.

And finally, there is the case where $NONE$ of the elements in A are in B (and thus $NONE$ of B is in A). This is seen in Figure 2.5. In this case, set (or category) A has nothing in common with set B. They are dissimilar.

Each of the Figures 2.2, 2.3, 2.4 and 2.5 illustrates what can be termed a true discovery (even if Figure 2.2 is inconclusive). This is very unlike what is found with single-category statistics. If it is possible to reach these types of conclusions with confidence, then important information is being added to the knowledge base. We move from knowing that some spiders are dangerous and some are not, to being able to say that the category of spiders defined as being black with red markings on the abdomen is a sub-category of those defined as

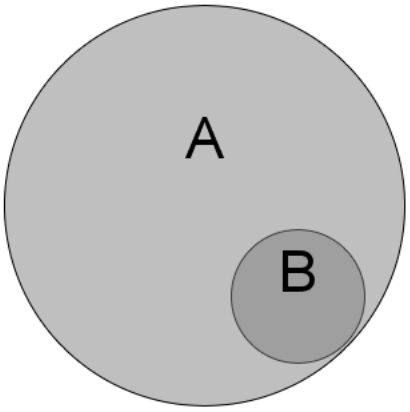


Figure 2.4: Venn Diagram 4: Sub-Category

dangerous. This is phenomenally useful information, and it can be put to good use. This of course is the goal of categorization.

Although it is desirable to make discoveries as illustrated above, in which we can use boolean terms such as *ALL* and *NONE*, the case is much more often that logically speaking *SOME* is the correct descriptor. This is simply because by definition *SOME* occupies the entirety of the range between *ALL* and *NONE*. However, because there is also great value in discoveries such as almost all black spiders with red on their abdomens are dangerous (very close to being a true sub-category), or that is is very rare to have a hurricane in April (very close to being dissimilar categories), categories are often considered statistically equal or sub-categories or dissimilar, even when there is some opportunity for error. In most fields of study, if there is confidence that the error rate will be less than five percent, i.e. the probability of misclassification is less than 0.05 ($\frac{1}{20}$), this is sufficient to consider a discovery to be significant (not random). In some specific cases a much smaller margin, such as $\frac{1}{100}$ or $\frac{1}{1000}$, is required. Whatever the case, it is rare that no error is permitted. Even our medicine is labeled to tell us that although it is considered safe (a sub-category), in rare instances it

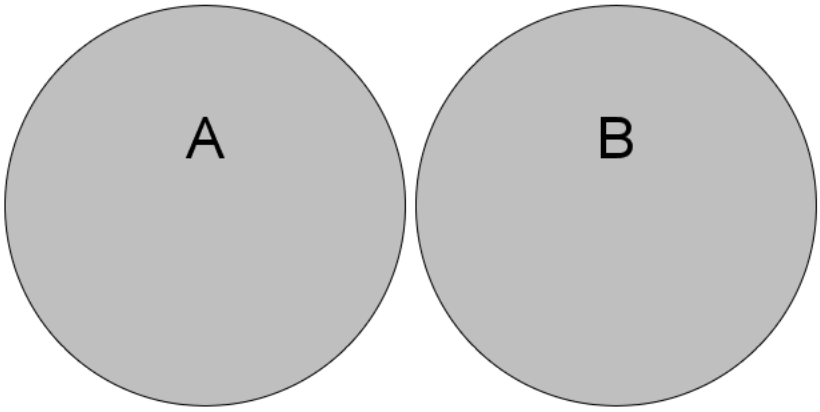


Figure 2.5: Venn Diagram 5: Dissimilar Categories

might cause various maladies and even death. Thus the norm is that some error is permissible, yet if the error rate exceeds an acceptable amount, which at a maximum is five percent, the discovery is then suspect. As we move into a discussion of TDC analysis in Chapter 5, this type of allowance will become critical. Figures 2.6, 2.7 and 2.8 are modifications of the above Venn diagrams to reflect 'statistical' discoveries. In each, the boundaries of the sets have been shifted slightly such that a very slight portion is outside its expected location.

2.2.3 Summary

We have touched on a number of ideas in the first part of this chapter. Before moving on to Sections 2.3 and 2.4, and adding context and example to the discussion, let us briefly revisit the main ideas presented in this section. I have a series of five in mind, each building on the previous.

First, category has been defined as concept, meaning that any given category represents an idea rather than a reality. Consequently, categories are not fixed but arbitrary. They are fluid and adaptable to the needs of those who create them. Second, every category

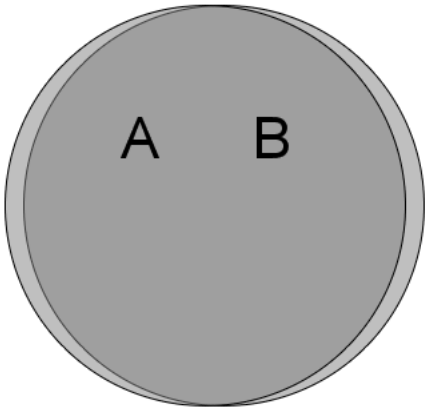


Figure 2.6: Venn Diagram 6: Statistically Equal Categories

carries with it a definition or set of tests, formal or not, that permit classification. Without the tests, the concept has little practical use. Third, for any given category, classification is a binary process. The decision is whether an item belongs within the bounds of the category, or without as part of the category complement. Fourth, rather than considering single categories, the root question is how one category compares with another. Comparison is what allows true discovery and adds to knowledge base. And finally, discoveries may be significant even with some error. Although it must be slight if we are to have any confidence in our discoveries, error is expected in applied theory. We shall return to these ideas throughout the following sections and chapters.

2.3 Empirical Data and Category

From a purely theoretical standpoint, it would be much nicer at this point to make a clean transition into a discussion of linguistic category. However, the primary focus of the chapters to follow is on data, be they the actual texts or the quantitative data which they yield. The

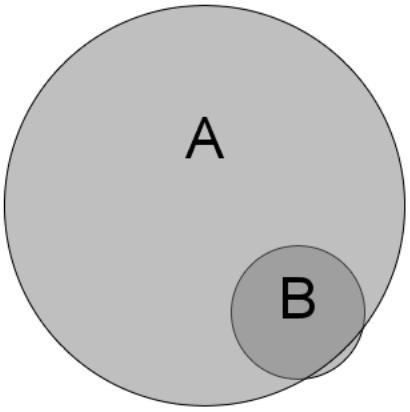


Figure 2.7: Venn Diagram 7: Statistical Sub-Category

task given is to distill roughly 30 million pages of real, natural language into a coherent and tangible format that allows useful discoveries. Thus, the Tobacco Documents Project is very much in the realm of Applied Linguistics, and we cannot continue realistically without first addressing the increase in complexity introduced by empirical data. This complexity is not an insurmountable problem, but it is a bump in the road which can jostles our thoughts and knock us off course if not approached carefully.

I have stated above that I believe the field of Linguistics should indeed be in the realm of Science, and if this is the case, then our theories must be based on the principle of replication. That is, we are not allowed simply to rest on ideas and anecdotal evidence. Instead, once a hypothesis is formulated, it must be tested and retested until it fails, or until it proves itself reliable for replication. In other words, sound theories allow one either to replicate outcomes by controlling the variables, or to predict outcomes by analyzing the variables. It is at the point of replication that our work becomes reliable and useful, which is why some form of the Scientific Method is taught at such an early age. However, the testing which is required will always involve data gathered from observation, and as the emphasis begins to move from

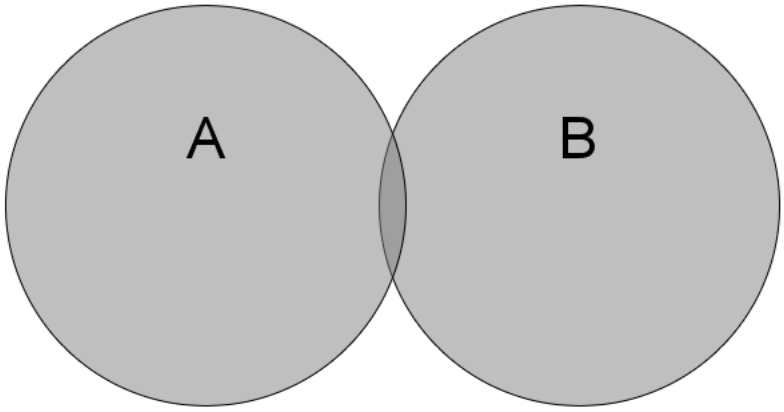


Figure 2.8: Venn Diagram 8: Statistically Dissimilar Categories

theory toward testing, it has a tendency to gather momentum and pendulum past the center point on towards an undo focus on the data themselves. If this happens, the end result is that the overemphasis on data causes problems in interpreting the meaning of categories.

The reason these problems develop is that with the introduction of empirical data we find ourselves on the edge of a slippery slope. The goal of testing is to gather sufficient data to measure the reliability of a given theory. Measurement, of course, always involves quantitative data in some form, even if we do nothing more than count qualitative responses. Quantitative data, however, because they involve numerical data, involve continuum. And continuum, as its name implies, is the antithesis of category. In other words, category denotes discrete divisions, while continuum requires that there be none. Any division of a continuum, any separation into parts, is in error from a mathematical or semantic perspective. Thus the end result of narrowly focusing on measurement and data is that the idea of category is weakened, and the entire process becomes confused.

As an example, let us return to the idea of color terms introduced in Section 2.2.1. Most would agree with the qualitative observations that *yellow* is the color of a ripe lemon, and

that *orange* is the color of a ripe orange. This gives us two categories: *yellow*, defined as *a color like that of a ripe lemon*, and *orange*, defined as *a color like that of a ripe orange*. And in fact, these two categories, which are subcategories of the category *color*, are employed regularly and with good results. However, when color is subjected to scientific scrutiny, one finds that *like a ripe lemon* is actually a poor definition of the physical stimulus interpreted by the eye and brain as *yellow*. Instead, what one finds is that the physical stimulus is actually a form of electromagnetic wave (or radiation) that falls on a continuum, called a spectrum, within a certain range of wavelengths which can be detected by the human eye. If these wavelength are measured quantitatively using a spectrograph rather than qualitatively using the eye and brain, they fall roughly in a range from 400 up to 700 nanometers (nm or 10^{-7} meters) in length. For the eye and Western brain, this would be from red to violet. On the spectrum of electromagnetic waves this is much shorter than radio waves, but much longer than gamma rays, just between infrared and ultraviolet. If we keep our categories of *yellow* and *orange* we find that there is now another way that they can be defined. Namely, *yellow* can be defined as the perception of waves which are nearest in length to 570nm, and *orange* can be defined as the perception of those nearest to 590nm in length (http://eosweb.larc.nasa.gov/EDDOCS/Wavelengths_for_Colors.html). This is certainly useful information.

If, however, our focus goes too far in the direction of data, we will reach the point of saying that there are no categories *yellow* and *orange* because they cannot be justified by the data. The reason for this, again, is that the measured lengths of electromagnetic waves form a continuum, thus no lines can be drawn separating *yellow* or *orange* from their respective complements (and in this case, the other category). If we did draw a line, the logical place to put it would be half way between *yellow* and *orange*, or at 580nm. If given a wavelength between 570nm and 590nm, we could then say that if it is shorter than or equal to 580nm it is *yellow* (and part of the complement to *orange*), or if it is longer, it is *orange* (and part of the complement to *yellow*). However, this is not perceptually or mathematically justified.

What we imply by such a division is that there is potentially a categorical difference caused by the least perceptible (or measurable) difference. That is, if it is possible to measure waves in nanometers to n decimal places, then a difference in wavelength of 10^{-n} nanometers could cause one wave to be classified differently from another if that difference occurred across the boundary. In the above case, the difference between 580.000nm and 580.001nm would be categorical assuming we can measure to the third decimal. These divisions become even less justified if we account for error rates.

Examples similar to color categories are numerous. If we focus on a measure of bacteria growth, can we justify saying a carton of milk is out-of-date at midnight, but not the second before? By measuring physical and mental development, can we verify that a child is instantaneously transformed into a categorically different being (a teenager) 410,248,800 seconds after birth (accounting for leap years), but is not able to drink alcohol responsibly until another 252,460,800 seconds have passed (less the number of seconds after midnight he/she was born)? Does one's measured education level suddenly change upon conferral of a degree? Is there a border between rough and smooth, or between hot and cold? The empirical/mathematical answer to all of these questions is, of course, no. None of these categories, which are so often employed in daily life, are justified by the data. Yet they are useful.

2.3.1 Apparent Paradox

What we seem to have is a paradox of sorts, a confusing incongruence between category and continuum. On one hand we have a desire for and incredible ability to create and use category, but on the other we find that empirical data rarely if ever justify the existence of our categories. This is similar to Michael Covington's basic description of language in the first chapter of *Natural Language Processing for Prolog Programmers*. Here he notes that humans essentially have 'digital' minds desiring to identify discrete parts, but live in an 'analog' world where data come to us in ambiguous forms. We are tasked with taking the ambiguous (continuous) and making it discrete (3). Our ultimate goal is an understanding of

how the individual (i.e. discrete) pieces fit together to form a whole, regardless of whether or not the individual pieces can truly be defined, and this is done through the construction of category-based hierarchies. To remain at the point of an abstract mathematical continuum prevents us from constructing these hierarchies, establishing relationships, and ultimately functioning correctly. We remain in a state of flux.

Where the problem arises is with encountering measurement data. The reason for this is that the data are exactly the opposite of what we desire, an abstract mathematical continuum rather than discrete pieces. If we become too focused on the data themselves, we compel ourselves (because it is our habit) to draw lines on the continuum. This, of course, is an error, but to draw lines where none exist is a great temptation. This is so much the case that in Moore and McCabe's *Introduction to the Practice of Statistics* the authors are compelled to give a stern and lengthy warning against using statistical measures as cutoff points (472). Their logic is focused more on error rates than on the implications of continuum, but it reaches the same conclusions. Because all measures come with a degree of reliability, and this is never 100 percent, any place a line is drawn on a continuum divides a population incorrectly. Theoretically, this is easy to see, yet the authors are required to give the warning because what we want to do with statistics, and what we regularly do even if we know better, is classify. We want to know if deviation from the norm is significant or NOT significant. A definite $MAYBE$, even if it is a mathematical reality, does not satisfy. Eventually, however, one comes to the realization that based on data no lines can be drawn. Essentially, what we have in hand does not match with what we have in our head, and unfortunately our reaction is often to throw out categories and attempt to work without them.

2.3.2 Resolution

The reality of the situation, however, is that a paradox does not exist. Rather, the problem is one of focus and a subsequent unjustified assumption. What has to be done is return to the definition of category presented in Section 2.2.1 and be reminded that categories are

concept rather than reality. They serve specific purposes in limited contexts by allowing complex data to be organized so that they can be effectively put to further use. In other words, categories are strictly qualitative in that we are interested in a small number of distinguishing characteristics or properties of the category elements that unify them into a single group. That is, common qualities. It is the common quality or a set of qualities that produces the category definition. However, categories say nothing about the true nature (the quiddity) of the items classified. To answer questions of quiddity, what a single item truly is, all qualities of the item must be examined. And if all qualities are considered, no groups can exist because *ALL* would include fundamental qualities such as specific time and place of creation, and no two items can match in both. In other words, quiddity requires *ALL*, *ALL* leads to continuum, and continuum denies category. In practice, categories are always a practical compromise in terms of quiddity because they are based on only a limited number of qualities which the category author determines to be the most salient. Thus categories reflect more about the perceptions of the author than the true nature of the category elements themselves. What this means is that a category's existence cannot be justified by its elements as the examination of elements is necessarily incomplete. They can only be justified by their definitions. Any category that is well defined is justified by its definition (as it reflects the needs of the author). It has no requirement for any additional justification. In terms of validation, it should be evaluated by how well it fits into larger paradigms, and by how universally functional it it. Although a large part of evaluating relationships can be done quantitatively, it is not the element data that do the initial defining.

What must be avoided is falling down the slippery slope and getting to the point of throwing the baby out with the bath water. When one goes to the additional trouble of collecting measurement data, the tendency is toward a myopic fixation on the data themselves. Yet because we ultimately desire strong category, rather than using the data as a quantitative means of comparing existing (qualitative) categories and making discoveries (refer to Section 2.2.2), fixation on data causes us to make a jump to drawing lines on the continuum

in an attempt to produce categories that are justified by the data. Yet because of the nature of continuum, this never exists, and the end result is further deconstruction of the categories.

The root error we encounter in this deconstruction process is equating a category with its elements. That is, by focusing on the data we shift our point of view from the idea that a category defines which elements it contains, to a category being defined by the elements it contains. The former is practical. However, the latter is a hopeless proposition in that the category would remain in a constant state of flux because there is no fixed definition (i.e. no boundaries and no tests for inclusion). In the case of color categories, it must be remembered that electromagnetic waves are not color, but simply electromagnetic waves. Thus, when we measure them, we are not collecting information about color, but about the waves themselves, which are physical entities. A color category such as *yellow* is not real, but a conceptual device that allows one to organize some of the multitude of information encountered daily.

The resolution, then, is to remember that the purpose of quantitative data are to permit comparison and evaluation of qualitative categories that already exist, but not to create them. It may very well be that the data cause us to redefine a category, as in changing the definition of *yellow* from *color like a ripe lemon* to *the perception of a wavelength near 570nm (like a ripe lemon)*. And it may be that data strengthen or weaken our belief in the usefulness of a category by allowing better understanding of what a category represents, as in *yellow* being a subjective evaluation of wavelength on a continuum. And it may even be that the data allow us to discover relationships between categories, as in learning that most Westerners define *yellow* similarly. But data do not confirm or deny category. Quantitative data are real, qualitative categories are not. No amount of counting can change the fact that I do not like yellow shirts or that Shakespeare's Countess Olivia does not like yellow stockings, cross-gartered or not. We have concepts of *yellow* (categories) which are independent of data. What data can do is provide a means for me to compare my concept (definition) of *yellow* to my concept of *orange* and make discoveries about my own perceptions. Or comparisons

can be made to Countess Olivia's concept of *yellow* to determine how similar her definition is to mine. There are many possibilities. What data cannot do, however, is delete[7] *yellow* as a category. We can use the measurement of wavelength to describe a set of perceptions, and subsequently predict how other wavelengths will be perceived and classified, but we cannot use wavelengths themselves to create color categories. Color categories are concepts used to organize an individual or group's perception of real events (photons striking the retina), but they are not the real events themselves. Wavelengths are not colors, but wavelengths are perceived as colors. Wavelengths are real, color categories, although useful, are not.

2.3.3 Ambiguity and Probabilities

Before moving on to Section 2.4 and the application of category to Linguistics, we first need to address probabilities, a specific case of mathematical continuum. The reason for this is that it is often the case that one is faced with the task of classification, but given poorly defined categories. The poor definitions could be simply from neglect, but more often they are the result of poor understanding (which is what motivates study to begin with). Whatever the case, poor definitions cause ambiguities. That is, the person doing the classification is unable to determine in which category an item belongs. As a tool, probabilities are often employed. With probabilities, one can look at similar past events and determine the most likely (probable) classification. This can be very useful.

Where we go wrong with probabilities (and other measurements) is not in using them to make decisions, because they are indescribably useful, but in accepting them as a model when in reality they are only a means for describing or predicting (not knowing) the outcome of a situation which we do not fully understand. That is, we try to use them for categorization rather than classification. The problem with this is that probabilities provide very limited understanding. Take for example the classic illustration of probability, the coin toss. By trial

[7]The question of what can remove a category from an individual is complex considering language. That is, for an individual to delete a category would require that the entire language culture also delete the category.

we know that the more times we flip a coin, the more the ratio of landing heads to landing tails approaches 1:1. However, this knowledge never allows us to determine the outcome of a single coin toss prior to the landing. On the other hand, the science of Physics tells us that if we know the environmental variables associated with a single toss, then we can determine the outcome before the landing. That is, if we know the shape and mass of the coin, its position in the hand, how much force is applied, where the force is applied, with what intensity the force is applied, how far above the landing site the event starts, the physical properties of the landing site, the density of the air, et cetera, then we can determine how the coin will land. In fact, because Physics is in the realm of Science, we know also that if we can replicate the environmental variables, we will always get the same outcome. In other words, there is no associated chance (probability $p = 1.0$). The problem is of course that we rarely find ourselves in the position to know or control the variables associated with the toss enough to determine the outcome, thus we use probabilities as a tool to describe and predict. However, models based on probabilities are necessarily incomplete in that allowing probability is an admission of a gap in understanding. And although it is a good thing to admit gaps in understanding, it is not good to use gaps in building theoretical models, however productive they might be for describing an event.

2.4 Applied (Linguistic) Category

If it is true that humans are predisposed to divide, then it is certainly more so for the *linguist* subcategory of humans. Language as a whole is unmanageable for study, so it is approached by linguists in pieces, sometimes very small pieces, with the hope that an understanding of the whole can be reached, oversimplified though it might be. In other words, our understanding of the whole is the summation of our understanding of the parts, or however much of them can be assimilated. This of course makes the process of division critical, simply because these earliest choices form the foundation for further study, as well as our perception and understanding of larger constructs. Just as the manner in which one divides an area into

dialects determines how the language of the larger region is viewed, how sounds are grouped into phonemes determines ones phonology, parts of speech determine syntax, and document classification determines the understanding of a corpus. Yet as much as linguists depend on category as we divide our workspace into manageable parts, it is infrequent that we bother to consider category before proceeding. In the preceding section we used the whimsical example definition of *looks like a fork* to illustrate a point, but this is not as ridiculous an example as it should be. Why is this language a Creole? or why is this word an adjective? or why is this a lower-middle-class speaker rather than an upper-working-class? Unfortunately, all too often the answer is that it looks or acts or sounds like one. The result of this categorization and classification methodology (or lack thereof) is of course that subsequent work is less reliable because the foundation is unknown.

More consideration of category is needed. Fortunately, all that has been discussed to this point applies without exception in the field of linguistics. Category remains *a concept of definable difference*. The change at this point, if it can be called such, is simply that we will no longer consider the concept and tests for *fork*, *spider*, and *yellow*, but instead look at ideas from linguistics, such as *register*, *noun*, and *basilect*. However, just as category is the same for linguistics (which means we can bypass further discussion of the basics), the opportunity for problems related to categorization is also the same, and it is this potential that will be the focus of the remainder of this section. In particular, we will examine several issues related to *concept*, and then follow with one particular problem associated with *definable difference*.

2.4.1 Applied Concept

In the introductory chapter of Jurafsky and Martin's *Speech and Language Processing* (2000) the authors note the 'rise in empiricism' during the latter part of the 1980's, and in particular they mention the rise in probabilistic models for speech and language processing (14). This came about, particularly in computational linguistics, with the realization that rule-based models of language, however appealing they might be theoretically, were not able

to account for much of the measurement data encountered in natural language processing. With probabilities, it was found that great headway could be made in tackling part-of-speech tagging, parsing, and other questions which regularly deal with ambiguities. This of course makes perfect sense in that when faced with ambiguity, logically the choice goes to the most probable answer, the one that has shown itself (empirically) to occur most often in the given environment. Thus many of the problems that plagued applied linguistics could be solved mathematically once reliable measurements were made. This involved a great deal more computation, but the algorithms themselves were often much simpler (and often more accurate), than rule-based versions which required one to account for every realization.

With the great successes of probabilistic processing, however, also came the development of a theoretical camp which was against (although perhaps not overtly) hierarchies of fixed categories in linguistics (i.e. system). Rather than there being thirteen standard vowels in English, discussion turned to a boundless continuum of vowel sounds, and rather than the traditional eight parts of speech and tree-style syntax, tagging algorithms began to have hundreds of possibilities and to calculate probabilities for all of them, ignoring syntax rules as they went. In other words the idea of category as it was being used at the time by the rule-base camps, became suspect. For some, the focus shifted to data, measurement, and continuum.

To be sure, the idea of empirical study and continuum which followed the emphasis on probabilities was not a new concept for linguistics. Referring back to Section 2.3, we know that whenever empirical data are collected, continuum is found, and there has been no shortage of empirical linguistic data. If one examines the phonetic transcriptions that McDavid and Loman produced while gathering data for the *Linguistic Atlas of the Middle and South-Atlantic States*, it is clear that the idea of continuum was obvious to them (Kretzschmar et al. 1993). Their vowel transcriptions are far from being selections from the 13 major vowel sounds, but instead are cryptic combinations of raising, lowering, fronting, and backing, and each to varying degrees, such that certainly the same sound was recorded with

different notation. When Lee Pederson, one of McDavid's students and perhaps one of the 'last true masters of impressionistic transcription' (Kretzschmar 2000) was asked in a seminar on transcription how to differentiate between vowels on a continuum (between a high-front vowel that is backed and lowered and a central vowel that is raised and fronted), he shrugged his shoulders and said essentially that there may be no difference between them, nonetheless they must be transcribed with one or the other as the root, which is most often the transcriber's pronunciation (Kretzschmar 2000). This of course is a return to Covington's paradoxical description of the language experience (see Section 2.3). Speech comes to us in analog form (as a continuum of sound waves) yet we must convert it to a digital form, phonemes and words (which are categories), in order to interpret it.

Going further back in time, another example is Gilliéron's description of the transition from French in Paris to Italian in Rome. Although there is certainly a difference between the language spoken in both locals, traveling village by village between the two he could not draw a line separating French from Italian. Each village could speak with the neighboring village in a common tongue, even across political borders (Gilliéron 1902–10, as relayed by W. Kretzschmar). And even further still, we can always refer to de Saussure's tree analogy for differentiating between synchronic and diachronic study (1916). That is, that synchronic studies are indiscriminate (and perhaps overlapping) horizontal slices of the tree narrowly focused on a specific data set, but when assembled together give us an understanding of the whole of language history. In other words, the whole of the continuum is studied in parts.

What was new in relation to probabilistic models of language was the computational aspect. What began in the mid 1980's was a monumental shift in computer hardware. This in turn precipitated a corresponding change in software, and almost overnight it became possible to measure with little restraint. Prior to this, limited computational resources necessitated a rule-based approach. It was very economical to read a rule into memory, evaluate it, flush it out of memory, and move on. However, with the advances in hardware it became possible to not only make thousands, if not millions, of calculations in a short time, but to also hold them

all in memory simultaneously and evaluate them as a whole. Thus, what was unmanageable before, mass measurement of language data, became not only manageable, but practical.[8]

Unfortunately, the usefulness of probabilities was not always kept in perspective, and this lead to an irony of sorts. As much as the idea of system (i.e. fixed category) was to be avoided, the truth is that the probabilistic methods themselves, which were being used to argue against system, resulted in fixed categories. For example, during the decision making process a probabilistic tagging algorithm may calculate the probability that a given character string X is part-of-speech Y for a million possibilities Y, but in the end X is classified as being some Y, and in almost all cases only one Y, which is the tag assigned. This of course, however it comes about, is classification. And if there is classification then there must have previously been categories (the parts of speech defined by the algorithm), and if there are categories then there is some conceptual paradigm or hierarchy in which they are organized (the author's theoretical model). In this way one returns down the slippery slope back to the beginning, and back to what one was attempting to avoid. Even more problematic is the fact that taggers are used to mark electronic text, and once the text is tagged, it is rarely re-tagged. Instead, it is passed on to others for study. Thus the theoretical model, categories, and classification of the tagging algorithm's author become a fixed system for the secondary user, who can only count and study what the tagger has previously classified.

This same type of interplay between category and continuum can be seen, but in the opposite direction, by examining Weinreich, Labov and Herzog's *Empirical Foundations for a Theory of Language Change* (1968). In the second and third sections of this work, the authors are resolved to the idea that structure and system in language are necessary givens. In Section 3, they return to the 'fundamental' question of the work, 'if a language must be structured in order to function efficiently, how does it function as the structure changes?' (150). Interestingly, the fundamental question for them is not the existence of fixed system.

[8]An excellent example of this type of mass measurement is Kretzschmar and Schneider's quantitative description of the *Linguistic Atlas of the Middle and South-Atlantic States* found in *An Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by the Numbers* (1996).

That system *does* exist is decided. Instead their fundamental question questions the problems introduced by empirical data, namely the difficulty in isolating a single, fixed system of speech. Their resolution to the continuum introduced by data is the concept of 'co-existent systems.' That is, an individual has many, fixed, rule-governed systems, each of which is employed in a specific situation or environment. The logical end to this argument is that an individual has an unlimited number of co-existent systems, one for each environment that causes a change in speech patterns. Unfortunately, unlimited co-existent systems is essentially the same as no system at all. That is, the more variation one has, the less system one has. Thus, in the opposite manner as the probabilistic camp, Weinreich, Labov and Herzog come full circle and return to the continuum they are trying to avoid.

Although named differently, the conflict that arises from the juxtaposition of category and continuum (or theory and empirical data) is common in linguistics, not unique to Weinreich, Labov and Herzog. In fact, these are some of the very first ideas taught in introductory courses. As an example, Michael Covington provides a very succinct (two page) but thorough introduction to language study in the first chapter of *Natural Language Processing for Prolog Programmers*. Here in a straightforward attempt to get non-linguists to begin thinking like linguists, he introduces and illustrates some fundamental concepts about language, one of which having to do with the idea that 'Everyone speaks his or her own language ... [and] every variety of English has definite rules of grammar, but the rules vary from dialect to dialect' (1994, 3). In other words, we live in a slurry of co-existent, overlapping systems. Each is a little different (the continuum), but still rule-governed (the system). Another idea Covington introduces is the fact that what one thinks about language is often markedly different from what one actually does with language. That is, the concept differs from the substance. This was of course recognized by Saussure in his distinction between *langue* and *parole* (1916), and later by Chomsky as the distinction between *competence* and *performance* (1957). Thus the issue for Weinreich, Labov and Herzog was not that empirical data were suddenly discovered, but that as dialectologists they were not able to separate themselves

from the continuum the data produced. For the theoretical camp it is easy enough to focus on the category of *langue* and largely ignore the data of *parole*. The dialectologist is not afforded this luxury. Dialect is discovered in the *parole*.

The two examples above illustrate a common issue in the study of linguistics, which is one of perspective more than anything else. Although they give the appearance of being the antithesis of each other, both stem from the same underlying idea. The common source is the assumption of category permanence. That is, rather than viewing categories as concepts that are created for specific purposes and free to change as needs change and understanding evolves (i.e. they are arbitrary), we make the assumption that the naming of a category implies a fixed and permanent reality.

Of course, there is little reason to make such an assumption. Outside of linguistics we have seen that there are few if any categories which have remained unchallenged since the dawn of time. Even the most basic categories, such as *day* and *night*, when they start and when they end, are not agreed upon. Inside linguistics we find this same lack of permanence. Even the theoretical linguists, who have as a goal the discovery of the most permanent language categories common to humans, have undergone multiple shifts in their category paradigm between Chomsky's introduction of Transformational Grammar (1957) and Minimalism (1995). Yet despite the evidence in support of category as concept, we humans (and linguists) have a tendency to assume that categories are reality. There are a number of possible reasons for this, two of which bear mentioning. First, at the lowest level, categories are a framework for handling very real items (i.e. the category elements), and this element reality tends to carry over to the category. In other words, at the very bottom of a category hierarchy, there is always a connection to real, tangible, physical items or events that the category author has encountered and desires to understand. Because these events have real qualities, and because we use these qualities to first form our categories and subsequently to classify, we assume that the categories must be as much of a reality and have as much permanence as the qualities from which they are created. Superficially, this make a lot of

sense. However, at a deeper level one finds that the connection to the real item or event is actually the perception of the category author, which is neither a comment on element quiddity nor a guarantee of permanence.

The other reason for assuming category reality has to do with the necessity of category. If our goal is to gain an understanding of the relationship between elements, then creating groups (categories) is truly a cognitive necessity. We are faced with far too much information to work without them. Yet because categories are such a necessity, our assumption is that they are also a necessity outside the realm of cognition, since we rarely cogitate on cognition itself. If we make this step, then we will generally continue down the slippery slope to category permanence. That is, if categories are necessary, then they must be real; if they are real, then there must be a correct form; and if there is a correct form, then it must be fixed (by nature). However we arrive at these conclusions, the unfortunate reality is that linguists often approach a topic of study with the view that the categories used, either those we author or those authored previously, are assumed to be fixed in the universe of all language. If we make this error, then the result is that we drift away from associating category with quality, and toward either of the two extremes illustrated above. That is, we tend to focus either on element quiddity as a reaction against the categories, or on category quiddity as a means of support.

If we find ourselves in the 'against' camp, it is because we assume that categories are intended to be considered real entities, and we do not agree with them. In this case, the most obvious means to refute them is to focus on measurement data, which will produce a continuum and allow boundaries to be challenged. That is, one can always argue that a given category does not reflect the quiddity of the elements it contains (see Section 2.3). However, because categories are necessary for cognition, they will be used inadvertently, which means that they will at best be poorly defined (as with many probabilistic taggers). With this approach, unbiased data are gathered, but there is no explicit framework in which to organize

them. The result, because of vague/undefined category, is that our work is unreliable for explanation. No one can know how to interpret the data.

This is analogous to being given the task of moving and being supplied with a truck and some pre-labeled moving boxes. However, because the box labels do not correctly represent the nature or room-distribution of the household items to be packed, we refuse to use them. Of course, we do realize it is impractical to transport items individually (although this would be quidditatively correct), so we pack in plastic grocery bags and load them into the truck. The work goes fast, but once at the new house, no one can remember which bag out of the hundreds packed has the TV remote. There is no logical organization to the bags. If there is a logical organization, then we have simply replaced the boxes with piles of bags, and have returned to what in the beginning we insisted should be avoided.

Conversely, if we find ourselves in the 'for' camp, it is because we believe that real categories do exist. Following the logical progression of this assumption, if categories are indeed real, then there must be a correct and true set that is valid in the universe of all language (perhaps yet to be discovered). The result of this line of reasoning is that one has the tendency to keep categories longer than is practical with the hope that the true set has already been discovered. However, this inadvertently produces bias in our work because we insist that any data we encounter fit into our existing category hierarchy (which we think is correct). Yet by forcing data into the existing paradigm both become meaningless. In other words, we bypass the idea that strong categories are the evolutionary product of evaluation according to empirical data, but instead gather biased data that supports our existing categories, making unjustified classifications in order to avoid modifying the existing system. Overall, the problem is acceptance of category before sufficient justification (the belief that categories are beyond concept), and the result is that our data become unreliable for evaluating our category hierarchy. They are biased.

Following the moving illustration, this is to insist that the supplied boxes are labeled with the correct categories, and that everything must be in the correct box prior to being

loaded onto the truck. Initially the work proceeds quickly, but then it falters and comes to a stop because the helpers are constantly having disagreements about classification, such as whether the hammer which was in the hall closet should be packed in the *kitchen* box (where commonly used items are often kept), or the *garage* box (where tools are kept). Each option could potentially violate the quiddity of the respective boxes, and unfortunately, there is no *hall closet* box. In the end, we are forced to either A) ignore the hammer because it is not really terribly important (just leave it in the closet), or B) rename it so that it can fit nicely in a box with other items with similar names (a hammer is really more of a *utensil* than a tool, so it will fit nicely in the *kitchen* box with the other utensils). Of course, we have no clear justification for either decision other than our insistence that the quiddity of the boxes not be violated.

The resolution of these issues is simply to return to the definition of category and remind ourselves that the naming of a category is the defining of a concept, not the description of a reality. Essentially, category is a method for handling Covington's paradox from Section 2.3, the idea that we have digital minds in an analog world. Categories serve as intellectual containers for data. They allow us to organize and process the continuum of data from very real but individually unmanageable language events into manageable but unreal perception-based groups. Thus we should approach the study of language with no idea of category permanence. Perhaps at some point in the future we will discover universal language categories. This is a fine goal and certainly would be the greatest linguistic discovery to date, yet in the meantime we must prevent ourselves from being fixated on system, but instead permit and expect category change as we gather data.

This is essential to learning. In fact, we can only say that we have truly 'learned' when our acquisition of knowledge reaches the point that it causes our category paradigm to become unstable and forces us to rebuild, resort and rename. Those great moments when we have life-changing epiphanies that cause us to stand up and shout 'Aha!' are always precipitated by realizing that we have narrow-mindedly continued far longer than the evidence allowed

with something in the wrong category, which is why our next statement is often 'I can't believe I didn't see that before.' When Detective Thorn came to the realization that 'Soylent Green is people!' (Fleischer 1973), what he was experiencing was a monumental shift in his category paradigm. He had gathered sufficient knowledge to force re-categorization (or at least re-classification). As much as he would have preferred to hold on, he was no longer able to keep his old world view. His category paradigm was irreparably changed by the data. As linguists, however, we should avoid such epiphanies. We need solid categories, but we cannot afford to remain fixed, like tectonic plates, until we suddenly give way in an earthquake. The damage is always too great, and the recovery too long. Our task is to stay in the middle and not pendulum out to the extremes of denying the necessity of category or, in the other direction, insisting on the correctness of category. System is a goal, not a method. Empirical study is a method, not a goal. We have to accept that any category is a compromise in terms of quiddity. It is never a reality, and it can never represent the true nature of all the elements it contains. In this way we maximize the usefulness of category by proceeding with acknowledged and well-defined categories, and by not holding on to those that are unproductive. This is what makes our data maximally interpretable and unbiased. If our moving boxes come pre-labeled with room names, then we can try them out knowing that the labels were given not because they model the universal organization of household goods, but because they model the common house and have been found useful for organizing household items. If they meet our needs, so be it. If they do not, then we can take a black marker and make the appropriate changes. If we are able keep our ideas about category permanence in perspective, then we can expect a smooth and steady progression of understanding.

2.4.2 Applied Definition

In the fall of 1998 I was able to attend an NWAVE-27 presentation of a paper by Peter Patrick titled 'Testing the Creole Continuum'[9] in which several language features of Creole speakers

[9]This paper was later incorporated into *Urban Jamaican Creole* (Patrick 1999).

from Kingston, Jamaica were described. Patrick's focus was an analysis of the linguistic transition from one group to the next across the spectrum of Jamaican Creole speakers. He concluded that

> there appears to be no clear dividing line in the grammar between the mesolect and the acrolect... but the absence of such [grammatical] knowledge boldly marks off basilectal speakers. Thus the situation in Jamaica most closely resembles Bailey's [(1974)] notion of gradatum: continuous variation within a wide mesolect, but a sharp boundary on the lower end, between it and the basilect. (Patrick 1998)

This was determined by the analysis of 4 linguistic features from the speech of 15 subjects described as 'all these speakers were drawn from the middle levels of Jamaican society, and... their speech is also intermediate.' The boundary between mesolect and basilect (or using the terms above, between the category *mesolect* and its complement *NOT mesolect*) is derived from the one speaker, Dinah, who did not use regular verb inflections with any detectable consistency. This makes her speech sharply divided by from the other 14 by her grammar.

As much as I enjoyed the presentation and the paper itself, particularly in relation to the social implications of the idiolects, I must admit that I am unsure how to contend with the data provided. The reason for this is that it is unclear to me how the speaker groups (categories) were determined. My confusion is centered around the term 'middle,' which is the common feature (the 'level') of all the speakers. Although it is used regularly, it is never formally defined. I would like to think that Patrick did in fact have a definition for the category, and consequently a series of tests for inclusion; however, in examining the biographical data provided for each of the speakers, it is difficult to find a connection that allows the key speaker Dinah, a Working-Class-5 domestic servant with three years of education, to be in the same *middle-level* category as the speaker George, a Middle-Strata-2 civil servant with 16 years of education. This is made even more complex by the note that 'their speech is also intermediate' which leads one to assume that the *middle-level* category

is defined, at least partially, by linguistic features. However, this cannot truly be known from the information or data given.

The problem, of course, is that we must know how the categories are defined in order to properly interpret the data provided and evaluate the conclusions. If it is the case that the subjects were randomly selected from a fixed social stratum, and that the terms *mesolect* and *basilect* have fixed definitions, then the data illustrate some interesting phenomena about the distribution and variety of linguistic features within the stratum, and it would be worthwhile to continue the study (i.e. test to see if the trend holds true as additional subject data are added). However, if indeed linguistic features were used, even partially, to define the categories, then the opposite is true because the argument becomes circular. That is, noting linguistic differences between groups that are defined by linguistic differences has much less value. It may well be that the differences noted are the very same ones used to select the subjects, which would be elevating the complement.

If we return to the definition of category presented above in Section 2.2.1, namely that category is *a concept of definable difference*, we see that there is a strict relationship between *concept* and *definable difference*, specifically that the former depends on the latter. In other words, categories are not created until distinctions can be made. While this seems rather straightforward, consider the relationship in reverse: if a category is being used, then a distinction has been made. This view is a bit more perplexing. What this means is that if a group of speakers (a category) is created, there is some manner (a definition) for distinguishing between category and complement. The trick for the researcher to know what the distinctions are so that he/she can properly interpret the data and inform the reader. The difficulty, of course, is that the true definition of the category may not be straightforward, but instead require considerable time and effort to delineate. As well, the researcher writes from the disadvantaged position of being too familiar with the study and related data, which tends to cause one to handle definitions in a cursory manner. The result is that it is generally easy to say that subjects A and B are similar or different, but difficult to explain specifi-

cally what makes them such without a great deal of thought about what salient feature or set of features truly tips the scale. What we often do instead when faced with situations such as this is produce somewhat ad hoc definitions which sound correct according to the jargon of the field, but have little to do with the true definitions (the tests actually used for classification). Unfortunately, these hasty definitions serve more as a red herring than as explanations, leading both the researcher and the reader away from the true definition. I am convinced that Patrick had a specific set of tests he used for selecting his speakers. He must because the group exists. However, the data and description do not convince me that the tests are directly related to social class as he proposes. Consequently, I do not know how to interpret his conclusions.

Not to leave Patrick exposed by himself, I would venture to say that we all are guilty of using poorly-defined (or undefined) categories and offering them to the reader. I have tried hard to not to do so in this work, but undoubtedly some can be found, and even in some of the foundational studies from the most recognized names in Sociolinguistics the categories are not as well defined as they might be. Neither William Labov in describing his 1966 study in New York City nor Walt Wolfram in his 1969 study in Detroit, both referenced by Patrick as a model, provide the detail necessary to replicate their studies, particularly when it comes to the criteria for selecting informants.

Quite often the resolution for undefined or poorly defined categories is simply a matter of discipline. If the desire is to make data maximally useful, then we must make the effort to sort out the tedious details and insure that the concepts discussed are sufficiently defined to allow evaluation of the data (and replication of the study). This is for the benefit of both the researcher and the reader, particularly when multiple categories are involved, and by itself would solve many problems. However, linguists often find themselves, as did Patrick, in a perplexing mix of categories with overlapping definitions, such as *middle level*, *intermediate in speech*, and *mesolect*. In these cases, diligence may not be sufficient to prevent elevating the complement. A change of method is necessary.

In his 1994 article *Analytical Procedures and Three Types of Dialect* William Kretzschmar expounds on Ferdinand de Saussure's (1916) thoughts on defining dialect. What Kretzschmar argues is that for the dialectologist there are fundamentally two approaches to defining a dialect (the third type being an abstraction of the second). The first approach yields what Kretzschmar calls an *attributive* dialect. With the attributive approach, the dialectologist begins with a 'predefine[d]. . . locality or category of speakers and seek[s] to describe the dialect features of that locality or category.' In other words, one starts with a bounded group, one whose members are known, and then describes its speech, what features it does or does not contain. For example, one could start with a geo-political region, as did Pederson for the *Linguistic Atlas of the Gulf States* (1988), and then describe the range of features encountered throughout the region; or as Labov did in his 1966 New York study, begin with socio-economic divisions and then note the existence of particular features encountered in the speech within each division. The end result in both cases is that the dialect of the group, however the group is defined, is equated with a set of features found within the group's speech. That is to say, a geographically-bounded group first, then a feature set.

The second dialect type Kretzschmar puts forth is the *blind* dialect. With a blind approach, the dialectologist begins with a wide-ranging survey of features and then proceeds to divide the survey area into regions (dialects) based on the distribution of survey features. An example of this is found in Kurath's *A Word Geography of the Eastern United States* (1949) where he divides the eastern seaboard into dialect areas according to the distribution of particular features sets encountered in a survey of the entire area. Another example, although a bit removed from dialectology (but not very far), is Douglas Biber's 1988 study of document registers. Just as Kurath, Biber began with sets of features (the output of multidimensional analysis of co-occurrence) and then matched it to qualitative registers. Thus the blind dialect approach theoretically is the converse of the attributive, approaching the same problem from the opposite direction: bounded sets of features first, then geographical boundaries. Kretzschmar realizes of course that group boundaries are never created wholly

outside the influence of the prior knowledge and biases which come from an individual's experience, knowledge, cultural and theoretical history. That is, dialectologists rarely place isoglosses on a map that do not correspond to some non-linguistic feature, be it conscious or not, and even if we use a political boundary to define a group, it is only because we have been taught to recognize them. Thus Labov's concept of socio-economic boundaries is certainly based on prior observations (features), and Kurath's choice of features is likely influenced by his knowledge of immigrant settlement patterns (boundaries) in the Eastern United States (Kretzschmar 1994, 9). Kretzschmar accounts for these influences with what he terms a *derived* dialect, which is the third of the three types from the title (10).

What one finds with Kretzschmar's description is that the relationship being described as dialect, be it attributive, blind or derived, is between the *natural* and the dialect *feature*. The natural is typically an area or group with non-linguistic boundaries, while the dialect feature has linguistic boundaries. For the dialectologist, the task is to match the natural with the dialect feature, which is why all of the above approaches share this same root function. What the dialectologist wants, the theoretical goal so to speak, is to have dialects comprised of groups with identifiable natural boundaries as well as identifiable linguistic boundaries (the feature sets). If we convert to the terminology from Section 2.2.1, a dialect can be equated with the discovery of two statistically equal categories, one that is linguistically-based, and another that is a non-linguistic. Thus, dialect has both a non-linguistic and linguistic portion.

In this light, Kretzschmar's observations, although made in the context of dialectology, are far more wide reaching. What they are actually doing is defining a method for discovery which circumvents the danger of elevating the complement. That is, if our method requires that discoveries be produced by the comparison of a clearly linguistic category with a purely non-linguistic category, then there is little opportunity for confusing category and complement. As well, providing meaningful category definitions is much easier simply because the compared categories are diverse. Thus an issue which is so often a stumbling block in linguistics is easily resolved. In Patrick's case, his study could be much stronger if the selection of speakers were

shown to be completely non-linguistic, such as by education level. In doing this, he would immediately provide himself with a easy-to-define category, and at the same time insure that all linguistic differences between a given group and its complement are true discoveries.

2.4.3 Summary

From the inception of the TDC as a norm of tobacco-industry discourse (specifically for comparison), through sampling and archiving, and eventually throughout analysis, the project has been and is an examination of categories and how they relate to each other. This being the case, it is essential that the categories used throughout the process be clearly defined according to the earlier discussion. It is equally important that that the given categories not violate the principles discussed in this section. They should not carry any implication of fixedness, and it should be clear to the reader that the compliments are not being elevated.

In relation to category permanence, in a personal conversation with Douglas Biber in 2003 about his 1988 study of register types, he said that he chose seven registers because it was the best fit to the data available, not as an absolute. 'I could have used three or five and had similar results,' was his explanation. My impression of the conversation was that he regretted that the idea of seven registers became a point of debate over the last twenty years in that the debate on category detracted from his main argument, the utility of statistical analysis in corpus study. Although not anticipating similar attention, in order to avoid similar misinterpretation there has been a concerted effort to present the data in the next chapters (as well as in the TDC Toolkit and website) with no implication of category permanence. The belief of the author and Project Investigators is that the major categories used for the construction and analysis of the TDC were well chosen given the design and intent of the corpus. However, there is no belief that they are the only categories that could or should be used. In fact, the expectation is that other researchers will develop and apply new analysis methods to the TDC, requiring new categories. Even in the course of the TDC Project there was a general evolution of our own categories. This is evident in Chapter 3

given the changes in the sampling domain, the general redefining of categories between the Core and Quota Samples, and the removal of the *public-health* category from the inventory of categories used for quotas. In Chapter 5 this is illustrated in the changes related to date-based categories which originally were decades, but were redone as half-decades, and then finally as rolling 5-year spans across the full date range. Each of these changes resulted in new arrangements of the data (i.e. new categories) and provided new perspective.

In relation to elevating the complement, what will be seen in Chapters 3 and 4 is that to a large extent the categories in the TDC are derived from pre-existing categories. That is, as much as practical the Project Investigators based the TDC categories on existent classification data, either from the document itself or as metadata from previous classifications by other investigators. This being the case, the great majority of TDC categories, from the sampling domain down to marking individual text events, are wholly non-linguistic. In other words, the categories are heavily dependent on document sets produced as a result of litigation, document sources, dates, and intended audience, not linguistic structures. In this manner, all linguistic discoveries made are in relation to non-linguistic categories, which minimizes the opportunity for elevating the complement.

2.5 General Summary

At the beginning of this chapter it was noted that to a large extent the success of the TDC Project is based on the categories used. In this respect, the purpose of this chapter is to provide a means for evaluating the categories used in the TDC, and specifically those discussed in the next chapters. This should give the reader an advantage over the author in that he/she can approach the data with an *a priori* awareness of category importance, knowing that category decisions effect the outcome of every sub-task in the construction and analysis of a corpus. This is not to say that the author was unaware of category at the onset of the project, but simply that their true significance was not fully realized until afterward. The 'categories used' that allowed the successful outcome of the project were in

fact the result of the author and Investigators' insistence on clear definitions and rigorous testing. I believe that in the following chapters the reader will see that the categories are quite strong. Their definitions were useful both in theory and application, clearly delimiting concepts while having a low margin of error and disagreement during classification. As the discussion moves to the a description of the specifics of TDC categories there is certainly room for debate about what categories may have better served the purposes of the project (which means that there is investigation yet to be done); however, the categories themselves are straightforward and well defined.

Chapter 3

Corpus Construction: Sampling Procedures

3.1 Introduction

Once the Investigators of the Tobacco Documents Project realized the necessity of creating a reference corpus, a plan was immediately set forth for its construction which emphasized rigorous sampling methods. This plan was initially formalized by Dr. Kretzschmar in his report entitled 'Sampling Plan For Creation Of Corpora For The Tobacco Documents Grant' which is presented in its entirety in Appendix A. Dr. Kretzschmar proposed that the sampling for the reference corpus (referred to below as the Tobacco Documents Corpus or TDC) take place in essentially two stages: an initial limited sample from the entire set of documents in order to estimate the prevalence and range of document types; and secondarily, the final representative sample based on quotas derived from the data provided by the initial sample. In other words, first determine what document types exist in significant proportions in the entire document set, and then determine a quota-based sampling procedure that best matches the proposed goal of the Tobacco Documents Project. In Dr. Kretzschmar's words,

> We should first draw a limited sample from the entire body of TDs [tobacco documents], so that we can determine the best classification of text types and estimate their proportions within the overall body of texts. We should next create a reference corpus of about 500,000 words from those text types that we consider relevant to (i.e. subject to) rhetorical manipulation; this corpus will be the result of a sample of all relevant TDs, whether or not they are thought to contain any manipulation. (Appendix A)

Through the remainder of this section, the procedures used and decisions made while enacting 'Part 1: Limited Sample of TDs' and 'Part 2: Reference Sample/Corpus' of Dr. Kretzschmar's sampling plan will be presented in detail. The primary goal is to provide sufficient description to allow both the theoretical and practical representativeness of the Tobacco Documents Corpus to be known. However, secondarily the goal is to provide some insight onto our procedures in order to assist others in designing their own sampling procedures for similar projects. I begin with the description of the limited sample. This initial sample became known during the project as the 'core' sample because of its metaphorical nature of drilling down through a mountain of documents as if collecting a geological core in order to determine the type and extent of unseen rock. Building on the Core Sample, the discussion will next move to the reference (or representative) sample, which is referred to simply as the metaphorically-disappointing Quota Sample. The final sample to be discussed became known as the Supplemental Sample. Although not part of Dr. Kretzschmar's original sampling plan, once text-type ratios for the Quota Sample were determined, we realized that several rhetorically important text types, although proportionally represented, would be too infrequent in the Quota Sample documents to allow realistic study. To remedy this we created a small additional quota-based corpus which specifically targeted these text-types.

In relation to 'Part 3: Parallel Corpora of Manipulated Documents' of Dr. Kretzschmar's sampling plan, I will not specifically discuss the details of text selection and archiving for the corpus of manipulated documents (known as the Rhetorical Corpus) which was constructed during the course of the Tobacco Documents Project. In this section, the focus will remain on the construction and analysis of the Tobacco Documents Corpus, which is the reference corpus to which the Rhetorical Corpus was and is to be compared. Although the assembly and study of the Rhetorical Corpus was the primary focus of the Tobacco Documents Project overall, by Kretzschmar's instructions for selecting the manipulated-documents, 'All documents in these parallel corpora... [were] intentionally selected according to their contents' (from Appendix A). Thus 'sampling' as it is described in this section does not apply. For

more information about the Rhetorical Corpus document set, please refer to the work of Cati Brown (2004, 2006), who served as the primary author, archivist, and researcher for the set.

Before proceeding with the discussion of the specifics of sampling, there are three administrative items to note that come into play in the following sections. First, in October 2001 the UGA Tobacco Documents Project investigators and consultants met in Atlanta, Georgia to finalize plans for the study. Of particular interest to the topic of this work, during the course of the meeting, limited initial findings based on preliminary sampling of the Tobacco Documents were presented to the group for evaluation. Based on these findings, as well as a discussion of the procedures used for gathering the data, changes were made to the specifications (definition) for the reference corpus as well as the sampling procedures prescribed for assembly. In the following two sections, I will make reference to procedures and data from before the meeting, as well as those produced afterward which include the prescribed changes.

A second administrative item to note is the use of 'Bates' numbers. In the course of litigation, when a defendant provides documents to the court or plaintiff during disclosure, each page entered into the court record is assigned a unique, sequential identifier. These identifiers, which are generally alpha-numeric sequences, are commonly referred to as Bates Number. My understanding is that this comes from the widespread use of the Bates Letterpress Numbering Machine for stamping documents with sequential numbers. Whatever the case, each page from each document released by the tobacco industry as a result of the Master Settlement Agreement should have been assigned (stamped with) a unique Bates Number, although there are some exceptions. Following the common practice in the study and archiving of tobacco documents, we identify specific documents using the beginning Bates Number for the range of numbers assigned to the pages of a given document. Specifically, TDC electronic files are named according to the beginning document Bates Number with a file extension denoting file content type. For example, the third item from the first round of sampling for the Quota Sample was a ten-page document beginning with Bates Number

514564391 and ending with Bates Number 514564400. In the TDC archives, this document is represented by the files 514564391.pdf (which is the document image), 514564391.xml (which is the document text with markup) and perhaps 514564391.txt (which would be extracted document text).

Finally, it should be noted that in the discussions to follow 'random' refers to a sequence of numbers produced by either of two methods: 1) from the output of the function `random.random()` or similar of the Python programming language, which is fully described in Section 6.4 of the *Python Library Reference* (van Rossum 2008); or 2) from numbering a set of cards with the available digits, shuffling, and blindly selecting a card from the shuffled set, repeating the previous steps as needed (with replacements).

3.2 Sampling Domain

According to the provisions of the Master Settlement Agreement (or MSA), the tobacco industry is required to continue releasing documents to the public until at least 2010 (NAAG 1998). Currently there are roughly seven million documents that have been released through the industry websites, which is approximately double the amount available in 1998. This being the case, the industry documents as a whole would be classified as a monitor corpus (Sinclair 1991), meaning it is not fixed in size or content, but dynamic, designed to grow with the release of additional documents in order to monitor change over time. While this is certainly valuable in it own right, it is problematic for the creation of a reference corpus such as the TDC. In his discussion of monitor corpora in *An Introduction to Corpus Linguistics*, Graeme Kennedy notes that the dynamic nature of this type of corpus generally renders it unsuitable for comparative studies given the lack of fixed descriptive statistics (1998, 60–62). That is, all frequencies, thus all proportions, are temporary pending the addition of new data. Likewise, not having fixed counts prevents one from sampling, which is the task at hand, because the domain is open-ended. As was seen in Chapter 2, when the domain is unspecified there is no means to determine if any set or subset is representative or complete.

What is needed for sampling are well bounded (clearly defined) sets, and in this direction the Project Investigators established the domain for sampling as those documents found in the NAAG Snapshot and Bliley document sets. These well-known sets were chosen based on their representativeness of the larger collection of industry documents (see Section 3.3.3 below) and the fact that they were predefined (fixed) according to non-linguistic features.

At the onset of the Tobacco Documents Project, the NAAG Snapshot represented the bulk of available tobacco-industry documents, approximately 3.5 million total. These documents were produced as a result of the Master Settlement Agreement. In terms of origin, the Legacy Tobacco Document Library website (see citation for Legacy) describes the set as such:

> As part of this agreement [the MSA], the industry agreed to release documents on their own websites, and to provide a "digital snapshot" of these sites as they existed in July 1999. Initially, data was provided by the individual tobacco companies to the National Association of Attorneys General (NAAG).

Thus the size and content the NAAG Snapshot are fixed according to the provisions of the MSA, and also by the nature of the electronic product provided to the National Association of Attorneys General by the tobacco industry.

In comparison, the Bliley documents represent a much smaller set of documents with a somewhat more complex origin. It is described by the Legacy Tobacco Document Library website in the following manner:

> The Bliley collection is a special collection comprised of documents that defendants in State of Minnesota v. Philip Morris, et al, claimed were privileged or otherwise protected from disclosure. A Special Master appointed by the Court determined that the documents were not privileged or were subject to the crime fraud exception to attorney-client privilege; the Court subsequently adopted those findings and ruled that the documents be produced. While the companies disputed

> that ruling in court, Congressman Thomas Bliley, the Chairman of the House Commerce Committee, subpoenaed the documents in question and soon after receipt posted them on the Commerce Committee web site.

In a manner similar to the NAAG Snapshot, the size and content of the Bliley set is also fixed by the nature of the physical product obtained by Congressman Bliley.

Although the MSA requires the tobacco manufacturers to maintain websites for the release of their documents, early in the production the individual websites proved difficult to use. Most had limited search capabilities, there were no connections between the websites, and in one case special software was required to view documents. In order to facilitate research, several third-party websites began working to consolidate and convert documents from the NAAG Snapshot, the industry websites, and a number of smaller collections (such as the Bliley documents) to a single location and format. This allowed the user to search and review all available documents using the same platform. At the time of the initial sampling for the TDC, Fall 2001, the Tobacco Documents Online (TDO) website (http://tobaccodocuments.org) provided the most complete and accessible archive of tobacco documents. Another reliable Internet archive that became available later in the course of the project is the University of San Francisco, Legacy Tobacco Documents Library (i.e. the Legacy website, http://legacy.library.ucsf.edu). This site was considered a suitable alternative to the TDO website for sampling procedures for the NAAG Snapshot. However, throughout the project the TDO website served as the primary document resource.

The count of NAAG Snapshot documents available on the TDO website during sampling was 3,357,441 and remained fixed during that period. At the time that the initial Core Sample was made, the Bliley documents had not been fully integrated into the TDO or Legacy websites (or any other Internet archive known to the author), and consequently they were not included in early sampling and count data. In October 2001, a subset of the Bliley documents was made available on the TDO website, and at this point their counts were added to the Core Sample (see Section 3.3.3). The subset consisted of 33,003 documents, 3,138 fewer

documents than the 36,141 documents now available at TDO. This difference comes from the absence of 2,635 documents from the American Tobacco Company, 502 documents from the Council for Tobacco Research, and 1 document from Philip Morris. The 33,003 count remained fixed throughout the sampling process. The Bliley documents were not added to the Legacy website until July of 2005, which was after the UGA Tobacco Documents Project had ended. Currently the Legacy website provides access to 28,945 Bliley documents.

In the U.S. District Court, District of Columbia ruling *U.S. v. Philip Morris INC.* (USDC 2002) the Bliley set is said to contain approximately 39,000 documents, which matched the estimate of documents listed on the Congressman Bliley website (which is no longer available, see Bliley). A far as we were able to determine through discussions with the TDO site administrator, the discrepancy of roughly 3,000 documents (6,000 at the time of sampling) is not the result of any systematic exclusion of document types. Rather, the missing documents belong to smaller subsets of documents which have not yet been included on the TDO website because of time and funding constraints.

3.3 Core Sample

The generation of the Core Sample is guided by the first half of 'Part 1: Limited Sample of TDs' of Dr. Kretzschmar's sampling plan, which outlines the necessity and purpose of the initial sample.

> The extant set of TDs comprises millions of documents, ranging in length from just a few words to hundreds of pages. It will clearly not be possible to inspect every word of every document. Yet we do need to know what kinds of documents exist in the set of TDs, and more specifically, what kinds of documents relevant to the grant exist in the set of TDs. Further, we need to know the extent of those documents, both the quantity of relevant documents and how long they tend to be. We cannot create a valid sample of relevant documents without this information. We should therefore sample the body of TDs according to a fixed

> random sampling frame, a procedure that gives every document in the collection an equal chance of selection. (from Appendix A)

Essentially, the task is to generate a limited, stratified random sample of the tobacco documents, and subsequently to classify the documents it contains into a number of project-relevant categories. Here we will discuss the sampling. Classification will be presented later in the discussion of quotas for the various strata in the Quota Sample.

3.3.1 Determining Sample Size

As an initial look into the tobacco documents, I randomly selected a month (March) and a year-final digit (0), and then gathered data for the month and year for each decade in the NAAG Snapshot document set. This was done by querying the search engine on the Tobacco Documents Online (TDO) website. The data are provided in Table 3.1. Page numbers were estimated by determining the average number of pages per document for the first 100 pages of a given stratum. It was estimated that the 21,964 documents in the sample contained approximately 90,000 pages, an average of 4.15 pages per document, most having a token[1] count between 25–250 per page.

Table 3.1: Initial Sample of Snapshot Documents.

Month	Year	Document Count	Estimated Pages
March	1930	13	21
March	1940	87	388
March	1950	169	1,500
March	1960	864	2,100
March	1970	2,509	3,100
March	1980	6,605	34,000
March	1990	11,717	50,000
Total		21,964	

[1] A *token* is defined as being the smallest meaningful constituent of the analyzable text produced as the output of a *tokenizing* procedure. In simple terms, tokens can be though of as the *words*, *numbers*, and *word-number* combinations. However, the allowable structures are actually defined by the tokenizing procedure, which serves as the category definition for *token*. Please refer to Chapter 5 for a more detailed discussion of tokens and tokenizing.

Extrapolating these data to the whole of NAAG Snapshot text by multiplying by twelve to expand to full years, and then by ten to expand to decades, provided initial estimates of 2.6 million documents, containing 10.9 million pages.[2] Documents averaged just over 4 pages in length, and assuming an average of 250 tokens per page, averaged about 1000 tokens. The documents also varied widely in count and length for each decade, progressing with time at an almost exponential rate (which would not be wholly offset by industry growth). This suggests a considerable change in communicative habit as well as variability in text type.

Based on this initial data, Dr. Kretzschmar suggested a stratified, random sample of 200 to 300 documents from the whole of the NAAG Snapshot set. He proposed five strata based on decade: 1900–1959, 1960–1669, 1970–1979, 1980–1989, and 1990–1999. The first four decades of the century were grouped into the same stratum because of the relatively low number of documents they contain. For each stratum, document collection was to be limited by randomly selecting a year (0–9) and month, and then drawing a sample of one percent (1.0%) using a random or fixed, sequential procedure, adjusting the percentage as necessary to insure the resultant set was between the 200 to 300 document target. He estimated this procedure would yield approximately 220 documents from the data in Table 3.1.

3.3.2 Sampling Procedures

Following the plan of Dr. Kretzschmar, the documents from the NAAG Snapshot were initially divided into five strata by decade. These are defined using the date format allowed by the TDO website: `yyyymmdd`. The strata are as follows: 1950 (all documents from 19000101 to 19591231), 1960 (all documents from 19600101 to 19691231), 1970 (all documents from 19700101 to 19791231), 1980 (all documents from 19800101 to 19891231), and 1990 (all documents from 19900101 to 19991231).

[2]At a later date, more accurate counts provided by archiving entities were made available. The Legacy Tobacco Documents Library suggests the NAAG Snapshot consists of approximately 3.5 million documents totaling over 30 million material pages (Legacy).

During the initial work on the Core Sample several discoveries were made. First, it was learned that searching the TDO archive for the date 19000000 returned 62,494 documents that were not marked for year, month or day. Given that the entire decade of 1910 yielded only 835 documents, and based on the format of the document images themselves, it seemed improbable that these documents originated in the 1900 decade. Instead, it was determined that unknown year, month or day data were archived as 00 in the date sequence. To account for the misdated 19000000 documents, which later were found to represent nearly two percent of the NAAG Snapshot, another stratum was added to the sample: 19xx (all documents dated 19000000). It was also discovered that the above date issue affected queries within decades. For example, to search for documents from the year 1951 it was necessary to begin with the date 19510000 rather than 19510101. This is because 19510000 accounts for unknown months. This same principle must be applied to month queries as well. This caused the amendment of strata definitions (date ranges) to the following: 1950 (all documents from 19000101 to 19591231), 1960 (all documents from 19600000 to 19691231), 1970 (all documents from 19700000 to 19791231), 1980 (all documents from 19800000 to 19891231), and 1990 (all documents from 19900000 to 19991231). Finally, it was discovered that only seven documents with dates prior to 19000000 were available in the NAAG Snapshot. These were left out of all further sampling.

Continuing the plan of Dr. Kretzschmar, for each stratum a year and month were randomly selected. Searching the TDO archive for the given dates yielded the data presented in Table 3.2. The exception to this is, of course, the 19xx stratum, which has no year and month data. To account for this, the stratum as a whole was carried forward to the next step of the process.

Based on the Table 3.2 data, it was decided by the Investigators that the goal for the Core Sample would be 250 documents, the median of the suggested goal of between 200 to 300 documents. Additionally, it was decided that at least ten documents should be selected to represent each stratum, but still maintaining similar stratum-to-total document ratios. This

Table 3.2: Core Sample: Initial Yields

Strata	Total Docs	% Total	Year	Month	Yield
19xx	62,494	1.86	na	na	62,494
1950	103,574	3.09	8	1	1,193
1960	223,544	6.66	0	4	1,136
1970	660,223	19.66	9	12	5,895
1980	1,318,813	39.28	1	1	7,185
1990	988,793	29.45	6	6	1,679
Total	3,357,441	100.00			

yielded the data in Table 3.3. It can be seen in these data that in order to meet the minimum of ten documents per stratum, two documents were added to the 1900–1950 stratum and five to the 19xx stratum. The result is 258 documents for the six strata.

Table 3.3: Core Sample: Secondary Yields.

Decade	Year	Month	Docs	% of Goal	Needed	Taken
19xx	na	na	62,494	1.86	5	10
1950	8	1	1,193	3.09	8	10
1960	0	4	1,136	6.66	17	17
1970	9	12	5,895	19.66	49	49
1980	1	11	7,185	39.28	98	98
1990	6	6	1,679	29.45	74	74
Totals				100.00	251	258

The actual selection of documents from the sets determined by each of the strata year-month limiters was done using a fixed-sequence process, thus insuring the opportunity for selection was equal across the entire span of the given set. From the available documents in each stratum set, every Nth document was selected until the required number of documents was obtained, where N is the last two digits of the decade-year combination. Using the 1970 decade for demonstration purposes, the TDO archive was searched for all documents from 19791200 to 19791231. From the 5,895 returned documents, every 79th document was

selected, beginning with the first document (i.e. 1st, 80th, 159th, 238th. . . 3,793rd), until the required number of documents was reached. In cases where the end of the set of available documents was reached prior to selecting the required number of documents, the process returned to the beginning of the set and began with the 2nd document rather than the first, then sampling every Nth document. This was done for the strata 1960, 1980, and 1990. In cases where a document selected in the sampling procedure above was identical to a previously selected document in the same stratum, the next document that was not scheduled to be sampled was used, preferably Nth+1, but in the case of year-month sets that required multiple passes the Nth+(*total passes*) was the target document. In the event that the alternate document had been selected previously, the Nth+2 document (and so on) was used. For the 19xx stratum, since the year was not known, the available documents were searched as a whole by selecting every Nth document, where N was the random four-digit number 5323. Because the group was large, this four-digit number provided the opportunity for 10 documents to be selected from the entire set without overrunning, in a manner similar to selecting decades and months to ensure coverage of the entire document set.

Although the above selection process may appear complex at first, it actually was designed to complement the limitations imposed by the TDO website, which was our primary avenue into the tobacco documents. That is, had we been sampling from our own archive or database, we could have simply randomized a list of documents from each stratum, and then selected the first N documents, where N is the required number of documents for the stratum. However, this was not an option. Instead, we had to work within the confines of the query types allowed by the TDO website. In particular, we were able to specify the number of documents returned per page of the results for a given query. For example, in querying for the 1970 stratum, we were able to set the number of documents returned to 79 (the N for 1970 stratum). Thus our process allowed us to simply and efficiently integrate with the TDO web site by setting the documents-returned variable to insure that the target document was

always the first document returned per results page. This was a great aide in assembling the sample.

3.3.3 Finalizing the Sample

Following a meeting of the UGA Tobacco Documents Project investigators and consultants in October of 2001, it was formally decided that the sampling domain of the Tobacco Documents Corpus would be limited to the documents found in the NAAG Snapshot and the Bliley set which by that time we had determined comprised approximately 3.4 million documents. The two primary reasons for this were that 1) based on Dr. Hirschhorn's prior work with tobacco documents, these sets were believed to be representative of the entire body of tobacco documents, and 2) both sets were clearly defined in terms of which documents they included. It was also decided at that time that the Core Sample should be a stratified random sample of 0.01 percent (1/10000) of the above sets, 340 documents, which is the nearest percentile (rounded to a multiple of 0.1) to the minimum number of documents thought necessary to construct a representative sample (as proposed by Kretzschmar). The earlier sample of 258 documents represented 0.008 percent. Following standard formulas for determining sample size, a sample of 340 documents has a maximum expected error of $+/-5.3$ percent (at 50 percent) with 95 percent confidence (Moore and McCabe 1999, 583).

The major addition to the Core Sample at this point, apart from an increase in size, was the inclusion of the Bliley documents. By October 2001, the majority of the Bliley document set had been added to the TDO website. A total of 33,003 documents were available for search from six industry groups, as shown in Table 3.4.

Table 3.5 is an augmented version of Table 3.2 showing the initial yields. It includes data from the Bliley document set and raises the target sample size to 340 documents. These data were used to revise the data from Table 3.4 and produce the final counts for each stratum of the Core Sample, as seen in Table 3.6. Again, a minimum of ten documents was selected

Table 3.4: Core Sample: Bliley Document Counts by Industry Group.

Industry Group	Document Count
R. J. Reynolds	18,370
Philip Morris	7,114
Brown and Williamson	4,104
Lorillard	2,338
Tobacco Institute	1,076
Council for Tobacco Research	1
Total	33,003

for each stratum. This raised the final targets for the Core Sample to seven strata totaling 349 documents.

Table 3.5: Core Sample: Initial Yields (Augmented).

Strata	Total Docs	% Total	Year	Month	Yield
19xx	62,494	1.84	na	na	62,494
Bliley	33,003	0.97	na	na	33,003
1950	103,574	3.06	8	1	1,193
1960	223,544	6.59	0	4	1,136
1970	660,223	19.47	9	12	5,895
1980	1,318,813	38.90	1	1	7,185
1990	988,793	29.17	6	6	1,679
Total	3,390,444	100.00			

Using the target counts from Table 3.6, the Core Sample was reconstructed following the procedures described above for the initial sample. This being the case, for the known-date strata (not 19xx) the final Core Sample included the documents from the initial sample. As well, multiple sampling passes were required for the 1960, 1980, and 1990 strata. The only exception to the above procedures was in determining the variable N used for selecting documents from the Bliley and 19xx strata. In the final Core Sample, this was done by randomly selecting N from the range of integers from 0 to X, where X is *the total available documents* divided by *the number of documents needed*. While still allowing the opportunity

Table 3.6: Core Sample: Final Yields.

Decade	Year	Month	Docs	% of Goal	Needed	Taken
19xx	na	na	62494	1.84	6	10
Bliley	na	na	33003	0.97	3	10
1950	8	1	1193	3.06	10	10
1960	0	4	1136	6.59	22	22
1970	9	12	5895	19.47	66	66
1980	1	11	7185	38.90	132	132
1990	6	6	1679	29.17	99	99
Totals				100.00	338	349

for any document in the given set to be selected, this change insured that sampling would be complete in one pass through the set (a definite improvement in procedure). For the Bliley set, N was the integer 1536, and for the 19xx stratum, N was the integer 2580.

3.3.4 Results

After the completion of the sampling, the images for each document were downloaded from the TDO website and archived. In most cases, only TIFF (Tagged Image File Format) images of the individual pages were available. In cases where PDF (Portable Document Format) archives were available, they were used instead of the individual TIFF images because they packaged all document pages into a single file. The downloaded files were named based on the Bates Number of the first page of the document, replacing spaces with an underscore character and adding the appropriate file extension (*.pdf* or *.tif*). Once archived electronically, the images were printed and bound for ease of review. The 349 documents totaled 1,818 pages, with an estimated token count range from 45,000 to 450,000 tokens given the 25 to 250 per page token count from the initial review of documents. No actual count of tokens was made. For more specific file information, including Bates-Number ranges and page counts for each document in the Core Sample, refer to Appendix B, *Document Metadata*, Section B.2.

3.4 Quota Sample

The generation of the Quota Sample is guided by 'Part 2: Reference Sample/Corpus' of Dr. Kretzschmar's sampling plan, which outlines the necessity and purpose of the sample. According to Dr. Kretzschmar,

> The purpose of the Reference Sample/Corpus is to create a control set of TDs from among those in which manipulation could have occurred (but did not necessarily occur), from which we can estimate the general frequency of occurrence of linguistic characteristics of interest in the analysis of rhetorical manipulation. Because many of these characteristics may occur with low frequency, this corpus must be large enough to ensure that the characteristics are represented. On the other hand, the corpus must not be so large that its creation overruns the resources in the grant to create it: a corpus of about 500,000 words appears to be as much as the resources of the grant might handle. (Appendix A)

Given the above description, accomplishing this task required a marked departure from the procedures used for the Core Sample. Because the Quota Sample was to be representative of documents 'in which manipulation could have occurred' (and be detected) rather than all documents in the TDs, the procedures were necessarily much less random in that random selection could occur only in the context of narrowly defined document-type strata determined to be useful to the overall goals of the study (i.e. the sampling domain was modified). Defining these strata necessitated adding several steps to the process prior to the actual sampling. In particular, additional care was needed for estimating the number of documents required to reach the target corpus size, and a thorough analysis of the Core Sample documents was required in order to establish the proper quotas for each of the strata. In this section, the procedures for these additional, preliminary steps will be discussed first, followed by a description of the sampling itself, and then a brief description of a 'check-and-balance' procedure used to increase confidence in the procedures. It should also be noted that unless

specified otherwise, all requirements and procedures discussed below were prescribed after the October 2001 meeting.

3.4.1 Estimating Document Counts

The initial estimation of the number of documents required for a 500,000-token corpus was 500 based on an estimated average of 1,000 tokens per document (from Appendix A). However, cursory review of the Core Sample documents brought this estimate into question, and it was determined that a better estimate was needed prior to continuing the work in order to insure the feasibility of constructing the reference corpus. Because we discovered early on that the poor quality of document images precluded reliable reading of documents using Optical Character Recognition (OCR) methods (see Section 4.3 below), we decided to make a quick estimate the overall document token rate by examining a random stratum from the Core Sample. Using the 49 documents from the 1970 stratum of the Core Sample, which was seen as the median of the strata, the approximate token count for each document was found by estimating the proportion of page space occupied by text; determining the line spacing of the text; calculating the number of 'full' pages of text ($page\ count * proportion$); approximating an average number of tokens per line and lines per page; calculating the number of tokens per document ($full\ pages * tokens\ per\ line * lines\ per\ page$); and finally, reducing token counts to a maximum value of 2,000 (according to the sampling plan). Once this was accomplished, an average for the entire stratum (all 49 documents) was calculated. The result was an estimate of 420 tokens per document for the 1970 stratum of the Core Sample. Extrapolating these data to the goal of 500,000 tokens, produced an estimate of 1,190 documents. Taking into consideration that for the Quota Sample only documents with 50 or more tokens of analyzable text were to be collected (see below), and that collection procedures would permit some long documents to contain more than 2,000 tokens depending on where paragraph breaks occur (see procedure below), 1,190 was seen as a maximum value. That is, each of these factors works to raise the average per-document token count, which

in turn lowers the number of documents required. This being the case, it was determined that the collection, archiving, and conversion of this number of documents, as a worst-case scenario, would still be possible within the limits of the project budget. As will be discussed below, the final analysis of collected TDC documents places the average per-document token count at approximately 660.

3.4.2 Determining File-Type Ratios

For a preliminary investigation of document type, the 1970 stratum was again chosen, and those 49 documents from the this stratum of the initial Core Sample (prior to the October 2001 meeting) were classified by the author into categories according to five areas of interest provided by the Project Investigators (which parallel those outlined in Dr. Kretzschmar's sampling plan). These areas of interest, and the subsequent categories, were defined as follows (noting that the terminology to follow does not conform to that of Chapter 2, which was codified at a later date):

- Public Health: Significant for Public Health (PH), or *NOT* significant for Public Health (NPH), exclusive of each other. Documents were classified as PH if they specifically mentioned health or health-related issues, or if they referred to pathogens other than tobacco itself. Otherwise documents were classified as NPH. Granted, as a linguist I was not qualified to make this judgment, but I did so for the initial estimations.

- Audience: Industry-Internal Audience (IN), or Industry-External Audience (EX), exclusive of each other. Documents were classified as IN if they were addressed to persons or groups within or hired by the company from which the document originated, or if they were correspondence between tobacco companies. Otherwise documents were classified as EX.

- Addressee: Personal Addressee (PA), or Multiple Addressee (MA), exclusive of each other. Documents were classified as PA if the count of addressees plus the count of

carbon-copy recipients totaled 5 or less persons and did not include named groups or departments (as per Dr. Kretzschmar's proposal). Otherwise documents were classified as MA.

- Style: Letter/Memo (LM), Report (RT), or Public Release (PR), exclusive of each other. Documents intended for public release, such as advertisements or press releases (not embedded in reports), but not in letter format, were classified as PR. Documents focused primarily on reporting information, particularly the results of research, tabular and numerical data, or description, regardless of document heading, but not intended for public release, were classified as RT. Documents delimiting policy, procedure, and other administrative tasks, or those in letter form (formal or personal), but not RP, were classified as LM.

- Language: English (EN). Documents written primarily in English (more than 50 percent) were classified as EN.

These classifications resulted in the counts of occurrences for the individual categories shown in Table 3.7. The combination of the above categories (less EN, which was assumed to be a requirement for inclusion in the TDC) yielded the possibility of 24 document types (that is $2 * 2 * 2 * 3 = 24$). The total counts of these document types was then extrapolated to the full initial Core Sample to approximate the quotas needed to model the document set. This information, which is represented in Table 3.8, was presented to the Investigators at the October 2001 meeting.

Table 3.7: Core Sample: Initial Classification of 1970 Stratum Documents.

Category	PH	NPH	IN	EX	PA	MA	LM	RT	PR	EN
Count	19	30	42	7	13	36	18	30	1	48
Stratum%	39	61	86	14	27	73	37	61	2	98

Table 3.8: Core Sample: Binary Table of Document Types.

Type	PH	NPH	IN	EX	PA	MA	LM	RT	PR	Tokens	% 49	Quota
1	0	1	0	1	0	1	0	0	1	0	0	0
2	0	1	0	1	0	1	0	1	0	1	2.04	5
3	0	1	0	1	0	1	1	0	0	0	0	0
4	0	1	0	1	1	0	0	0	1	0	0	0
5	0	1	0	1	1	0	0	1	0	0	0	0
6	0	1	0	1	1	0	1	0	0	2	4.08	10
7	0	1	1	0	0	1	0	0	1	1	2.04	5
8	0	1	1	0	0	1	0	1	0	16	32.65	76
9	0	1	1	0	0	1	1	0	0	5	10.2	24
10	0	1	1	0	1	0	0	0	1	0	0	0
11	0	1	1	0	1	0	0	1	0	1	2.04	5
12	0	1	1	0	1	0	1	0	0	4	8.16	19
13	1	0	0	1	0	1	0	0	1	0	0	0
14	1	0	0	1	0	1	0	1	0	4	8.16	19
15	1	0	0	1	0	1	1	0	0	0	0	0
16	1	0	0	1	1	0	0	0	1	0	0	0
17	1	0	0	1	1	0	0	1	0	0	0	0
18	1	0	0	1	1	0	1	0	0	0	0	0
19	1	0	1	0	0	1	0	0	1	0	0	0
20	1	0	1	0	0	1	0	1	0	5	10.2	24
21	1	0	1	0	0	1	1	0	0	4	8.16	19
22	1	0	1	0	1	0	0	0	1	0	0	0
23	1	0	1	0	1	0	0	1	0	3	6.12	14
24	1	0	1	0	1	0	1	0	0	3	6.12	14
									Totals	49	99.97	234

Following the October 2001 meeting, a number of changes/decisions were made in relation to establishing quotas for the representative sample. These can be grouped into the following six areas:

1. Limits: As mentioned in Section 3.3 above, the sampling domain of the Tobacco Documents Corpus was formally limited to the documents found in the NAAG Snapshot and the Bliley set which by that time we had determined comprised approximately

3.4 million documents. The two primary reasons for this were that 1) based on Dr. Hirschhorn's prior work with tobacco documents, these sets were believed to be representative of the entire body of tobacco documents, and 2) both sets were clearly defined in terms of which documents they included.

2. Core Sample Size: Also mentioned in Section 3.3 above, it was decided that the Core Sample should be a stratified random sample of 0.01 percent (1/10000) of the above sets, 340 documents, which is the nearest percentile (rounded to a multiple of 0.1) to the minimum number of documents thought necessary to construct a representative sample (as proposed by Kretzschmar).

3. Primary Categories: It was decided that the Core Sample documents would be classified according to the following four primary categories which reflect the major interests of the Tobacco Documents Project Investigators: *internal source*, *internal audience*, *named addressee*, and *Public Health* (all with corresponding complements). These categories are defined below.

4. Secondary Categories: In addition to the four primary categories, it was decided that the Core Sample documents would be classified according to the following six secondary, feature-based categories: *form*, the document is a form; *image*, the document is an image; *English*, the document is written in English; *marginalia*, the document contains marginalia; *editing*, the document contains editing notes; and *short*, the document is short, not having sufficient text for rhetorical analysis. These were added as a measure of document-type diversity, and apart from those described in Item 5 below, were not used in determining quotas. These will be defined below.

5. Mandatory Document Types: It was decided that all TDC documents would be members of the primary categories *internal source* and *Public Health*, of the secondary category *English*, and not a member of the secondary category *short*. In other words,

all documents in the Quota Sample should be internal source, have Public Health significance, be primarily English, and not be short. Quotas will be derived from the remaining two primary categories: *internal audience* and *named addressee* (and their complements).

6. Quota Percentages: Finally, it was formally decided that quotas would be based on the percentages represented in the Core Sample, classified as per items 3 and 4 above, and adjusted to allow each stratum to represent a minimum of ten percent[3] of the total documents.

In terms of the above primary and secondary categories, the project investigators reached consensus on the definitions presented below after three rounds of proposals, testing on the initial Core Sample, feedback, and revision. In the following definitions, the quoted portions are the final working definitions taken directly from the authors notes (dated February 2, 2002). Again, the quoted portions in the following items do not conform to the terminology found in Chapter 2, which was codified at a later date. Specifically, in relation to Chapter 2 terminology, pairs of exclusive categories below refer to a category/complement pair. This is noted in each of the items below; however, subsequent discussion and tables will use the original terminology found in the author's notes.

- Primary Category: *Source*: 'industry internal/industry external (exclusive) - industry internal is defined as any document that 1) originates within the tobacco industry structure or area of control whether it be a tobacco company, subcontractor, or funded organization (public or private), or 2) is produced by any individual or organization that engages in the production, distribution or sale of tobacco or tobacco products,

[3]Although '10 percent' was decided (and recorded in the project notes), this seems to have been largely ignored going forward, with no recorded explanation. My suspicion is that because it was decided early on that a supplemental corpus would be advantageous, which served to augment the strata with low percentages, it was also decided that we should maintain the representativeness of the TDC to the Core Sample rather than artificially increase counts. However, I have no proof of this at this time. There is no substitute for keeping good procedural records, and little recovery from a lack of it.

to include vendors at all levels. This should be checked against an authority list of known industry internal sources if possible.' The major revision to the definition of the *source* category was an expansion of *internal* to include all for-profit/for-hire entities directly involved in the research, growing, processing, distribution, and sales of tobacco products. This stems from the prevalence of vendor and research-related material found in the tobacco documents, which although not specifically sourced from within the seven industry entities named in the Master Settlement Agreement, were generated by entities supported by the tobacco industry. In Chapter 2 terminology, the category is *internal source* with the complement *NOT internal source*.

- Primary Category: *Audience*: 'industry internal/industry external (exclusive) - following the definition given above for industry internal. Also, advertisements included as a document with no accompanying report or commentary will be considered external audience. That is, we make the assumption, in the absence of other indicators, that advertisements are intended for public release. However, any additional copy or text included in a document with an advertisement image is to be considered an indication of continued evaluation or editing, meaning the document should be classified as internal.' The major revision to the definition of the *audience* category (apart from revisions of the definition for the *source* category above) is the additional specifications concerning advertisements. There were a number of questions raised during initial trials related to advertising copy and images, whether it was for internal or external audience. That is, advertising by its nature is external, yet the tobacco documents contain a great deal of advertising documents that were never released to the public. This required additional clarification. In Chapter 2 terminology, the category is *internal audience* with the complement *NOT internal audience*.

- Primary Category: *Addressee*: 'named individual(s)/unnamed individuals (exclusive) - named individual is defined as any document that is addressed to a specific person, a named individual, regardless of the number of other addressees (this is a change from

the earlier personal/multiple addressee distinction which was based on the number of named addressees), to include individuals addressed by initials or titles if from the content of the document or an authority list it can be determined that the initials or titles represent a person and not a department. Carbon copy or 'Copies To:' names are not included, nor are names of groups, nor are standardized 'Distribute To:' stamps, boxes or forms containing only initials (it has not yet been proven that these initials represent any individual).' The major revision to the definition of *addressee* was the move from *personal* to *named*. Our original intention was to distinguish between documents intended for wide company/industry distribution and those intended for only a few individuals. A definition based on number of individuals named proved difficult to interpret in light of the document set. What was found was a large variety of methods to designate document recipients, both in the original document and in marginalia (formal and informal). In particular, questions were raised about documents originally intended for small audiences which were later distributed to much larger audiences, as determined by stamped distribution boxes (marginalia) denoting wide distribution and attached cover sheets with large numbers of 'carbon-copy' recipients. The simpler requirement of one identified person resolved most of these issues. In terms of minor revisions, a stipulation about more non-traditional address was added because of the occurrence of communications in which individuals are addressed using initials or titles. As well, the stamped distribution boxes common in the 1990s decade, although many included specific names, were excluded as a test for the *named* category because of their formulaic nature. In Chapter 2 terminology, the category is *named addressee* with the complement *NOT named addressee*.

- Primary Category: *Public Health*: 'significant/not significant (exclusive) - Dr. Hirschhorn will have to define this one.' There were no revisions/changes of this definition (from the author's perspective). However, as the classification progressed Dr. Hirschhorn determined that this category was difficult to define clearly, and eventually he con-

cluded that any document meeting the mandatory requirements for inclusion in the TDC, point five above, is significant to Public Health at some level. That is, any document classified as *internal source*, *English*, and *NOT short*, would also be classified as *Public Health*. Thus, in the end the category was not considered in determining quotas (although the realization is certainly significant on a larger scale), and my task was greatly simplified. In Chapter 2 terminology, the category is *Public Health* with the complement *NOT Public Health*.

- Secondary Category: *Form*: 'Yes if the document consists of bullet-type captions and pre-defined text fields (templates) with spatial constraints that limit text input (for example, printed boxes), and that by the printing style or quality or marginal information indicate that the document is regularly duplicated or produced apart from the place and time that data were entered. Legal documents and contracts are excluded (because it is difficult to determine the amount of standardization in computer generated forms). Also excluded are data presented in a tabular or columnated manner.' The major revision to the definition of *form* was the specific exclusion of fill-in-the-blank legal documents and tabular data. Both of these document types are often standardized/formulaic text, but by the mid-1980's most we being generated by computer, and it is difficult for the reader to determine the amount of formulaic text verses original text. That is, document processing moved from documents being printed and then filled in by hand or typewriter, to documents being filled in by computer and then printed. Thus, earlier documents are easy to classify, but later ones are difficult without considerable research. To avoid confusion and inconsistency, these two common document types were excluded from the category. In Chapter 2 terminology, the category is *form* with the complement *NOT form*.

- Secondary Category: *Image*: 'Yes if the document consists only of images and image captions of less than 50 words.' There were no revisions/changes of this definition. Note, however, that text within an image was not considered *continuous* with non-image text

in classifications of *short* (see below). In Chapter 2 terminology, the category is *image* with the complement *NOT image*.

- Secondary Category: *English*: 'Yes if the document is primarily written in English, more than 50 percent.' There were no revisions/changes of this definition. In Chapter 2 terminology, the category is *English* with the complement *NOT English*.

- Secondary Category: *Editing*: 'Yes if the document image has hand-written or typed marginal or intra-text marks or comments added by a reader or editor after the completion of the original document that indicate a need for revision.' There were no revisions/changes of this definition. Note, however, that this applies only to editing marks or notes on/in a given document. This should not be confused with editing as it relates to multiple drafts in the Parallel Corpora of Manipulated Documents (the rhetorical corpus) described in Dr. Kretzschmar's sampling plan (from Appendix A). In Chapter 2 terminology, the category is *editing* with the complement *NOT editing*.

- Secondary Category: *Marginalia*: 'Yes if the document image has any marginal or intra-text mark or comment added by a reader after the completion of the original document, but not for editing purposes, to include personal notes, filing notes, evaluations, stamps, check marks, initials, etc., but excluding signatures following closing salutations. Hand written documents are not considered to have marginalia unless the rater can determine that marks or writing were added after the completion of the document.' The major revision to the definition of *marginalia* was a general broadening of permitted types to include all intentional, non-editing, secondary marks on the document. This was necessary given the high frequency of secondary marks, and wide variety of types. The reasoning was that a move in the other direction would have required a large rule set to evaluate the marks and would not have added to the overall analysis (although it would have added to its complexity). The obvious exception, although not noted above, was Bates Numbers. Had we included these stamps, which were actually tertiary marks

required by non-tobacco-industry entities, the category would have become meaningless to us as essentially ALL documents include them. In Chapter 2 terminology, the category is *marginalia* with the complement *NOT marginalia*.

- Secondary Category: *Short*: 'Yes if the document (image or text) has FEWER than 50 words of continuous, non-template text. Continuous is defined (by Dr. Rubin) as "cohering by virtue of syntactic or semantic cohesive ties linking text into topical units. Continuous text is not [separated by any box boundaries] or other graphic device that attempts to segregate one unit of text from its neighboring units. Thus, text that appears in different cells in a table is not continuous, even if pertaining to a similar topic." Template text (for defining non-template) is replicated or standardized prose such as form entry descriptors or explanations, headings, titles, standardized contracts or legal explanations, etc., quite often created by non industry sources.' The major revision to the overall definition of the category *short* was the definition of *continuous* added by Dr. Rubin. Early attempts at classification were confused by stylistic variations in format which indicate divisions in the text to varying degrees. Some examples are section, paragraph and topic divisions/headings; enumerated and bulleted paragraphs; lists and outlines; and line spacing. The intent of the above definition was to allow most styles of text layout, and thus focus attention on word count. In practice, because of the nature of the document set, text was classified as non-continuous only when it contained text boxes, as in forms. In Chapter 2 terminology, the category is *short* with the complement *NOT short*.

Having defined the categories, the 349 documents from the final Core Sample were examined and classified by the author, with the assistance of Dr. Hirschhorn (in relation to *Public Health*) and Catherine Brown. This yielded the tabulated results shown in Table 3.9. The full and final results set, including individual classifications of each document, can be viewed in Appendix B, Section B.2. All of the tabulated data to follow comes from the Appendix B data set.

Table 3.9: Core Sample: Initial Classification.

	19xx	Bliley	1950	1960	1970	1980	1990	Totals
Total Docs.	10	10	10	22	66	132	99	349
Form	2	0	0	2	8	18	19	49
Image	0	0	2	1	1	0	0	4
English	10	9	10	22	63	130	99	343
Editing	3	1	0	2	3	3	5	17
Marginalia	4	9	5	12	34	73	39	176
Short	3	2	2	4	20	37	33	101
Nm. Addressee	0	7	3	13	29	62	32	146
Int. Audience	8	9	6	20	53	109	88	293
Int. Source	8	9	8	20	55	108	93	301
Public Health	9	10	10	22	61	126	96	334

In order to verify the initial classification, fifty documents from the 349 Core Sample documents were randomly selected and given to Dr. Hirschhorn to be classified according to the three primary categories *internal source*, *internal audience*, and *named addressee* (he had already classified the entire set using the fourth, *Public Health*). Dr. Hirschhorn's classifications were then compared to those of the author and Catherine Brown. Again, all classification data for the specific documents selected for the comparison can be found in Appendix B, Section B.2.

Of the 150 comparisons made (three classifications for each document), five discrepancies were found in five documents (only one type per document). The discrepancies, as well as the resolutions, were as follows (the quoted material was taken directly from the author's notes):

- Document 1970-15 (Bates Number 2501015916): Issue: Dr. Hirschhorn 'notes that the author, Woodfield, was commissioned by TI and PM for other pro-industry papers (see Bates Numbers 2040593032 and 5025823917) and should be considered internal. The document itself gives no indication of internal funding (it is a university research

report on the economics of smoking in Australia). As well, the paper seems to give all indication that it was for public release from the university ([Dr. Hirschhorn and the author] both classified it as such), which supports the non-internal source.' Resolution: Not changed, citing that the classification 'Remains as external because of the nature of the document which appears to be released by the university to the public. We have no indication at this time that this [document] was funded by tobacco (doesn't include the same indicators as other similar but internal docs).'

- Document 1970-40 (Bates Number 621094587): Issue: Dr. Hirschhorn 'classified this as not named. [The author] classifies this as named because the report, which is a form, has a blank in the top left corner that is labeled 'requested by' and has the name D. V. Cantrell in it. [The author] suspects that this is a report addressed to D. V. Cantrell. See also doc. 1970-26 [(Bates Number 659055253)] of the same type.' Resolution: Changed to *NOT named addressee*, citing that the investigators 'Decided that it was not named as the 'requested by' name is also in the c.c. list. Thus, not addressed specifically to D. V. Cantrell. Also changed doc. 1970-26 [(Bates Number 659055253)].'

- Document 1980-119 (Bates Number 660113863): Issue: Dr. Hirschhorn 'classified this as general/unnamed audience. However the addressee, Donna Sengelaub, is named two times in the header of the report.' Resolution: Not changed.

- Document 1990-40 (Bates Number 2063588303): Issue: Obvious error on the part of the author. Resolution: Changed to *named addressee*, citing that 'This was an error on the part of [the author]. The addressee, Cliff Lilly, is named.'

- Document 1990-95 (Bates Number 2061878709): Issue: Dr. Hirschhorn 'notes that the Restaurant Association is 'in cahoots' with TI.' Resolution: Changed to *internal audience* citing that at a later date '[Dr. Hirschhorn] found the funding link.'

Table 3.10 provides revised (final) classification data for the 349 Core Sample documents that incorporate the above changes. The count for the 1970 stratum *named addressee* category was lowered by two; the 1990 stratum *named addressee* category was raised by one; and the 1990 stratum *internal audience* category was raised by one. Variations from Table 3.9 are noted with asterisks, of which there are five: three are the result of classification changes, and two from consequential changes in totals.

Table 3.10: Core Sample: Final Classification.

	19xx	Bliley	1950	1960	1970	1980	1990	Totals
Total Docs.	10	10	10	22	66	132	99	349
Form	2	0	0	2	8	18	19	49
Image	0	0	2	1	1	0	0	4
English	10	9	10	22	63	130	99	343
Editing	3	1	0	2	3	3	5	17
Marginalia	4	9	5	12	34	73	39	176
Short	3	2	2	4	20	37	33	101
Nm. Addressee	0	7	3	13	*27	62	*33	*145
Int. Audience	8	9	6	20	53	109	*89	*294
Int. Source	8	9	8	20	55	108	93	301
Public Health	9	10	10	22	61	126	96	334

3.4.3 Establishing Quotas

Once the classification of the Core Sample was complete, establishing quotas for a representative sample was rather straightforward. Following the decisions on mandatory document types made after the October 2001 meeting (see Item 4 and 5 above in Section 3.4.2), the first step was to eliminate all unnecessary documents from the data. That is, all documents in the Core Sample that were classified as *NOT English*, *short*, *NOT internal source*, or *NOT Public Health* could be excluded from consideration because they did not meet the minimum requirements for inclusion in the Quota Sample. This reduced the set of docu-

ments in play from the 349 of the full Core Sample to 202 documents[4] considered *usable* for deriving quotas.

The second step was to remove the clutter of unnecessary categories from the matrix. Because all the remaining documents were necessarily *English*, *NOT short*, *internal source*, and *Public Health*, these categories could be ignored. As well, the secondary categories, which were not to be used in determining quotas, could also be ignored. This reduced the number of categories from the ten originally evaluated to two: *named addressee* and *internal audience*. The end result is presented in Table 3.11, a set of 202 *usable* documents along with revised counts per stratum for the two remaining categories. These counts, all of which came from the data in Appendix B.2, were reduced by the removal of *NOT usable* documents.

Table 3.11: Core Sample: Reduced Classification.

	19xx	Bliley	1950	1960	1970	1980	1990	Totals
Usable	5	7	6	16	36	71	61	202
Nm. Addressee	0	6	2	9	15	39	21	92
Int. Audience	5	7	4	14	34	70	61	195

The final step was to rearrange the data from the remaining documents and category counts into quotas. Given the two remaining categories, each of which had a binary classification (yes or no), there was the possibility for four document types. That is, any given document could be *named addressee* OR *NOT named addressee*, AND *internal audience* OR *NOT internal audience*. This is summarized in Table 3.12, where '1' denotes YES, and a '0' denotes a NO or NOT. Also provided in the table is the title that will use below to refer to

[4]The actual number of documents classified as *usable* is 203. By an oversight, this was originally determined to be 202. The count of *usable* documents in the 1990 stratum is 62 rather than the 61 given in Table 3.11. This error reduced the final document total from 812 to 808 (0.50 percent) for the full Quota Sample, and from 248 to 244 (1.61 percent) in the 1990 stratum. In terms of differences in ratio, the 1990 stratum to full corpus ratio was reduced 1.11 percent. Because this error was not discovered until after the completion of sampling (and given the small variance), the original counts of *usable* documents are used in all discussions of quotas. This error is noted to avoid confusion between the data presented here and the Core Sample metadata in Appendix B.2.

the document type, where 'I' is 'internal' for *internal audience*, 'E' is 'external' for the complement *NOT internal audience*, 'N' is 'named' for *named addressee*, and 'U' is 'unnamed' for the complement *NOT named addressee*.

Table 3.12: Quota Sample: Binary Table of Document Types.

Type	Nm. Addressee	Int. Addressee	Title
3	1	1	NI
2	1	0	NE
1	0	1	UI
0	0	0	UE

Having the document types, the pertinent Core Sample classification data from Appendix B.2 could then be placed in a cross-tabs format to provide initial quotas for each stratum and document type, per 202 documents, as seen in Table 3.13. As an example, Table 3.13 can be interpreted as such: reading left to right in the fourth data row, per sample of 202 documents, of the *1960* stratum documents, 8 should be of type NI, 1 of type NE, 6 of type UI, and 1 of type UE, totaling 16 documents. Although these quotas could be converted into percentages in a straightforward manner, they were left as 'per 202' ratios because of the sampling procedures used for gathering documents (see Section 3.4.4 below).

Table 3.13: Quota Sample: Initial Quotas per 202 Documents.

Decade	NI	NE	UI	UE	Total
19xx	0	0	5	0	5
Bliley	6	0	1	0	7
1950	2	0	2	2	6
1960	8	1	6	1	16
1970	13	2	21	0	36
1980	38	1	32	0	71
1990	21	0	40	0	61
Total	88	4	107	3	202

The last change made to the quotas before finalizing them was to separate the 1950 stratum into two strata, the 1950 stratum (all documents from 19500000 to 19591231), and

the 1900 stratum (all documents from 19000101 to 19491231). This was done to insure that the earlier stratum, which represented fewer documents in the whole of Tobacco documents, would be represented in the TDC. Early attempts at sampling for the combined 1950 stratum showed that this was difficult. To make the separation, quotas were divided equally between the two new strata. The final quotas are presented in Table 3.14. These were used for all sampling procedures for the Quota Sample. For convenience in interpreting the data, percentages (in parentheses) are provided along with the 'per 202' ratios.

Table 3.14: Quota Sample: Final Quotas.

Decade	NI		NE		UI		UE		Total	
19xx	0	(0.00)	0	(0.00)	5	(2.48)	0	(0.00)	5	(2.48)
Bliley	6	(2.97)	0	(0.00)	1	(0.50)	0	(0.00)	7	(3.47)
1900	1	(0.50)	0	(0.00)	1	(0.50)	1	(0.50)	3	(1.49)
1950	1	(0.50)	0	(0.00)	1	(0.50)	1	(0.50)	3	(1.49)
1960	8	(3.96)	1	(0.50)	6	(2.97)	1	(0.50)	16	(7.92)
1970	13	(6.44)	2	(0.99)	21	(10.40)	0	(0.00)	36	(17.82)
1980	38	(18.81)	1	(0.50)	32	(15.84)	0	(0.00)	71	(35.15)
1990	21	(10.40)	0	(0.00)	40	(19.80)	0	(0.00)	61	(30.20)
Total	88	(43.56)	4	(1.98)	107	(52.97)	3	(1.49)	202	(100.00)

3.4.4 Sampling Procedures

Once the quotas were established, the underlying sampling procedures for the Quota Sample changed very little from those used for gathering the Core Sample (see Section 3.3.2 above) and the initial specifications of Dr. Kretzschmar, which are as follows:

> After establishment of text type quotas, selection of particular documents will be randomized by a fixed, sequential sampling frame. For the Reference Sample/Corpus, the sampling procedure outlined in Part 1 should be repeated; however, for Part 2 only documents which fit the established quotas should be included. For instance, if the Nth document cannot be included in the sample, either because it does not fit a quota category or because the quota category

> that it its is already full, then the investigator will move on to the next nth document in sequence. (from Appendix A)

Of the adaptations made in transition from the Core Sample procedures to those of the Quota Sample, the most notable was the inclusion of two additional layers of selection: the mandatory document types (i.e. the document must be *English*, *NOT short*, *internal source*, and *Public Health*), and the quota document types from Table 3.12 (NI, NE, UI, and UE). With the Core Sample, the quota of documents to be selected for each decade-based strata was filled by randomly selecting documents from the assigned date range. For the Quota Sample this was also the case; however, once the initial random selection was made, a given document was required to match all mandatory document types, and to match any unfilled quota document type. The obvious consequence of these additional layers is that the ratio of documents examined to documents selected increased significantly from the 1:1 ratio of the Core Sample. This is because any document, once selected randomly from a date range, could be rejected by either of the subsequent procedures. In fact, subsequent rejection for being the wrong document type occurred at a rate of roughly 42 percent (this will be discussed further in Section 3.4.6 below).

In order to maintain the correct ratios between quotas, documents were sampled (gathered) in sets of 202, which is the sum of all quotas found in Table 3.14 and represents the number of *usable* documents in the Core Sample. This fixed-set method was necessary given that some of the quotas are 1, which affords no opportunity to collect documents in anything less than multiples of 202 and still maintain the given ratios. As well, it provided some efficiency to the process. As each set was gathered, the documents selected were immediately started in the archival process, which included conversion from images to a plain ASCII text format. This allowed accurate running counts of usable words/tokens to be maintained, and in turn allowed us to determine if additional sets were needed. Based on earlier estimates we expected to need as many as 1,190 documents to reach the goal of 500,000 tokens for the Quota Sample (see Section 3.4.1 above). This would have required collecting 1,212 doc-

uments (six multiples of 202) to maintain ratios. Had sampling been completed as a single set based on these estimates, we would have incurred a document overage of approximately 50 percent (see Section 3.4.5 below).

For the general strata (decades 1950–1990) sampling was accomplished in the following manner: for each document to be collected a starting point was chosen by randomly selecting a year (from 0 to 9), a month (0 to 12), and a day (0 to 31). The 0's in the month and day were necessary to insure that undated or partially dated (denoted by 0's in the date) documents were included. Once the start date was chosen, the NAAG Snapshot was searched for documents having that specific date. This was done online using either the TDO or Legacy websites. Of the total number of documents returned, a random number N was selected such that the Nth returned document from the random-date search became the start document for the document-type-and-quota search. Beginning with this start document and advancing to the succeeding documents in order, the first document encountered that met the following four criteria was selected:

1. the document was generated on the given date. In some cases, documents were returned from the search because they contained the given date as a character string in the document itself, perhaps as a number, rather than in the document metadata. That is, the website search algorithms matched strings in both the document metadata and document text. All documents were checked to insure the correct document generation date.

2. the document passed all mandatory document-type requirements. Namely, that the document was *English*, *NOT short*, *internal source*, and *Public Health*).

3. the document matched one of the remaining (unfilled) quota types (from Table 3.12: NI, NE, UI, or UE).

4. the document had not been previously selected. Each document was checked against a running list of Bates Numbers to prevent duplication within the full sample.

In the event that the above procedure failed to yield a start document or any usable document prior to reaching the end of the document set returned by the date search, the entire procedure was repeated beginning with the selection of another random date. This was done until all general strata quotas were filled for the given sampling set. The notable exception to this is that if 0 was selected for the month, the day was set to the wildcard '*' to insure the retrieval of the entire set of 0-month documents for the given year.

In several cases, the sampled documents did not have stamped Bates Numbers on the document image, which prevented standardized file naming and tracking (described below). In these cases, these documents were assigned a temporary tracking number until the Bates Number issue could be resolved. For all such cases, the resolution was to locate an alternate document based on the 'alias' Bates Number provided by the TDO website. This alias was a reference to an identical document image (a copy of the same document) found in another location.

There were also several cases in which documents were collected but later found to be of the wrong document type, to be a duplicate document, or to consist predominantly of illegible text. For these documents, replacements were made as if the original sampling were taking place.

For the 1900–1949 stratum, the same general procedures as above were used with the exceptions that the year was randomly selected from 0–49, and that the combination of year 0 and month 0 was not allowed. This prevented mixing documents with the 19xx stratum. In cases where the random date selection produced a 0-0 combination for year-month, the process was restarted.

For the 19xx stratum, the same general procedures as above were used with the exception that the start document was determined by random selection of a number between 1 and 62,494 (the total of 19xx documents).

Finally, for the Bliley stratum, the same general procedures as above were used with the exceptions that the year was randomly selected from the full century range 0 to 99 and, as

with the 19xx stratum, the combination of year 0 and month 0 was not allowed. With the Bliley stratum, searches were made using only the TDO website. As mentioned above, at the time of sampling the Legacy website contained only the NAAG Snapshot documents.

3.4.5 Results

Sampling for the Quota Sample was completed after gathering four of the six expected sampling sets, bringing the total document count for the Quota Sample to 808 documents. As mentioned above, document conversion was occurring simultaneously with sampling. This allowed accurate token counts to be maintained as the sampling sets were completed, which in turn allowed the sampling to be stopped once the total token count bypassed the goal of 500,000 set by Kretzschmar. The current official count of tokens in the Quota Sample is 543,959. However, this number can vary widely depending on which text is extracted from the XML archive and differences in tokenizing[5] procedures.

As documents were sampled, the images for each were downloaded from the TDO or Legacy websites and archived. By this point (mid-2002) most documents were available as PDF archives. This was the preferred format given that all the document page images were packaged into a single file. However, in some instances only TIFF images of the individual pages were available. Downloaded files were named based on the Bates Number of the first page of the document, replacing spaces with an underscore character and adding the appropriate file extension (*.pdf* or *.tif*). They were then stored in folders named according to the sampling set (Q1 through Q4) and the decade strata. For more specific file information for each of the 808 docunets in the Quota Sample (Bates-Number ranges, sampling set, page counts, token counts, et cetera), refer to Appendix B, *Document Metadata*, Section B.3.

[5]Refer to Chapter 5 for a more detailed discussion of tokens and tokenizing.

3.4.6 Sample Rejection Rates

Given that A) the Core Sample was random from within each stratum, and that B) the quotas for the Quota Sample were derived directly from the Core Sample document-type ratios, the expectation is that the set of all documents examined while assembling the Quota Sample would maintain similar document-type ratios to that of the Core Sample. In other words, given a Core Sample with 202 usable documents produced by examining 349 total documents and rejecting 147 (42.12 percent), in order to produce a Quota Sample of 808 usable documents, one should expect to examine 1,396 total documents and reject 588 (also 42.12 percent). Had I had the foresight to keep accurate records of document rejections, this would have been an excellent measure for evaluating the Quota Sample collection procedures. That is, because the ratios (quotas) were fixed, and because they were based on a random sample of the same domain, any significant variation in overall rates of document rejection would indicate a procedural problem. Unfortunately, my focus early on was the usable document rather than the unusable, and proper records were not kept. However, for other purposes,[6] we did track the number of document pages skipped or rejected prior to accepting a document, and we were able to use these data to good effect for making the above comparison.

For a total of 808 documents collected (the usable documents) for the Quota Sample, records indicate by a non-zero count of pages skipped that for 586 of those documents there was at least one document rejected. This provides a minimum of 1,394 document examined, and puts the rejection rate at 42.04 percent. However, no records were kept for 22 of the usable documents. If an expected 42 percent of the unrecorded iterations (roughly 9) is added to the 586 recorded rejections, it yields 1,403 total with 595 rejections, which is a rate of 42.41 percent. Even at the extreme possibility of 1,416 total with 608 rejections, if all the unrecorded iterations began by rejecting a document, the rate would be 42.94 percent, which is still within a percent of the rate in the Core Sample.

[6] I have no idea at this point what those purposes were.

It is important to keep in mind that the true rejection rates will be higher given that a sampling iteration may have rejected several documents before accepting one. In the majority of cases the record of pages skipped is low, five or less, which indicates a single document skipped (rejected), thus our belief is that the rejection rates would be only slightly higher. Although we have no way of proving this, by applying a standard formula for comparing ratios (Moore and McCabe 1999, 583), significant deviation from the expected value is not reached until the number of rejected document reaches 681 out of 1,489 documents total. This reflects an error rate of 14.45 percent above the expected rejection rate, which to us seemed unlikely. Overall, we were pleased with the rates and believe they provided an additional level of confidence in the sampling procedures. All of the above data are summarized in Table 3.15.

Table 3.15: Quota Sample: Comparison of Sample Rejection Rates.

	Usable	Rejected	Total	Rate
Core Sample	202	147	349	42.12%
Quota Sample (known)	808	586	1394	42.04%
Quota Sample (expected)	808	595	1403	42.41%
Quota Sample (extreme)	808	608	1416	42.94%
Point of Significance	808	681	1489	45.74%

3.5 Supplemental Sample

By the middle of June, 2002, the Quota Sample had progressed enough that we were able to determine that there would be insufficient numbers of external-audience documents (both named and unnamed, NE and UE) to allow adequate comparison of potentially-deceptive documents to non-deceptive. That is, the expectation was that deception would occur in documents addressed to external audiences, yet only four percent of the sample (32 out of 808) were classified as NE or UE. These low counts meant that it would be difficult to make valid comparisons of documents from the Rhetorical Sample with those in the Quota Sample

in order to determine if differences were the result of deceptive practices or the result of a shift in style associated with the change in audience.

To remedy this, a second, smaller quota-based sample was constructed which contained only external-audience documents. When the plan for this corpus, known as the Supplemental Sample, was finalized at the end of June, 2002, it was determined that there were sufficient resources to gather and process a corpus of roughly 50,000 tokens. The plan itself was simply to construct the new corpus in the same manner as the Quota Sample with the exception of an additional external-audience-only requirement.

3.5.1 Determining Quotas

Based on Quota Sample data which placed the average token count per document at roughly 660, the Supplemental Sample size was fixed at 100 documents to be gathered in a single sampling iteration. Aside from being a nice round number for formulating quotas, a goal of 100 documents provided a 20 percent cushion over the target of 50,000 tokens given that external documents showed a tendency to be shorter than internal.

The quotas per decade strata were derived from the percentages found in Table 3.5, *Core Sample: Initial Yields (Augmented)*, which indicated the portion of total documents represented by each stratum. These were balanced so that each stratum was allowed at least one document, and adjusted to whole numbers so that they equated to the count of documents needed for the 100-document set. These adjusted strata quotas were then divided equally between the *Named* and *Unnamed* (NE and UE) categories, which approximates the distributions found in Table 3.14, *Quota Sample: Final Quotas*. The exception is for the strata *Bliley* and *19xx* which allowed only a single document. These were divided one to each category, *Bliley* being given to NE, and *19xx* given to UE. These data are summarized in Table 3.16.

Table 3.16: Supplemental Sample: Final Quotas.

Stratum	Actual %	Adjusted %	NE	UE	Total
19xx	1.84	1	1	0	1
Bliley	0.97	1	0	1	1
1950	3.06	4	2	2	4
1960	6.59	6	3	3	6
1970	19.47	20	10	10	20
1980	38.90	38	19	19	38
1990	29.17	30	15	15	30
Total	100.00	100	50	50	100

3.5.2 Sampling Procedures

For the Supplemental Sample, which is also a quota-based sample, the sampling procedures were identical to those used for the Quota Sample. See Section 3.4.4 above for more detail.

3.5.3 Results

As documents were sampled, the images for each were downloaded from the TDO or Legacy websites and archived. By this point (mid-2002) most documents were available as PDF archives. This was the preferred format given that all the document page images were packaged into a single file. However, in some instances only TIFF images of the individual pages were available. Downloaded files were named based on the Bates Number of the first page of the document, replacing spaces with an underscore character and adding the appropriate file extension (*.pdf* or *.tif*). They were then stored in folders named according to the sampling set (S1) and the decade strata.

Once archiving and conversion were complete, it was determined that the goal of 50,000 tokens for the Supplemental Sample was not officially reached. The current official count of total tokens in the Supplemental Sample is 48,916, which is 1,084 tokens (2.17 percent) low. However, given that this number can vary widely depending on which text is extracted from

the XML archive and differences in tokenizing[7] procedures, no attempt was made to add additional documents. This would have required another full sampling set (100 documents) to maintain the prescribed quotas. For more specific file information, including Bates-Number ranges, page counts, and token counts, for each document of the 100 in the Supplemental Sample, refer to Appendix B, *Document Metadata*, Section B.4.

3.6 Replacements

Over the course of sampling, archiving, and validation of both the Quota and Supplemental Samples, there were several opportunities for sample documents to be reclassified and/or rejected based on the reexamination of their content. In a few cases, during archiving the document was found to be illegible to the extent that no analyzable text could be recovered, but more frequently it was discovered that documents had been misclassified during sampling, being assigned to the wrong decade or audience type. In the event that being illegible or being reclassified to a quota that was already filled made the document unusable, replacement documents were needed. In all cases, replacement documents were selected using the same procedure used for the original document (i.e. the original sampling was duplicated), regardless of the sample affected. In most cases, the documents requiring replacement were discovered prior to the end of sampling, and it was simply a matter of moving the document to an unfilled quota (if possible), or deleting the document from the sample and releasing the quota slot it held to be filled by another document.

Another notable replacement issue is the use of 'alias' Bates Numbers. With a number of documents selected during sampling for the Quota and Supplemental Samples, the document image did not have a Bates Number assigned to it. In all such cases that we encountered, the Tobacco Documents Online web site offered an alias Bates Number, which was the Bates Number assigned to an identical document found elsewhere in the NAAG Snapshot collection. After comparing the images and verifying that the alias document contained the

[7]Refer to Chapter 5 for a more detailed discussion of tokens and tokenizing.

same data as the original, the original document was replaced by the alias document. This provided a known Bates Number for archiving such that all sample documents could easily be located in other archives.

Chapter 4

Corpus Construction: Archiving Procedures

4.1 Introduction

At the onset of the Tobacco Document Project, the plan for building the various corpora of computer-analyzable texts was to collect documents from both the Minnesota repository (physicals) and the various internet repositories (computer images) and convert them in a semi-automated process to a machine-readable format using optical scanning devices and optical character recognition (OCR) software. However, once we began to examine the documents which had been made available by the tobacco industry, we quickly realized that this would not be possible. The condition and format of the documents themselves precluded any reliable automated conversion. What we found was that documents, both physical pages and electronic images, were generally of very poor quality in terms of resolution, often being copies of copies far removed from the originals. As well, we found that the text was often tilted on the page, that handwritten documents (or documents with handwritten additions) were common, that low-pin-count dot-matrix printing was prevalent, images were frequent, stamps and other marginalia were regular, format and font styles changed frequently, and so forth. In general, the documents were legible with the human eye, but a complete mess for computer reading.

The resolution to this problem, as one might expect, was to have human readers enter the data into the computer manually by keyboard. Although this was certainly a marked departure from intended procedure, which significantly reduced quantity because of the additional time and cost associated, it did effectively raise the quality (i.e. accuracy) of corpus documents to reliable levels. As well, it afforded a convenient opportunity to format the

documents and add the markup in preparation for permanent archiving. In the following sections a more detailed description of the issues related to text conversion is provided. We will begin with illustrations of the various the problems we encountered, and then move on to the procedures used for the solutions, namely the markup and keyboarding.

4.2 Document Examples

Before proceeding to the discussion of procedures for document conversion, a few document examples are necessary to provide some insight into the nature of the task. In other words, the entire concept of manual document conversion, let alone the complexity of the markup schema used, might well seem unnecessary having not examined some actual tobacco documents. What most have in mind when they imagine what a document might look like is something similar to this dissertation, a document that may not be rhetorically captivating but is certainly well formatted and uniform, having clearly delimited division, a single font style, and lines of text that run horizontally on the page. If this were the norm for tobacco documents, then indeed the procedures below would be unnecessary. OCR and machine formatting would be able to convert the images to text with a high degree of accuracy. But unfortunately this is not the case. The documents are, I have always suspected, in the worst possible condition that the courts would allow.

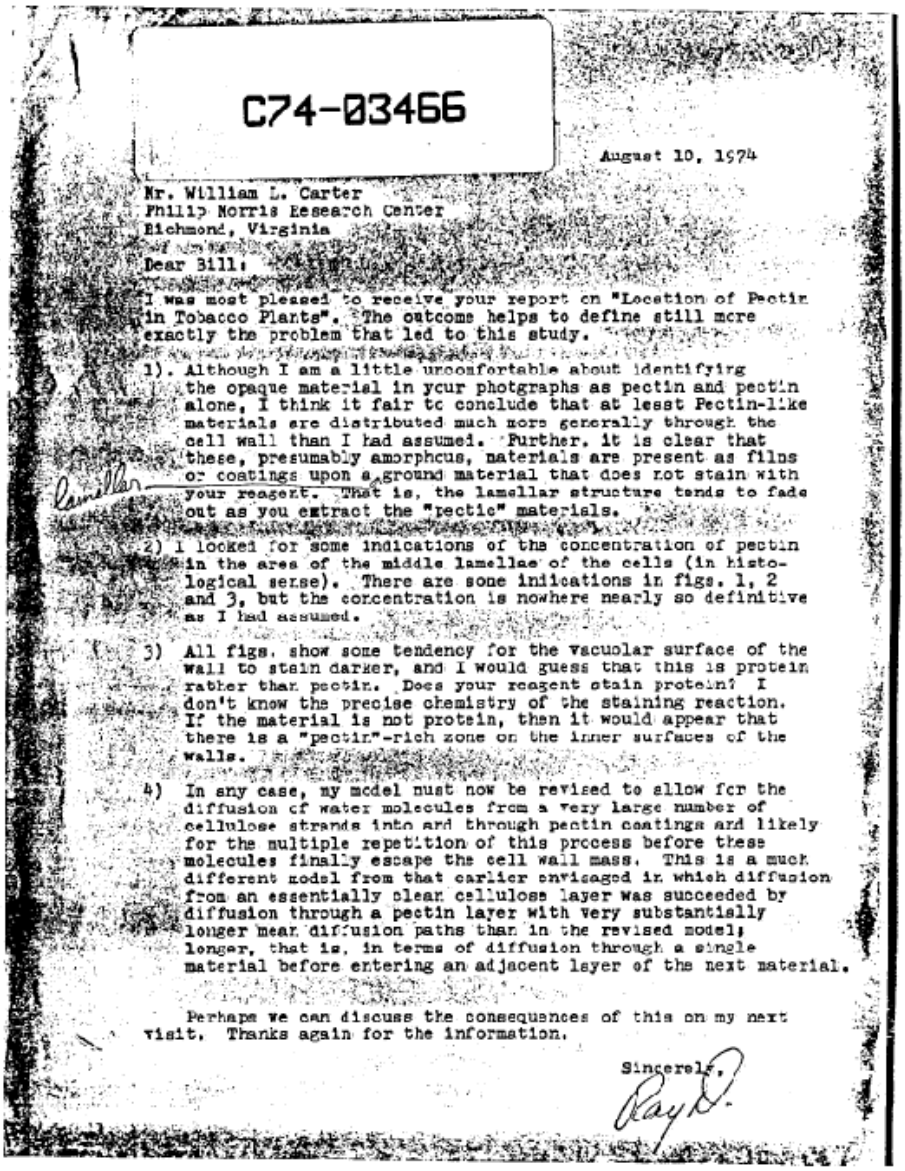
To illustrate what one might find in the tobacco documents, I have selected as examples ten pages from documents in the Quota Sample. Although none of the examples are the model of a 'typical' tobacco document, as a group they give a clear idea of the range of document types and forms found in the Quota Sample. These should serve well to acquaint the reader with document formats and features, and also as a reference for later discussion.

The first example, Figure 4.1, is a single page document from Philip Morris produced in December of 1974 (Bates Number 1000845352). Overall the document is well formatted and straight on the page (not excessively tilted in photocopying). However, there are a number of document attributes which complicate conversion to text. First, the document has poor

resolution, probably being a photocopy of the original. The text is not clear, but fuzzy and faded, and there are dark areas around the margins. While these lower the accuracy of OCR software, other features increase the complexity of the coding. For example, notice the big label/sticker on the top, the handwritten editing marks in the middle, and the signature at the end. In other words, these non-document additions cannot be accounted for in the archive with simple text. There must be a manner to indicate that '*lamellar*' is inserted between '*a*' and '*ground*', and to indicate that '*C74-03466*' is on a label pasted to the top of the document.

The next example, Figure 4.2, is the fifth page of an eight-page marketing report from Lorillard produced in January of 1992 (Bates Number 92043636). The most immediate irregular feature is that as a whole the text is tilted to the left. Actually, most of the text is tilted to the left two degrees, while the top right corner is tilted right (or down) seven degrees, a result of bending the page during copying. While many OCR programs assert that they can convert images tilted up to five degrees, our experience is that accuracy is greatly decreased. For example, here is the OCR output for this page that was copied from the Tobacco Documents Online website (TDO, http://tobaccodocuments.org) in November 2007:

```
Nedia'Ad4 ertising PraErarn A,n cxtcnsivt med:c progr'am wil3 arinounce and reWaree tht
hlcrit U{tirnamesseEe. This meciage posixions Ulctn,4 as a brand that will bring Merit's low
tar, f .rtat tas,e heritagc and qu:'ity to the Lowest scgment. Tht adveni6-,~ ,~ig carrmpaign
will be modern andnews- oriented,tostimu- late renewed intcrest ir, the entire Merit family.
Media - vehicics include magazinc;, ouYdoor, newspaper;, and ' - supplcments. DAMS: i The
U3tima mt3ia and Adtertising camp aign is designed to create alWareness and generittc
Cxcitcment about new , Uttirna as welt as the Merit brand ntume, 3191 thruugh b!92 DiKEC'T'
MAIL: A dirGCt, rnarketing progrium will bqin in rtud- March. This prograrn wil3 include: * A
maifing to srnokers of Carlton, Now, and cornpeti- tive full margin uitra low tar brdndc. The
miling will delivertwofrre packs of t3ttirrca as well asa continuity offer for adc3itional
inctnt:ves.  A m.edia offe: clrallenginb Carlton and Now trnokers to try U7tima. Iri ordor
to gerterate additional Cartton and Now names, we will offer two f-ee packs for one Car11 on
orNow pxck pro Df Qf put~chusc. The continuity offLr v.iil also lx- included in this
progrsun. In order to prevent "sticker shock" at the end of the off- laty-a prict prornotian, 2
carton stn:ffcr coupon offer will be inserttc' ir; ari caRun,, produced during the lmt weel:s
of the introduction. Consumcrs will br offered \$2.50 off- carton r,oupons by mail with an
Ulcrrna carton proof of purchase, ?SdE F.^LL."1'.'l3it'F'^nJ. f : ~}~V:1~ C1El:'? '^~.. Prr'_ ~-
Ck f 0 ~ ~~.~..= CY ~ rL2- - -i~ r.i.~ c~ue-:Gwr esTC k~saFtlFs'silVL'2L TD7n_h.r.G.Ow t~ c"A
_1 s .._.~_. .. . .~..--.-. PFa:aG . t30?
```



C74-03466

August 10, 1974

Mr. William L. Carter
Philip Morris Research Center
Richmond, Virginia

Dear Bill:

I was most pleased to receive your report on "Location of Pectin in Tobacco Plants". The outcome helps to define still more exactly the problem that led to this study.

1). Although I am a little uncomfortable about identifying the opaque material in your photgraphs as pectin and pectin alone, I think it fair to conclude that at least Pectin-like materials are distributed much more generally through the cell wall than I had assumed. Further, it is clear that these, presumably amorphous, materials are present as films or coatings upon a ground material that does not stain with your reagent. That is, the lamellar structure tends to fade out as you extract the "pectic" materials.

lamellar

2) I looked for some indications of the concentration of pectin in the area of the middle lamellae of the cells (in histological sense). There are some indications in figs. 1, 2 and 3, but the concentration is nowhere nearly so definitive as I had assumed.

3) All figs. show some tendency for the vacuolar surface of the wall to stain darker, and I would guess that this is protein rather than pectin. Does your reagent stain protein? I don't know the precise chemistry of the staining reaction. If the material is not protein, then it would appear that there is a "pectin"-rich zone on the inner surfaces of the walls.

4) In any case, my model must now be revised to allow for the diffusion of water molecules from a very large number of cellulose strands into and through pectin coatings and likely for the multiple repetition of this process before these molecules finally escape the cell wall mass. This is a much different model from that earlier envisaged in which diffusion from an essentially clear cellulose layer was succeeded by diffusion through a pectin layer with very substantially longer mean diffusion paths than in the revised model; longer, that is, in terms of diffusion through a single material before entering an adjacent layer of the next material.

Perhaps we can discuss the consequences of this on my next visit. Thanks again for the information.

Sincerely,

Ray D.

1000845352

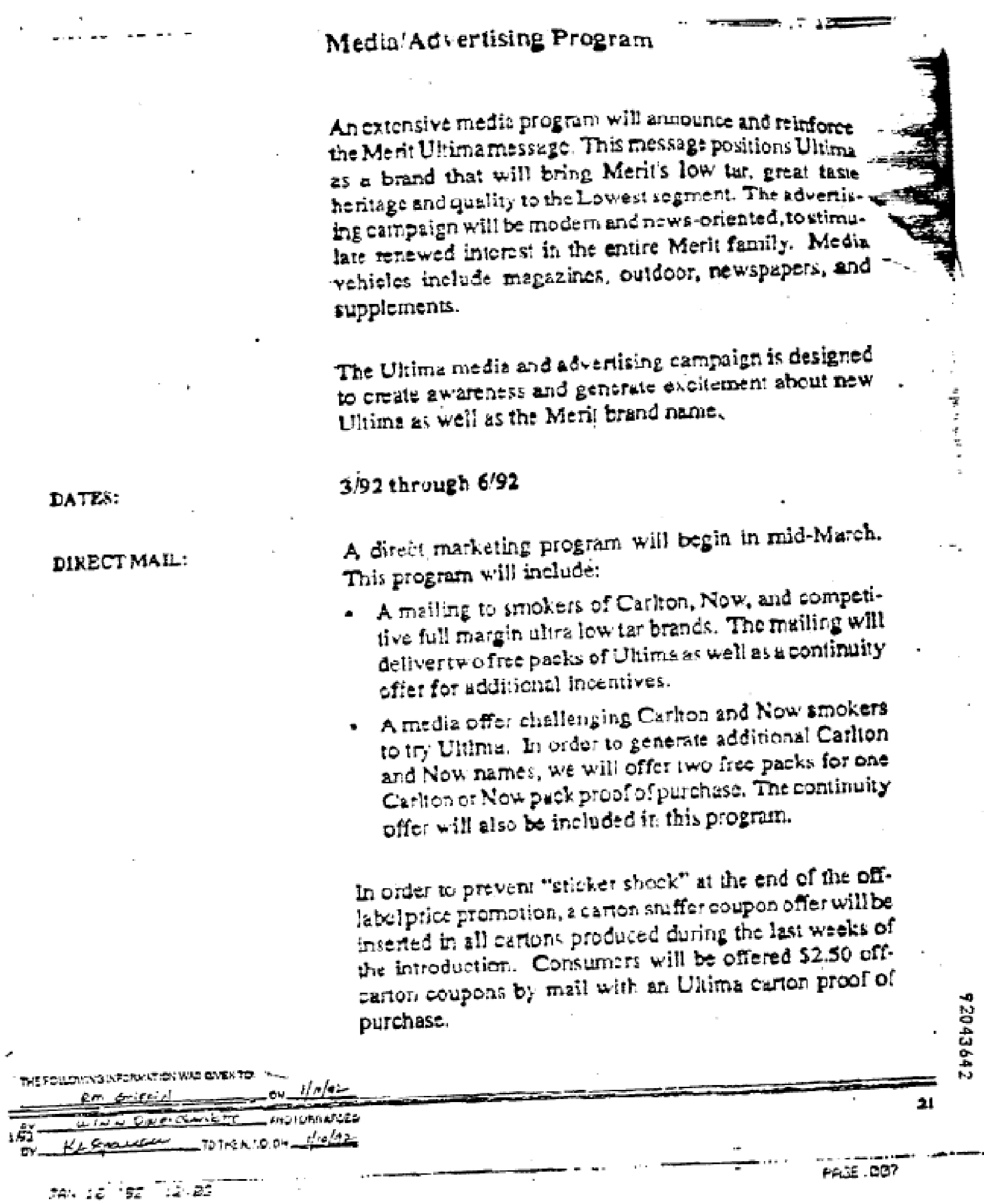
Figure 4.1: Document Example 1: Low Resolution Text

To say the least, this has less-than-desirable accuracy.[1] This is particularly the case in the top right corner where the tilting and dark margin result in '*announce and reinforce*', '*message positions Ultima*', and '*low tar, great taste*' becoming '*arinounce and reWaree*', '*meciage posixions Ulctn*', and '*low tar, f .rtat tas,e*'. In terms of coding, there are also a number of problematic document features which raise questions. For example, how is one to treat the stamp in the lower left corner, which has both machine and hand-produced text, and how should the facsimile machine (fax) transmission data at the bottom of the page be handled? And more simply, what should be done with the document title in the header at the top of the page? Does it fit structurally or rhetorically between the last paragraph on page 4 and the first on this page? Again, just as OCR programs do not do well with converting the text, text alone does not do well with preserving the structure of the document.

This next example, Figure 4.3, is an example (an illegible one) of the many odd items found in the Quota Sample. It is the second page of a two-page document from Lorillard produced in February of 1976 (Bates Number 03671878). Interestingly, this is a cover page for the first page in the document and contains the image of the first page ghosted in reverse from being stuck together. The only original data is found in the stamp on the lower right corner indicating it was received by C. H Judge on the fourth day of an illegible month in 1976. As a side note, this stamp image is ghosted in reverse on the first page of the document. The question raised is of course what to do with it in terms of coding. It is part of a series of Bates Numbers considered to be a document, so it must be recorded in some form.

The fourth example page, Figure 4.4, is the tenth page of ten in a document from Lorillard produced in 1979 (Bates Number 04233241). The document explains Lorilard's desire to change attitudes toward tobacco using the media. It is a typed document, but the sixth, eighth, and tenth pages are notes related to the document content. Bypassing any OCR dis-

[1]Not all of the non-ANSI characters in the data from the TDO website were rendered correctly when converted to the PDF format of this document. However, this has little effect on evaluating the OCR program's accuracy.



Media/Advertising Program

An extensive media program will announce and reinforce the Merit Ultima message. This message positions Ultima as a brand that will bring Merit's low tar, great taste heritage and quality to the Lowest segment. The advertising campaign will be modern and news-oriented, to stimulate renewed interest in the entire Merit family. Media vehicles include magazines, outdoor, newspapers, and supplements.

The Ultima media and advertising campaign is designed to create awareness and generate excitement about new Ultima as well as the Merit brand name.

DATES: 3/92 through 6/92

DIRECT MAIL: A direct marketing program will begin in mid-March. This program will include:

- A mailing to smokers of Carlton, Now, and competitive full margin ultra low tar brands. The mailing will deliver two free packs of Ultima as well as a continuity offer for additional incentives.
- A media offer challenging Carlton and Now smokers to try Ultima. In order to generate additional Carlton and Now names, we will offer two free packs for one Carlton or Now pack proof of purchase. The continuity offer will also be included in this program.

In order to prevent "sticker shock" at the end of the off-label price promotion, a carton stuffer coupon offer will be inserted in all cartons produced during the last weeks of the introduction. Consumers will be offered $2.50 off-carton coupons by mail with an Ultima carton proof of purchase.

92043642

21

THE FOLLOWING INFORMATION WAS GIVEN TO

PHOTOGRAPHED

TO THE N.Y.O. ON

PAGE .007

Figure 4.2: Document Example 2: Tilted Text

Figure 4.3: Document Example 3: Illegible Text

cussion, the question becomes how can inserted pages of handwritten, unorganized, partially-illegible notes be recorded systematically in the archive.

The fifth example page, Figure 4.5, is a single-page document from R. J. Reynolds produced in June of 1970 (Bates Number 503257271). Although G. S. Caluman has nice penmanship and format, this document certainly bypasses the capability of OCR programs. In addition it raises coding issues with features such as the illegible return address, the secondary writing at the top which is lined out and illegible, and inserts like the one on the fifth line of the first paragraph.

The sixth example page, Figure 4.6, is the sixth page of a twelve-page introductory speech for the annual meeting of the Council for Tobacco Research in 1970 (Bates Number HK1871057). The entire document is a handwritten outline.

The seventh example page, Figure 4.7, is the first of a two-page document from Philip Morris produced in December of 1991 (Bates Number 2044938392a). Overall the document is well formatted and legible. However, it has a large amount of marginalia that brings to light some of the sub-categories of secondary text. For example, there are initials indicating viewing or approval, comments about the document content, comments about extra-document events, questions related to the document content, questions directed to individuals, editing lineouts, editing inserts, underlining, and even a short list of trade show gifts. This gives some idea that a single marker for secondary text (marginalia) is not sufficient to record all types.

The eighth example page, Figure 4.8, is from what appears to be a journal article about smoking cocoa leaves as an alternative to tobacco. This is the second of a two-page document from the Council for Tobacco Research from August of 1979 (Bates Number 11277820). This was included as an example of text found within an image, which is another coding issue that must be addressed. In this case there is an advertisement for porous plug wrap that must be addressed. The data are part of the document, but clearly not rhetorically relevant

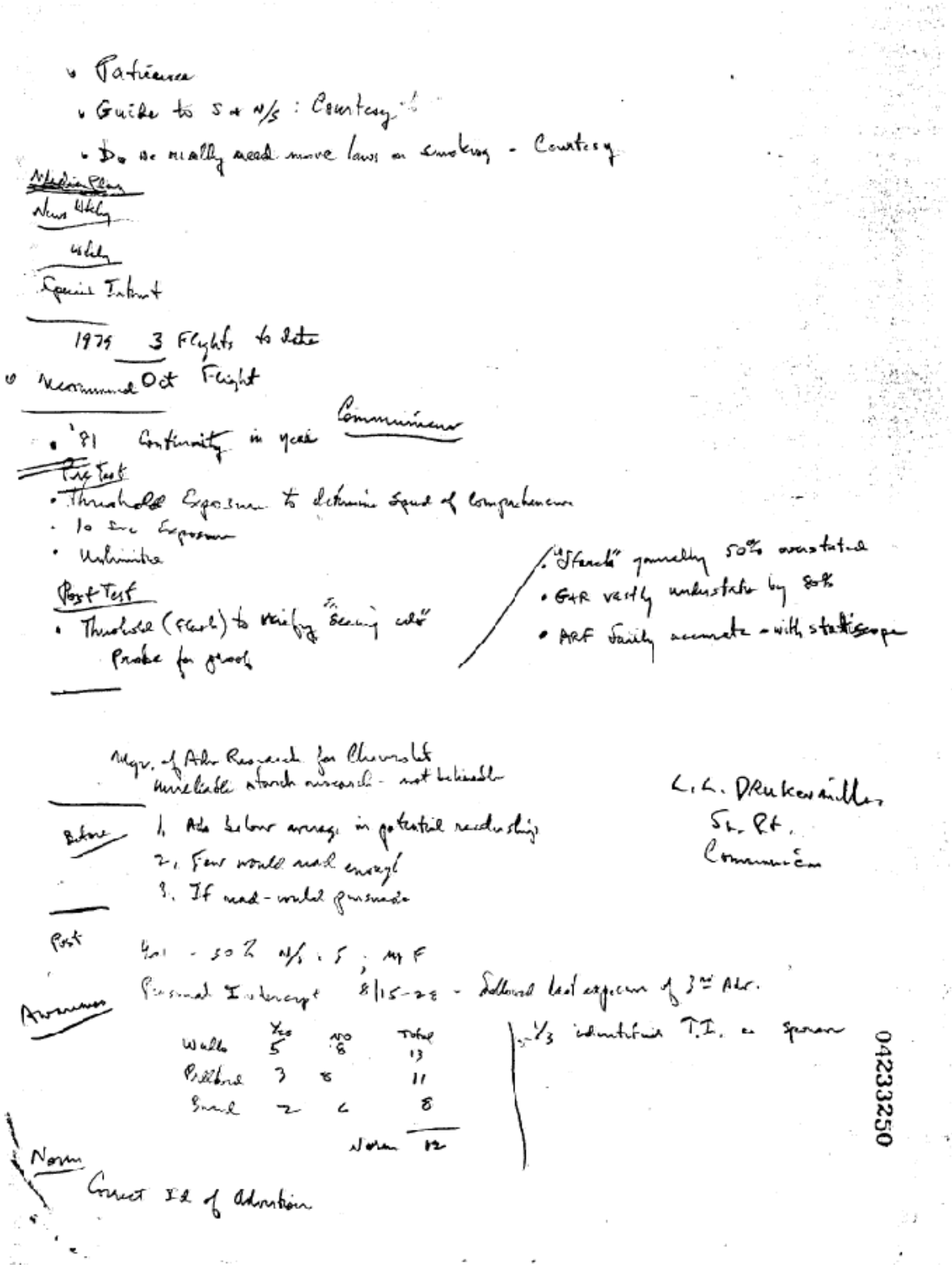
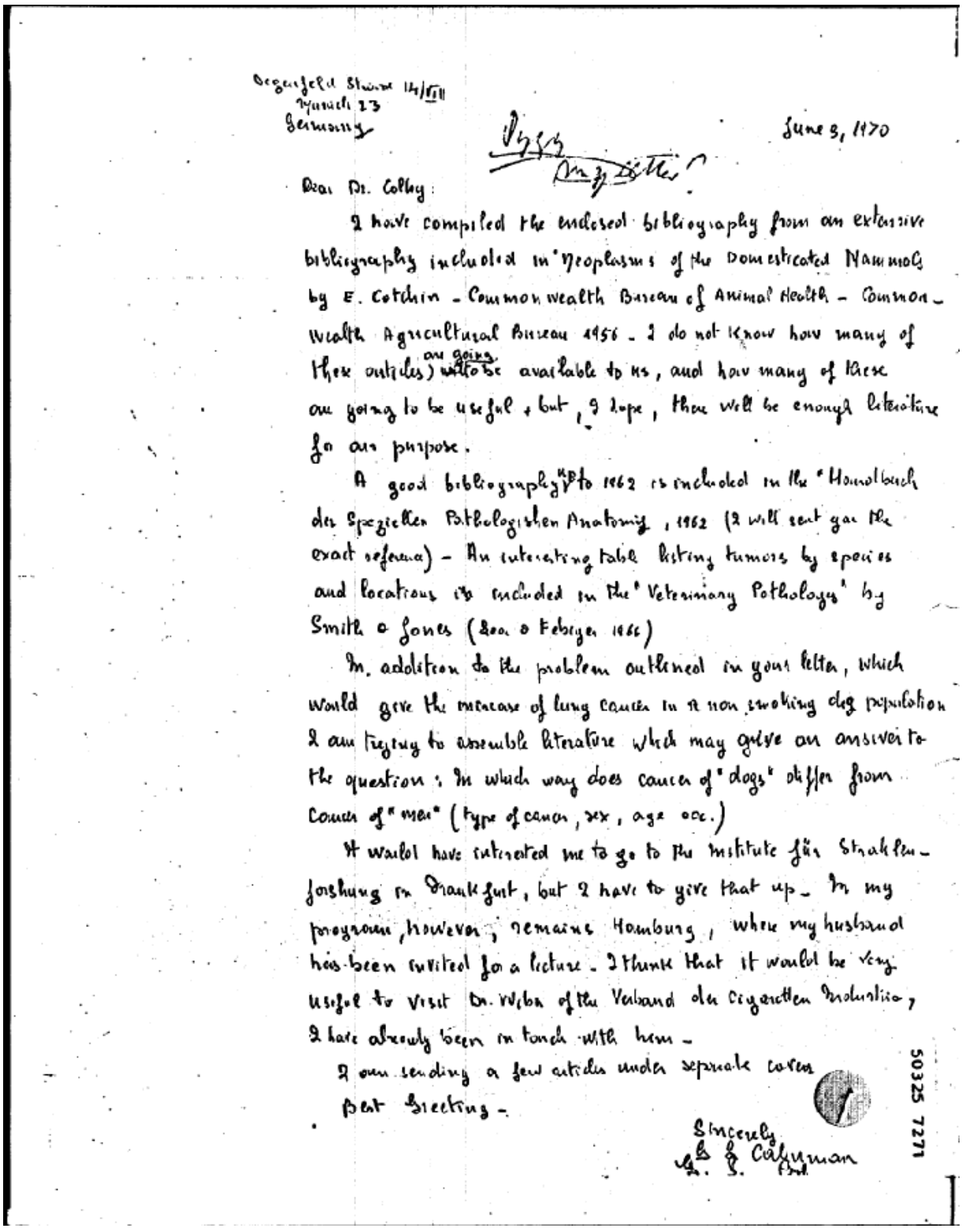


Figure 4.4: Document Example 4: Handwritten Notes



Degenfeld Straße 14/III
Munich 23
Germany

June 3, 1970

Dear Dr. Colby:

I have compiled the enclosed bibliography from an extensive bibliography included in "Neoplasms of the Domesticated Mammals by E. Cotchin - Commonwealth Bureau of Animal Health - Commonwealth Agricultural Bureau 1956 - I do not know how many of these articles) are going will be available to us, and how many of these are going to be useful + but, I hope, there will be enough literature for our purpose.

A good bibliography up to 1962 is included in the "Handbuch der Speziellen Pathologischen Anatomie, 1962 (I will sent you the exact reference) - An interesting table listing tumors by species and locations is included in the "Veterinary Pathology" by Smith a Jones (Jour. of Pathology 1956)

In addition to the problem outlined in your letter, which would give the increase of lung cancer in a non smoking dog population I am trying to assemble literature which may give an answer to the question: In which way does cancer of "dogs" differ from cancer of "man" (type of cancer, sex, age acc.)

It would have interested me to go to the Institute für Strahlenforschung in Frankfurt, but I have to give that up - In my programm, however, remains Hamburg, where my husband has been invited for a lecture - I think that it would be very useful to visit Dr. Weber of the Verband der Cigarettenindustrie, I have already been in touch with him -

I am sending a few articles under separate cover

Best Greeting -

Sincerely
G. S. Calhoun

50325 7271

Figure 4.5: Document Example 5: Handwritten Letter

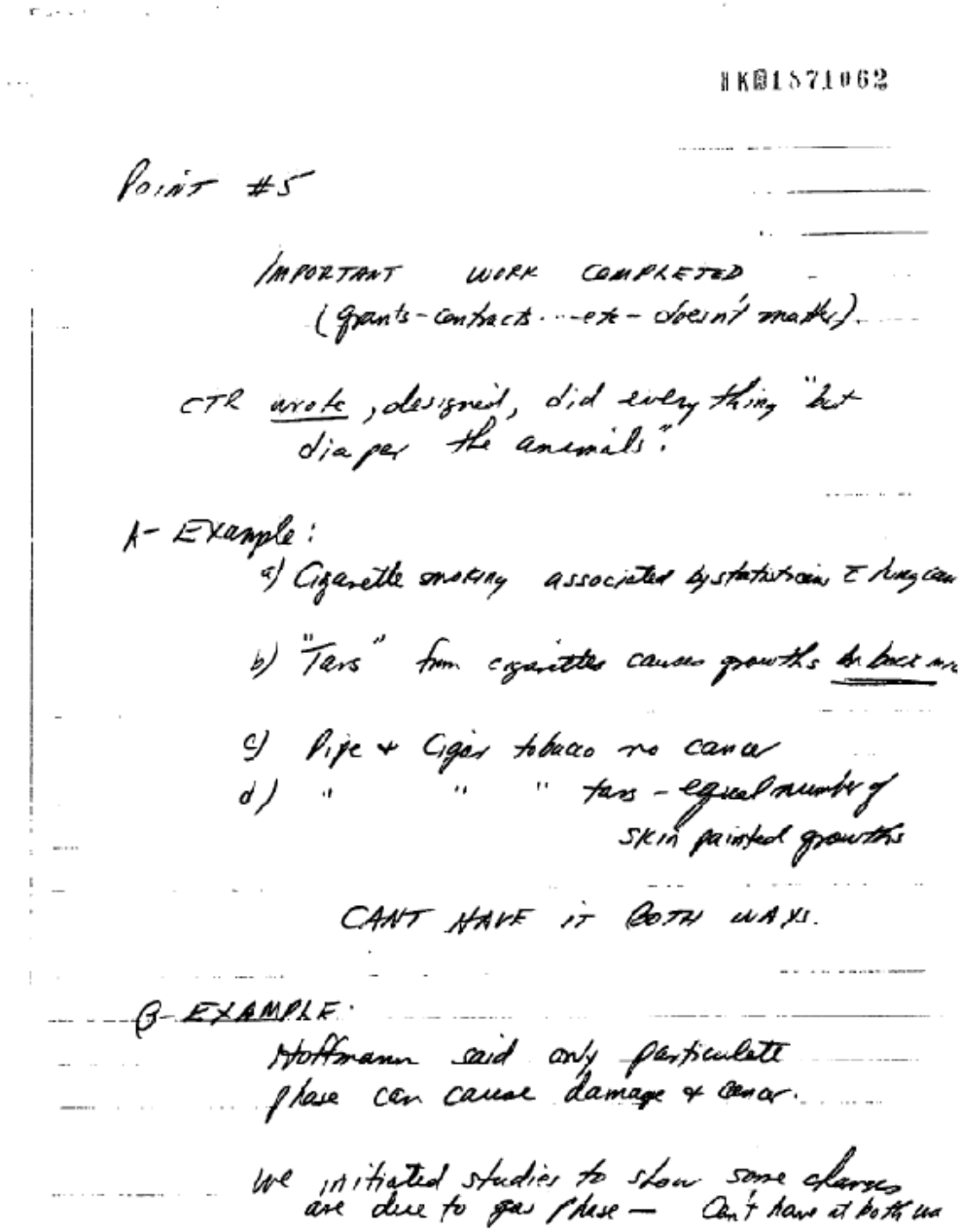


Figure 4.6: Document Example 6: Handwritten Document

12/13 2044938392
ADRIAN -
CAN/SHOULD I BE INVOLVED?
PLEASE ADVISE.

PHILIP MORRIS U.S.A. INTER-OFFICE CORRESPONDENCE JANE
120 PARK AVENUE, NEW YORK, N.Y. 10017

TO: Distribution Trade Gift CIP Event Mail DATE: December 11, 1991

FROM: Pam Gill PG

SUBJECT: MARLBORO MEDIUM 100'S LAUNCH PROGRAMS

Last week we held a meeting on Marlboro Medium 100's to generate additional consumer promotion ideas and to reaffirm our launch objectives. The following recaps the ideas that were discusses and which should be considered when developing the promotion plan recommendation.

Trade Gift -
- Branding iron
- Coffee pot w/ cups.
- Bell Bottle / Person wt.

OBJECTIVES
- Generate "quality" trial and conversion
- Build volume
- Help create awareness of Medium 100's/Added visibility

↳ main target is → ?

STRATEGY
- Field a unique bonus product retail trial offer behind Medium 100's and all Marlboro 100 MM non-menthol packings.
- Use multiple media offers against quality trial/conversion and volume objectives.
  - Request sampling against women 100's smokers
  - Trial offer against young adult 100's smokers
  - Trial offer behind Medium packings coupled with a Marlboro family volume building offer
- Deliver direct mail trial and continuity offers to key competitive 100's smokers.

ELEMENTS UNDER CONSIDERATION
- Execute B3G3F flip top box display promotion in Supermarkets. Offer on all Marlboro 100 MM non-menthol box product (Medium 100's, ~~Gold and Lights 100's~~). Promotion will provide support for Marlboro's Gold and Lights 100's and insulate these businesses from trial on the new packing. Aug 1 - start date

  Buy 3 G3F, D to buy 4 G2F

  QUESTION: Regarding trade portion of the promotion, what display sizes and payments are appropriate and do we need a dealer loader?
- Schedule "Denim Shirt" Free Pack ad in books that reach young adult 100's smokers. Free pack coupon will be good on Medium 100's and Kings. → Special Trial Offer ← competitive smoker - Marlboro - B3G2F; buy get free 5 pack offer
- Schedule "Campfire Coffee" request sampling ad in select women's books. Offer will be designed against quality trial and follow-up mailing will include continuity offer against conversion objectives. BRC

2044938392 A

Figure 4.7: Document Example 7: Marginalia

to the rest of the document. As well, the document is tilted four degrees and has a large *f* in the top left corner (from the previous page).

The ninth example page, Figure 4.9, was included as an example of text from within a form. In this case, it is the notary page in an articles of incorporation document from R. J. Reynolds produced in October of 1988 (Bates Number 508074644). Also included is a *BEST COPY* stamp in the margin. Stamps are a very common form of secondary text that must be recorded.

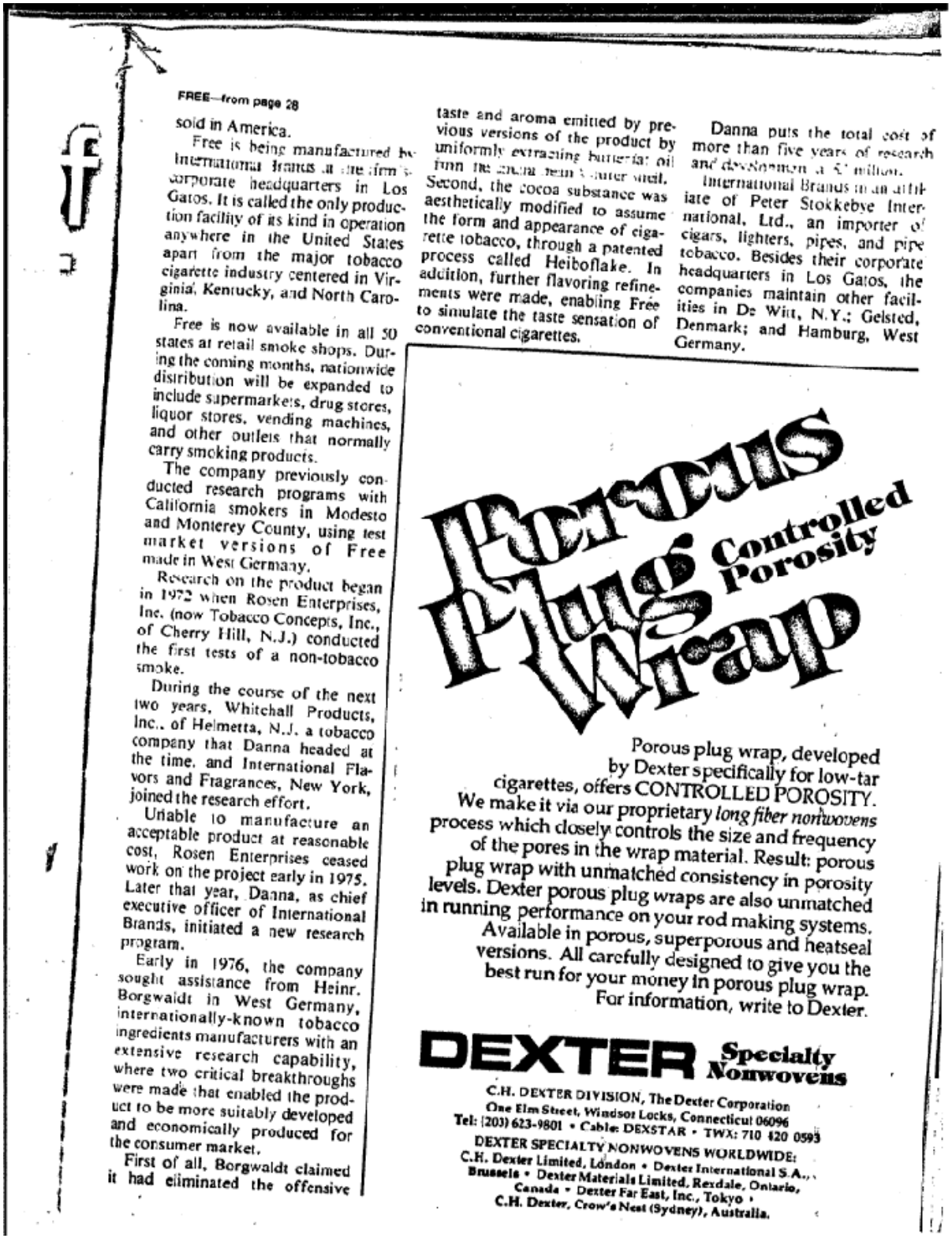
Finally, the tenth example page, Figure 4.10, is an example of text from within a table. This is the ninth of eleven pages in a marketing evaluation report from Philip Morris dated July of 1979 (Bates Number 2040260786). Other notable features are the ninety-degree rotation, the columnated numerical data, and the *TABLE 5* block at the bottom right.

In summary, the above examples illustrate the necessity of both manual conversion and a complex markup set as means to reliably and accurately preserve the document text and structure during archiving.

4.3 OCR Testing

As a means of verifying our assumption that the poor resolution of document images would significantly effect the quality of the final corpus, we conducted a limited experiment with OCR conversion by comparing the text data from manually converted documents to the text data produced from OCR processing of the same documents.

Having established the Quota Sample, twenty documents were randomly selected from across the four Sets (all 808 documents). These are shown in Table 4.1. Each of the sampled documents was converted into text and archived according to the standard TDC methods to include markup (described in the sections below). Following this, the analyzable text was extracted from the archive using a very general stylesheet that retained all document information apart from notes and descriptions (i.e data added by the coders). Data from all other markup tags, such as pretext, posttext, appendix, pre and postdoc, et cetera, were



FREE—from page 28

sold in America.

Free is being manufactured by International Brands at the firm's corporate headquarters in Los Gatos. It is called the only production facility of its kind in operation anywhere in the United States apart from the major tobacco cigarette industry centered in Virginia, Kentucky, and North Carolina.

Free is now available in all 50 states at retail smoke shops. During the coming months, nationwide distribution will be expanded to include supermarkets, drug stores, liquor stores, vending machines, and other outlets that normally carry smoking products.

The company previously conducted research programs with California smokers in Modesto and Monterey County, using test market versions of Free made in West Germany.

Research on the product began in 1972 when Rosen Enterprises, Inc. (now Tobacco Concepts, Inc., of Cherry Hill, N.J.) conducted the first tests of a non-tobacco smoke.

During the course of the next two years, Whitehall Products, Inc., of Helmetta, N.J. a tobacco company that Danna headed at the time, and International Flavors and Fragrances, New York, joined the research effort.

Unable to manufacture an acceptable product at reasonable cost, Rosen Enterprises ceased work on the project early in 1975. Later that year, Danna, as chief executive officer of International Brands, initiated a new research program.

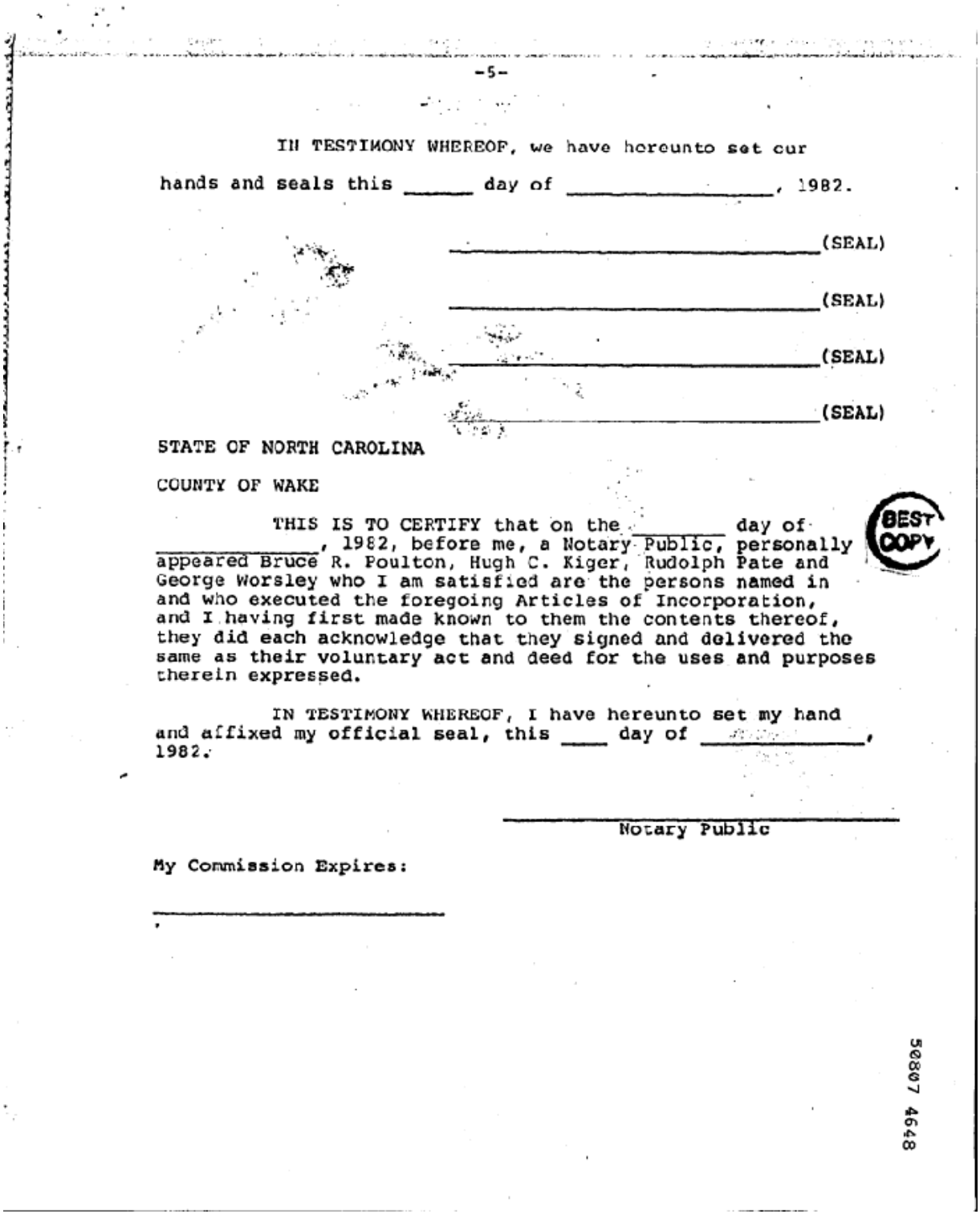
Early in 1976, the company sought assistance from Heinr. Borgwaldt in West Germany, internationally-known tobacco ingredients manufacturers with an extensive research capability, where two critical breakthroughs were made that enabled the product to be more suitably developed and economically produced for the consumer market.

First of all, Borgwaldt claimed it had eliminated the offensive taste and aroma emitted by previous versions of the product by uniformly extracting butteric oil from the cocoa bean's outer shell. Second, the cocoa substance was aesthetically modified to assume the form and appearance of cigarette tobacco, through a patented process called Heiboflake. In addition, further flavoring refinements were made, enabling Free to simulate the taste sensation of conventional cigarettes.

Danna puts the total cost of more than five years of research and development at $7 million.

International Brands is an affiliate of Peter Stokkebye International, Ltd., an importer of cigars, lighters, pipes, and pipe tobacco. Besides their corporate headquarters in Los Gatos, the companies maintain other facilities in De Witt, N.Y.; Gelsted, Denmark; and Hamburg, West Germany.

Figure 4.8: Document Example 8: Image Text



-5-

IN TESTIMONY WHEREOF, we have hereunto set our hands and seals this _____ day of _____, 1982.

_____(SEAL)

_____(SEAL)

_____(SEAL)

_____(SEAL)

STATE OF NORTH CAROLINA

COUNTY OF WAKE

THIS IS TO CERTIFY that on the _____ day of _____, 1982, before me, a Notary Public, personally appeared Bruce R. Poulton, Hugh C. Kiger, Rudolph Pate and George Worsley who I am satisfied are the persons named in and who executed the foregoing Articles of Incorporation, and I having first made known to them the contents thereof, they did each acknowledge that they signed and delivered the same as their voluntary act and deed for the uses and purposes therein expressed.

BEST COPY

IN TESTIMONY WHEREOF, I have hereunto set my hand and affixed my official seal, this ____ day of _____, 1982.

_____
Notary Public

My Commission Expires:

_____

50807 4648

Figure 4.9: Document Example 9: Form Text

NAME SELECTED AS FIRST OR SECOND MOST APPROPRIATE FOR CONCEPT

	CAMBRIDGE			SUMMIT			BELMONT			MAYFIELD			SANO		
	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female
N =	277	134	143	277	134	143	277	134	143	277	134	143	277	134	143
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
This is an ultra low-tar product for those who prefer the milder taste of an ultra low tar cigarette															
First Choice	16	18	15	27	28	27	13	10	17	23	22	24	21	23	18
Second Choice	16	19	14	25	23	27	21	22	21	24	23	24	13	13	13
First or Second Choice	32	37	29	52	51	54	34	32	38	47	45	48	34	36	31
This is an ultra low-tar cigarette that will offer the lowest tar delivery of any product on the market															
First Choice	15	13	16	23	26	20	16	10	20	19	17	20	28	33	23
Second Choice	18	19	16	27	27	28	18	16	20	22	21	22	15	16	13
First or Second Choice	33	32	32	50	53	48	34	26	40	41	38	42	43	49	36

Question: This (SHOW CONCEPT CARD P) is a description of one of the cigarettes we have been talking about. Which of the five brands do you think it is?

What other brand do you think it is?

This (SHOW CONCEPT CARD Q) is another description of the cigarettes we have been talking about. It could be the same name you mentioned before or a different one. Which of the five brands do you think it is?

TABLE 5

2040260794

Figure 4.10: Document Example 10: Table Text

kept. This became the TYPED text data for the comparison. A copy of the actual stylesheet used for this transformation is included in Appendix C.

Table 4.1: OCR Sample Documents.

Set	Bates Range	Year	Pages	Notes
Q1	501869639-9657	Bliley	19	
Q1	511466579-6579	1990	1	
Q1	516927839-7839	1990	1	
Q1	659032972-2973	1980	2	no usable OCR text
Q1	ATX030025431-5431	1980	1	OCR Error
Q1	ATX040827617-7619	1970	3	OCR Error
Q1	CTRSP-FILES023888-3888	1980	1	
Q2	03732265-2270	1970	6	
Q2	2045083228-3228	1980	1	
Q2	504337601-7603	1980	3	
Q2	512180224-0226	1980	3	
Q2	88208744-8770	1980	27	
Q2	89301401-1424	1970	4	
Q3	501772729-2741	1960	13	
Q3	ATX040895180-5180	1980	1	OCR Error
Q4	01148569-8569	1960	1	
Q4	11320767-0767	1980	1	no usable OCR text
Q4	2040940418-0418	1980	1	
Q4	680258078-8078	1960	1	
Q4	680279579-9581	1960	3	no usable OCR text

At the time of the above sampling and manual conversion, the TDO website had implemented an online OCR capability that produced analyzable text from the TIFF images released by the tobacco industry. There was considerable fanfare associated with this event because of the additional search capabilities it added to the TDO archive. According to the website's 'Search Help' page (at the time), the OCR process produced text from the document images that 'is often slightly incorrect, but close.' Given their confidence in the OCR process being used, we decided this would be a reasonable source for our OCR data. All of the twenty sample documents from Table 4.1 were located on the TDO website, and if available the OCR-produced text was collected. This became the OCR text data for the comparison.

All twenty documents from the sample were located on the TDO website; however, for three documents the website OCR function repeatedly produced no result, but instead returned an error message. Based on other results, it was assumed that the error was external to the actual OCR process. That is, we believed it was not that the OCR function could detect no text, but that for some reason the image could not be loaded for processing. These documents were removed from the sample, leaving seventeen documents for the comparison. For the remaining documents the OCR function did return data, although for three documents that clearly contained text data no text was returned, only a series of non-alphanumeric characters. In these cases, the documents were left in the sample because we attributed the problem directly to the OCR function (i.e. the document layout and poor image resolution prevented the recovery of the text).

The comparison was done by tokenizing the two data sets and performing token and type counts. For tokenizing, the data sets were converted to lower case, stripped of all non-alpha characters other than intra-word apostrophes (i.e. replaced by ASCII 32), and then divided into tokens by any sequence of whitespace characters. For example 'The cat's 1 hat.' would become `['the', ''cat's'', 'hat']`.[2]

For the TYPED text data, the seventeen files contained a total of 15,492 tokens comprised of 3,217 different types, which is a type-token ratio of 1 to 4.82 and an average of 911 tokens per document. The OCR text data contained a total of 23,372 tokens of 5,823 types, which is a type-token ratio 1 to 4.01 and an average of 1,375 tokens per document. This increase in both type and token counts for the OCR text data is made more significant in that mathematically we should expect just the opposite given that there were actually only fourteen documents with text compared to the seventeen of the TYPED sample. Yet instead of the expected fewer, we found roughly 51 percent more tokens and 81 percent more types.

In reviewing the results I found generally two reasons for the increased counts in the OCR text data. The first is that there is not an easy way to distinguish rhetorically significant

[2]Refer to Chapter 5 for a more detailed discussion of tokens and tokenizing.

text from the non-rhetorical types (i.e. tabular data, forms, et cetera) except by manual evaluation. Given this problem, even though most tabular data (non-alpha characters) were removed during tokenization, the OCR text data did contain some text from non-rhetorical sections of the document, while the TYPED text data did not, which would be expected to cause an increase in counts for those documents.

The second reason for the increase, and more important in terms of counts, was that there were a great many inaccurate 'word' renderings in the OCR text data. These errors were generally of three types, concatenations of multiple words (*accompanyinglstatement*), disjunctions of a single word (*cer*, *ing*, *tion*), and misspellings (*acqluisition*, *bolivla*, *ccncer*). Each of these error types added to the type count by generating unknown word types, and consequently increased the token count as well. In terms of numbers, of the 5,823 types in the OCR text data, 54 percent (3,146 types, see Appendix C) were not found in the TYPED text data. Even more, we later determined that 40 percent (2,351) could not found in the entire typed Quota Sample. In contrast, the TYPED text data contained virtually no rendering errors given that each file was spell checked and verified against the original document.

From the above data we estimated that at least 40 percent of the word types in the OCR text data were errors of some sort. Given this estimate we determined that while it is true that the addition of the OCR function to the TDO website did provide an increased search capability, the text rendering was clearly not accurate enough for our intended analysis. In fact, given such high error rates one must be wary of even using the OCR text data for searching because (although better than nothing) they are far from reliable. Of the 3,217 word types in the TYPED text data, roughly 17 percent (540 types, see Appendix C) could not be found in the OCR text data. Thus what was clearly represented in the document, words such as *carcinogenic*, *deficiency*, and *lungs*, could not be located.

4.4 Document Structure

Before moving to a description of the document conversion process, the actual keyboarding and associated decision making, one must first be familiar with the general structure and purpose of the TDC document archive. This is necessary given that the primary task in the process is not the manual labor of typing, but instead the cognitive task of making decisions related to the structure of the original document: what its rhetorical purpose is, and how it should be recorded. In other words, the archivist's most important and difficult task is to decide what the parts of the document are, and be able to properly organize and mark them for easy retrieval, which is not at all a simple task given the document examples in Section 4.2. The typing itself is secondary.

It was decided early in the project that the TDC would be archived using Extensible Markup Language, which is commonly known as XML. For those unfamiliar with XML (otherwise skip ahead four paragraphs), it is not actually a computer language, although this is implied by its name. Rather, it is a standardized syntax (style/format) for marking portions of a text in order to overtly name the items and establish hierarchical relationships to other marked portions of text, for whatever reason (i.e. according to the needs of the user). Although a true introduction to XML is beyond the capability of this work, a short example can go a long way toward understanding the remainder of this chapter and Chapter 5. For more in-depth information, a good place to begin is the W3Schools website (http://www.w3schools.com). In the meantime, consider the following short excerpt from an XML document:

```
<memorandum>
    <subject>Learning XML</subject>
    <author>John Doe</author>
    <addressee>
        <person>Jane Smith</person>
        <person>Jim Smith</person></addressee>
    <text>All personnel, without exception, <emph>must</emph>
          learn XML by Friday.  You will be given a test.</text>
</memorandum>
```

In XML, the markers that identify the document hierarchy and text events are commonly referred to as *tags* and are denoted by a tag name enclosed in angle brackets. The tags are used in pairs (the ending tag having a front slash at the beginning) which together enclose all data that belong to the given tag, including other tag pairs. For example, the `<memorandum> </memorandum>` tag pair above (called a parent tag) is the highest level tag. Structurally, it contains four parts, the child tags `<subject>`, `<author>`, `<addressee>` and `<text>`. The `<addressee>` tag is the parent to the two `<person>` tags, which are its child tags. Finally, the word 'must' within the `<text>` tag has been marked to indicate that it is an instance of *emph*, however that is defined. The result of this format for markup is that all of the text and peripheral data are named and established in a document hierarchy, such that the implicit conventions and assumptions found in documents become explicit objects which are retrievable by computer. This also resolves problems associated with complex document structures. All data found in a document can be retained as long as they are incorporated into a tag hierarchy that properly identifies their relationship to the other data. In fact, even non-document data (metadata), such as the archivist's classification data and notes, can be included if they are properly marked.

Another notable aspect of XML is that the names of the tags used and the hierarchical relationship between them are not fixed. For a given project, such as the TDC, the project authors are free to devise tag names, hierarchies, and definitions according to the needs of the project. As well, each tag can have a series of 'attributes' assigned to it. These name-value pairs are placed inside the angle bracket of the opening tag, as in `<memorandum class="mandatory" recipients="2">`, and provide the opportunity for recording additional specifications or data. Of course, because tags and attributes can be equated with categories, it is necessary to define them. In terms of structure, which is what can be validated by computer, the set of allowable tags (with attributes) and the hierarchical relationship between them are formally defined in a document called a Document-Type Definition or DTD, which like XML has a very specific syntax. On the other hand, because the

computer is less able to validate non-structural usage, XML has no requirement that tags be defined in relation to meaning. This is left to the project authors. For the TDC, both the tag hierarchy and tag meanings are defined in Section 4.4 below.

Finally, although using XML solves many problems related to document complexity, it actually creates problems in relation to corpus complexity. The corpus archives suddenly contain much more data than needed for any single study. However, because of the regular syntax of XML, and because of the explicit tag definitions provided by the DTD, these complexities can easily be managed by computers. That is, programs can be used to extract the desired data from the corpus and reformat them into an easily interpretable form. In fact, because XML has become the standard means of archiving in many fields, there are many pre-existing software applications for manipulating XML, some of which will be discussed in Section 4.5.4 below, and then later in Chapter 5.

Returning now to the TDC, the formal structure of the archived documents is governed by the XML tag set used during the coding of documents. This in turn is defined by the governing Document-Type Definition (DTD). If the goal is to produce document archives that have both regular structure and accurately classified data (or to understand the process description), being familiar with both the tag hierarchy permitted by the governing DTD and the intended use of each tag becomes a necessity. As discussed in Chapter 2, without knowing both well there can be no consistency between archives. This would render the XML useless as its only purpose is to allow the secondary user to reliably extract particular data types in an automated manner. Both issues will be addressed in this section.

4.4.1 XML Tag Selection

Of course, prior to any discussion of structure and purpose, the question of *which* XML tag set (or DTD) must be answered. Those familiar with XML will know that there are a number of established tag sets for archiving text documents. The most notable XML-based sets are the tags from the Text Encoding Initiative or TEI (Burnard 1995), a very large

and complex set, and XML-based Hypertext Markup Language (XHTML), which is much more simple. When we investigated these types of existing tag sets, we found that they were well suited for their intended use, which is primarily electronic archiving of well-formatted texts. With the TEI set the focus is on recording semantically-important meta elements. This is done both at the document level with tags for marking divisions such as sections, chapters, and paragraphs, and at the sub-paragraph level with tags for marking semantically significant elements such as titles and emphasis (which are often denoted by typesetting in conventional texts). As with TEI, XHTML and similar sets work at both the document and sub-paragraph level. However, the focus of XHTML is on specifying how elements should be displayed in web pages, particularly in relation to font style and spatial arrangement (i.e. typesetting specifications).

Although this is exactly what one would expect given the underlying motive of conservation, these design purposes are not parallel to those of the TDC Project, and consequently the tag sets do not lend themselves easily to our work. What makes the transfer of the tags difficult is the fact that the TDC Project is not motivated by the historical preservation and display of texts, but by rhetorical analysis. In other words, the primary interest is not in recording the structural layout and typesetting used in the original document so that it can can be displayed properly in an electronic environment, but in recording those portions of the document which are rhetorically significant or distinct.

In terms of structure, the main concern is locating the rhetorical center of the document, and subsequently recording the text data from the given document in relationship to that center, such that the rhetorically significant data can be extracted for analysis in an automated fashion. There is no desire to use the archive to display the document electronically in a format similar to the original document. Consequently, the structural value of knowing in which section, chapter, or paragraph a given passage of text appears is greatly reduced given our purposes, and it need not be strictly preserved. While at the same time, knowing a portion of text comes from a title, image, extended quote, or marginal note has high structural

value because it adds significantly to the rhetorical interpretation (wherever it happens to occur in the document). Thus tags denoting canonical document structure become much less desirable in favor of tags that record the target structures. Of course, delineating structure is made all the more complicated by the nature of the documents themselves, which in many case are actually a compilation of several documents (see below).

In terms of display the situation is similar. Conventional display tags denote the style of the text, yet our interest is much less in how the text appears on the page, and much more in what the appearance indicates. For example, various typesetting styles such as italics, boldface, underlining, quotation marks, and all capitals can be found in the Quota and Supplemental Samples denoting text emphasis. Emphasis, of course, is a rhetorically significant event and needs to be preserved. However, these same typesetting styles have also been found to denote titles, headings, names, quotations, formulas, and even standard text, all of which may have little value for rhetorical analysis. Thus, marking text in a document with a tag designed to denote typesetting, such as `<italics>`, is ambiguous when the corpus is to be analyzed rhetorically. There is no way to determine the significance of the text apart from manual re-analysis (and we should not forget that the purpose of markup is to avoid re-analysis).

All of the above discussion is not provided to say that tag sets like those from TEI could not be used, but that they are not easy to use outside the context for which they were designed. TEI certainly provides a very large, flexible, and configurable tag set for archiving and linguistic analysis of documents. However, in practice the tags are difficult to use because their flexibility comes from the addition of a complex set of tag attributes, and complexity leads to error in keyboarding. This would be especially true if the TEI tags from both the general guidelines and the linguistic guidelines were used simultaneously.

The final nail in the coffin for the idea of using a pre-existing tag set is the fact the converting (often called transforming) an XML archive from one tag set to another is a relatively simple process. By design, XML structures the archive using a fixed hierarchy and

syntax. Because of this, changing tag sets is simply a matter of replacing one tag-attribute combination with another. It can be accomplished easily using a program specifically designed for XML transformations (such as XSLT), regular expressions, or in some cases simple search and replace. Thus, if the need arose for the TDC archive to conform to the TEI standards, it would not be difficult to convert it.

Given the additional complications that would be inherited by using a predefined tag set, and knowing that it was a relatively simple matter to make a transformation to another tag set, the decision was made early in the project to devise a set of XML tags specifically for our purposes. Our goal was to make the tag set as straight-forward and simple as possible in order to reduce error during document encoding and the subsequent extraction of data from the archive. More specifically, we believed that archiving and retrieval of the text data would be aided by a tag set that 1) contained the minimum number of tags and tag attributes necessary to archive the data types of interest, 2) used clear (transparent) tag names and attributes with specific definitions, and 3) focused on rhetorical structure and features (analyzable text) rather than the conventions of typesetting.

The initial development a tag set for the TDC Project was done by creating a model XML document based on the overall analysis goals of the project in combination with the experience gained through study of the Core Sample documents. We began with a set of 44 tags focused on the types of data we hoped to analyze. After a series of experiments making test archives from Core Sample documents and discussing the results with the project's Principal Investigators, an acceptable model document containing 46 tags was constructed. This was then formalized through the creation of the TDC Project DTD.

Following the creation of the DTD, only two major revisions were made. The first revision came when we began to archive the Rhetorical Sample documents. What we found is that there were additional attributes needed to accommodate the extra classifications found in the sample. To prevent having to modify the existing archives for the Quota and Supplemental Samples, and to allow the use of the same basic DTD, the needed attributes were added into

what was then the root tag (`<ugat:document>`) and defined to be optional. These attributes are included in the descriptions to follow.

The second major modification came when we began to contemplate making the TDC Project data deliverable via the web. At this point we discovered the need to create a single XML archive that contained the data from multiple documents. To make this possible the parent tag `<tobaccodocs>` was added to the DTD. Subsequently, all archive files were modified using an automated process to make `<tobaccodocs>` the top-level tag. At the same time, the *ugat* namespace was removed from the `<document>` tag and the `<!DOCTYPE>` declaration was added to specify the DTD for validation.

The final result was a very functional set of 46 simple tags with a minimum of attributes. This set was used to archive both the Quota and Supplemental Samples, and all other UGA Tobacco Documents Project XML archives. Its simplicity allowed rapid training of archivists, ease of conversion, and provided a very low error rate (a 98 percent first-pass agreement between archivists on critical tags, see Section 4.5.5 for more details). This tag set will be described in detail in the following two sections, beginning with an overview of the tag set's hierarchical structure, and then continuing with a more detailed description of each tag and its attributes.

4.4.2 XML Tag Hierarchy

Although there are some who can easily read a Document Type Definition (DTD) and visualize the permissible structures, I have to admit that I am not one of them, and my suspicion is that most readers of this work will have a similar inability. For those who can read DTDs, or those who need a model as a basis for creating their own, the full DTD used for the TCD, with usage notes, is included in Appendix D.1. For the rest of us, I will describe the overall document structure permitted by the DTD in parts using a simple indented structure to represent the tag hierarchy. Also for simplicity, the tag attributes are not included here, and closing tags have been left out. As a reminder, XML tags come

in pairs, a start tag and an end tag of the same type. Data and/or 'children' start-end tag pairs are contained between the start and end tags of the 'parent' tag. For example: `<parent> some text data <child> more text data </child> </parent>`.

In total, there are 46 tags permitted by the DTD. At the highest level of a TDC archive document is the `<tobaccodocs>` tag which is the parent to all other tags. Its purpose is to permit multiple document archives to be grouped together in a single archive. This is a purely technical construct and makes little difference to the archivist. The real parent tag of the archive is the `<document>` tag, and this has only two 'children' tags: `<metadata>` and `<docdata>`. The `<metadata>` tag contains all data that are associated with the original document, but not specifically part of the original, such as Bates-Number ranges, the number of pages, the archivist, et cetera. The `<docdata>` tag contains the actual text data from the original document or in limited cases a description when the data cannot be easily represented as text. Putting all the above together, at this point we have a structure as such:

```
<tobaccodocs>
    <document>
        <metadata>
            <bates_start>
            <bates_end>
            <uga>
            <external>
        <docdata>
            <note>
            <predoc>
            <maindoc>
            <postdoc>
            <xdoc>
```

In terms of metadata, space is reserved specifically for the start and end Bates Numbers, and items of particular interest to the Tobacco Documents Project (the `<uga>` tag). There is also a tag to house any undefined metadata collected from other tobacco documents archives (the `<external>` tag). The `<uga>` tag is the parent to eight tags which contain only simple text data that describe basic features of the original document and who created and/or validated the document archive. It generates the following structure:

```
<uga>
    <note>
    <date>
    <pages>
    <words>
    <section>
    <encoded_by>
    <verified_by>
    <image_file>
```

A regular feature of tobacco industry documents which was not illustrated in Section 4.2 is the fact that a single range of Bates Numbers, which is considered in the repositories to be a 'document' unto itself, may actually contain several structurally and rhetorically separate parts (true documents), such as the draft of an advertisement with a cover letter of editing instructions attached to the front and several pages of notes and comments in the rear, each authored by a separate person. To account for this, the children tags of the `<docdata>` tag (which has all true document data) are basically four: `<predoc>`, `<maindoc>`, `<postdoc>`, and `<xdoc>`. The additional `<note>` tag shows up in a lot of places to allow the archivist to describe any odd constructs, so it will not be addressed each time. In the case that a range of Bates Numbers (pages) for a given 'document' actually contained more than one structurally or rhetorically distinct document, the archivist would select the single document from within the range that was believed to the be the primary focus for the whole set. These data would be archived within the `<maindoc>` tag, with only one `<maindoc>` tag allowed per archive. Any document from the given Bates-Number range that preceded the `<maindoc>` pages was archived within a `<predoc>` tag, and any document from the Bates-Number range that followed the `<maindoc>` pages was archived within a `<postdoc>` tag. The `<xdoc>` tag was included as an extra position for unclassified attachments to the main document but was rarely used. Multiple `<predoc>` and `<postdoc>` tags were allowed as needed. In this way, the `<maindoc>` tag is always present and considered the focus of the archive. The others often are not present, but in some cases have multiple instances.

Summarizing the structure thus far, an archive has two parts, one containing extra-document data and calculations (<metadata>), and one containing the document data, or descriptions of that data (<docdata>). Within the document data portion of the archive, the focus is on one or more rhetorically and structurally coherent sets of pages from the given range of pages which represent the general idea of 'document' in a much truer form. These sets of pages are archived in the tags <predoc>, <maindoc>, or <postdoc> as necessary.

The next step is to account for the general form of a canonical document. For archiving purposes, we determined that four distinctions would account for the rhetorical form of most tobacco documents. In general, documents have 1) data such as titles, headers, distribution lists, and addresses that precede the main body of text, 2) the main body of text (for analysis), 3) data such as salutations, signature blocks, and legal notices that follow the main body of text, and 4) appendices which are separate from the main body of text but both structurally and rhetorically connected. To account for these document parts, each of the document types (the <predoc>, <maindoc>, <postdoc>, and <xdoc> tags) were given the four children tags <pretext>, <text>, <posttext>, and <appendix>. Using <maindoc> as an example (because it is the rhetorical focus), the next level of structure is as such:

```
<maindoc>
    <note>
    <pretext>
    <text>
    <posttext>
    <appendix>
    <part>
```

At this level, the primary and required tag is of course the <text> tag. The other text-type tags, <pretext> and <posttext>, are optional, but when they occur have an internal structure similar to that of the <text> tag. The <appendix> tag has an internal structure identical to the <docdata> tag (see below). The <part> tag is a special tag used with large documents in which the token count of analyzable text exceeds the limit set by the sampling plan. In these cases, instead of placing all the document text in a single <text> tag, the data are sampled in three 'parts' and placed in <text> tags within <part> tags. This sampling and

archiving of large documents will be discussed further in Section 4.5, but when implemented creates an additional level of structure within the <maindoc> as such:

```
<maindoc>
    <note>
    <pretext>
    <part>
        <text>
    <part>
        <text>
    <part>
        <text>
    <posttext>
    <appendix>
    <part>
```

The <text> tag is unique in that it marks the beginning of a transition in tag purpose from rhetorical structure to rhetorical intent, and the change from tags that only contain child tags to tags that actually contain document data (potentially analyzable text). In essence, it holds the child tags that delineate the final (lowest level) rhetorical structure[3] of the analyzable running text of a document. In practice, we found it necessary to account for the following: the running text itself as it is divided into paragraph-type structures (<p>), interruptions in the running text associated with page breaks (<page>), non-page-break interruptions (<npb>), and various figures that may be inserted in the running text (<image>, <table>, <form>, <xitem>). This generates the following structure:

```
<text>
    <page>
    <p>
    <npb>
    <image>
    <table>
    <form>
    <xitem>
    <note>
```

It is at this point, having moved at least six levels into the structure to the <p> tag, that we reach the document text which is generally considered to be of rhetorical value and

[3]There are tags to denote events such as footnotes which could be considered structural.

analyzable, and here is where the full transition to rhetorical intent has been made. The <p> tag contains the actual text of the document as well as child tags that mark portions of that text according to various attributes/features that make them rhetorically distinct from surrounding text. Thus the <p> tag may contain PCDATA or the following:

```
<p>
    <footend>
    <anc>
    <h>
    <quote>
    <margin>
    <insert>
    <lineout>
    <marked>
    <emph>
    <symbol>
    <illegible>
    <formula>
    <note>
```

Practically speaking, the child tags of the <p> tag are at the full depth of the XML tag hierarchy as they generally contain nothing more than text data. However, each of the child tags of the <p> tag can also contain various child tags. In general terms, the children of the <p> tag all have each other as children, or even the <p> tag itself. These were rarely used, but the option does exist for constructions such as the following, which would be used for a paragraph-type block of text which contained quote made of several blocks of text, one having marginalia with marked text:

```
<text type="text">
    <p>
        <quote>
            <p></p>
            <p>
                <margin>
                    <marked></marked>
                </margin>
            </p>
            <p></p>
        </quote>
    </p>
</text>
```

This being the case, theoretically speaking there is actually no bottom to the hierarchy because the DTD allows recursion. In other words, there are instances where A contains B and B contains A, which forms a loop. This happens at a number of levels. The highest level loop occurs with appendices to a document, which are considered to have the same general structure as a document. This being the case, the `<appendix>` tag has the same children as the `<docdata>` tag:

```
<appendix>
    <note>
    <predoc>
    <maindoc>
    <postdoc>
    <xdoc>
```

In theory, this can lead from `<appendix>`, to `<maindoc>`, to `<appendix>`, although this never occurs in the TDC archive. A less obvious loop comes from the fact that images, forms, and tables, which occur within text, may also contain text. Loops can also be created within the `<p>` tag given that most of its children have each other, or even the `<p>` tag, as children. However, even though these loops are theoretically possible, recursion is limited by the data themselves, which would become too complex to comprehend after only a few iterations. This allows these recursive structures to be retained for the benefit of simplicity (no new tags or attributes are required) without fear of adding a computational pitfall (as long as no one tries to print the full structure).

As a final illustration to help put all of the above into a more tangible form, Figure 4.11 is an actual XML archive of a relatively simple, single-page, handwritten document from the Quota Sample (Bates Number 503257271). The original document image can be seen in Figure 4.5. Beginning at the `<tobaccodocs>` tag, the most significant illustration of this example is that although the possible structures are essentially unlimited, the actual structure is quite simple. Just as with natural speech, the depth and complexity of the hierarchy is governed by the comprehension of the author and reader. In this case, there is only the

single <maindoc>, and within it a maximum depth of three levels represented by various illegible items within marginalia and marked-out text.

```
<?xml version="1.0" ?>
<!DOCTYPE tobaccodocs SYSTEM "http://www.uga.edu/tobaccodocs/tobacco.dtd">

<tobaccodocs>
    <document sample="quota" decade="1970" isource="rjr" class="ni">
        <metadata>
            <bates_start>503257271</bates_start>
            <bates_end>503257271</bates_end>
            <uga>
                <note>handwritten letter, some names in German</note>
                <date>19700603</date>
                <pages>1</pages>
                <words>1389</words>
                <section amount="all"></section>
                <encoded_by>Anastasia Wright</encoded_by>
                <verified_by>NONE</verified_by></uga>
            <external></external> </metadata>
        <docdata>
            <maindoc type="text">
                <pretext>
                    <illegible>Return Address </illegible>June 3, 1970
                    <margin type="comment">
                        <illegible></illegible></margin>Dear Dr. Colby: </pretext>
                <text type="text">
                <p/>I have compiled the enclosed bibliography from an extensive
                bibliography included in Neoplasms of the Domesticated Mammals
                by E. Cotchin - Commonwealth Bureau of Animal Health - Commonwealth
                Agricultural Bureau 1956 - I do not know how many of these articles
                    <lineout>
                        <illegible></illegible></lineout>
                    <insert>are going </insert> to be available to us, and how
                many of these are going to be useful, but, I hope, there will
                be enough literature for our purpose.
                <p/> A good bibliography
                    <insert>up</insert> to 1962 is included in the
                    <illegible>"Homolbuch der Speziellen Pathologishen </illegible>
                Anatomy, 1962 (I will send you the exact reference) - An interesting
                table listing tumors by species and locations is included in
                the "Veterinary Pathology" by Smith
                    <illegible>&</illegible> Jones (
                    <illegible>Seo. & Febiyer</illegible> 1966)
                <p/> In addition to the problem outlined in your letter, which
                would give the increase of lung cancer in a non smoking dog
                population I am trying to assemble literature which may give
                an answer to the question: In which way does cancer of "dogs"
                differ from cancer of "men" (type of cancer, sex, age, etc.)
                <p/> It would have interested me to go to the Institute
                    <illegible>Jr Staoklen Jorshung </illegible>in Frankfurt,
                but I have to give that up. In my program, however, remains
                Hamburg, where my husband has been invited for a lecture. I
                think that it would be very useful to visit Dr.
                    <illegible>Weber</illegible> of the
                    <illegible>Verband der Cigaretten Industria,</illegible>
                I have already been in touch with him.
                <p/> I am sending a few articles under separate covers.</text>
                <posttext>Best Greeting - Sincerely
                    <illegible>
                        <note>Signature: G. S. Caluman </note> S. S Post.</illegible>
                </posttext> </maindoc> </docdata> </document></tobaccodocs>
```

Figure 4.11: Document Example 11: XML

4.4.3 XML Tag Descriptions

Now that the general structure of a document archive has been established, the discussion can now focus more directly on the intended purpose of each of the 46 tags allowed by the TDC's DTD. Of course, there is no distinct line that can be drawn between structure and purpose, so this section, just as with the previous, will discuss both. However, here the attention will be more on function than hierarchy.

The tags below are ordered in a logical sequence of how they might be encountered based on general document order and hierarchy depth, similar to how they were presented in the section above. For each tag there is 1) a general description of the tag and its purpose/use, 2) a description of all tag attributes, 3) a list of possible parent tags, and 4) a list of allowable data types (text data and/or child tags). For easier reference, this will be done in an identical format across all the tag descriptions. Also note that the definitions presented here evolved over the course of the project as we gained experience working with the text.

1. `<tobaccodocs>` Description: This required tag is the root or highest-level tag for all TDC XML documents. Its main purpose is to provide a means for including more than one `<document>` in a single archive. Although this tag is used for every archive, its usefulness can best be seen in the TDC Toolkit (http://www.tobaccodocs.uga.edu/TDC) where the `<tobaccodocs>` tag is the container for all XML output from the corpus-generating scripts. That is, the output is a single XML archive with up to 1113 `<document>` tags (all Quota, Supplemental, and Rhetorical Sample[4] documents). Parent Tag: None. Attributes: None. Required Data: The `<document>` tag, at least one. Optional Data: None.

2. `<document>` Description: This required tag is the parent or highest-level tag for a single document archive as defined by a given Bates-Number range. It contains all data for

[4]The DTD used for archiving the Quota and Supplemental Samples is used for all UGA Tobacco Project documents, to include the Rhetorical Sample. Although not specific to the topic of this work, these features of the DTD are included and described here.

the document. Attributes: There are three required attributes (*decade*, *isource*, *class*), and four optional attributes allowed for documents in the Rhetorical Sample (*rcase*, *rclass*, *raudience*, *rnumber*). The required attributes all relate to sampling as described in Section 3.4. The *decade* is the decade-based sampling stratum of the document: *1950*, *1960*, *1970*, *1980*, *1990*, *19xx*, or *Bliley*. The *isourse* is the initials of the industry source of the document: *atc* (American Tobacco), *bw* (Brown and Williamson), *ctr* (Council for Tobacco Research), *ll* (Lorillard), *pm* (Philip Morris), *rjr* (R. J. Reynolds), or *ti* (Tobacco Institute). The *class* is according to the document-type classification: *ni* (named-internal audience), *ne* (named-external audience), *ui* (unnamed-internal audience), *ue* (unnamed-external audience). The *rcase* is a specified case number that allows for matching the given document with other Rhetorical Sample documents. For example, in multi-draft documents all drafts are given the same case number. The *rclass* is the Rhetorical Sample class or document type: *A* for cross-audience cases, and *D* for cross-draft cases. The *raudience* is a Rhetorical Sample case sub-specification for cross-audience documents: *internal* (internal audience), or *external* (external audience). The *rnumber* is the draft version of the document for Rhetorical Sample cross-draft documents: *first* for the first draft, *last* for the last draft, or a draft number (1–10). Parent Tag: The `<tobaccodoc>` tag. Required Data: The `<metadata>` and `<docdata>` tags, one each. Optional Data: None.

3. `<note>` Description: This optional tag provides the archivist with a place to record any type of note about the current tag or tag content necessary for clarification or description. The `<note>` tag is found as an optional child of most tags. Attributes: None. Parent Tag: The `<uga>`, `<docdata>`, `<maindoc>`, `<predoc>`, `<postdoc>`, `<xdoc>`, `<pretext>`, `<text>`, `<part>`, `<posttext>`, `<appendix>`, `<description>`, `<caption>`, `<page>`, `<p>`, `<h>`, `<quote>`, `<emph>`, `<marked>`, `<illegible>`, `<formula>`, `<symbol>`,

<footend>, <insert>, <lineout>, or <margin> tag. Required Data: None. Optional Data: Archivist's notes as PCDATA.[5]

4. <metadata> Description: This required tag provides a location to record all data related to the document as a whole but not found on the original document image (prior to Bates Numbering), or not directed specifically toward a document element, such as an archivist note or description. Attributes: None. Parent Tag: The <document> tag. Required Data: The <bates_end>, <bates_start>, <external>, and <uga> tags, one each. Optional Data: None.

5. <bates_start> Description: This required tag provides a location for the starting Bates Number of the defined Bates-Number range for the given document (i.e. the Bates Number of the first page of the document). For a single-page document, this will be the same as the <bates_end> tag data. Attributes: None. Parent Tag: The <metadata> tag. Required Data: A Bates Number as PCDATA. Optional Data: None.

6. <bates_end> Description: This required tag provides a location for the end Bates Number of the defined Bates-Number range for the given document (i.e. the Bates Number of the last page of the document). For a single-page document, this will be the same as the <bates_start> tag data. Attributes: None. Parent Tag: The <metadata> tag. Required Data: A Bates Number as PCDATA. Optional Data: None.

7. <uga> Description: This required tag provides a location for all document metadata supplied by the University of Georgia (UGA) Tobacco Documents Project other than Bates Numbers. All other metadata are found in the <external> tag. Attributes: None. Parent Tag: The <metadata> tag. Required Data: The <date>, <pages>, <words>, <section>, <encoded_by>, <verified_by>, <image_file> tags, one each. The tags must be present in the archive; however, they are not required to contain data (i.e.

[5]General non-XML text data are referred to as PCDATA, which is shorthand for *parsed character data*.

if none are available). The <image_file> tag was a late addition to the DTD, and although required, was not fully implemented into all TDC archives prior to the author leaving the project. Optional Data: The <note> tag.

8. <external> Description: This required tag provides a location for any and all metadata not meeting the requirements for inclusion in the <uga> tag. It is included to contain various summaries and synopses related to the document but authored externally. Although a required tag, it may contain no data. Attributes: None. Parent Tag: The <metadata> tag. Required Data: None. Optional Data: No restrictions (ANY).

9. <date> Description: This required tag provides a location for the date of document origin (not a filing date) following the standard date indexing order used by the UGA TDC Project: YYYYMMDD, all Arabic digits, using *00* as a marker of unknown year, month, or day. Attributes: None. Parent Tag: The <uga> tag. Required Data: A date as PCDATA. Optional Data: None.

10. <pages> Description: This required tag provides a location for the page count of the given document. All Arabic digits (i.e. no commas). Entered, updated, and/or verified by an automated process. Attributes: None. Parent Tag: The <uga> tag. Required Data: A number as PCDATA. Optional Data: None.

11. <words> Description: This required tag provides a location for the word (total token) count of the given document. All Arabic digits (i.e. no commas). Entered and verified by an automated process. Attributes: None. Parent Tag: The <uga> tag. Required Data: A number as PCDATA. Optional Data: None.

12. <section> Description: This required tag provides an indication of whether the <docdata> tag contains all text data from the given document or, in the case of large documents, contains a sample (only part of the text). Attributes: There is one required attribute: *amount*. The allowed values are *all* if all text is recorded, or *part* if a sample

is used. Parent Tag: The <uga> tag. Required Data: None. Optional Data: Unspecified PCDATA.

13. <encoded_by> Description: This required tag provides a location to record the document archivist's name. This was used for assigning work and for verification processes. Attributes: None. Parent Tag: The <uga> tag. Required Data: The archivist's name as PCDATA. Optional Data: None.

14. <verified_by> Description: This required tag provides a location to record the name of the archivist who verified (checked) the document coding for XML validation, spelling, hierarchy, and tag choices. Attributes: None. Parent Tag: The <uga> tag. Required Data: The verifying archivist's name as PCDATA if available. Optional Data: None.

15. <image_file> Description: This required tag provides a location to record the permanent URL and name of the document image file (PDF, TIFF, PNG, etc.) from which the data were collected. This was a late addition to the DTD, and although required, was not fully implemented into all TDC archives prior to the author leaving the project. Attributes: None. Parent Tag: The <uga> tag. Required Data: The image file path/name as PCDATA. Optional Data: None.

16. <docdata> Description: This required tag is the parent or highest-level tag of all data recovered directly from the original document other than the Bates Numbers. Attributes: None. Parent Tag: The <document> tag. Required Data: The <maindoc> tag, one each. Optional Data: Any of the following tags as needed: <note>, <predoc>, <postdoc>, or <xdoc>.

17. <maindoc> Description: This required tag provides a location to record data from the main (major, primary) document from within a given range of Bates Numbers. That is, in the case that a single range of Bates Numbers provided by the tobacco industry as a single 'document' actually contains several structurally and rhetorically separate

parts (true documents), the archivist must choose the document of central focus and record it here. Only one <maindoc> is allowed. This is a required tag, even in the event that only a single document exists within the Bates-Number range. The remaining documents are recorded using the <predoc>, <postdoc>, <xdoc>, and <note> tags as needed. Attributes: There is one required attribute: *type*. The allowed values are *text* (normal running text, possibly with embedded elements), *form* (the document is a form containing analyzable text), *image* (the document is an image containing analyzable text), *table* (the document is a table containing analyzable text), or *xtype* (the document is of unknown type, but contains analyzable text). Parent Tag: The <docdata> or <appendix> tag. Required Data: The <text> tag, or in the case of a large document the <text> tag is replaced with a series of three <part> tags. Optional Data: Any of the following tags as needed: <note>, <pretext>, <posttext>, or <appendix>.

18. <predoc> Description: This optional tag is defined by the location of the <maindoc> tag data. It provides a location to record data from any single structurally and rhetorically separate document occurring prior to the <maindoc> data in a given range of Bates Numbers provided by the tobacco industry as a single document. For example, this tag would be used to record data from a cover letter or introductory document to the main document. There can be multiple <predoc> tags as necessary to record distinct documents. Also see the <maindoc> tag for additional information. Attributes: There is one required attribute: *type*. The allowed values are *text* (normal running text, possibly with embedded elements), *form* (the document is a form containing analyzable text), *image* (the document is an image containing analyzable text), *table* (the document is a table containing analyzable text), or *xtype* (the document is of unknown type, but contains analyzable text). Parent Tag: The <docdata> or <appendix> tag. Required Data: The <text> tag, or in the case of a large document the <text> tag is replaced with a series of three <part> tags. Optional Data: Any of the following tags as needed: <note>, <pretext>, <posttext>, or <appendix>.

19. <postdoc> Description: This optional tag is defined by the location of the <maindoc> tag data. It provides a location to record data from any single structurally and rhetorically separate document (not an appendix) occurring after the <maindoc> data in a given range of Bates Numbers provided by the tobacco industry as a single document. For example, this tag would be used to record data from an attachment following the main document. There can be multiple <postdoc> tags as necessary to record distinct documents. Also see the <maindoc> tag for additional information. Attributes: There is one required attribute: *type*. The allowed values are *text* (normal running text, possibly with embedded elements), *form* (the document is a form containing analyzable text), *image* (the document is an image containing analyzable text), *table* (the document is a table containing analyzable text), or *xtype* (the document is of unknown type, but contains analyzable text). Parent Tag: The <docdata> or <appendix> tag. Required Data: The <text> tag, or in the case of a large document the <text> tag is replaced with a series of three <part> tags. Optional Data: Any of the following tags as needed: <note>, <pretext>, <posttext>, or <appendix>.

20. <xdoc> Description: Following the definitions of the <maindoc>, <predoc>, and <postdoc> tags, this optional tag provides a location to record data from any single structurally and rhetorically separate document (not an appendix) in a given range of Bates Numbers provided by the tobacco industry as a single document that cannot be logically placed in one of the above tags. There is no expectation that this tag be used, but multiple instances are allowed as necessary. In practice, this tag was used twice in the Quota Sample, once to mark an opinion poll attached to the <maindoc>, and once to identify a short attached periodical. Also see the <maindoc>, <predoc>, and <postdoc> tags for additional information. Attributes: There is one required attribute: *type*. The allowed values are *text* (normal running text, possibly with embedded elements), *form* (the document is a form containing analyzable text), *image* (the document is an image containing analyzable text), *table* (the document is

a table containing analyzable text), or *xtype* (the document is of unknown type, but contains analyzable text). Parent Tag: The <docdata> or <appendix> tag. Required Data: The <text> tag, or in the case of a large document the <text> tag is replaced with a series of three <part> tags. Optional Data: Any of the following tags as needed: <note>, <pretext>, <posttext>, or <appendix>.

21. <pretext> Description: This optional tag is defined by the location of the <text> or <part> tag. It provides a location to record all non-analyzable text that precedes the <text> tag data. In general, this tag contains all text data with associated items and descriptions up to the title of a document, or up to and including the opening salutation of a correspondence. See the <text> tag for additional information. Attributes: None. Parent Tag: The <maindoc>, <predoc>, <postdoc>, or <xdoc> tag. Required Data: None required. PCDATA is expected. Optional Data: Document text as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <h>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <anc>, <insert>, <lineout>, or <margin>. Note that the <p> tag is not an option in the <pretext> tag.

22. <text> Description: This tag provides a location to record the analyzable text of a document, image, table, form or other item. 'Analyzable' is defined in this context as a body of text having one or more blocks containing 50 or more words of continuous non-template data, where 'continuous' refers to text data 'cohering by virtue of syntactic or semantic cohesive ties linking text into topical units. Continuous text is not [separated by any box boundaries] or other graphic device that attempts to segregate one unit of text from its neighboring units. Thus, text that appears in different cells in a table is not continuous, even if pertaining to a similar topic' (Dr. Rubin). This tag is required for all documents. It contains the data beginning with the title or following the opening salutation of a correspondence, to the end of the text body or up to closing salutations if no postscripts exist (otherwise continuing through the postscripts). The

exception to this is with long documents where the <text> tag is replaced by a series of three <part> tags. This tag is optional in images, forms, tables and other items, in which case it is used as needed to record analyzable text. Attributes: There is one required attribute: *type*. The allowed values are *text* (analyzable text from a document, possibly containing embedded items, but not from within any items), *form* (analyzable text from within a form), *image* (analyzable text from within an image), *table* (analyzable text from within an image), or *xtype* (analyzable text from an unclassified item). Parent Tag: The <maindoc>, <predoc>, <postdoc>, <xdoc>, <part>, <image>, <form>, <table>, <xitem>, or <caption> tag. Required Data: None required. The <p> tag is expected. Optional Data: Any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <npb>, or <p>. Note that unlike the <pretext> and <postext> tags, PCDATA is not allowed in the <text> tag. It must be contained in the one of the allowable child tags (generally the <p> tag).

23. <part> Description: This tag is allowed in a document tag only as a series of three (i.e. three <part> tags in a row). It is used to introduce an additional level of structure into the tag hierarchy between a document level tag, such as <maindoc>, and the <text> tag. Because the full content of long documents is not recorded, this additional level is necessary to record data related to sampling. See Section 4.5 below for more details on sampling and archiving large documents. Attributes: There are two required: *section* and *bates*. The allowed values for *section* are *begin* (sampled from the beginning of the document), *middle* (sampled from the middle of the document), and *end* (sampled from the end of the document). The allowed value for *bates* is PCDATA representing the Bates Number of the document page on which the sample begins. Parent Tag: The <maindoc>, <predoc>, <postdoc>, or <xdoc> tag. Required Data: The <text> tag. Optional Data: The <note> tag as needed.

24. <posttext> Description: This optional tag is defined by the location of the <text> or <part> tag. It provides a location to record all non-analyzable text that follows

the <text> tag data. In general, this tag contains all text data with associated items and descriptions following the main body of text, including closing salutations of correspondences if no postscript data are present. See the <text> tag for additional information. Attributes: None. Parent Tag: The <maindoc>, <predoc>, <postdoc>, or <xdoc> tag. Required Data: None required. PCDATA is expected. Optional Data: Document text as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <h>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, <lineout>, or <margin>. Note that the <p> tag is not an option in <posttext>.

25. <appendix> Description: This optional tag provides a location to record data from appendices of a given document (i.e. within a <maindoc>, <predoc>, <postdoc> or <xdoc> tag). Data in the <appendix> tag differ from <postdoc> tag data in that they are both rhetorically and structurally connected (often specifically named as an appendix) to the given document. Attributes: None. Parent Tag: The <maindoc>, <predoc>, <postdoc>, or <xdoc> tag. Required Data: The <maindoc> tag, one each. Optional Data: Any of the following tags as needed: <note>, <predoc>, <postdoc>, or <xdoc>.

26. <image> Description: This optional tag provides a location to record data from or about an image found within text. An image is defined as any graphical item (photos, artwork, charts) distinct from the running analyzable text by a means other than typesetting (i.e. not simply a decorative font and format). At a minimum, a description of the image must be provided. However, the image caption is also recorded when present, and the image text is recorded if analyzable (see the <text> for more information). Attributes: None. Parent Tag: The <pretext>, <text>, <posttext>, <page>, <quote>, <emph>, <marked>, <insert>, <lineout>, or <margin> tag. Required Data: The <description> tag. Optional Data: The <text> or <caption> tags.

27. <form> Description: This optional tag provides a location to record data from or about a form found within text. A form is defined as a portion of a document distinct from the running analyzable text which has bullet-type captions and pre-defined text fields (templates) with spacial constraints that limit text input (for example, printed boxes). For consistency, legal documents and contracts are excluded because it is difficult to determine the amount of standardization in computer-generated forms. At a minimum, a description of the form must be provided. However, the form caption is also recorded when present, and the form text is recorded if analyzable (see the <text> for more information). Attributes: None. Parent Tag: The <pretext>, <text>, <posttext>, <page>, <quote>, <emph>, <marked>, <insert>, <lineout>, or <margin> tag. Required Data: The <description> tag. Optional Data: The <text> or <caption> tags.

28. <table> Description: This optional tag provides a location to record data from or about a tabular display found within text. A table is defined as a structure with a columnated format designed to display data (primarily numerical). At a minimum, a description of the table must be provided. However, the table caption is also recorded when present, and the table text is recorded if analyzable (see the <text> for more information). Attributes: None. Parent Tag: The <pretext>, <text>, <posttext>, <page>, <quote>, <emph>, <marked>, <insert>, <lineout>, or <margin> tag. Required Data: The <description> tag. Optional Data: The <text> or <caption> tags.

29. <xitem> Description: This optional tag provides a location to record data from or about an unclassified item found within text. At a minimum, a description of the item must be provided. However, the item caption is also recorded when present, and the item text is recorded if analyzable (see the <text> for more information). Attributes: None. Parent Tag: The <pretext>, <text>, <posttext>, <page>, <quote>, <emph>, <marked>, <insert>, <lineout>, or <margin> tag. Required Data: The <description> tag. Optional Data: The <text> or <caption> tags.

30. <description> Description: This required tag provides a location to describe the contents of images, forms, tables, formulas, and other items. Attributes: None. Parent Tag: The <image>, <form>, <table>, <xitem>, or <formula> tag. Required Data: None required. PCDATA is expected. Optional Data: A description of the item content as PCDATA, and/or the <note> tag.

31. <caption> Description: This optional tag provides a location to record a caption associated with an image, form, table of other item. Attributes: None. Parent Tag: The <image>, <form>, <table>, or <xitem> tag. Required Data: None required. PCDATA is expected. Optional Data: Caption text as PCDATA, and/or the <note> or <text> tag.

32. <page> Description: This optional tag provides a location to record all non-analyzable text data associated with page changes, such as page numbers, footers, headers, dates, et cetera. That is, it holds all data causing interruptions of running/continuous text that can be attributed to page breaks. Attributes: None. Parent Tag: The <pretext>, <text>, <posttext>, <page>, <quote>, <emph>, <marked>, <illegible>, or <lineout> tag. Required Data: None required. PCDATA is expected. Optional Data: Document data as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <h>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <anc>, <insert>, <lineout>, or <margin>.

33. <npb> Description: This optional tag is used to record any noteworthy non-paragraph break within text. Examples include lines, large white space areas, text boxes, and form boundaries. Originally this tag was designed to be an empty tag of the form <npb/>, but it may appear in the standard two-part form <npb>PCDATA</npb>. Attributes: None. Parent Tag: The <text> or <footend> tag. Required Data: None. Optional Data: None intended, but it may contain non-document text PCDATA.

34. Description: This tag provides a location to record regular document text data. Originally this tag was designed to be an empty tag[6] of the form , but it may appear in the standard two-part form PCDATA . By the DTD this is an optional tag. However, by direction to the archivists, all text (paragraphs or individual lines) begin with the tag, and therefore the tag is the first data (child tag) in any <text> or <footend> tag. Attributes: None. Parent Tag: The <text>, <quote>, <emph>, <marked>, <illegible>, <footend>, or <lineout> tag. Required Data: None required. PCDATA is expected. Optional Data: Given the option to use this tag in a two-part form, it allows any document text as PCDATA, and/or any of the following tags as needed: <note>, <h>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <footend>, <anc>, <insert>, <lineout>, or <margin>.

35. <h> Description: This optional tag provides a method to denote headers in the text data, where a header is defined as any text that names a subsequent section of text (i.e provides a title) regardless of font or typesetting. The focus here is on intent or purpose rather than format. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, , <footend>, or <insert> tag. Required Data: None required. PCDATA is expected. Optional Data: The header/title text as PCDATA, and/or any of the following tags as needed: <note>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, <lineout>, or <margin>.

36. <quote> Description: This optional tag provides a method to denote quoted material in the text data, where quoted is defined as any text identified by the document author as being quoted/taken directly from another source regardless of the punctuation, font, or typesetting. The focus here is on intent or purpose rather than format.

[6]Although convenient for HTML, using the empty-tag format to mark paragraphs was, quite frankly, a mistake. Once we began the process of extracting text from the archive we realized the necessity of precisely marked data. Using the two-part format to enclose all separate paragraphs and lines would have provided the necessary accuracy. Our intent was to repair this mistake in an automated manner, but this proved difficult and to date has not been done. Refer to Chapter 6.2 for more detail.

This tag is not used to mark text in quotation marks unless it is a true quote. Quotation marks are regularly used to mark events, such as a foreign word or non-standard usage, which should not be marked using the <quote> tag. Attributes: None. Parent Tag: The <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is expected. Optional Data: The quoted text as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <p>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, <lineout>, or <margin>.

37. <emph> Description: This optional tag provides a method to denote text emphasis conveyed by typesetting or font, regardless of the type (bold, italics, capitals, underlining, et cetera). The focus here is on intent or purpose rather than format. To use this tag, the marked text must be set apart from the surrounding text BY typesetting FOR the purpose of emphasis. This tag should not be used as the <h> tag for headers (which are by design different from the surrounding text), but may be used inside of a header if needed. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is expected. Optional Data: The emphasized text as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <p>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, <lineout>, or <margin>.

38. <marked> Description: This optional tag provides a method to denote text that has been marked for emphasis by a reader or editor after the completion of the original document, such as circled or underlined text. The key feature here is that the original document has a secondary mark which was made by hand (i.e with a pencil, pen, or highlight) after the completion of the document. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is

expected. Optional Data: The marked text as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <p>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <anc>, <insert>, <lineout>, or <margin>.

39. <illegible> Description: This optional tag provides a method to denote the existence of text that is illegible to the archivist because of physical issues with the document (i.e. poor image quality, bad penmanship, overwriting, et cetera). The archivist's best guess or best partial rendering is placed inside the tag. The tag should span only that portion of the text data where the archivist is uncertain about the physical rendering. This tag was not used to mark non-English or chemical-biological terminology if the rendering was certain. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is expected. Optional Data: The archivist best attempt at rendering the text as PCDATA, and/or the <note>, <page>, or <p> tags.

40. <formula> Description: This optional tag provides a method to denote the existence of a chemical or mathematical formula that cannot be represented correctly using ANSI characters. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is expected. Optional Data: An approximate representation of the formula as PCDATA, and/or the <note> or <description> tags.

41. <symbol> Description: This optional tag provides a method to denote non-image, non-ANSI characters or symbols found in a line of text. For example, this might be used to record a check mark, star, bullet, circle, or other mathematical or scientific symbols. This is not to be used to record images or formulas. This is intended to mark single characters or very small character clusters. Attributes: None. Parent Tag: The

<pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is expected. Optional Data: A description of the symbol as PCDATA, and/or the <note> tag.

42. <footend> Description: This optional tag provides a method/location to record data from footnotes or endnotes. Attributes: There are two required attributes: *type* and *num*. The allowed values for the *type* attribute are *comment* (the note contains analyzable text), and *ref* (the note contains non-analyzable bibliographic data). If the *comment* value is used, the text data from the note are entered as tag content. If the *ref* value is used, the tag is left with no PCDATA. The *num* attribute records the note's reference or anchor number as PCDATA. Parent Tag: The <p> tag. Required Data: None. Optional Data: The tag content as PCDATA, and/or any of the following tags as needed: <note>, <table>, <xitem>, <npb>, <p>, <h>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, <lineout>, or <margin>.

43. <anc> Description: This optional tag provides a method/location to record footnote or endnote anchors in the text (often seen as superscript numbers or numbers in parentheses). Originally this tag was designed to be an empty tag of the form <anc/> where the anchor number is recorded as an attribute, but it may appear in the standard two-part form <anc></anc>. Attributes: There is one required attribute: *num*. This attribute corresponds to the *num* attribute of the <footend> tag. It records the note's reference or anchor number as PCDATA. Parent Tag: The <pretext>, <page>, <p>, or <marked> tag. Required Data: None. Optional Data: Undefined PCDATA.

44. <insert> Description: This optional tag provides a method/location to record data added to the text during editing. The key feature here is that the original document has secondary text additions made by hand (i.e with a pencil or pen) after the completion of the document. This differs from marginalia by rhetorical purpose in that

marginalia is a comment about the document text while an insertion is an addition to the document text. Small insertions are often marked with a caret or wedge (∧) symbol at the insertion point, but larger insertions are regularly found written in the margins, perhaps with an arrow leading to the insertion point. The focus here is on intent or purpose rather than format or location. The <insert> tag should be placed in the text at the point of insertion if it is known. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, <lineout>, or <margin> tag. Required Data: None required. PCDATA is expected. Optional Data: The insertion as PCDATA, and/or any of the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <h>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, <lineout>, or <margin>.

45. <lineout> Description: This optional tag provides a method to denote that text data have been marked out (for removal) during editing. The key feature here is that the original document has a secondary mark made by hand (i.e with a pencil or pen) after the completion of the document that indicates the text should be removed in subsequent drafts. The focus here is on intent or purpose rather than format, which may be a line through the text, a series of X marks, or even a circle. Quite often the lined out text will be illegible and need to be marked as such. Attributes: None. Parent Tag: The <pretext>, <posttext>, <page>, <p>, <h>, <quote>, <emph>, <marked>, <footend>, <insert>, or <margin> tag. Required Data: None required. PCDATA or the <illegible> tag is expected. Optional Data: The marked text as PCDATA, and/or the following tags as needed: <note>, <image>, <form>, <table>, <xitem>, <page>, <p>, <quote>, <emph>, <marked>, <illegible>, <formula>, <symbol>, <insert>, or <margin>.

46. <margin> Description: This optional tag provides a method/location to record non-editing data (marginalia) added to the original document, generally handwritten. The key feature here is that the original document has secondary marks, stamps or stickers

added after the completion of the document, but prior to Bates Numbering. Marginalia may be any comment about the document, filing information, routing information, initials, et cetera, as long as it was not part of the original document and not allowable data for the `<marked>`, `<insert>`, or `<lineout>` tags. Marginalia does not contain editing or editing marks, but it may contain comments about or suggestions for editing. Bates Numbers are excluded because they are non-tobacco-industry additions. The `<margin>` tag should be inserted in an appropriate location nearest to where it applies (by text content or document location). Attributes: There is one required attribute: *type*. The allowed values are *comment* (comments made about the document text), and *label* (non-commet marginalia used for filing, routing and tracking). Parent Tag: The `<pretext>`, `<posttext>`, `<page>`, `<p>`, `<h>`, `<quote>`, `<emph>`, `<marked>`, `<footend>`, `<insert>`, or `<lineout>` tag. Required Data: None required. PCDATA is expected. Optional Data: The marginalia as PCDATA, and/or the following tags as needed: `<note>`, `<image>`, `<form>`, `<table>`, `<xitem>`, `<quote>`, `<emph>`, `<marked>`, `<illegible>`, `<formula>`, `<symbol>`, `<insert>`, or `<lineout>`.

4.5 Document Conversion

When described on paper, the physical process of document conversion that the archivists goes through in the daily routine of reading and typing is actually a rather simple set of steps. However, the reason that the people who have their hands on the keyboards have the title *archivists* instead of *typists* is that the bulk of the work they do is not the physical typing (which could be done overseas at a much lower rate) but the cognitive exercise of making decisions about the text and how it should be archived. In the end, the decisions they make as they go through the day will determine the coding accuracy (and ultimately the usefulness) of the resultant corpus. Thus the major task in document conversion when classification is involved is preparing (or maturing) the archivists, getting them to the point where they can consistently make and record reliable classifications.

With the Tobacco Documents Corpus this maturation process generally consisted of four parts: understanding the general structure and purpose of the archive document, learning the general steps of the conversion process, knowing how to address special circumstances, and incorporating (or recycling) experience with the documents into the procedures. Although in practice there was considerable overlap between the four, they are, I believe, a reasonable and understandable division of the overall process of document conversion. Consequently, they will be used as a guideline in the subsequent sections.

I will make the assumption here that for the reader the first part, general structure and purpose of the archive and tag set, has been described sufficiently in Section 4.4 above and does not need to be readdressed at this point. However, I will reiterate this point: the implementation of all conversion procedures hinges on the archivists understanding the purpose behind the XML being used. For the TDC there is a particular non-canonical mindset that must underpin the procedures if they are to be productive, and this proved to be the major obstacle. Our archivists, although very experienced, came from a library setting where historical preservation was the focus. They were not easily convinced that the XML being used for the TDC did not have that same purpose. But once they were convinced, the other three parts fell into place rather nicely.

Moving forward we can begin with a description of general procedures, then move to special circumstances, both for intra-document events and for large documents, and finally to a description of validation, review, and revision procedures.

4.5.1 General Procedures

The following are the general procedures used by the archivists during document conversion of both the Quota and Supplemental Samples:

1. Select document. As one might expect, the first event in the conversion process was to select a document. This was done in two parts, with no specific chronological order. Part A) As the document files were collected during sampling, they were placed in

folders (directories) on the master archive computer according to the sampling iteration and decade group. Individual archivists were assigned to convert all the documents in specific folders by the project supervisor. That is, no folder would have more than one archivist assigned to it. The archivist was generally free to choose any document in the folder which had not been previously converted. Converted files were stored in the same folder as the document image file and had similar names, making it easy to determine if the archive had been made. Part B) Archivists were required to keep a record of each file assigned to them and the current file status (generally in a spreadsheet). The archivist was free to select any file from the file record which had not been previously converted. Regardless of the method of selection, the archivist was required to cross-verify file status between the actual archive folder and the file record. Also, as a general policy (not strictly enforced), archivists were asked not to begin the conversion of a document image if it could not be completed during the current work period, such that no archive was left partially complete. In the event that a file was left incomplete, the archive file extension was changed from *.xml* to *.inc*. This was an assistance in keeping records, but a necessity for running the automated checks discussed below.

2. Open the document image. During sampling, document images were collected as either TIFF or PDF image files (refer to Chapter 3 for more details). These files could be opened using a number of common viewers available on the archivists workstations. The choice of viewer was left to the discretion of the archivist. In all cases, however, the image viewer was a separate program from the XML editor.

3. Open the editor and template file. Archivists were required to use a validating XML editor, meaning an editor that compared the open file to the DTD used by the TDC Project and provided an indication of whether or not the file was structurally correct or *valid* according to the defined tag set and hierarchy. After some experimenting, the archivists settled on the XMetaL brand of editing tools (formerly a QuadSoft product,

currently owned by JustSystems, http://na.justsystems.com). This proprietary software made full use of the DTD by not allowing the archivist to enter non-valid data, and by displaying and auto-inserting the allowable tags using drop-down menus. Once the editor was started, a template file containing a minimal valid tag structure was opened. Each archivist used their own template file, but in general they were similar to the template file shown in Figure 4.12.

```
<?xml version="1.0" ?>
<!DOCTYPE tobaccodocs SYSTEM "http://www.uga.edu/tobaccodocs/tobacco.dtd">
<tobaccodocs>
    <document sample="" decade="19" isource="" class="">
        <metadata>
            <bates_start></bates_start>
            <bates_end></bates_end>
            <uga>
                <note></note>
                <date></date>
                <pages> </pages>
                <words> </words>
                <section amount="all"></section>
                <encoded_by></encoded_by>
                <verified_by> </verified_by>
                    </uga><external></external></metadata>
        <docdata>
            <maindoc type="text">
                <pretext></pretext>
                <text type="text"></text>
                <posttext></posttext>
                    </maindoc></docdata></document></tobaccodocs>
```

Figure 4.12: Example 12: XML Template File

4. Add the required metadata. The first data added to the archive were the metadata necessary to identify the file. These data came from the file itself and from the sampling data. The required data were the *sample*, *decade*, *isource*, and *class* attribute values for the `<document>` tag; the `<bates_start>` and `<bates_end>` tags; and the `<uga>` tag children `<date>`, `<section>`, `<pages>`, and `<encoded_by>`. The remaining metadata tags were left empty at this time pending later procedures.

5. Rename and save the archive file. At this point the new archive would have been valid and could be saved as a permanent file (i.e. renamed from the template file name). Following the file naming procedures used during sampling, the archive file was named by the starting Bates Number of the document with all spaces replaced

by an underscore character, given the *.xml* extension, and placed into the same folder as the image file. Theoretically, the XML archive should have the same name as the image file apart from the extension. If this was not the case, the image file name was changed to the start Bates Number of the document to match the archive file name. In this manner, sorting the folder by name would place the archive next to the image for simple tracking and indexing. As mentioned above, in the event that a file was left incomplete (which was discouraged), the archive file extension was changed from *.xml* to *.inc*.

6. Determine the document structure. The final step before entering data was to scan the document and established the general structure. In other words, the archivist read through the document and decided which of the major structural tags would be used, i.e. were there `<predoc>` or `<postdoc>` elements, appendices, `<pretext>` or `<posttext>` elements, images, forms, tables, et cetera. As well, the archivist had to determine if the document met the requirements to be be classified as a large document, in which case additional planning, structure and sampling would be required (see the *Large Document Procesures* section below).

7. Enter the data. Once the major document structure was planned and in place, the archivist then added the actual document data to the archive, placing them as necessary into the structural tags (planned in the previous step), and at the same time marking as necessary using the rhetorical tags. Although this seems at first glance to be the simplest procedure, the complexity of the documents made it the most difficult. New data environments were constantly being discovered.

8. Validate the XML. Once the initial document conversion was complete, the archivist ran the validating tool of the XML editor and checked for any structural errors.

Once this was done, the document structure was manually reevaluated for correctness of usage (i.e. were the various parts of the document correctly classified and tagged/marked).

9. Validate spelling. As a final document check, a spell-check tool was run on the document (most editors have spell checking as an option). The purpose of spell checking was not to correct misspellings, but to assist in verifying that the spelling in the archive file matched the spelling in the image file. Misspelled words found in the archive file were manually compared to the image file and corrected as necessary. Unfortunately, there was no way to make the comparison in the opposite direction. Although we could quickly verify that misspellings in the archive file were also misspelled in the image file, we could not easily verify that the correct spellings in the archive were also correct in the image.

10. Update the master archive. Once an archive file was complete and validated, it was copied to the master archive repository on the control computer (the project supervisor's computer, control number 547885). This was generally done at the end of a work session. Twice a week the master archive would be copied to each archivist work station. This kept all computers up-to-date and also served as a simple backup system.

11. Update the file record. The final task in the archivist's daily routine was to update their file record with the new file status.

12. Automated additions and formatting. Once the archivist moved an archive to the master repository it would be formatted and updated using computer a script (a small computer program). The formatting procedure converted all the archives into an indented format based on the tag hierarchy, much like the one seen in Figure 4.11. This provided nothing more than visual uniformity across the entire corpus when viewing the archive files. The updating procedure completed the count data which was left out of the metadata during archiving, namely the data for the `<words>` tag. Word

counts were produced by extracting the analyzable text from the children tags of the `<document>` tag using an Extensible Stylesheet Language (XSL) transformation. The stylesheet used for this procedure is available in Appendix D.2. Once extracted, the text data were simply divided by whitespace into tokens and counted,[7] and the count was added as data to the `<words>` tag.

4.5.2 Specific Procedures

Over the full course of conversion for the Quota and Supplemental Samples a number of questions were raised concerning specific data environments which had not been considered previously, or if they had, not with the specificity necessary to create a reliable archive. In each case, the project supervisor, Principal Investigator, and lead archivist decided on a course of action. This was implemented going forward, and as much as possible for previously archived documents. However, some environments (such as those not marked) were difficult or impossible to locate in the existing archives, and time and cost constraints prevented going back through the document images.

In many cases questions and confusions could be resolved by simply reminding ourselves of our tag set's purpose. The following is a quote taken directly from the archivist notes. It was added once questions began to arise (which was just about the exact moment we started archiving).

> The tag set used for this project is designed to preserve rhetorically significant items, not for historical or archival purposes. Thus, for example, while we are interested in the existence of headers, we are not interested in the intricacies of their typesetting. Nor, although we are interested in preserving text that receives emphasis, are we interested in how emphasis is marked (underline, bold, italics).

[7] This tokenization method is very simplistic in that it does not account for capitalization, punctuation or other features of running text. However, this is permissible and produced reliable counts because 1) only token counts were made at this stage, and 2) only data known to be running text were used. During subsequent analysis of the text data (see Chapter 5), where type counts became important, more sophisticated tokenization methods were used.

However, there were times when specific instructions were needed. It would have been better if we had more narrowly focused our attention on text environments during the planning stages of the project and had standing instructions for the archivists at the beginning of the project. However, we were not aware of the necessity of this level of specificity (our error and loss is the readers' gain). Instead, we did as much as possible to correct the existing archives, and added to the archivists' notes as the process progressed. The following are the directions and explanations for specific environments (edited for clarity and similarity of format), given in the order questions were raised. Each of them caused both retraining of the archivists and checks and/or revisions of the previously archived documents. Keep in mind that the definitions given for tags in the previous sections were not as well developed at the beginning of the process, meaning the items below are less redundant than they might seem.

1. Signatures and initials. If signatures or initials are included within the `<text>` tag, then they should have the following format wherever they occur:
   '`<note> signature: John Doe </note>`' where John Doe is the archivists interpretation of the signature.

2. Handwritten text. If text data are handwritten, use the `<note>` tag to record this fact, but otherwise consider it to be regular text if it is part of the original document. Comment marginalia is expected to be handwritten, so the note is not needed for `<marginalia type="comment">` tags.

3. Capitals. If text is written in all capitals, record it in all capitals.

4. Spaces. XML tags do not count as spaces. Be sure to use spaces as they occur between words in the document image.

5. Quotes. The `<quote>` tag is only for quotations having a typesetting format that indicates that the text is from another source, regardless of the format. 'Odd' words marked with quotation marks should be recorded just as they are in the document. As well, the

quotation marks of quoted text should be left intact in the XML (inside the <quote> tag).

6. Special characters. Most special characters can be inserted using ASCII or ANSI characters because our output is text, but the &, <, and > must be inserted as &, <, and > to keep the parser from getting confused.

7. Non-ASCII or non-ANSI characters. Use the <symbol> tag and describe them as concisely as possible.

8. Paragraph breaks. Always start any line or block of text and always mark breaks with either <p/> or <npb/>. The <p/> tag is needed any time a carriage return or series of returns is indicated. The <npb> tag is for any break other than a standard paragraph break (that is, for lines and other graphics and for large spaces).

9. Emphasis versus marked. The <marked> tag means done by hand, where the <emph> tag means done by typesetting. The <emph> tag should not be used for typesetting that indicates things like headers, or for 'quoted' things. The <emph> tag is designed to mark rhetorically significant emphasis (unusual typesetting used to draw attention to particular text). Conventional typesetting to mark headers, titles, quotes, et cetera, should not be considered for the <emph> tag. Marked text is text which has been circled, underlined, highlighted, et cetera, by hand after the completion of the original document.

10. Text boxes. Use the <npb> tag to mark the beginning and end of text boxes, and use the <note> tag to describe the text box if it is odd for some reason. Do not make a text box into an image.

11. Notes. Anything odd, confusing or questionable should get a note. We don't want to have to go back to the original document.

12. Page breaks. The <page> tag should include only items that are directly associated with, or caused by, a page break, not paragraph headers or titles unless they are repeated at the beginning of a new page and associated with text or typesetting that indicates such.

13. Footnotes and end notes. Footnote or end note superscripts should be put in the <anc> tag as an attribute. The notes themselves should be put (each individual note) in a <footend> tag with an attribute that links it to its appropriate <anc> tag. Place the <footend> tag in the most appropriate location compared to the physical layout of the document. However, do not break a paragraph to insert it. Move to the next or preceding paragraph break.

14. Bibliographies. Put the title of bibliographies in the <posttext> tag with a note indicating the number of entries and other relevant data. Do not list all the entries. Non-rhetorical.

15. Formulas. If a formula is within the normal typesetting of a sentence, or used as a common name, no tag is needed. Type it as normal text. The <formula> tag is for describing complex formulas that are difficult to represent with ASCII characters.

16. Illegible documents. If more than 25 percent of a document is illegible, record its Bates Number and give it to the project supervisor to convey to the Minnesota Repository archivist. She will look it up in the archives to see if a better copy is available.

17. Determining doc-type: <predoc>, <maindoc>, <postdoc>. The difference between doc-type is determined by the nature of the individual 'documents' found within a Bates-Number range (the range of Bates Numbers given as a document in the Snapshot set). Remember that using different doc-type tags, like <maindoc> and <postdoc>, means that the range of Bates Numbers contains what you believe to be two separate documents. If this is not the case, the different text-type tags should be used instead.

18. Punctuation with tags. As much as possible, all punctuation in the document should be included in the XML archive regardless of the tags used to indicate the rhetorical nature of the text.

19. Pretext. From the beginning of the document up to the beginning of analyzable text, to include greetings in letters and memos, but not including a title or header that precedes the initial block of analyzable text, and not including `<predoc>` data.

20. Text. Analyzable text, following the greeting in a letter or memo, but including any titles or headers directly preceding the initial block of analyzable text.

21. Posttext. All that follows the final block of analyzable text, to include salutations, references, carbon copy declarations, et cetera, but not including appendices or end notes which should be in their appropriate tags (some displacement from the physical layout of the document is permitted), and not including `<postdoc>` data.

22. Captions. Include captions in their entirety.

23. Spelling. Spelling should conform to that of the document. Use the spell checker to check your own spelling.

24. Insert. The `<insert>` tag is for editing only.

25. Page order. If in a document of ordered Bates Numbers there is an obvious case of the original document pages being out or order, and if that improper order effects the rhetorical continuity of the analyzable text, and if the rhetorical continuity can be restored by rearranging the pages to what clearly represents the author's intended order, then it should be done. Record this in a note near the beginning of the rearrangement. If, however, rearranging the data is complicated and/or offers no rhetorical advantage, leave the pages ordered by Bates Number.

26. Duplicates. Only one document with a given range of Bates Numbers should be in the data set. Within the same document, duplicate pages should not be entered. As much

as possible, duplicate documents with different Bates Numbers should not be entered into the data set (but it will be hard to find this).

4.5.3 Large Document Procedures

From the earliest stages of planning it was known that it would be necessary to make special provisions for long documents encountered in the sampling. We could not simply exclude them as it would mean possibly excluding a text type specific to long documents, and our goal was a corpus representative of the tobacco documents as a whole. On the other hand, we could not fully include them given that the sampling could have easily produced several large documents that comprised the majority of the corpus text, biasing the corpus in the opposite direction. As a remedy, Dr. Kretzschmar suggested the following in his Sampling Plan:

> Overall, the number of documents in the Reference Sample/Corpus should be about 500. We will accept entire documents of up to 2000 words, and a 2000-word segment of documents larger than that size... Selection of a 2000-word segment from a longer document will also be accomplished by a randomized process. 2000-word segments should to the extent possible consist of coherent sections of text. The investigator will begin the segment at a heading, subheading, or paragraph break in the text, and end the segment at the completion of the paragraph or other text unit closest to but not exceeding the 2000-word limit. The investigator will draw a number from 1 to 4, and then will select the first available heading, subheading, or paragraph of text in the quartile of the document corresponding to the number drawn, to begin the 2000-word segment. (from Appendix A)

Of course, what we were not aware of in the beginning was the structural complexity of the documents themselves. Consequently, although based on Dr. Kretzschmar's proposal, our large document procedures had been adjusted somewhat prior to implementing. Overall, any doc-type element (`<maindoc>`, `<predoc>`, `<postdoc>`, or `<xdoc>`) with more than 2,500

words of <text> data was considered a large/long document. With these long elements, a total of approximately 2,000 words were entered into the archive. These were sampled from three parts of the <text> data, its beginning, middle, and end. All total, this was done for 48 documents in the Quota Sample and three in the Supplemental Sample.

The following are the actual procedures taken from the archivists' notes (edited for clarity and similarity of format):

1. Introduction. To keep long documents from having undue representation in the overall corpus, our goal is that no more than 2,000 words of analyzable text be entered from any one document (more realistically, between 2,000 and 2,500). Here are the procedures for determining what text to enter for the <maindoc> tag:

2. Estimate word count. Single-spaced, 10-point type yields 400 to 500 words per page, so any document with a <maindoc> element of 4 or more pages of text should be examined to see if it contains more than 2,000 words of analyzable text. This is done by estimating the average number of words per line, and the average number of lines per page, and multiplying the product of these with the number of pages of text. If the resulting estimation of words is near 2,000, but definitely not over 2,500, then the <maindoc> tag data may be entered as normal. Otherwise, continue with the large document procedures.

3. Archive structure. For <maindoc> elements with more than 2,500 words of analyzable text, text will be gathered from three locations: beginning, middle, and end. This is to be recorded in the following tag structure:

```
<maindoc type="text">
    <pretext></pretext>
    <part section="begin" bates="number">
        <text type="text">PCDATA here</text></part>
    <part section="middle" bates="number">
        <text type="text">PCDATA here</text></part>
    <part section="end" bates="number">
        <text type="text">PCDATA here</text></part>
```

```
    <posttext></posttext>
</maindoc>
```

Notice that three <part> tags with corresponding section attributes take the place of the <text> tag normally found in the <maindoc> tag (three are required if the <text> tag is missing). Data are entered, instead, within a <text> tag inside the <part> tags. Also, the tag attributes are required. The *bates* attribute is the Bates Number for the page where the sample begins.

4. Sampling counts. 2000/3 = 666.66 or 667 which is rounded to 700. Gather 700 words from each of the above locations (approximately 2,100 words) using the procedures below.

5. Beginning text. To gather data from the beginning of the <maindoc> element, estimate the number of lines of text needed to yield 700 words. From the beginning of the <maindoc> analyzable text, enter the estimated number of lines of text in the <part section="begin"> tag, plus any lines needed to reach a paragraph break (that is, end at a paragraph break).

6. Middle text. To gather data from the middle of the <maindoc> element, estimate the mid position of the <maindoc> text, and the number of lines needed for 350 words of text (half of the 700 from above). From the middle position in the text, count back the number of lines needed for 350 words. From this position, move back (toward the start of the document) to the nearest paragraph break, but not overlapping the data entered as beginning text. Enter the estimated number of lines required to yield 700 words in the <part section="middle"> tag plus any lines needed to reach a paragraph break (that is, begin and end with a paragraph break).

7. End text. To gather data from the end of the <maindoc> element, estimate the number of lines of text needed to yield 700 words. From the end of the <maindoc> text, count back the required number of lines. From this position, move back (toward the start of

the document) to the nearest paragraph break, but not overlapping the data entered as middle text. Enter the text from this point to the end of the `<maindoc>` text into the `<part section="end">` tag.

8. Appendices. If a large/long `<maindoc>` element has an appendix, sample the `<maindoc>` element and not the appendix. However, include the `<appendix>` tag and a description of the content of the appendix. You will have to do this using a `<note>` tag in the following format:

   ```
   <appendix><text><note>put note here</note></text></appendix>
   ```

   If a normal `<maindoc>` element (not long) has a long appendix, enter the `<maindoc>` element as normal, to include `<pretext>` and `<posttext>` as needed, beginning from the start of analyzable text, to the point that 2000 total words of analyzable text is reach. Here, 'total' means analyzable text from both the `<maindoc>` text and `<appendix>` text. Be sure to include `<posttext>` data even if the entirety of the analyzable text is not entered.

9. Other doc-type elements. `<predoc>`, `<postdoc>`, or `<xdoc>` elements (which are essentially separate documents from the `<maindoc>` even if included in a 'document' defined by Bates Numbers) are sampled as if they were `<maindoc>` data. That is, you may have several doc-type elements in a set of Bates Numbers that have more than 2000 words. Sample and record each doc-type element. This means that an archive may up to 2,500 words per doc-type element.

4.5.4 Validation Procedures

Early in the project the plan was that each archive file would be validated (checked for structure and content) by a second archivist once the initial conversion of files was completed. The intent was to do a whole-file validation in which the secondary archivist would essentially

duplicate the process of the first archivist apart from data entry, and then resolve any discrepancies found. The expectation was that the overall reliability of the corpus would greatly increase given the second pass. However, as we began the conversion process it quickly became apparent that time and budgetary constraints would not permit whole-file validation of every document. As an alternative, we relied on a number of partial procedures for secondary review of the archive files. These were primarily of two types, those done by the archivists, and those done in an automated or semi-automated manner using computer scripts (small computer programs written to perform narrowly-defined tasks).

In terms of archivists' validation, at the core of all questions related to 'correct' document structure and tag content is whether or not the archive conforms to the Document Type Definition (DTD). This is because the DTD prescribes the allowable tags, the tag hierarchy, and the tag content as it relates to structure (i.e. the DTD has no bearing on the language itself, only the cataloging of the orthography). This being the case, the most significant tool/method for checking structure is a validating XML editor. This is easily overlooked because it is a normal part of XML processing, but using a validating editor to create and maintain an archive guarantees that the archive structure conforms to that of the DTD. In the context of an entire corpus, this provides a fixed root structure for every archive document, without exception, which for reliability is an invaluable advancement. Thus the primary validation procedure for the archivist was simply to use the validation feature of the XML editor. It provided a basis for all other checks.

The other validation done by the archivist was a limited form of whole-file validation. During the course of the archiving, the project supervisor would periodically select documents which had been converted by the junior archivists and pass them to the lead archivist for review. If questions were raised in relation to tagging choices, the Principal Investigator, project supervisor, and lead archivist would decide the correct course of action and provide additional training to the archivists. As well, archivists would be asked to return to their own work (all converted documents) and make revisions as needed. In most cases, we were

able to assist the archivists in making revisions by locating potential errors and/or correcting them in a semi-automated manner using computer scripts. For example, when the definition for the `<quote>` tag was revised, a script was used to locate all instances of the tag so that they could be examined and repaired. If simple mechanical errors were found during secondary validation, a similar processes was used to correct them: additional training and semi-automated revisions. We were able to go through this type of secondary validation with 33 documents in the Quota Sample and two in the Supplemental Sample. Although the numbers were not high, the results were good, showing a general decrease in error rates across time. That is, the later checks produced fewer discrepancies than the earlier.

Although the validation done by the archivists formed the basis for all secondary checks, the bulk of the validation work we were able to do on the TCD archives was accomplished with the help of computer scripts. Even with whole-file validation, repairs to the archive files were generally made with the assistance of scripts used to locate error-prone data. Overall, the scripts took advantage of the fact that XML prescribes a very regular structure denoted by fixed sequences of characters. Because of this, it was a relatively simple matter to use a computer to locate specific data types in individual files across the entire corpus, and then present them to the user in a very focused format. In other words, the computer did the bulk of the labor by locating the data and not requiring the archivist to reread the entire corpus. Once located, they could quickly be checked for regular format, compared to other data, replaced, and/or displayed as needed. In this manner, the entire set of archives, for both the Quota and Supplemental Corpus, could be reviewed and repaired in a matter of minutes rather than days or weeks, and with a much higher degree of accuracy.

Over the course of the project there were quite a number of scripts written to address and correct specific issues as they became known, and there were often several versions of the same script. Rather than describe each script, which would become overly tedious (given that this chapter already seems to have been written by Dogberry), I will describe the basic script functions, which can be placed into a small number of groups, namely six:

1. Mechanical checks. Perhaps the simplest function of the scripts used for validation was to correct 'mechanical' issues, meaning those items not related to content, items such as spaces, punctuation, special characters, and XML declarations. In particular, there were some problems early in the archiving with the archivist leaving out spaces that preceded the beginning location of a tag (i.e. assuming the tag would separate two words). In this case a script was used to locate all tags which were not preceded by a whitespace character and, at the discretion of the reviewer, insert a needed space (some tags, such as `<illegible>`, may occur inside a word). Similarly, scripts were used to locate non-ASCII characters in the archives that were causing errors during XSL transformations and replace then with XML escape sequences; to insert and modify XML tags, attributes, and declarations when revisions were made to the DTD; and to check and repair the placement of certain data, such as moving quotation marks inside of associated `<quote>` tags.

2. File path and name checks. According to both the sampling and archiving procedures, both the archive file name and storage directory path have significance, being derived from the starting Bates Number, the sampling stratum, and the sampling set, respectively. This being the case, checks were regularly made to insure these conventions were maintained. In particular, comparisons were made between the file name and the `<start_bates>` child tag of the `<metadata>` tag (which should match), and between the storage directory and the *decade* attribute of the `<document>` tag (which should be parallel). As well, checks were made to insure that the archive file had a companion image file with the same name. For these types of checks, discrepancies were recorded in an error log and then checked and repaired manually.

3. Intra-file data checks. Another function of scripts which was used regularly for validation took advantage of the fact that there were a number of checks that could be made by examining the path, name, and content of the archive file itself. For example, the page count found in the `<pages>` child tag of the `<metadata>` could be validated by

counting the number of <page> tags found as children of the <docdata> tag, and by calculating the difference between the <bates_start> tag and the <bates_end> tag. All three should match. When inconsistencies were found, a log was printed noting the error, the file, the file path, and the archivist's name. Once the error log was produced the archivist would return to the file and make corrections as necessary (there were some cases where inconsistencies related to Bates Numbering were correct because of inserted or deleted pages). Following a similar procedure, internal checks were also made between the *decade* attribute of the <document> tag and the <metadata><uga><date> tag (first three digits of the year should match), and between the *amount* attribute of the <metadata><uga><section> tag and <part> tags in the <docdata> tag. These types of checks and repairs were made repeatedly during the archiving process until all discrepancies could be accounted for.

4. Extra-file checks. Scripts were also used to compare the archive metadata to outside data sources. In particular, following the completion of the preliminary conversion of Quota and Supplemental Samples, the values of the *decade*, *isource*, and *class* attributes of the <document> tag for each document were compared to the data gathered during sampling. Just as with the intra-file checks, a log of discrepancies was produced and used as a guide for making the necessary corrections. In general, noted discrepancies were repaired by returning to the document image, reevaluating the classifications, and updating the attribute values as needed.

5. Verifying quotas. Once the metadata had been validated within the file and against the sampling data, it was then possible to determine if the sampling frame had been followed. This was done by classifying each archived document according to its <document> tag attributes and then making counts of these classifications for each storage directory. Because the storage directories paralleled the decade strata in the sampling frame, which in turn matched the *decade* attribute, it was possible to reverse the process and verify that the established quotas had been met. In other words, the

quotas assigned by the sampling frame should have matched the class counts in the storage folders. Although this may seem somewhat circular, what prevents it from being so is that the archivist was not required to assign the same class or decade to a document that was given during sampling, meaning that theoretically each document was evaluated at least twice, once in sampling and once in archiving. As an example, in the Supplemental Sample archive, document Bates Number 510976569 was removed from the 1980 stratum, and document Bates Number 2047319752 was added to the 1990 stratum. Both documents were of the class *UE*. However, during the archiving and validation processes the decade of document 510976569 was determined to be 1980 rather than 1990, leaving an overage in in one decade stratum and a shortage in the other.

6. Tag review. The most labor intensive validating process used on the completed archives was an exhaustive tag review. This was the final review of the sample archives that we were able to complete. For each tag type in the project DTD that allowed PCDATA, the content of each realization of the tag was extracted from every file in the Quota and Supplemental Sample archives and printed to a single text file. Also included in the output file was a small amount of surrounding text, as well as necessary file reference data for each entry. The format of the text file itself, and the fact that the target data were separated from the other file text, allowed the archivists to focus attention on the tags in question and rapidly scan through the entries to locate potential errors. This was much like scanning through a Key-Word-In-Context display looking for environmental variations around a word, although the entries were often much longer. For example, Figure 4.13 shows a partial printout from the tag review file for the `<quote>` tag. Using the printout, the data lines could be quickly evaluated (notice that all the start tags are aligned on the page) for mechanical errors such as placing quotations marks outside the tag, or for content errors such as using the `<quote>` tag to mark emphasis. As errors were found, the file reference data could be used to locate and correct them manually in

the archive. In the example, it can be seen that file 570320909.xml appears to have an error given that there is nothing indicating why the <quote> tag was used, no quotation marks or notes. After returning to the document image, it was determined that this was in fact an error. In the final archive file, quotation marks are included inside the <quote> tag. Similar items can be see in file atx040209930.xml, which had a total of 34 <quote> tags. The second quote from this file is missing the closing quotation marks, the fifth needed to be checked to be sure it was not a form of emphasis, and the sixth has an oddly placed space prior to the closing quote. Each was checked against the document image and corrected as necessary.

```
FILE: 1005133325.xml   - n is sobering: <p/> <quote>"My previous concern about this study
has been confirmed by the finally published article, which so completely fails to bear out
the claims announced at the American Cancer Society's press conference last
February."</quote></text> <posttext

FILE: 570320909.xml    - pport the view that <quote>listeriosis is not so much a rare disease
as <page>-3-</page> a rarely recognised one</quote>. Mair considers

FILE: atx040209930.xml - ociety, complained: <quote>"We may be somewhat prejudiced, but
cigarette smoking is given very light treatment as a problem; we believe it should have been
listed as one of the major health problems in this report."</quote> <p/> <quote>"Lig

FILE: atx040209930.xml -  </emph> is quoted: <quote>"...The whole society has a stake in the
health of the individual...the non-smokers subsidize those who smoke. </quote> The letter of tr

FILE: atx040209930.xml - tal on page 11 says <quote>"studies show that youngsters who once
urged their parents not to smoke have themselves become cigarette smokers."</quote> <p/>SMOKING
is d

FILE: atx040209930.xml - ph on page 19 shows <quote>"rates of first heart attack"</quote>
according to amo

FILE: atx040209930.xml - gar and cholesterol <quote>"take a large toll"</quote> from the middle-

FILE: atx040209930.xml - es pointed out that <quote> "Degree of use of coffee, tea and tobacco
is largely self-selected, not randomly selected, resulting in biased usage subgroups that
present special problems in hypothesis testing. "</quote> <p/> In more col
```

Figure 4.13: Example 13: Tag Validation Printout

The use of scripts for validation allowed us to reach a very high level of consistency, near 100 percent for the types of structures that could be checked. This was a great advance over using archivist validation alone, and it added significantly to the reliability of the resultant corpus. This is particularly the case when one considers the number of errors encountered and repaired, and the fact that most would have remained in the corpus without the use of

scripts. Our first major run of a validation script produced over 200 inconsistencies between the file metadata, file naming, file path structure, and sampling data. As well, the mechanical fixes produced an unspecified number of repairs that would have caused the concatenation of a large number of words if left in the corpus. And finally, the tag review scripts were an overwhelming success in improving not only the consistency of tag structure, but also the consistency of content.

It must be kept in mind, however, that although the scripts used for validation increased the overall reliability of the corpora, script-assisted validation in and of itself is incomplete in a number of ways and is often misleading. Of particular concern in our case is that scripts tend to shift the focus of validation from content to structure. Scripts are particularly well suited for locating and evaluating character sequences (i.e structural and mechanical issues), and the temptation is to consider an archive validated once these types of things are checked and corrected. The problem with this is that the primary focus of the Tobacco Document Project DTD is capturing the rhetorical and/or pragmatic value of the document content, which apart from the tag review (which was script-assisted) was largely ignored during the script-validation processes. Another misleading aspect of our script-assisted validation is that it was primarily discrepancy-based. In other words, potential errors were located because an expectation was violated, something that should have matched did not. While this was certainly a reasonable means for detecting errors, the fact is that most of the data remained unchecked because either there was no simple means to set up the required expectation, or the expected value was false. For example, scripts were used detect the existence of text in the `<maindoc><text>` tag, but there was no means to describe what text should have been there, and consequently no means to determine if it had been correctly entered. In tags with less complex data that did have expected values, such as the *decade* attribute, the fact that the tag content matched the expected value was no assurance that either was correct and not in need of review. These situations, which comprised the majority of possible checks, were necessarily overlooked by the scripts. In the end, what we can say is that the scripts

were a great asset, and they contributed greatly to the overall reliability of the archives. What we cannot say is that we were satisfied with our validation process. There were too many data types that we were not able to check. The optimal case would be to use scripts in conjunction with whole-file validation by the archivists; scripts to validate structure and mechanics and help the archivists focus their attention on specific data types, and archivists to evaluate the full document content.

4.5.5 Coding Reliability

Once the archiving process had been implemented and was functioning regularly (i.e. following the first release of the project DTD and initial archivist training), it was decided that a formal study of inter-coder reliability should be conducted to establish a baseline measurement of consistency (agreement). We had no idea at that point how the coding practices of the archivists compared to a DTD-based ideal or to each other. We felt this would be necessary to evaluate the resultant archives and also to properly interpret future comparisons. At the time the study was conducted (it was completed July 12, 2002), there were two archivists working on document conversion and between them approximately 60 documents had been converted.

To make the comparison, the last ten documents converted by the lead archivist (A1) were selected and assigned to the second archivist (A2) to be independently re-coded. The documents were the following: 2001216666.xml, 2025027475.xml, 2047405735.xml, 2051011027.xml, ATX040085814.xml, ATX080011761.xml, HK0042150.xml, SF0823867.xml, TIMN0109658.xml, and SF0823867.xml.

In general terms, once the recoded set was complete, the project supervisor compared the XML data from both sets to the original document images. From these comparisons, the project DTD, the tag explanations, and the archivist instructions, a third set of XML documents was produced: the ‘standard’ set. This set represented what the project supervisor

believed to be the best fit of the project XML to the test documents. At this point, the A1 and A2 XML documents were compared to the standard set, and the differences were noted.

More specifically, only the <docdata> tag and its children were examined for the comparison, (although the <metadata> tag content was examined for errors related to automated entries). The reason for this is that the <metadata> tag content was fixed by definition and not dependent on decisions made by the archivists. Using a script, the <docdata> content from the three sets of XML files was reduced to lists of opening tags to facilitate comparison. The <p> tag children of the <pretext> tag were excluded from these list because by definition they are optional in this location. This was also the case for the <note> tags, but in all locations.

Each of the three document sets consisted of approximately 42,000 characters containing 4,500 words of analyzable text. For the standard set, the <docdata> tags and their children totaled 337 tags (tag pairs), 74 of which occurred outside of a <text> tag, leaving 263 text-related tags.

At this point all differences between the archivists' files (A1 and A2) and the standard set were counted. A count of one 'difference' was made for each deviation from the standard set – a missing tag, an extra tag, a misused tag, a tag order differences, or a tag attribute differences – with the exception that the attributes and children tags of a missing or extra parent tag were not counted given that the difference (or error) was in the inclusion or exclusion of the parent tag. For example, if a archivist included a <table> tag when it was not called for, the difference count increased by only one. That is, it did not increase by one for every tag in the standard set that occurred in place of the <table> tag, and by one for every child tag of the <table> tag in the archivist's document.

The counts for the noted differences are summarized in Table 4.2, and corresponding percentages of the total count given in Table 4.3. The rows correspond to archivists A1 and A2. The column labels defined as follows:

- TD. Total differences counted. Any difference between the standard set and the given archivist's set, a missing tag, an extra tag, a misused tag, a tag order differences, or a tag attribute differences.

- TX. Text tag differences. A subset of the total differences (TD) that occurred as a `<text>` tag or its child. In other words, all tags directly related to the analyzable text of a document.

- TXR. Text recovery differences. A subset of the total differences (TD), also a subset of the text tag differences (TX), which represented differences that caused a change in the type or amount of text recovered for analysis using XSLT and the text extraction stylesheet (as it was defined at the time of the study). In other words, these differences directly effected our ability to recover and analyze text.

- P. Paragraph tags. A subset of the total differences (TD) that were caused by inclusion or exclusion of the `<p>` or `<npb>` tags. Generally not significant for for text analysis. This is noted separately because it accounted for over 30 percent of TX differences.

- NTXR. Non-text-recovery differences. A subset of the total differences (TD) that should not cause any significant changes in the analysis of text (in any tag). This includes differences in tags such as the `<p>` and `<npb>` tags (P above), the `<page>` tag, additional `<illegible>` tags, missing `<posttext>` tags when no children tags or PCDATA were needed, slight order differences involving `<image>` tags with no `<text>` tag children, the combination of `<margin>` and `<illegible>` tags, `<margin type="label">` tags, slight order differences involving `<margin>` tags, and form blanks represented by the `<symbol>` tag.

Looking at the averages, overall differences (TD) were at 11.2 percent, which is 88.8 percent agreement with the standard. Within the `<text>` tag (TX), agreement with the standard increases to 89.3 percent. Removing paragraph-related differences ($TX - P$) increased

Table 4.2: Coder Reliability: Difference: Raw Count

	TD	TX	TXR	P	NTX
A1	33	28	4	13	8
A2	43	28	1	5	17
Average	38	28	2.5	9	12.5

Table 4.3: Coder Reliability: Difference: Percent of Total

	TD	TX	TXR	P	NTX	1-NTX
A1	9.8	10.7	1.7	4.9	3.0	97.0
A2	12.6	10.7	0.4	1.9	6.5	93.5
Average	11.2	10.7	1.1	3.4	4.8	95.2

agreement to 92.7 percent, and removing all differences that would not be significant in text analysis ($TX - NS$) moves the average agreement to 94.1 percent. Finally, considering only those differences critical to text recovery (TXR), agreement was an average of 98.9 percent.

Given that this was the first formal analysis of coding, the Project Investigators were pleased with the results, particularly in relation to the recovery of analyzable text. However, in order to reduce future difference rates, the study results were used to provide additional training to the archivists, to include returning to the test documents and making corrections so that they matched the standard set. As well, the study results were used in the design of several validation scripts to identify potential problem areas.

Chapter 5

Corpus Analysis: Using the Data

5.1 Introduction

The previous two chapters go into considerable detail describing the chain of processes that were involved in creating the Tobacco Documents Corpus (TDC). The intent was to show not only that data were gathered using fixed and principled methods, but also that were was meticulously categorized and archived. Yet even with more than 35,000 words the description is incomplete in conveying the amount of time and the collaborative effort put into the TDC during the three-year span of the project. Certainly, this labor can be justified on a number of theoretical levels. If for no other reason, the maturation of thought resulting from the process is sufficient to warrant it. However, in terms of practical value, the question of whether or not the work is justified still remains. Can the TDC and associated data be put to good use? or was the process simply an intellectual exercise? This applies not only to the TDC as a whole, but also to the individual processes. Great effort was put into the corpus to insure clear and reasonable representation, accurate classification, consistent format, and accurate transcription, and at this point it is fitting to ask whether or not the effort was worthwhile. In other words, does the TDC lend itself well to subsequent study?

I believe that the answer to this question is a straightforward 'yes,' that the TDC is a well-defined and reasonable representation of tobacco-industry documents, accurately rendered, with useful categories encoded into the archive, and that is does lend itself to further study. However, the argument and evidence for this conclusion cannot be based exclusively on the previous chapters, nor on the examination of the archives themselves. Instead, it must come from experience using the archives, which is to say that the proof of the pudding is in

the eating. Certainly the pudding recipe is necessary for understanding and replication, but success can be judged only by a taste, which for the TDC comes as a bit of analysis.

From the conception of the TDC, the Project Investigators believed that by establishing a norm of tobacco-industry documents and applying proven methods of analysis, the reader could quickly develop a general understanding of the content and trends in tobacco-industry documents, but without the extensive reading that is normally required. Of course, this is essentially the premise for this dissertation (see Chapter 1.2). To this end, early in the process we began to develop ideas for the distribution and presentation of not only the TDC archives, but also the results of analyses. As the project evolved, this presentation of the combination of archives and analyses became known as the 'UGA Tobacco Document Corpus and Toolkit,' and it is now the primary means of communicating the content of the TDC. It is available online as part of the UGA Tobacco Documents Project website (http://www.tobaccodocs.uga.edu/TDC) and has also been converted to a stand-alone application on CD-ROM (a copy of which should be located inside the back cover of this dissertation). While it would not serve the purposes of the current work to describe the inner workings of the Toolkit itself, careful description of the principles and processes behind the different analyses available in the Toolkit does make sense. It can readily provide the 'taste' of analysis necessary to illustrate the usefulness of the TDC as a practical tool.

As might be expected, computers were used extensively in the processes described below to handle all repetitive and mathematical events. This was done primarily to save time given that in corpus linguistics quantitative analysis typically involves counting and calculation beyond the capacity of a human. However, secondarily it provides a guarantee of consistency knowing that the computer will follow the algorithms tirelessly and without deviation. Of course, the computer's efficiency is dependent on the algorithm which defines the process being executed, meaning that the added speed and accuracy have no value apart from well-formed algorithms (unless one finds value in tirelessly making exactly the same mistake at a high rate of speed). In a similar manner, the researcher's efficiency in interpreting the process

output is dependent on understanding the algorithm which defines it. The importance of this chapter, then, is to clearly explain the process, first to allow the reader to determine that the algorithms are well formed, and then to properly interpret the data produced. The focus will be on the various processes (steps) necessary to move from the file archives to the presentation of data. In particular, the text data must be extracted from the archives, tokenized, parsed into constituents, counted, analyzed, and then displayed. Each of these steps will be discussed in the sections to follow, and as much as possible, examples from the Toolkit will be used for illustration.

As a point of clarification, *analysis* as it is used in this chapter actually refers to quantitative analysis, which is the manipulation and presentation of count data in order to assist the user with subsequent qualitative analysis. In other words, the goal here is to gather, process, and present data to the user in a format that allows rapid assimilation and understanding of corpus content. At this point, the user can engage in whatever qualitative analyses are necessary (i.e make judgments about the data). The hope is that the quantitative analysis will lead the user to a more informed position from which to make these qualitative judgments. It should also be noted that the sum of procedures described below is the method we used to construct a specific product for presentation via the Toolkit. While I believe that together the processes involved represent a reasonable and complete method for obtaining the project goal, there is no implication that these processes are the only (or best) method for obtaining and analyzing data, although they can serve well as a model for other studies.

5.2 Text Extraction

The first step in the overall process of data analysis is to gather text from the XML archives. One of the great advantages of using XML for archival is that it allows the archivists to record a great deal more data than the text of the document itself, such as document structure, classification data, and notes/descriptions provided by the archivists. However, as noted in Chapter 4.4, this increase in data is accompanied by an increase in the complexity of corpus

management in that it would be counterproductive to include these non-document data, as well as the tags that mark them, in the analysis. They would pollute the statistics with items that have high occurrences but little linguistic value. As well, there are portions of the document text itself, such as the `<pretext>` and `<posttext>` data, which have little rhetorical value, that for similar reasons do not need to be included in the general analysis. Thus, the process must begin with the extraction of the desired text data from the archive.

Because XML has a regular and defined structure both in terms of syntax and hierarchy, and because all the TDC archives were validated against the project DTD, using computers to automate the process of locating tags in the archive, and subsequently reading the text data from those tags, is relatively straightforward and can be approached in a number of different ways. Often the simplest method is to write a computer script to extract text from specific tags, just as was done during tag validation (see Section 4.5.4). However, in order to insure that our procedures were well defined and easily replicated, we chose to use Extensible Stylesheet Language (XSL) methods as the basis for all TDC text extraction because it represents a standard for working with XML files. As such, it prescribes a fixed syntax for defining how the content of the various XML tags in an archive should be processed and displayed. This formal description, the *stylesheet* itself, can theoretically be applied to any XML document, by any XSL transformation (XSLT) program, and return the same results. This means that the stylesheet we used to extract text for general analysis can be used by others to duplicate our process, regardless of the XSLT program used. The only caveat to this is that most XSL transformations are used for Internet services in which data are extracted from XML archives and sent directly to the network as a character string (i.e. the data have no permanence). In this respect, modifications must be made to most existing XSLT processes in order to save the character string as a text archive. We accomplished this by embedding various XSLT engines into Python scripts (which captured and saved the output text). The *Pyana* (http://sourceforge.net/projects/pyana), *libxml2* (http://xmlsoft.org/),

and *Sablotron* (http://www.gingerall.org/sablotron.html) engines all worked well for this purpose.

The major advantage of using XSL and XSLT is that it allows one to focus on more important issues. The real question at this point in the process is not *how* text is to be extracted, the XSLT engine takes care of that, but *what* text is to be extracted. In other words, the engine is much less important than the stylesheet itself, which not only defines the text output, but ultimately governs all the analysis to follow. That is, if the wrong text is extracted, then there is much less value in subsequent analyses. Using XSL allows/forces more careful consideration of the text-output definition. The major disadvantage of using XSL and XSLT, however, is that the entire process is governed by (and only by) the XML structures as defined by the project DTD. What this means is that because XSL is a fixed process, the only data that can be extracted are those that have previously (and properly) been encoded in the XML. Essentially, this is the true test of the applied/practical value of the XML schema. The hope is that the tags were initially created to compliment the Project Investigators' interests, and that useful data can be extracted.

For the TDC, the general text-extraction stylesheet was developed according to the project Investigators interest in the analysis of rhetorically significant 'continuous, non-template text' as a means for establishing a norm of communication (as defined in Chapter 3.4.2). To this end, the stylesheet was simple in design. Basically, all continuous, non-template text originating in the document itself was extracted from the document archive. This included all document text from the `<text>` and `<part>` tags of all `<predoc>`, `<maindoc>`, `<postdoc>`, `<xdoc>` and `<appendix>` tags. Excluded were all template or non-document text and data, to include data from the `<metadata>` tag at the document level, data from the `<pretext>` and `<posttext>` tags at the paragraph level, and data from tags such as `<note>` and `<description>` at the sub-paragraph level. Exclusion was accomplished in the transformation by substituting a space character for the tag content. The notable exception to this was the `<formula>` tag, which was used during archiving to mark descrip-

tions chemical formulas. Because the description represented document text, but itself did not originate in the document, the string `form_ula` was inserted in its place as a marker. This allowed subsequent processes (or readers) to locate and count chemical formulas in the data. See Chapter 4.4 for more specifics on tag content.

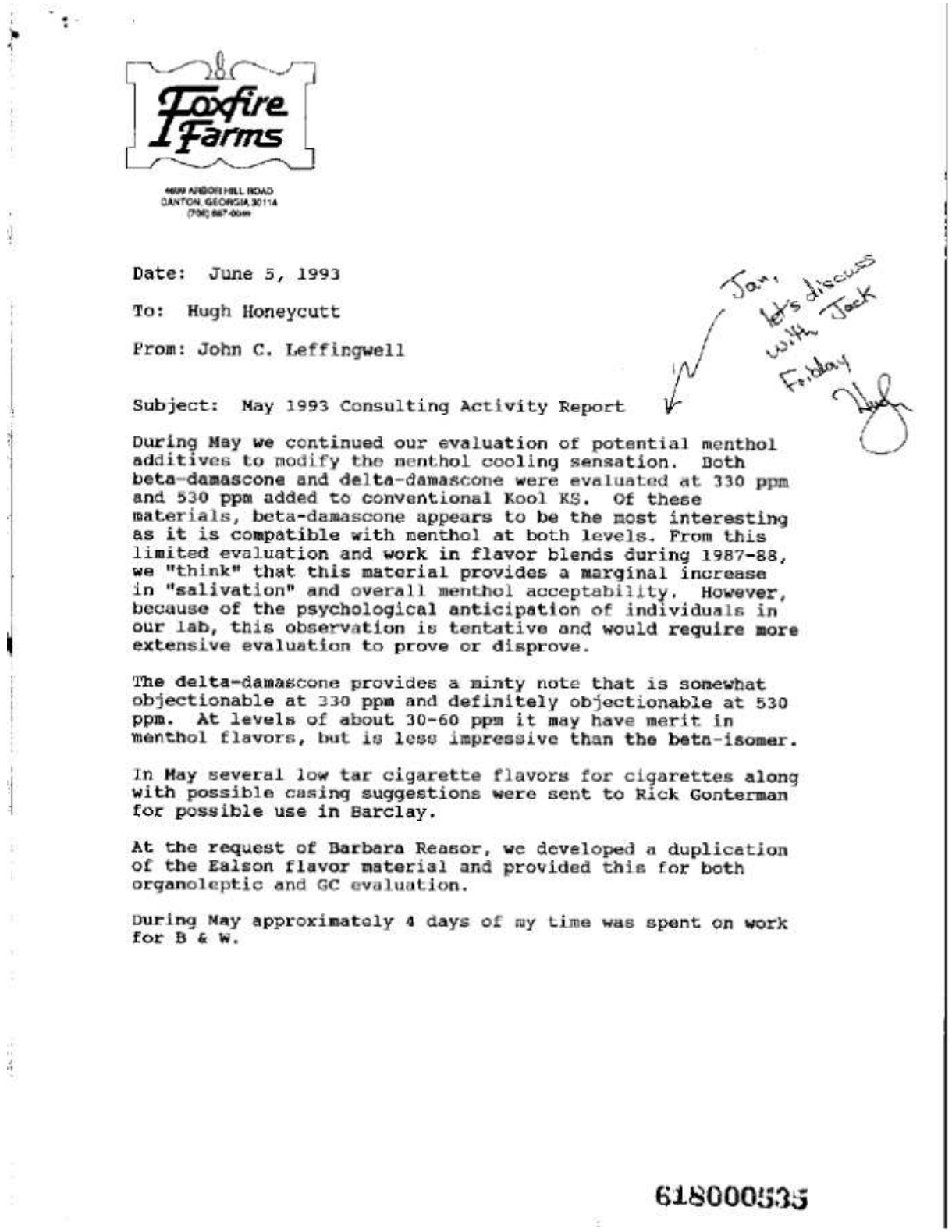
The general text-extraction stylesheet used for the TDC is provided for the reader in Appendix E.1. Those familiar with XSL will notice that this stylesheet is markedly similar to the stylesheet used for extracting text for token counts during archiving (see Chapter 4.5.1 and Appendix D.2). This is indeed the case. The counting stylesheet served as a template and testbed for the general stylesheet. When analysis began, the counting stylesheet had been in use for several months and had been through several revisions, and we were quite familiar with its output. The only notable difference between the two is that with the general stylesheet, classification data are extracted from the archive `<metadata>` tag and used to form a header that is added to the output file as the first line. This was used to identify the extracted text and to allow the archives to be re-grouped in order to form various sub-corpora. In all cases, this header was disregarded during text analysis.

The overall procedure for the extraction process was fully automated and accomplished as follows. For each archive in a given sample, the XML data were read and passed to the XSLT engine. The text string returned from the engine was 'cleaned' and written to a text file with the same name as the XML archive, but having a *txt* extension. Text cleaning was done in three parts. First, special-character sequences were replaced. Although we tried to minimize the use of multiple-byte character encodings during archiving, there were at least two scenarios that caused them to be recorded in the archives. XML requires the special sequences `"`, `&`, `>`, and `<` to represent the code characters `"`, `&`, `>`, and `<`, and the default for our archival software was to use a Unicode sequence for any non-ASCII character (the setting was not always turned off). These sequences were replaced with their corresponding single-character values. The second text-cleaning task was to regularize whitespace characters (space, tab, newline, et cetera). All sequences of whitespace characters

were first replaced with a single space (ASCII 32). It was determined prior to text extraction that analysis would not make use of whitespace except in parsing/tokenizing, so it was not necessary to preserve the original whitespace during extraction. Finally, simply to make file viewing easier, the text string was 'hard wrapped' to approximately 72 characters by replacing the nearest space character with a newline sequences (ASCII 10-13).

As an illustration of the above, Figure 5.1 is the image of a short document from the Quota Sample, 1990 stratum (Bates Number 618000535). This document was chosen as an example because, apart from being short, it contains a number of key items, notably `<pretext>` data, an arrow symbol, and marginalia. The corresponding XML archive is shown in Figure 5.2, and the extracted and cleaned text data are in Figure 5.3. Obviously, the data in the first line of the text archive between the `<note>` tags are the header. Again, these data were not analyzed with the text, but used to identify the text passage's origin and as an aide in forming sub-corpora. The remainder of the data are the analyzable text. Notice the absence of the `<metadata>` tag data (which did not originate in the document), the `<pretext>` tag data (which are not continuous, non-template data), and the `<symbol>` tag data (which are a description added by the archivist, not document data). However, the text from the `<margin>` tag is included (because it originated in the document), although it is a bit out of context. A similar text archive file was produced for each XML archive in the Quota and Supplemental Samples, a total of 908 files.

As a final note, I would like to reiterate that the files produced by the above process are not the only or correct representation of extracted text data. Although general and able to be used for a wide range of subsequent analyses, they were produced to supply specific data to a specific process. Had the analysis process or the project Investigators' interest been different, the extraction would have been different as well.



Foxfire Farms

4699 ARBOR HILL ROAD
DANTON, GEORGIA 30114
(706) 667-0000

Date: June 5, 1993

To: Hugh Honeycutt

From: John C. Leffingwell

Subject: May 1993 Consulting Activity Report

Jan, let's discuss with Jack Friday

During May we continued our evaluation of potential menthol additives to modify the menthol cooling sensation. Both beta-damascone and delta-damascone were evaluated at 330 ppm and 530 ppm added to conventional Kool KS. Of these materials, beta-damascone appears to be the most interesting as it is compatible with menthol at both levels. From this limited evaluation and work in flavor blends during 1987-88, we "think" that this material provides a marginal increase in "salivation" and overall menthol acceptability. However, because of the psychological anticipation of individuals in our lab, this observation is tentative and would require more extensive evaluation to prove or disprove.

The delta-damascone provides a minty note that is somewhat objectionable at 330 ppm and definitely objectionable at 530 ppm. At levels of about 30-60 ppm it may have merit in menthol flavors, but is less impressive than the beta-isomer.

In May several low tar cigarette flavors for cigarettes along with possible casing suggestions were sent to Rick Gonterman for possible use in Barclay.

At the request of Barbara Reasor, we developed a duplication of the Ealson flavor material and provided this for both organoleptic and GC evaluation.

During May approximately 4 days of my time was spent on work for B & W.

618000535

Figure 5.1: Example Document 618000535: Image

```
<?xml version="1.0" ?>
<!DOCTYPE tobaccodocs SYSTEM "http://www.uga.edu/tobaccodocs/tobacco.dtd">
<tobaccodocs>
    <document sample="quota" decade="1990" isource="bw" class="ni">
        <metadata>
            <bates_start>618000535</bates_start>
            <bates_end>618000535</bates_end>
            <uga>
                <note></note>
                <date>19930605</date>
                <pages>1</pages>
                <words>213</words>
                <section amount="all"> </section>
                <encoded_by>Anastasia Wright </encoded_by>
                <verified_by>NONE</verified_by> </uga>
            <external> </external> </metadata>
        <docdata>
            <maindoc type="text">
                <pretext>
                    <image>
                        <description>Foxfire Farms emblem. </description></image>4609
                ARBOR HILL ROAD CANTON, GEORGIA 30114 (706) 687-0099 Date: June
                5, 1993 To: Hugh Honeycutt From: John C. Leffingwell Subject:
                May 1993 Consulting Activity Report </pretext>
                <text type="text">During May we continued our evaluation of potential
                    <margin type="comment">
                        <symbol>arrow </symbol>Jan, let's discuss with Jack Friday
                    Hugh </margin>menthol additives to modify the menthol cooling
                sensation. Both beta-damascone and delta-damascone were evaluated
                at 330 ppm and 530 ppm added to conventional Kool KS. Of these
                materials, beta-damascone appears to be the most interesting
                as it is compatible with menthol at both levels. From this limited
                evaluation and work in flavor blends during 1987-88, we "think"
                that this material provides a marginal increase in "salivation"
                and overall menthol acceptability. However, because of the psychological
                anticipation of individuals in our lab, this observation is
                tentative and would require more extensive evaluation to prove
                or disprove.
                <p/>The delta-damascone provides a minty note that is somewhat
                objectionable at 330 ppm and definitely objectionable at 530
                ppm. At levels of about 30-60 ppm it may have merit in menthol
                flavors, but is less impressive than the beta-isomer.
                <p/>In May several low tar cigarette flavors for cigarettes along
                with possible casing suggestions were sent to Rick Gonterman
                for possible use in barclay.
                <p/>At the request of Barbara Reasor, we developed a duplication
                of the Ealson Flavor material and provided this for both organoleptic
                and GC evaluation.
                <p/>During May approximately 4 days of my time was spent on work
                for B &amp;amp; W. </text>
                <posttext></posttext> </maindoc> </docdata> </document></tobaccodocs>
```

Figure 5.2: Example Document 618000535: XML

```
<note> sample="quota" bates="618000535" isource="bw" decade="1990" class="ni" date="19930605" </note>
During May we continued our evaluation of potential Jan, let's discuss
with Jack Friday Hugh menthol additives to modify the menthol cooling
sensation. Both beta-damascone and delta-damascone were evaluated at
330 ppm and 530 ppm added to conventional Kool KS. Of these materials,
beta-damascone appears to be the most interesting as it is compatible
with menthol at both levels. From this limited evaluation and work in
flavor blends during 1987-88, we "think" that this material provides a
marginal increase in "salivation" and overall menthol acceptability.
However, because of the psychological anticipation of individuals in
our lab, this observation is tentative and would require more extensive
evaluation to prove or disprove. The delta-damascone provides a minty
note that is somewhat objectionable at 330 ppm and definitely
objectionable at 530 ppm. At levels of about 30-60 ppm it may have
merit in menthol flavors, but is less impressive than the beta-isomer.
In May several low tar cigarette flavors for cigarettes along with
possible casing suggestions were sent to Rick Gonterman for possible
use in barclay. At the request of Barbara Reasor, we developed a
duplication of the Ealson Flavor material and provided this for both
organoleptic and GC evaluation. During May approximately 4 days of my
time was spent on work for B & W.
```

Figure 5.3: Example Document 618000535: Extracted Text

5.3 Tokenizing

The next step in the progression toward analysis is to reduce the text to its smallest significant constituents in preparation for parsing and counting. In the case of the TDC, as with most corpus studies, the decision was made that the 'word' would be that smallest part. This certainly makes sense as the project is focused on high-level topics, such as content and rhetorical style. However, it is generally taken for granted that language is comprised of words, and consequently the lowly word often goes undefined. Yet just as with text extraction, the procedures for identifying words cascade down to subsequent events, which again means that this step governs all those that follow. If it is not clear what a word represents, then the meaning of any analyses or statistics derived from the words will also be unclear. In other words, to properly interpret any results, even simple counts, we must know whether or not `cancer` has the same value as `Cancer`, `CANCER`, `cancer's` or `canc.`, or perhaps even `cancers`. To this end, the procedures used to identify words in our analysis of the TDC will be detailed below.

To begin, the text constituents being located are generally referred to as *tokens* rather than words, being the output of a defined process of *tokenization*. The term *word* is somewhat misleading in that most text contains a variety of non-word character sequences which are not without value and are generally not excluded from word counts and analyses. For example, the extracted text from the example document in the previous section contains the sequences `ppm`, `1987-88`, `B`, `&`, and `W`; and jumping ahead to Table 5.3, one seas that of the roughly 2,000,000 tokens in the combined text of the Brown and the Freiburg-Brown Corpora (to be defined in Chapter 5.6.2), approximately 22,000 tokens are from non-word sequences (numbers and alpha-numeric sequences). Thus, the canonical definition of emphword, if the term were used, would have to be twisted somewhat to fit the data. On the other hand, the term *token* carries little semantic baggage, allowing the inclusion of any character sequence, word or not. Tokens are simply what the tokenization process returns, being defined strictly by the process algorithm. This being said, the present task is to explain the tokenization algorithm used in the analysis of the TDC data to the extent that its output is predictable to the reader.

The script used for tokenizing the TDC text data is actually very simple, the process being accomplished in four steps, each building on the previous: number handling, character translation, apostrophe handling, and splitting. Because the Unicode sequences are handled during extraction, all characters in the tokenizing input and output are represented as single bytes in the ordinal range 0 to 255, which is the 128 ASCII characters plus the additional 128 ANSI characters. The actual Python code used for tokenizing the TDC data can be found in Appendix E.2.

The input to the tokenizing script is a 'string' of characters, which is actually just a series of codes that represent the characters that would be seen if the string were printed on a computer screen. Once the tokenizer receives the string, it processes it using the following steps:

1. Number Handling - The first step in tokenizing is handling numbers. In general, numbers are conserved (kept as tokens) because of their importance in Tobacco-Industry discourse. The only number conversion that takes place at this point is that digit characters separated by a single comma, period, or colon character are joined by removing the separating character. This is done to keep closely-associated digit sequences together, such as decimal numbers, long numbers written with commas, times, and ratios. This is accomplished using a regular expression substitution on the character string.

2. Character Translation - Once selected number are joined, the entire string is subjected to a translation process in which each character is replaced by another according to a translation table. The exact substitutions can be found by examining the translation table in the tokenizing script, or a simplified version found in Appendix E.2. However, in general terms the results are straightforward. All alpha characters (letters) are replaced by their lowercase counterpart, numeric characters are replaced by themselves (no apparent change), and all non-alphanumeric characters (punctuation, whitespace, symbols) except apostrophes are replaced by a space (ASCII 32).

3. Apostrophe Handling - In order to avoid breaking words containing apostrophes, the ASCII 39 character (') is replaced by itself in the previous step, leaving the string unchanged in terms of apostrophes. However, at this point any sequence of an apostrophe preceding or following a space character (either an ASCII 32,39 or 39,32 sequence) are replaced by a single space using a string handling function native to Python. Practically speaking, this removes superfluous apostrophes. Only those that are within an alphanumeric sequence are left.

4. Splitting - The final step is dividing the string into parts. This was done by a Python function aptly named `split()` which divides a string into parts (tokens) based on whitespace characters, removing the whitespace in the process such that all tokens contain only alphanumeric and apostrophe characters.

The output from the tokenizing script is an ordered 'list' of strings, each representing a token. Essentially, the string provided to the tokenizer is normalized, divided into smaller stings by punctuation and whitespace, and returned in the same order it was received. As simple examples, the string '*Jan, let's discuss with Jack Friday.*' is tokenized to the list ['jan', "let's", 'discuss', 'with', 'jack', 'friday'], the string '*4609 ARBOR HILL ROAD CANTON, GEORGIA 30114 (706) 687-0099*' becomes ['4609', 'arbor', 'hill', 'road', 'canton', 'georgia', '30114', '706', '687', '0099'], and '*my time was spent on work for B & W.*' becomes ['my', 'time', 'was', 'spent', 'on', 'work', 'for', 'b', 'w']. These examples were taken from the example document shown in Figure 5.1. The full list of tokens (219 in all) from the extracted text from Figure 5.3 is shown in Figure 5.4.

```
['during', 'may', 'we', 'continued', 'our', 'evaluation', 'of', 'potential', 'jan', "let's", 'discuss',
'with', 'jack', 'friday', 'hugh', 'menthol', 'additives', 'to', 'modify', 'the', 'menthol', 'cooling',
'sensation', 'both', 'beta', 'damascone', 'and', 'delta', 'damascone', 'were', 'evaluated', 'at', '330',
'ppm', 'and', '530', 'ppm', 'added', 'to', 'conventional', 'kool', 'ks', 'of', 'these', 'materials',
'beta', 'damascone', 'appears', 'to', 'be', 'the', 'most', 'interesting', 'as', 'it', 'is', 'compatible',
'with', 'menthol', 'at', 'both', 'levels', 'from', 'this', 'limited', 'evaluation', 'and', 'work', 'in',
'flavor', 'blends', 'during', '1987', '88', 'we', 'think', 'that', 'this', 'material', 'provides', 'a',
'marginal', 'increase', 'in', 'salivation', 'and', 'overall', 'menthol', 'acceptability', 'however',
'because', 'of', 'the', 'psychological', 'anticipation', 'of', 'individuals', 'in', 'our', 'lab', 'this',
'observation', 'is', 'tentative', 'and', 'would', 'require', 'more', 'extensive', 'evaluation', 'to',
'prove', 'or', 'disprove', 'the', 'delta', 'damascone', 'provides', 'a', 'minty', 'note', 'that', 'is',
'somewhat', 'objectionable', 'at', '330', 'ppm', 'and', 'definitely', 'objectionable', 'at', '530', 'ppm',
'at', 'levels', 'of', 'about', '30', '60', 'ppm', 'it', 'may', 'have', 'merit', 'in', 'menthol', 'flavors',
'but', 'is', 'less', 'impressive', 'than', 'the', 'beta', 'isomer', 'in', 'may', 'several', 'low', 'tar',
'cigarette', 'flavors', 'for', 'cigarettes', 'along', 'with', 'possible', 'casing', 'suggestions', 'were',
'sent', 'to', 'rick', 'gonterman', 'for', 'possible', 'use', 'in', 'barclay', 'at', 'the', 'request', 'of',
'barbara', 'reasor', 'we', 'developed', 'a', 'duplication', 'of', 'the', 'ealson', 'flavor', 'material',
'and', 'provided', 'this', 'for', 'both', 'organoleptic', 'and', 'gc', 'evaluation', 'during', 'may',
'approximately', '4', 'days', 'of', 'my', 'time', 'was', 'spent', 'on', 'work', 'for', 'b', 'w']
```

Figure 5.4: Example Document 618000535: Tokenized Data

As a more involved example, Figure 5.5 is an interlinear display of some tokenized text from a 1985 document from the Quota Sample (Bates Number 2501659008). The input string is the top line of each line pair, and the tokenized list is the bottom. Careful study of this example should provide a clear understanding of how most character sequences are handled by the TDC tokenizer. Notice in particular the overall conversion to lowercase and loss of punctuation, the conservation of apostrophes in line 2, the removal of a superfluous

apostrophe in line 4, the combined numbers in lines 8, 11, and 12, and the separation of alphanumeric character sequences in lines 4, 9, 11 and 12.

```
 1: In view of its apparently tempting economic and organoleptic qualities, it is
  ['in','view','of','its','apparently','tempting','economic','and','organoleptic','qualities','it','is',

 2: worth reviewing diethylene glycol's toxicity profile. Human experience Insight
   'worth','reviewing','diethylene', "glycol's", 'toxicity','profile','human','experience','insight',

 3: into the lethality of diethylene glycol (DEG) was gained when in 1937 a new
   'into','the','lethality','of','diethylene','glycol','deg','was','gained','when','in','1937','a','new',

 4: and previously untried elixir of sulphanilamide' preparation made by the S.E.
   'and','previously','untried','elixir','of','sulphanilamide','preparation','made','by','the','s','e',

 5: Massengill Co. containing 72% DEG killed 105 people (Calvery & Klumpp, 5th. med. J.,
   'massengill','co','containing','72','deg','killed','105','people','calvery','klumpp','5th','med','j',

 6: Nashville 1939, 32, 1105). The lowest total dose of the Massengill elixir
   'nashville','1939','32','1105','the','lowest','total','dose','of','the','massengill','elixir',

 7: reported to cause death in the children involved, the youngest aged only 7
   'reported','to','cause','death','in','the','children','involved','the','youngest','aged','only','7',

 8: months, was 5 ml (3.6 ml DEG); total dose in the adults that died ranged
   'months','was','5','ml','36','ml','deg','total','dose','in','the','adults','that','died','ranged',

 9: from 20-240 ml of the elixir (14-170 ml DEG). No dose-response data are
   'from','20','240','ml','of','the','elixir','14','170','ml','deg','no','dose','response','data','are',

10: available, but a cumulative dose of 14 ml DEG in an adult weighing 60 kg
   'available','but','a','cumulative','dose','of','14','ml','deg','in','an','adult','weighing','60','kg',

11: in equivelant to a total intake of about 0.23 ml/kg; the average fatal dose
   'in','equivelant','to','a','total','intake','of','about','023','ml','kg','the','average','fatal','dose',

12: in adults was about 71 ml DEG, or 1.2 ml/kg (i.e. about 1.3 g/kg ).
   'in','adults','was','about','71','ml','deg','or','12','ml','kg','i','e','about','13','g','kg',

13: Some of the survivors tolerated much higher doses.
   'some','of','the','survivors','tolerated','much','higher','doses']
```

Figure 5.5: Example Document 2501659008: Inter-Linear Display

As a final note, it bears mentioning again that the tokenizing method presented here is a method of obtaining analyzable data, but certainly not the only or even the best method. It suited the needs of our analysis, given the nature the our archive, which was accurately rendered. In particular, the tokenizer presented here is simple and straightforward. However, the fact that it is simple also means that it may not be well-suited for general use where more irregular data are expected (as with OCR data), although it can serve as a model.

5.4 Parsing

The next major step towards analysis is to identify the structures in the tokenized text that are to be counted, which is generally known as parsing. This is necessary given that quantitative analysis, as simple and redundant as it might sound, is based on count data. Although the tendency in parsing is to focus on the input string or output structure, the key element in parsing is actually the rule set used to identify structures. In other words, parsing is a hunt for known structures (not unknown ones), and it is the rule set that defines what is known. If the rule set is not well defined, then the output and all subsequent analyses are unreliable because they are equally undefined.

In Linguistics, parsing is most often associated with formal grammars (which are rule sets). For example, the following is a simple set of grammar rules having a form similar to Definite-Clause Grammar (DCG) rules, which in turn are not unlike Phrase-Structure Rules:

```
s   --> np,vp.
np  --> [d],[ap],n,[pp].
np  --> pro,[pp].
pp  --> p,np.
ap  --> [adv],adj.
vp  --> v,[np].

adj --> [higher].
adv --> [much].
d   --> [];[the].
n   --> [survivors];[doses].
p   --> [of].
pro --> [some].
v   --> [tolerated].
```

As an illustration of parsing, these rules can be loaded into the online version of the Student PARSing Environment II (SPARSE II, Darwin 2001), and when given the list of tokens ['some', 'of', 'the', 'survivors', 'tolerated', 'much', 'higher', 'doses'], a parsed output is returned (shown in Figure 5.6). What can be seen in the output 'tree' is that the input represents a known structure *s*, which has as constituents the known structures *np* and *vp*, and so forth until the token level is reached, tokens being the lowest level of known structure (defined by the tokenizing algorithm). Of course, the rules are not fixed, and as they are

changed the expectation is that either the parse results change, because a different structure has been identified, or the parse fails because no known structure is found. Both are successful outcomes. Problems only arise when structures are misidentified.

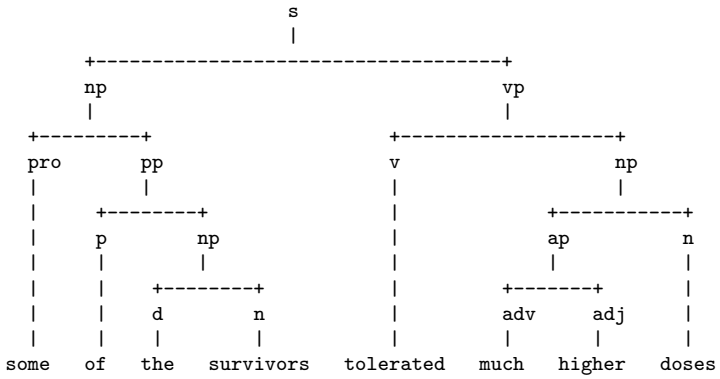


Figure 5.6: Parsing Example: SPARSE II Display

Past studies have shown that there are any number of linguistic structures which might be counted. However, these same studies have also shown that clearly defining these entities (specifying the parsing rules) is often problematic, particularly as the structural complexity increases. An example of this is Douglas Biber's 1988 study of text registers. In my opinion, this study is monumental as a proof-of-concept for the application of statistical analysis to corpus study, and it has forever changed the field. However, it also suffers in that many of the 67 phrasal structures counted for the analysis are very complex and not well-defined in the report (Chapter 4). This leads to a decrease in the reliability of the resultant data. In this case, the problem is not that the constituents were wholly undefined, or that the definitions were inconsistently applied (the analysis was done by computer, which requires the former and insures the latter), but more that the reader is left unsure what the definitions represent, which means there is no clear way to judge the reliability of the parsing. One does not know the extent to which multiple definitions were given for the same structure (causing overcounting), or which structures were overlooked (causing undercounting). This is not to say that Biber was negligent in what he did, not at all, but that the task of reliably parsing 67 different structures is overwhelming. Even the simple set of DCG rules above is questionable

in terms of counting. If the interest is in the occurrence of *np*, is it correct to count *the survivors* at two levels of the hierarchy? In other words, is an *np* with an embedded *np* equivalent to a simple *np* with no embedded structure? Probably not, but maybe, depending on the purpose of the exercise. Ultimately, it is a lack of specificity in relation to parsing that hinders the reliability (or perhaps interpretability) of the results, and just as with text extraction and tokenizing, all that follows this step is governed by it. Certainly this is true for procedure, but also in the understanding of procedure. This is again a return to category, which must be well defined to allow reliable classification.

In their work on statistical taxonomy, Sneath and Sokol (1973) address similar issues. Although not dealing specifically with parsing, they return regularly to the importance of defining specifically what is to be counted. Their recommendation, as a means to avoid as many of the definition pitfalls as possible, is to always use the most transparent and least controversial marker for count data (147). Their emphasis is that the researcher must know what the marker represents (i.e. it must be definable), which is sound advice. What cannot be clearly defined, cannot be reliably counted.

In the case of the TDC, the purpose behind the analysis of the corpus relates directly to the original proposal of this work: to develop a straightforward and reliable method for approaching the document set that allows the researcher to quickly obtain a general understanding of the corpus document types, content, and events. In other words, the focus of analysis is not the study of structure (grammar), but instead the understanding of content, and from early in the project we believed that this understanding could be accomplished through the analysis of tokens, with little or no traditional linguistic structure. In other words, our 'most transparent and least controversial marker for count data' is the token. Initially, our thought was that only individual tokens would be used. However, we quickly realized that the addition of a small amount of context dramatically increased the interpretability of the data, and we began to examine simple token combinations as well. Bringing a short

end to a long explanation, the counts and analyses presented below are based on two simple, non-linguistic structures: tokens and collocations.

Just as one would expect, a token is defined as any single item (token/string) from the output list of the tokenizing script described in Section 5.3 and shown in Section E.2. This being the case, parsing tokens consists only of cataloging the tokens from the tokenizer output. Collocations are slightly more complex. Within the ordered list of tokens produced by the tokenizer, a collocation is defined as any node-collocate pair, where the *node* is any token from the list, and the *collocate* is any token that occurs as the first, second, or third list element to the right of the node, such that any list of N tokens contains $3N - 6$ collocations. For example, the list ['the','cat','in','the','hat'], which contains 5 tokens, also contains $15 - 6 = 9$ collocations: 'the /3 cat', 'the /3 in', 'the /3 the', 'cat /3 in', 'cat /3 the', 'cat /3 hat', 'in /3 the', 'in /3 hat', and 'the /3 hat' (where the */3* denotes that the collocate is within three places of the node). Note in particular that order is retained, such that 'the /3 in' is differentiated from 'in /3 the'. A maximum distance of three was chosen based on observation and prior studies of phrasal verbs (a form of collocation) that show a distance of 3 places from the node is outside the third standard deviation, noting a rapid decrease in semantic relation between words beyond this point (Gray and Darwin 2001). Given that the definition of a collocation is more mechanical (mathematical) than linguistic, parsing is straightforward and can be validated by the checksum $3N - 6$.

The TDC Toolkit actually offers data on two additional structures: 2-grams and 3-grams. Whereas a token is also an n-gram of length 1 (a 1-gram), these additional structures can be defined very simply as a continuous, ordered sequence of two or three tokens respectively. They differ from collocations in that the structures imply adjacency, although the sequence of tokens in any bigram or trigram would also produce collocations. The checksum of counts for any n-gram of length n given a list of N tokens is $N-(n-1)$ grams. For the sake of space, and given that tokens and collocations are sufficient to illustrate the various processes/steps in the analysis of TDC data, these are not discussed further in this work. These structures

do, however, add context to individual tokens in a manner not seen with collocations, and the reader is encouraged to experiment with these structures by using the Toolkit itself.

The end result is that by using minimal structures (the least complex allowable for the desired analysis) the typical problems associated with parsing, which relate to well formed definitions, are bypassed. The parsing itself is not debatable, although the choice of structures may be. There are, as mentioned above, many other structures that might have been examined, and could still be, but as I hope to demonstrate below, the rebuttal is simply that these were sufficient.

5.5 Counting and Count Data

Although it may seem a bit odd to have a section on counting, once the process is considered a number of questions arise. It has already been decided *what* is to be counted (tokens and collocations), but still unexplained are the issues of *how* they are to be counted. In particular, there are questions related to counting methods and the interpretation of count data; how traditional text boundaries are handled; and how the various corpora and sub-corpora are assembled from the individual archives. The importance of this, once again, is that these procedures govern those that follow, and understanding them is necessary to interpret the final data.

5.5.1 Types and Tokens

To begin, a distinction must be made between *type* and *token* in relation to counts. For counting, a *type* is a unique, defined idea, while a *token* is an occurrence of a type. Another way to put it is that a *type* is a category (as defined in Chapter 2), and the *token* is the classified item, the thing which can be found and counted. For example, in the list ['some', 'of', 'the', 'survivors', 'tolerated', 'much', 'higher', 'doses'], if the type is *word*, a category defined as an alpha-character sequence, then there are eight tokens, the eight items which fit the definition. Likewise, if the type is *number*, defined as a digit sequence, then there are

no tokens. However, if the type is *np*, defined by the DCG rules from the previous section, the above parse indicates that there are three tokens (i.e. three instances of the type *np*), and this is where it can get a bit confusing as the term *token* is not defined the same for counting as it is in tokenizing and parsing.

In tokenizing, the term *token* refers both to the category, which is defined by the tokenizing algorithm, and to the items in the category, which are contained in the output. That is, tokens are of the type or category *token*. Although this is confusing, what causes the most trouble is that tokenizing carries the implication that the algorithm (definition) produce output that represents the smallest significant constituents of the input. This is not the case in counting where a token is an occurrence of any type, no matter how complex. Yet because tokenizing, parsing, and counting go hand-in-hand, one often finds ambiguous terminology, such as *token types* and *total tokens*. This bears clarification.

In this work, unless noted otherwise *token* refers only to an element from the output of the token parser. Practically speaking this is the same as the output of the tokenizer, but theoretically tokenizer output is unparsed. By specifying parser output, *token* is a parallel term to *collocation*, which likewise refers to an element from the output of the collocation parser. When referring to both tokens and collocations, the term *item* is frequently used. As would be expected in counting, the term *type* is used when referring to a category of tokens or collocations. In all cases, differences in type are based on differences in the character strings (items) in the parser output. That is, items made of identical character strings are of the same type. When the plural *types* is used, it refers to the count of unique character strings in the parser output. For example, the parsed token output [‘the’,‘cat’,‘in’,‘the’,‘hat’] contains five tokens of four types. The ‘the’ tokens at indexes 0 and 3 are of the same type.

5.5.2 Frequency Counts and File Counts

Another clarification to make in relation to counts has to do with count method. For each token or collocation type examined there were two count methods employed, and conse-

quently at least two statistics provided. The first is referred to here as a *frequency* count. All item-related statistics denoted as *frequency* or *freq* data are based on the count of all occurrences of the given item type in the entire corpus under consideration. Originally this was the only method of counting. However, early in the testing of analysis methods we realized that frequency counts alone allowed the opportunity for a few archives with high counts of a certain types to bias the overall results. In other words, there was no measure to provide an indication of a type's distribution across the set of documents in the corpus. To remedy this we began to make file counts as well, which is the second of the two measurements provided for each item. All item-related statistics denoted as *file* data are based on the count of files in the given corpus that contain at least one occurrence of the type in question. By working with counts of rate and distribution, the researcher is provided with a much-improved understanding of a type's occurrence.

5.5.3 Text Boundaries

Examining the output of the text extraction process and the tokenizer, it becomes obvious that the majority of traditional text boundaries (i.e. documents, chapters, paragraphs, sentences, clauses) are not considered during analysis. Text extraction removes all the major divisions marked by whitespace, and tokenizing removes the minor divisions marked by punctuation, which means that these divisions do not enter into parsing. Again, given the nature of the analysis we did not believe that maintaining these boundaries would add significantly to the results or their interpretation. The issue that remains unanswered is whether or not the collocation parser maintains a boundary between archives. In other words, can the node of a collocation come from the end of one archive, and the collocate come from the beginning of another. The answer is no. Each archive is processed separately. The reason for this is that the value of a collocation is that it records the proximity of the node to the collocate, which adds a degree of context. Because the interest is in semantics rather than syntax, and given the short distance allowed between a node and collocate, intra-document boundaries

have low value. That is, given the node 'mass,' knowing whether it is in close proximity to a collocate such as 'marketing' or 'spectrometry' adds a great deal more to the interpretation than knowing whether or not the collocate occurs in the same clause or sentence. However, given that the Quota Sample is a random sample and that documents were selected for processing in a random manner, there is no guarantee of any semantic connectedness between the end of one document and the beginning of another, except at a very abstract level. Thus parsing collocations across document boundaries makes no sense.

5.5.4 Document Sets

The final issue to address in *how* counts are made is the question of document sets. Obviously, coming to terms with a corpus requires that counts be made on a corpus (i.e. a set of documents). The question then is what sets? More specifically, of the 908 archives in the Quota and Supplemental Samples, how are the files grouped to form corpora and sub-corpora for study. Theoretically, the answer is that the archives can be grouped according to the needs of the researcher, using any identifiable feature, tag, or metadata classification. For example, a corpus could be made of all the archives that contain the token type 'the,' which would be comprised of over 99 percent of the sample files. Or a corpus could be made of all the archives that contain `<formula>` tag data, which would be only a few files. The boundary of what might be done is defined only by the researcher's ability to identify features and/or classifications in the archive. Practically speaking however, limits are imposed by a number of factors external to the archives themselves, time and budget in particular, and choices have to be made. For the TDC Project as a whole, but particularly for use with the Toolkit, the corpora and sub-corpora assembled for analysis were based on the major categories used during the sampling process (see Chapter 3), and the document metadata collected during archiving (see Chapter 4).

Referring back to Section 5.2 above, during the process of extraction a header line containing document metadata was added to the extracted text of each archive. This was done

specifically for constructing category-based sub-corpora for analysis. For example, the following is the header line from the extracted text of the example document in Figure 5.3:

```
<note> sample="quota" bates="618000535" isource="bw" decade="1990" class="ni" date="19930605" </note>
```

Along with the document's starting Bates Number and full date, the header contains four meta-classifications: the root sample, the industry source, the decade strata, and the document class. From these data, for each major corpus (based on the two root samples *quota* and *supplemental*), 18 sub-corpora were assembled for general analysis (36 total). This was done by grouping together documents with like classifications, each being a sub-set of the given major corpus (i.e. sampling with replacements). There were seven sub-corpora based on industry source (atc, bw, ctr, ll, pm, rjr, ti), seven based on *decade* (1950, 1960, 1970, 1980, 1990, 19xx, Bliley), and four based on *class* (ni, ne, ui, ue). The primary analyses done on the TDC and available via the Toolkit used these corpora and/or combinations of such. There are also analyses based on date metadata; however, these were not done specifically through the formation of sub-corpora for study.

For the purpose of illustration in this particular work, additional restrictions need to be made. For the remainder of this chapter, the major illustrations of TDC data are based on the major corpus comprised of all extracted text from the Quota Sample, and five decade-based sub-corpora comprised of the extracted text from the five decade strata (of the Quota Sample) with defined date ranges (1950, 1960, 1970, 1980, 1990). These particular five document sets were chosen for illustration because they provide a multidimensional view of the TDC which cannot be demonstrated as clearly with other sets. Aside from being sufficient to illustrate the major types of analyses used for presentation of the TDC, they also offer insight into the history of TDC documents (and supposedly the industry as well) in that they provide a diachronic view, which is a very tangible concept. Keep in mind, however, that all that is demonstrated using these sets, could also be demonstrated with the others, even parallels to the diachronic.

5.5.5 Count Data

It is at this point, having defined both the *what* and the *how* of counting, that actual count data can be provided. As an initial example, Table 5.1 provides a look at the counts of the top 50 token and collocation types from the extracted and tokenized text of example document 618000535. Careful study of this example in comparison to the tokenized text in Figure 5.4 should provide a clear understanding of how items are counted. In this case, the file count for all items is 1, given the data are from a single file. Overall, the text yielded 219 tokens of 132 types, and 651 collocations of 629 types (which matches the checksum).

Understanding the count procedures, we can now move to count data for the example corpora. Table 5.2 provides a range of descriptive statistics (basically counts) for the Quota corpus (the major corpus) and the five decade-based sub-corpora. In combination with the sampling specifications from Chapter 3, these data should provide a high-level view of the example corpora in terms of size and extent. Keep in mind that the sub-corpora are subsets of the Quota corpus such that all sub-corpora text is also in the Quota corpus. This is referred to as sampling with replacements. However, the Quota corpus, being the general corpus, also contains the 19xx and Bliley documents, so the sum of sub-corpora values for a given statistic are not expected to equal the corresponding value from the Quota corpus. In the table, *tokens* and *collocations* are defined as expected. *Files* refer to the extracted text files. *Words* are defined as alpha-character sequences with optional apostrophes, *numbers* are numeric-character sequences, and *others* are any remaining sequences (alphanumeric-character sequences). *Totals* are the count of the item in the corpus, *types* are the same as defined above, and *ratio* is the ratio of items per type.

In terms of actual count data for corpus items (tokens and collocations), additional limitations must be put in place as it is not practical to display all count data gathered for the various example corpora. In most cases the full data sets produced during counting contain tens-of-thousands of lines of data each. As an example, Table 5.2 indicates that there were 26,232 token types in the Quota Sample. Using a single sort method such as by-

Table 5.1: Example Document 618000535: Descriptive Statistics.

Rank	Token-Type	Count	Rank	Collocation-Type	Count
1	of	8	1	at /3 ppm	3
2	and	8	2	to /3 the	2
3	the	7	3	the /3 of	2
4	in	6	4	provides /3 a	2
5	at	6	5	ppm /3 and	2
6	to	5	6	of /3 the	2
7	ppm	5	7	objectionable /3 ppm	2
8	menthol	5	8	objectionable /3 at	2
9	this	4	9	menthol /3 both	2
10	may	4	10	during /3 we	2
11	is	4	11	during /3 may	2
12	for	4	12	delta /3 damascone	2
13	evaluation	4	13	both /3 and	2
14	damascone	4	14	beta /3 damascone	2
15	with	3	15	at /3 of	2
16	we	3	16	at /3 levels	2
17	during	3	17	at /3 and	2
18	both	3	18	at /3 330	2
19	beta	3	19	530 /3 ppm	2
20	a	3	20	330 /3 ppm	2
21	work	2	21	330 /3 and	2
22	were	2	22	would /3 require	1
23	that	2	23	would /3 more	1
24	provides	2	24	would /3 extensive	1
25	possible	2	25	work /3 w	1
26	our	2	26	work /3 in	1
27	objectionable	2	27	work /3 for	1
28	material	2	28	work /3 flavor	1
29	levels	2	29	work /3 blends	1
30	it	2	30	work /3 b	1
31	flavors	2	31	with /3 suggestions	1
32	flavor	2	32	with /3 possible	1
33	delta	2	33	with /3 menthol	1
34	530	2	34	with /3 jack	1
35	330	2	35	with /3 hugh	1
36	would	1	36	with /3 friday	1
37	was	1	37	with /3 casing	1
38	w	1	38	with /3 both	1
39	use	1	39	with /3 at	1
40	time	1	40	were /3 to	1
41	think	1	41	were /3 sent	1
42	these	1	42	were /3 rick	1
43	than	1	43	were /3 evaluated	1
44	tentative	1	44	were /3 at	1
45	tar	1	45	were /3 330	1
46	suggestions	1	46	we /3 this	1
47	spent	1	47	we /3 think	1
48	somewhat	1	48	we /3 that	1
49	several	1	49	we /3 our	1
50	sent	1	50	we /3 evaluation	1

Table 5.2: Example Corpora: Descriptive Statistics.

Measurement	Quota	1950	1960	1970	1980	1990
Files: total	808	24	64	144	284	244
Files: 0-9 tokens	0	0	0	0	0	0
Files: 10-99 tokens	94	2	10	21	29	26
Files: 100-999 tokens	533	16	39	99	191	161
Files: 1K-10K tokens	181	6	15	24	64	57
Files: 10K+ tokens	0	0	0	0	0	0
Tokens: total	543,959	15,809	43,874	82,123	188,259	172,721
Tokens: max/file	8,877	2,191	2,684	2,631	3,191	8,877
Tokens: avg/file	673.22	658.71	685.53	570.3	662.88	707.87
Words: total	519,312	15,453	42,280	78,466	179,539	163,422
Words: avg/file	642.71	643.88	660.63	544.9	632.18	669.76
Numbers: total	22,351	338	1,441	3,408	7,808	8,387
Numbers: avg/file	27.66	14.08	22.52	23.67	27.49	34.37
Others: total	2,296	18	153	249	912	912
Others: avg/file	2.84	0.75	2.39	1.73	3.21	3.74
Tokens: total	543,959	15,809	43,874	82,123	188,259	172,721
Tokens: types	26,232	3,176	5,572	9,201	14,669	14,687
Tokens: ratio	20.74	4.98	7.87	8.93	12.83	11.76
Collocations: total	1,627,029	47,283	131,238	245,505	563,073	516,699
Collocations: types	763,541	32,966	80,697	152,374	317,308	295,474
Collocations: ratio	2.13	1.43	1.63	1.61	1.78	1.75

frequency, at one item per line it would require over 400 pages to display the data. This being the case, an effort was made to reduce the data to a presentable amount, yet still provide enough information to the reader to allow insight into the content of the TDC sufficient to test the proposal in Chapter 1. Toward this end, Appendix E.4 contains the following four tables of reduced/limited count data for the Quota Sample:

1. Top 500 Tokens Ranked by Frequency Count - Page 309
2. Top 500 Tokens Ranked by File Count - Page 316
3. Top 500 Collocations Ranked by Frequency Count - Page 324
4. Top 500 Collocations Ranked by File Count - Page 331

The primary reduction of data in these tables comes from from the fact that only the count data for the full Quota Sample are given, not data from all example sub-corpora listed

in Table 5.2. Although the most obvious role of the displays is to illustrate the type and extent of data collected during counting, an equally important role is to begin the process of familiarizing the reader with the content of the TDC as a whole. In this respect, the Quota Sample offers a unique view which would not be available using one of the sub-corpora, and which compliments the data in the Toolkit. Both the sub-corpora and Toolkit (which works with sub-corpora) are used to locate trends and events within the TDC. Although this does require count data from the full Quota Sample, the Toolkit does not provide the specific data to the user. Thus, an understanding of the TDC as a whole is reached somewhat circuitously. In contrast, using the Quota Sample data addresses the TDC as a whole directly and accurately, adding a great deal of complementary insight/perspective to the Toolkit analyses. This would not be the case if count data from one of the sub-corpora were used for the illustration.

In examining raw count data one quickly discovers that in relation to understanding the general content of a corpus, the most valuable data are on the high end of the scale (i.e. the data with the higher counts). That is, one wants to see what content words are most frequent or in the most files, which is a general indication of topic. This being the case, the second major reduction of data is that only the counts for the top (highest count) 500 items are provided. Not wanting to contradict Chapter 2, 500 is admittedly an arbitrary number, being a compromise between value added and space available. More data were preferred, but there are sufficient data to serve the purposes of illustration and familiarization discussed above.

The final reduction of the count data was done in order to improve the quality of the information presented in relation to goals of the project. Overwhelmingly, the highest ranking collocations tend to contain function words as the node or collocate, or both. For example, the ten most frequent collocations in the Quota Sample are the following: [‘the /3 of’, ‘of /3 the’, ‘to /3 the’, ‘in /3 the’, ‘the /3 the’, ‘and /3 the’, ‘a /3 of’, ‘for /3 the’, ‘the /3 and’, ‘of /3 and’]. Given that the interest here is more on content, collocations containing ‘noise’

words in either the node or collocate position were eliminated from the display. Thus the top or bottom N collocations are the N most extreme that did not contain a noise word. For this process, noise words are any of the following list, which are the top 30 non-content tokens from the Quota Sample: [‘the’, ‘of’, ‘to’, ‘and’, ‘in’, ‘a’, ‘for’, ‘is’, ‘be’, ‘that’, ‘on’, ‘with’, ‘this’, ‘as’, ‘are’, ‘by’, ‘will’, ‘was’, ‘have’, ‘from’, ‘or’, ‘at’, ‘it’, ‘not’, ‘were’, ‘we’, ‘i’, ‘an’, ‘has’, ‘you’].

The end result is a reasonable set of data that serves well as both an illustration of count data, and as a means for familiarizing oneself with the TDC as a whole. Careful study of the data, as simple as they are, will lend a great amount of insight into the content of the TDC. For example, in the frequency sort of tokens types in Section E.4.1, the highest ranked content token is in fact *tobacco* on line 24 with a frequency count of 2,079. This is followed closely by *smoking*, *cigarette*, *smokers*, *smoke*, *new*, and *more* at line 44, which together do well to approximate the theme of tobacco-industry documents. Thus even not knowing the source of the corpus documents or having read any documents, in viewing the data one is immediately informed of the general corpus content. Another interesting fact that becomes apparent rapidly is that there is an odd mixing of item types related to marketing, such as *brand* at line 69, *products* at line 87, *market* at line 106, and *sales* at line 126, with those related to public health, such as *health* at line 112, *cancer* at line 143, and *exposure* at line 158. As study of the documents continues, one will find that this too is a marker of tobacco industry discourse, a strange interplay of market research with cancer research.

One should also take note of the difference in perspective provided by examining both frequency and file counts (see Section 5.5.2). Using the same types as in the previous examples, sorting by file count rather than frequency provides additional information (see Section E.4.2). As with the frequency sort, *tobacco* is the highest ranked token type that would not be considered a function word, being found in approximately 45 percent of files (see line 37). This strengthens the argument that it is a major feature of the corpus, being the highest content item both by frequency and file count. In other words, not only do we know that the

type occurs most frequently, we also know that the occurrences are well distributed across the corpus, not from a few files with high counts. As well, file counts add important information to the above example involving marketing and public health. In terms of frequency, *market* and *cancer* run close together with occurrences of 521 and 412 respectively, *market* being ahead by roughly 27 percent. However, examining file counts, *market* is ahead of *cancer* 153 to 85, which is about 80 percent. Thus *market*, with an average of 3.35 occurrences in every 5.28th file, is much more widely distributed in the TDC than *cancer*, which averages 4.85 occurrences in every 9.5th file.

It bears mentioning at this point that when the token type counts are ordered from high to low as they are in the Appendix E.4 tables, their plots are strongly hyperbolic, particularly with the frequency count. This is shown in Figures 5.7 and 5.8. For reference, if present in the top 500 types, the location of the items from the following list were added to both plots: [‘the’, ‘is’, ‘tobacco’, ‘brand’, ‘cancer’, ‘lung’, ‘women’, ‘disease’, ‘market’]. Although not strictly meeting the approximation defined by Zipf’s Law, which paraphrased is that given the maximum count m for any type, the count of the N most frequent token will be m/N (Zipf 1949), the plots do follow the expectations outlined by Kretzschmar and Tamasi (2003). While to most this is not as entertaining as comparing *market* to *cancer*, it does confirm that the Quota Sample data conform to a well-established norm for language count data. Keep in mind that a set of random integers ordered from high to low has an expected slope of -1, which is a straight line, meaning that if the use of token types were similarly random, for the 26,232 token types in the Quota Sample, any plot of the top 500 should decrease by only 1.91 percent of the maximum value. This is certainly not the case in Figures 5.7 and 5.8. In other words, the hyperbolic nature of the plots is an indication that the Quota Sample is not grossly misconstructed. In fact, jumping ahead to Figure 5.10, one sees that compared to similar data from the combined text of the Brown and the Freiburg-Brown Corpora (to be defined in Section 5.6.2), the plots are nearly indistinguishable.

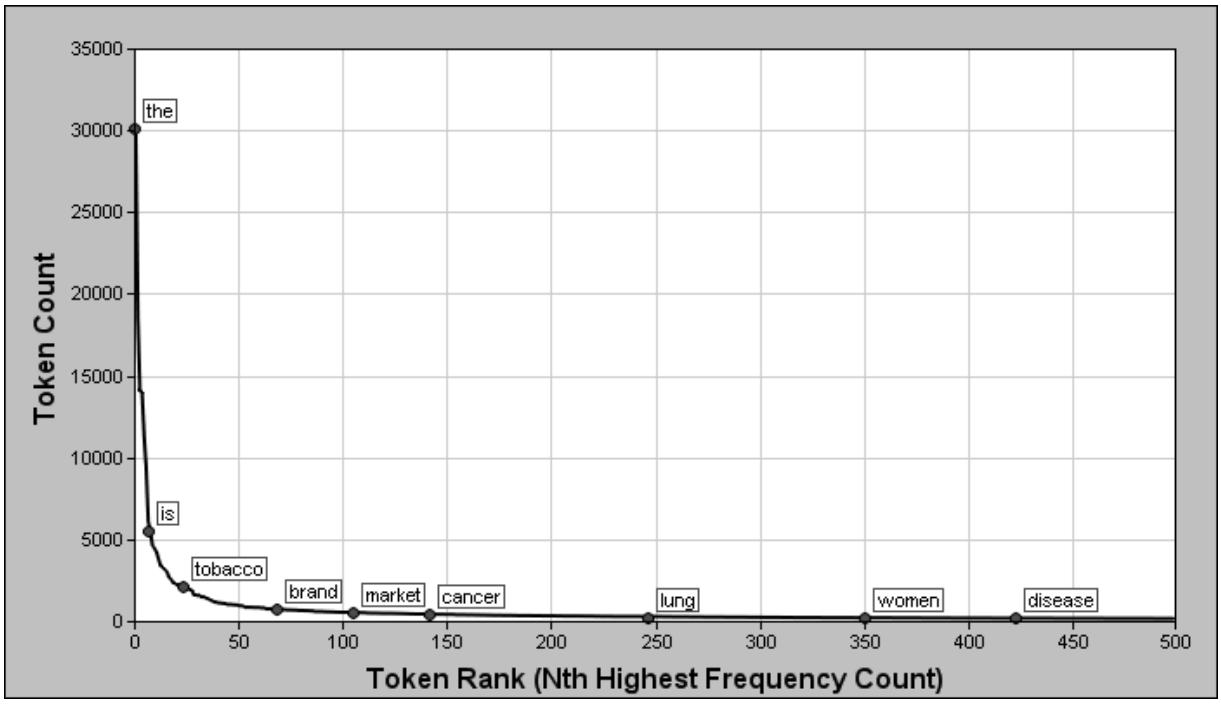


Figure 5.7: Quota Sample: Token Distribution Ranked by Frequency Count

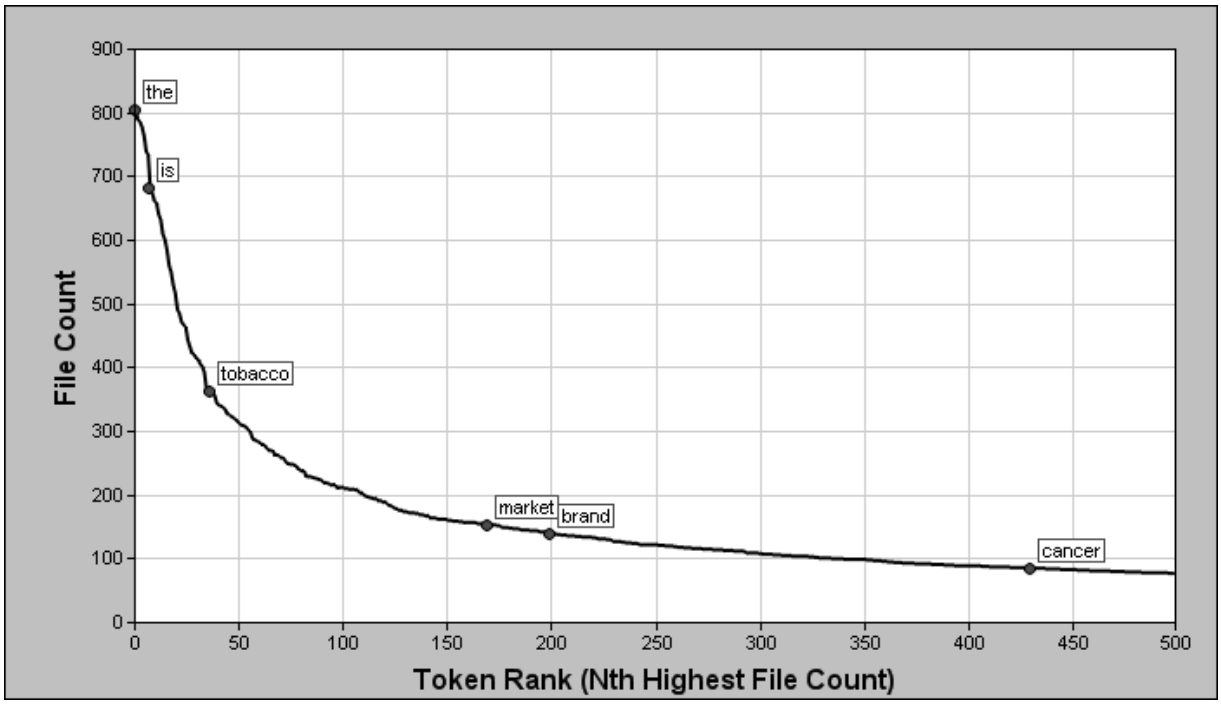


Figure 5.8: Quota Sample: Token Distribution Ranked by File Count

There are a great many other examples from the basic count data presented in the Appendix E.4 tables that might be discussed. I continue to be amazed at the quantity of information that can be gleaned from such displays. However, additional discovery must be left to the reader. As a final note, however, is should be stated that all that has been discussed above in relation to token types applies equally to collocation types, but with the additional context provided for the node by the collocate. That is, *cancer* becomes *lung /3 cancer*, *smoking /3 cancer*, and *cancer /3 research*; and *market* becomes *test /3 market*, *share /3 market*, and *market /3 share*. This again changes the perspective, adding new data to the equation and re-arranging the paradigm.

5.6 Analysis Methods

The great advantage to having spent so much time meticulously (or perhaps tediously) working through the steps necessary for analysis, not just in this chapter, but beginning with defining category, working through defining a corpus, and then defining the means for harvesting data from that corpus, is that the description of the actual analysis is straightforward. The categories are known, and the counts are known. What remains is to unify them to produce meaningful data that moves us forward in answering questions about corpus content, but apart from reading substantial amounts of the text. This of course was the original goal for the TDC, which has not changed. Likewise, the original means for obtaining the goal has not changed. From the concept of the TDC forward, the key idea has been and remains comparison.

In Chapter 1, the proposal (or purpose) for constructing the TDC was to establish a norm of written tobacco-industry discourse, and norms of course are specifically for comparison. The belief was that through comparison to a norm insight could be gained into the nature of various experimental documents and sub-corpora. One of the main points in Chapter 2 is that for any count or descriptive statistic for a category, meaning is derived only by comparison to a similar statistic from another category. By itself, the statistic is useless

(see Section 2.2.2). In the chapters on corpus construction and archiving, both in sampling (choosing which documents) and in archiving (choosing which text), the unifying theme is insuring that the TDC is representative of the NAAG Snapshot documents, which again is a question of comparison. And even in this chapter, as the item counts began to be presented in Section 5.5.5 and in the tables in Appendix E.4, the value we derive from them is by comparison, either one item to another in the same data set, or as a set to another set, such as a familiar or archetype set. The fact that the count of the token type *compared* in the Quota Sample is 157 has no value by itself, only in its relation to other counts from other corpora. Thus, one can rightfully conclude at this point that analysis is comparison in some form, and the question becomes not what to do (this is inherent), but how it should be done. In other words, what things are to be compared, and by what method.

The simple answer to what things are to be compared is corpora, or more specifically, two corpora. That is, the purpose of the TDC is to be a norm to which another corpus can be compared, so it stands to reason that this would be the case. In very gross terms, this is true. Given a known corpus as a norm (the reference corpus), and given an unknown corpus for study (an experimental corpus), the question being asked is what distinguishes the experimental corpus from the normal corpus (i.e. what, if anything, marks it as different). However, this comparison cannot be made directly. In quantitative analysis of corpora, the only things that can be compared directly are counts, and a corpus as a whole has no count other than 1 corpus (which makes for a boring comparison). Thus the corpus must be broken down to the point that useful counts can be made, which could be anywhere along the chain of events described in this chapter: separation into files, text extraction, tokenization, and parsing. Once broken into parts, the constituents can be counted, the counts compared, and then the results reassembled (quantitatively or qualitatively) to make higher-level evaluations. Even the simplest of corpus comparisons, such as the number of words or files, implies this process of deconstruction and reconstruction.

Another issue inherent in comparing corpora that hinders direct comparison, even of the counts, is that they generally are different sizes. If this is the case then the raw count of a constituent in one corpus cannot be directly compared to another as the totals from which the counts were taken are not equal. In other words, a count of X constituents in one corpus would represent a different proportion of the total n than the same count X in another corpus where the total is $2n$. The simplest remedy for this it to convert all counts into a proportion $p = \frac{X}{n}$, where X is the count out of the total possible constituents n. Although generally expressed as a decimal number from 0.0 to 1.0, in essence this procedure establishes a common denominator, making the numerators directly comparable.

At the very base level, then, the comparison we had in mind as the TDC was prepared for analysis is between two proportions based on the counts of a given constituent in two corpora, one the experimental and another the reference. The question being asked is whether the observed proportion p_1 in the experimental corpus is outside the expected range of the normal proportion p_2 established by the reference corpus. If it is, then the constituent can be viewed as a marker of the experimental corpus. In other words, the experimental corpus is marked by a disproportional (i.e. unexpected, either high or low) occurrence of the constituent compared to the reference corpus. This was driven by the hypothesis that the combination of such markers, being the distilled result of millions of comparisons, would provide valuable insight into the content of the experimental corpus. In statistical terms, which are somewhat the opposite of the Linguist's perspective, the root task is to prove the null hypothesis $H_0 : p_1 = p_2$, which is that the observed proportion is not significantly different from the normal proportion. If this cannot be proven, then significant variation is assumed.

An excellent example of this concept can be found if we jump ahead to the compared plots in Figure 5.11. These are the plots of the sorted token-type counts for two corpora. This was prepared in the same manner as the plot in Figure 5.7 above, but with the counts converted to percentages of the total token counts so that two data sets could be displayed

on the same graph. As well, this figure is a 'zoomed' view of the plot data from Figure 5.10, which is difficult to read because of the similarity between the data sets. The item of interest in this plot is the token type *women*. This type is shown on both plot lines, but in noticeably different locations. In the Brown-Frown corpus (to be defined below), *women* occurs at a higher rate than in the Quota-Sample corpus. If the Brown-Frown corpus is considered the reference, then one can say that there is a disproportionally low occurrence of the *women* token type in the Quota-Sample corpus (the experimental corpus). Figure 5.12 gives a similar indication, but in this case in relation to file count. The *women* token type is seen in the top 500 on the reference corpus plot, but not in the experimental because of its lower rate of usage. The question that remains, of course, is whether or not the differences can be considered significant. As it turns out, they are. Here is the actual comparison data for the token type *women* taken from the table of data similar to that in Appendix E.5.1:

Token	Freq-Z	Freq-V	File-Z	File-V
women	-3.555	1	-11.539	1

The meaning of each datum will be explained below, but the four statistics for *women* indicate that the Quota-Sample corpus is marked by a disproportionally low use of the type, both in frequency and file count.

5.6.1 The Comparison Statistic

I must make the assumption in this section that the reader has a basic understanding of Statistics, and in particular the meaning and implication of z-scores. Unlike XML, an overview of Statistics is outside the capacity of this work and would probably do more harm than good. This is particularly the case given that the z-statistic presented below does not use 'standard' methods for determining the mean and standard deviation necessary for calculating z-scores.

Continuing with the discussion, once one settles on the idea of using differences in constituent proportion across corpora as a means of comparison, the choice of statistic is severely

limited. In fact, there is only one common method for testing the proposed null hypothesis. It is generally referred to simply by description, as in 'a significance test for comparing two proportions.' In Linguistics, the method is not often employed, but it is mentioned briefly by L. Davis in 'Statistics in Dialectology' (1990). It is much more common in general Statistics and is routinely included in introductory texts along with more complete descriptions. In function, given two sample populations with corresponding proportions, the statistic returns a z-score based on the size of the samples and the difference in the proportions by estimating the mean and standard deviation from the pooled data. In Moore and McCabe's 'Introduction to the Practice of Statistics,' the procedure is described in the following manner. Given a Population 1, having a sample of size n_1, a count of successful trials X_1, and a sample proportion $\hat{p}_1$, and given a second Population 2, having a sample of size n_2, a count of successful trials X_2, and a sample proportion $\hat{p}_2$,

To test the hypothesis

$$H_0 : p_1 = p_2$$

compute the z statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_p}$$

where

$$s_p = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

and

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

(594).

According to Moore and McCabe, the P-value for testing H_0 against the alternative hypothesis $H_a : p_1 \neq p_2$ using this statistic is $2P(Z \geq |z|)$, taken from a table of standard normal probabilities, where Z is any normal random Z-value (594). This being the case, if the acceptable probability of error is set at $p \leq 0.05$, significance is reach when $z \geq |1.96|$. At $p \leq 0.01$, it is not reached until $z \geq |2.58|$.

Although this significance test will return a z-score for any set of properly-formed proportions (non-negative with a positive sample size n), not all returned z-scores are reliable for estimating the extent of deviation or the probability of error. Although Davis poses no requirements for use, Moore and McCabe note that the test becomes less reliable as proportions become more extreme. That is, as input proportions move toward 0.0 or 1.0, results become less interpretable. Given that this particular test is based on the normal approximation to the binomial distribution, Moore and McCabe recommend its use only when $n_1\hat{p}$, $n_1(1-\hat{p})$, $n_2\hat{p}$, and $n_2(1-\hat{p})$ are all 5 or greater (594). To account for this requirement, all z-scores in this work and in the Toolkit are accompanied by a *v-score*, which is a binary indicator of validity (reliability), either a 1 to indicate that the comparison met the above requirements, or a 0 to indicate that it did not.

In practice, the calculation of both z-scores and v-scores was handled by a single Python function (shown in Figure 5.9). By passing in two sets of count data, the resultant z-score and v-score are returned as a tuple. More specifically, the input is four integers ($x1$, $n1$, $x2$, $n2$) where $x1$ and $n1$ are the count of successful trials and total trials (i.e. sample size) which form the first proportion, and $x2$ and $n2$ are those are those which form the second. The order of proportions passed to the script is important because the output z-score has a sign value that indicates that the first proportion ($x1/n1$) is higher (positive) or lower (negative) than the second proportion ($x2/n2$). It is also important that the full value of the counts be provided for both proportions because both effect the resultant z- and v-scores. As an example, the output of a comparison between a first proportion 0.001 and a second proportion 0.002 will vary depending on the size of the samples, which can be seen in the following queries:

```
>>> prop_compare(   1,   1000,   4,   2000) returns ( -0.63298323803456369, 0)
>>> prop_compare(  10,  10000,  40,  20000) returns ( -2.0016687528977446,  1)
>>> prop_compare( 100, 100000, 400, 200000) returns ( -6.3298323803456364,  1)
>>> prop_compare(1000,1000000,4000,2000000) returns (-20.016687528977446,   1)
>>> prop_compare(4000,2000000,1000,1000000) returns ( 20.016687528977446,   1)
```

All of the queries contain the same proportions (0.001 and 0.002), yet they all return a different tuple of scores (z-score, v-score). Notably, as the sample sizes increase, the z-scores become more extreme, indicating that the probability of error is decreasing. At the same time the v-scores are more likely to be 1. Also notice that in the last query the proportions data were re-ordered, which causes the z-score to become positive.

```
import math
def prop_compare(x1,n1,x2,n2):
    "x1=obs_freq,n1=obs_total,x2=norm_freq,n2=norm_total"

    # required: n values > 0
    # required: x values >= 0

    # floats
    n1,n2 = float(n1),float(n2)

    # get proportions
    p1 = x1/n1
    p2 = x2/n2
    p3 = (x1+x2)/(n1+n2)

    # check validity
    valid = 0
    if n1*p3 >= 5 and n1*(1-p3) >= 5 and n2*p3 >= 5 and n2*(1-p3) >= 5:
            valid = 1

    # get standard deviation
    sd = math.sqrt((p3*(1-p3))*((1/n1)+(1/n2)))

    # get z statistic
    z = 0.0
    if sd:
        z = (p1-p2)/sd

    # return z-score and v-score
    return z,valid
```

Figure 5.9: Proportion Comparison Formula: Python Code

It was noted at the beginning of this chapter that the purpose of quantitative analysis is to assist the user in making qualitative judgments, and this does not change with significance testing. By design tests of significance provides the user with a probability that H_0 is true as an aid for making qualitative evaluations. This being the case, in selecting a z-score value as a cutoff for accepting H_a, what one is actually doing is qualitatively defining the category *significant* to contain data that pass a quantitative test. In other words, the acceptable probability of error used to establish the cutoff is a qualitative decision based on the needs of the user. For example, if the z-score requirement were combined with a v-score requirement,

the category *significant*, which is designed to contain all data that are valuable to the user, might be defined by the set of tests $z \geq |1.96|$ *AND* $v = 1$, each quantitative in nature, but qualitatively selected. It may just as well be defined by the single test $z \geq |3.27|$ (which is $p \leq 0.001$) or any other. However it is defined, *significant* is a very functional category considering the quantity of data generated in corpus analyses. In most cases, all of the data cannot be examined, so quantitative tests to aid in selecting the documents that are desirable for reviewed make sense. As well, if it is acknowledged that the category is qualitatively defined, then there is no violation of Moore and McCabe's prohibition of using statistics to define categories (472), which is described in Chapter 2.3.

In Linguistics, the minimal acceptable probability of error for significance is generally defined as $p \leq 0.05$. In practical terms, this means that the difference between two proportions is not considered meaningful unless the chance that the difference was caused by accidents in sampling is less than or equal to 5 percent, or 1 in 20. As stated above, for the given statistic this is reached when $z \geq |1.96|$ (i.e. when the z-score is more than 1.96 or less than -1.96). Thus for our purposes, *significant* is defined as the set of comparisons for which $z \geq |1.96|$ using the Python function in Figure 5.9, where z is the z-score returned. Likewise, but as a separate category, *valid* is defined as the set of comparisons for which $v = 1$ using the same Python function, where v is the v-score returned.

In terms of application, the primary statistic in TDC analysis for both frequency and file comparisons is the Python function in Figure 5.9. Of the four input variables ($x1$, $n1$, $x2$, $n2$), $x1$ and $n1$ are the observed count of the given type and the sum of all types counts from the experimental corpus, respectively. Likewise, $x2$ and $n2$ are the normal count of the given type and the sum of all types counts from the reference corpus. In this order, the sign of the returned z-score indicates the direction of disproportion for the experimental corpus. For example, the data for the token type *women* displayed above were obtained from the following two queries in which the experimental corpus is the Quota-Sample corpus and the reverence corpus is the Brown-Frown corpus (defined below):

```
>>> prop_compare(192, 543959, 954, 2042345) returns (-3.5546619900838072, 1)
>>> prop_compare( 43,    808, 256,    1000) returns (-11.538635733353548, 1)
```

In the first query, 192 is the frequency count of the token type *women* in the experimental corpus, 543959 is the total count of all tokens in the experimental corpus, 954 is the frequency count of the token type *women* in the reference corpus, and 2042345 is the total count of all tokens in the reference corpus. Based on the returned data, this comparison is both *significant* and *valid* according to the above definitions, and it indicates that the frequency of the token type *women* is disproportionately low (negative sign) in the experimental corpus in relation to the reference corpus. For the second query, 43 is the file count of the token type *woman* in the experimental corpus, 808 is the total count of files in the experimental corpus, 256 is the file count of the token type *woman* in the reference corpus, and 1000 is the total count of files in the reference corpus. Based on the returned data, this comparison is also *significant* and *valid* according to the above definitions. As well, it too indicates that the file count of the token type *women* is disproportionately low in the experimental corpus in relation to the reference corpus.

In most cases, these same types of comparisons would be made for every token (1-gram) type, bigram type, trigram type, and collocation type in the experimental corpus. As well, this would be done for the items from the reference corpus that were not in the experimental corpus by setting the $x1$ value to 0. This would cause the item to be included in the output data being shown with a disproportionately low occurrence. By holding the population sizes constant the resultant z-scores are comparable to the z-scores from similar items in terms of disproportion. For example, keeping the same experimental and reference corpora as above, the frequency and file count for the token type *men* can be compared using the following queries:

```
>>> prop_compare(121, 543959, 1269, 2042345) returns (-11.280347401630365, 1)
>>> prop_compare( 48,    808,  441,    1000) returns (-18.160578184283249, 1)
```

Notice that the $n1$ and $n2$ values are the same as they were for the comparisons of *women*. These results indicate a similar low disproportional use of the token type *men* in the experimental corpus in relation to the reference, but more extreme. If however the population sizes change, as with a change of experimental corpus or a change to an item that is defined differently (such as a collocation, which will have a different total type count), the resultant data are less directly comparable.

As a summary of the above, we can return briefly to the terminology from Chapter 2 and restate the purpose and means for comparison in less statistical/technical terms. In a broad sense, the task at hand is to determine if there are multiple categories whose definitions produce the same set of elements. One category, which is given, is defined by non-linguistic features, while the other, which must be discovered, is defined by linguistic features. The categories are, of course, the various corpora and sub-corpora under investigation. For demonstration, the Quota Sample and the five decade-based sub-corpora with known dates have been selected as examples. These corpora have been produced according to and are defined by non-linguistic features, in this case date ranges. What we want to determine is if there is also a set of linguistic features that produces the same document set. If found, this would be the discovery of the linguistic category. The optimal result would be that two categories A and B, having distinctly different definitions, define the same elements such that $A = B$. This would mean that new information about the corpus has been discovered, that documents know to be related by non-linguistic features are discovered also to be related by linguistic features. Practically speaking, we are content with less strictly-defined discoveries. For example, using the Toolkit it can be shown that documents from the 1960s decade have a disproportionately high use of tokens associated with *cancer*. Thus a set of documents known to be related by date, are found to also be related by topic.

The means for this type of discovery is to compare the corpus under study (the experimental corpus) to a reference corpus and look for linguistic variation. Any significant variation from the norm becomes a linguistic marker of the experimental corpus, and the com-

bination of these markers becomes the linguistic definition of the new category. The actual comparison, however, is not one corpus to another, but one count to another, for each countable item in the compared corpora, which in this case are all tokens and collocations as defined in Section 5.4. That is, for each corpus, for each item, the frequency count and file count are compared to similar counts in the reference corpus. If the comparison indicates that the item occurs at a significantly disproportional rate, either high or low, that item becomes a marked feature of the experimental corpus. Thus the corpus comparison is actually the combined results of thousands, if not millions, of item comparisons.

5.6.2 Comparison Examples: Quota Sample

The temptation at this point is to move directly to the Toolkit output as a source of example data. For the moment, however, the discussion will focus on data not directly available in the Toolkit as this gives the reader an advantage in at least two ways. The first has to do with the nature of the Toolkit displays. As noted above, the number of comparisons made in generating the data set used by the toolkit is quite large. Simply comparing the collocation types in the decade-based sub-corpora to those in the Quota Sample corpus involves $1627029 * 5$ or 8,135,154 z-scores, plus an equal number of v-scores. Adding to this the token comparisons, and then multiplying by the number of sub-corpus groups (i.e. source-based, audience-based, et cetera), the amount quickly becomes overwhelming, both in terms of display and assimilation by the user. As a remedy for this, the Toolkit was designed as an easily understandable front end to the data to facilitate interpretation and assimilation of the TDC content. It functions by focusing on narrow topics, compiling necessary data, and presenting those data to the user in a simple graphical format. While this has proven itself as an effective means of rapidly assimilating TDC content, by necessity it obscures a large part of the data. It is not that the data are completely unavailable in the Toolkit, but that they are presented as bits and pieces, one item or small group of items at a time. However, just as with the tables of count data discussed in Section 5.5.5, placing the individual datum

together in a sorted table provides a different perspective, allowing the rapid development of ideas and relationships in a manner not permitted by the Toolkit. Thus it provides a complementary view, one which I believe is valuable to the reader.

The second advantage in not going directly to the Toolkit is the need to examine the TDC in relation to other corpora. If a corpus is to serve as a reference, it should to some extent be familiar to the researcher in relation to another known corpus. In other words, before becoming a reference, a corpus should first be studied as the experimental. With the Toolkit, the reference corpus to which all sub-corpora are compared is the Quota Sample, or more specifically, the corpus consisting of the 808 documents[1] in the quota sample, extracted and tokenized according to the procedures outlined earlier in this chapter. However, within the Toolkit itself, the Quota-Sample corpus cannot be studied directly as a whole. Instead, one's knowledge of the full Quota Sample is derived from the content and trends discovered in the study of the various sub-corpora. Although this does allow users to quickly familiarize themselves with the full TDC document set, the additional perspective gained through comparing the Quota Sample corpus to another known corpus is also a valuable complement to the Toolkit data.

In order to provide these additional perspectives to the reader, a non-tobacco-industry corpus was selected to serve as a reference and comparisons were made using the Quota-Sample corpus as the experimental corpus. The resultant data were compiled and made available for the reader in Appendix E.5. In the next several paragraphs this process and the resultant data will be detailed following the format used for the description of count data found in Section 5.5.5.

[1]The data set for the Toolkit was actually compiled using 812 documents rather than 808. This was an oversight on the part of the author. The error amounts to a 0.5 percent increase in file count, which does have an effect on the z-scores in the output of the comparisons. The most extreme case with a v-score of 1 is the comparison of the file count for 'to' between the 1980 decade corpus and the Quota Sample corpus, in which case the difference in z-score is $0.888 - 0.472 = 0.416$. This is only an issue when comparing the data in Appendix E.5 to that of the Toolkit. Within each realm the data are consistent and reliable.

The corpus chosen as the reference is actually a combination two corpora: the 'Standard Sample of Present-Day American English,' commonly known as the Brown Corpus (Francis and Kucera 1964) and the 'Freiburg-Brown Corpus of American English,' referred to as the Frown Corpus (Hundt, Sand, and Skandera 1999). The Brown Corpus was chosen simply because it is by far the most well known electronic corpus with a size similar to that of the TDC. However, The Brown Corpus contains only text from the calendar year 1961, which limits how comparable it is to the TDC which has texts from across more than five decades. To widen the time span, the Frown Corpus was added to the Brown, being a model of the Brown Corpus constructed with texts from the 1990s. This combined corpus, referred to below as the Brown-Frown corpus, was then prepared for analysis in the exact manner detailed in the earlier sections of this chapter. The basic descriptive statistics for both the Brown-Frown and Quota-Sample corpora are shown in Table 5.3.

Table 5.3: Comparison Corpora: Descriptive Statistics.

Measurement	Brown-Frown	Quota
Files: total	1,000	808
Files: 0-9 tokens	0	0
Files: 10-99 tokens	0	94
Files: 100-999 tokens	0	533
Files: 1K-10K tokens	1,000	181
Files: 10K+ tokens	0	0
Tokens: total	2,042,345	543,959
Tokens: max/file	2301	8,877
Tokens: avg/file	2,042.35	673.22
Words: total	2,020,267	519,312
Words: avg/file	2,020.27	642.71
Numbers: total	19,968	22,351
Numbers: avg/file	19.97	27.66
Others: total	2,110	2,296
Others: avg/file	2.11	2.84
Tokens: total	2,042,345	543,959
Tokens: types	65,721	26,232
Tokens: ratio	31.076	20.736
Collocations: total	6,121,035	1,627,029
Collocations: types	2,467,173	763,541
Collocations: ratio	2.481	2.131

It is worth noting that there is a considerable disparity in the average token count per file between the two corpora. This is a result of the overall structure of the Brown-Frown

corpus. Individually, both the Brown and Frown corpora were constructed from 500 files of approximately 2,000 tokens each. This structure is somewhat artificial in that the chunks of 2,000 tokens were extracted from larger texts and do not necessarily represent a unified text span (i.e. a true file). Based on Quota-Sample statistics, one would expect the structure to have been reversed, such that there are 2,000 files of 500 tokens each, given that the average file length in the Quota-Sample corpus is much closer to 500 than to 2,000.

The effect of the file-size disparity can be seen in Figures 5.10 through 5.12, which are plots of ordered token data constructed in a manner similar to Figures 5.7 and 5.8 above, the major difference being that here the raw counts have been converted to percentages so that the data from both the Brown-Frown and Quota-Sample corpora, which have different totals for tokens and files, can be included on the same plot. In terms of frequency counts, Figure 5.10 indicates a remarkable degree of normalcy for the Quota-Sample data compared to Brown-Frown data. In fact, it is not until the data are amplified in Figure 5.11 that a difference between the corpora can be seen. While this is not a guarantee that either corpus is well formed, it does indicate that both, in a very similar manner, exhibit a well documented characteristic of English text. In contrast to this, Figure 5.12 illustrates a notable difference in the distribution of token types across the file sets. The plot of the Quota-Sample data are roughly as expected, having a steep negative slope from the onset. The plot of the Brown-Frown data, however, is noticeably flatter, even S-shaped to an extent (actually a reversed S). That is, there are an unexpected number of token types that have a distribution at or near 100 percent. This is a result of the high ratio of token types to files seen in the corpus. Based on the Table 5.3 data, this is 65.7:1 for the Brown-Frown corpus, compared to 32.5:1 for the Quota-Sample. What this means is that any given token type is expected to occur in a higher percentage of files in the Brown-Frown corpus because there are proportionally fewer files. This fact should be kept in mind when interpreting the file count comparisons.

At this point we can move to an examination of the comparison data. However, because the full sets of z-scores were quite large, roughly six million lines, an effort was made to

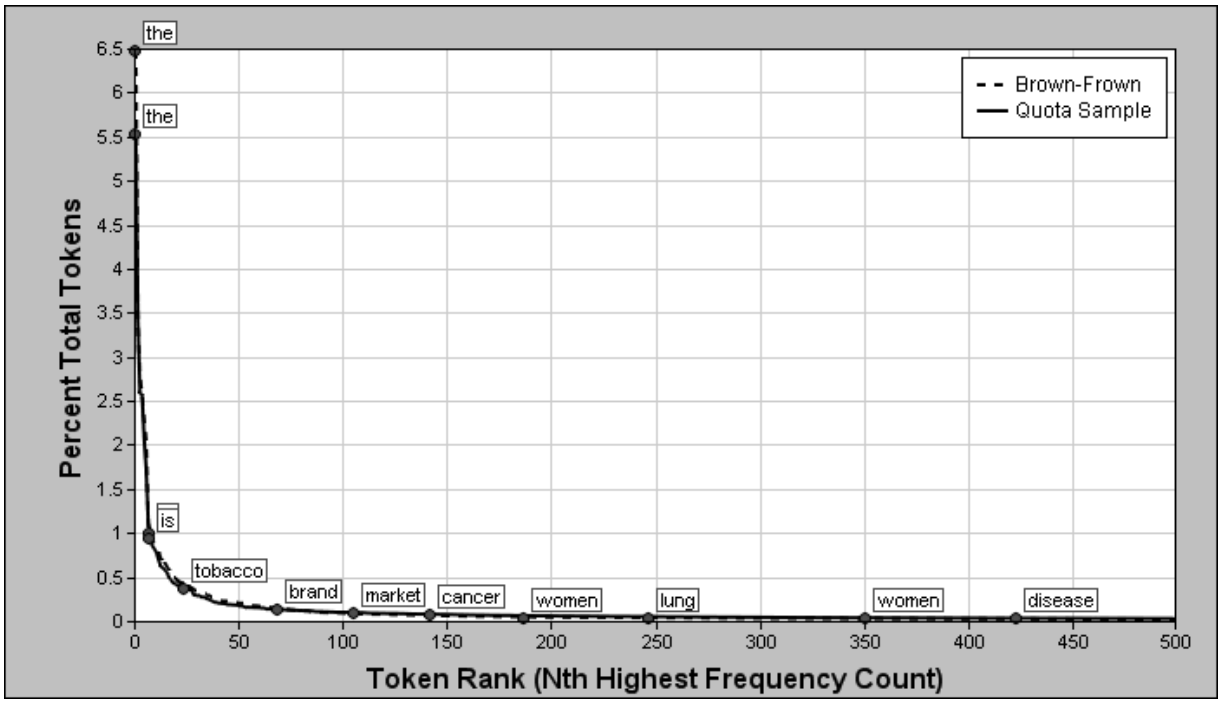


Figure 5.10: Comparison Corpora: Token Distribution Ranked by Frequency Percent - I

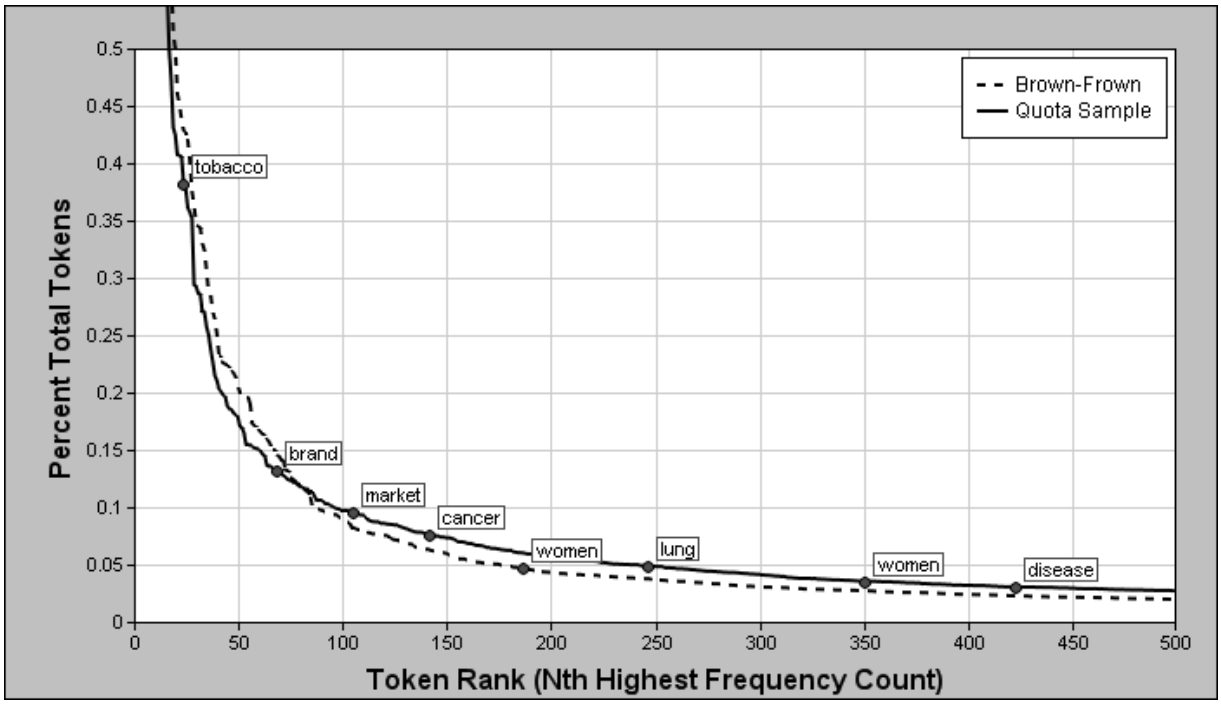


Figure 5.11: Comparison Corpora: Token Distribution Ranked by Frequency Percent - II

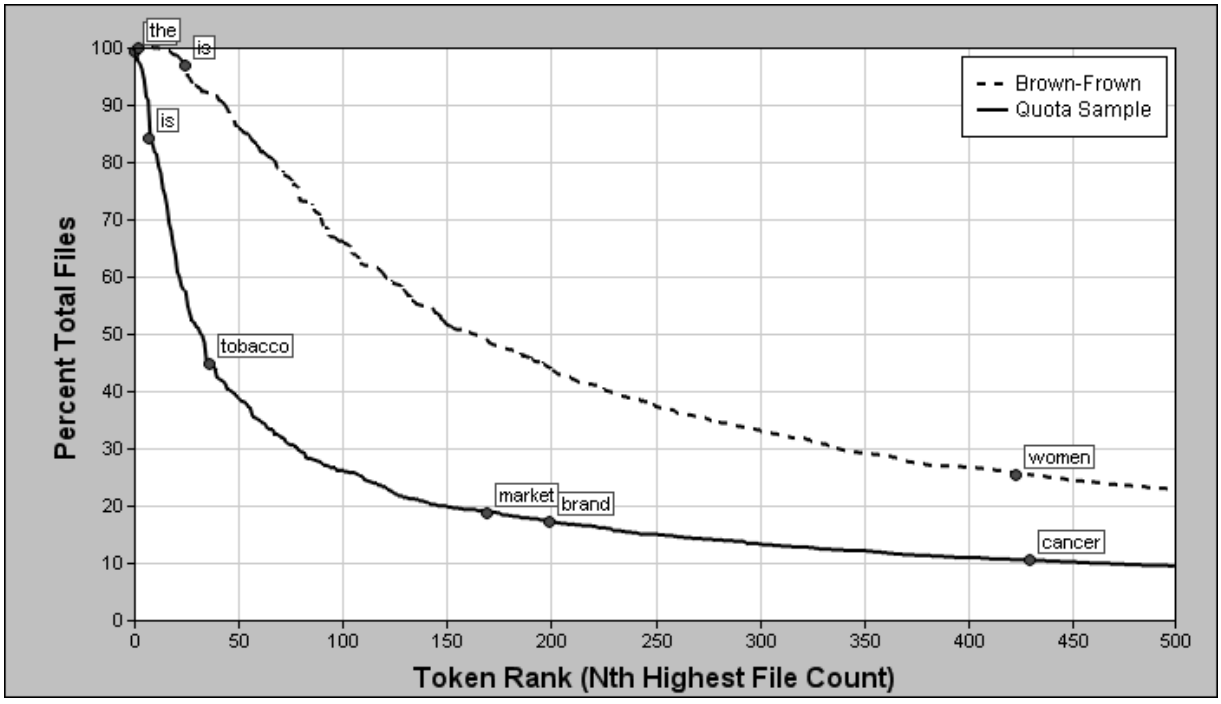


Figure 5.12: Comparison Corpora: Token Distribution Ranked by File Percent

reduce the data to a presentable amount, yet still provide enough information to allow insight into the content of the TDC sufficient to test the proposal in Chapter 1. Toward this end, Appendix E.5 contains the following four tables of reduced/limited data derived from comparisons made between the Brown-Frown and the Quota-Sample corpora:

Following the count-data examples, the major data reductions in these tables were made by limiting the sets, allowing a maximum of 500 lines, and placing limits on allowable collocation types. Please refer back to Section 5.5.5 for more details on these. Further reductions were

made by excluding any item type that was not both *significant* and *valid* for both the frequency and file comparison, according to the definitions of these categories given in the previous section. In this way, only the items with the strongest indication of disproportion are included in the display.

There were also some reductions that resulted from the method used to divide the 500 lines between the two tails of the output. Unlike count data, with z-scores the most valuable information is found at both ends of the range of scores, not just the high end. In other words, knowing that an item like *women* has a disproportionately low use in the experimental corpus (see Figure 5.11), can be as valuable as knowing that *tobacco* has a disproportionally high use. They are both marked features of the experimental corpus and need to be displayed. However, of the two tails, the high end tends to have more valuable information. This being the case, the display is divided 4:1 in favor of high-disproportion items, providing up to 400 items from the high end, and up to 100 items from the low end. The caveat to this is that all high-end items were required to have positive z-scores for the given sort (by frequency or file z-score), and conversely all low-end items were required to have negative z-scores. In some cases (collocations), once this stage was reach there were not 400 items with positive z-scores remaining. This final reduction, along with the previous four, produces sets in which all items are strong markers of the experimental corpus (the Quota-Sample corpus) in comparison to the reference.

The question that may be asked at this point is how this type of analysis is more productive than the displays of ranked counts, which can obviously provide substantial insight into the content of the corpus. The answer to this becomes obvious when the data are compared. For example, the top ten items from the table 'Top 500 Tokens Ranked by Frequency Count' from Section E.4.1 are the following:

Rank	Token	Freq	%Total	Files	%Total
1	the	30090	5.5317	804	99.505
2	of	18596	3.4186	786	97.2772
3	to	14095	2.5912	790	97.7723
4	and	14029	2.5791	778	96.2871
5	in	11401	2.0959	764	94.5545

6	a	9393	1.7268	740	91.5842
7	for	6137	1.1282	734	90.8416
8	is	5513	1.0135	681	84.2822
9	be	4686	0.8615	673	83.2921
10	that	4482	0.824	632	78.2178

For comparison, the top ten items from the table 'Top 400 and Bottom 100 Tokens Ranked by Frequency Z-score' from Section E.5.1 are the following:

Rank	Token	Freq-Z	Freq-V	File-Z	File-V
1	tobacco	87.323	1	22.208	1
2	smoking	73.206	1	16.985	1
3	cigarette	67.372	1	17.038	1
4	smokers	66.307	1	16.008	1
5	smoke	59.818	1	12.31	1
6	cigarettes	57.965	1	17.204	1
7	nicotine	55.68	1	13.952	1
8	brand	49.293	1	10.179	1
9	product	46.628	1	9.333	1
10	filter	45.818	1	11.761	1

Obviously, there is value added. The first set of data, the count set, presents all function words, while the second set, based on z-scores, is all content words. The difference, in a nutshell, is that although *the* and the other function words have high frequencies in the experimental corpus, proportionally they are normal when compared to the reference (i.e. H_0 is proven). Since the task is to find markers, these are removed from the results.

Concentrating on the high-end data, what one finds is exactly what the uninformed would expect: company/product names and cigarette/smoking terms. However, one also finds data that only the informed would expect: sales, marketing, strategy, planning, research, development, testing, chemical name, and cancer terms. Of course, being on the informed side of the spectrum in terms of tobacco documents it is easy to overlook the significance of the discovery permitted by the data. In other words, we already know what is there from reading, so discovering it in the data is less eventful. What one must remember is that all of this, and even more, can be discovered in the data without having read a single document. For example, who would have the foresight to predict that Quota-Sample corpus is marked by a lack of personal pronouns? Yet, across all four comparison data sets in the appendix personal pronouns are consistently on the low-end. This makes sense in hindsight, once it is

seen in the data, but how long would one have to read documents to discover this (if it could be discovered by reading).

Once one comes to terms with what these data represent, which the examples above should bring forth, the tables in Appendix E.5 become self-explanatory and easily interpretable. The reader is encouraged to spend a few minutes scanning through each one. Although sometimes subtle, each offers a different perspective, but they all work toward a common understanding of the Quota-Sample corpus as whole.

5.6.3 Comparison Examples: TDC Toolkit

The justification to include examples from the Toolkit, just as with the Quota-Sample examples in the previous section, is that they add a perspective to the data that earlier examples do not provide. The examples provided below were generated using the PLOT and PEAK functions of the Toolkit, which are two of several tools that can be used to manipulate and display TDC data.

In terms of procedures, the Toolkit analyses shown here use the same core processes and algorithms that produced the Quota-Sample data in Appendix E.3 and discussed in the previous sections. However, there are several higher-level processes that differ notably from previous examples. First, the Quota-Sample corpus is the reference for all comparisons. No Brown-Frown corpus data are used in the Toolkit. This being the case, all sub-corpora are treated as experimental corpora. Thus the z-score sign values in the results are in the direction of the disproportion in the sub-corpora. Second, comparisons are not made en masse as they were for the Quota-Sample. Instead, the user provides a limited list of tokens or collocations as input, and a display is prepared using only the data for the given items. And finally, input items are processed in a combined form as if they were a single item. In other words, the counts of input items from the experimental corpus are first combined, and then the combined value is compared to a similar combined value from the reference corpus, producing a single z-score.

In terms of display, the most notable difference is that the primary medium is a graph (a plot) rather than a table of figures. As well, the data from multiple sub-corpora (predetermined groups of similar sub-corpora) are presented together on the same display. This allows the researcher to view data for the given items across a range of similar sub-corpora in order to locate trends in the item usage. For example, Figure 5.13 is a plot of the frequency z-scores for the token types *cancer*, *cancers*, *carcinoma*, and *carcinomas* across the five decade-based sub-corpora.

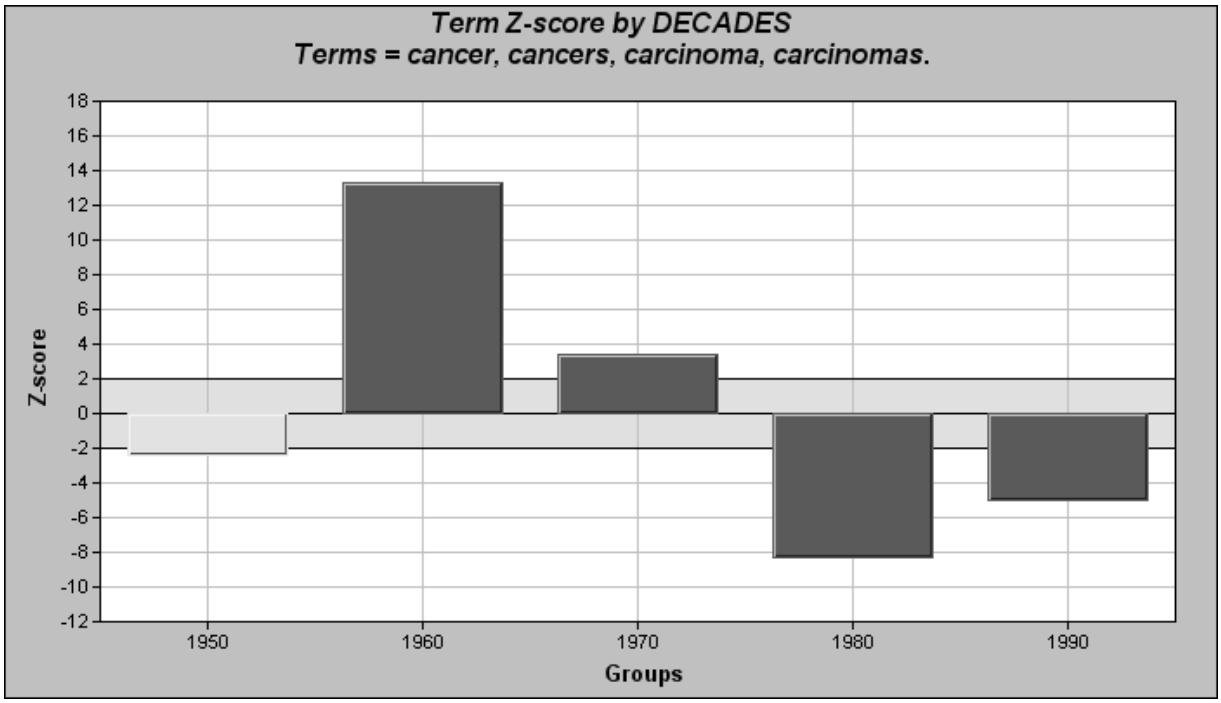


Figure 5.13: PLOT Tool, Frequency Z-score, Cancer Lemma, Decades

Each of the decades found along the X axis is represented by a bar on the graph. The heights of the bars are determined by the z-scores returned from the individual comparisons of the decade sub-corpora to the Quota-Sample corpus. The values are found on the Y axis. Any value outside the gray-shaded area is significant according to the definition in Section 5.6.1. The color of the bar is an indication of the overall reliability of the particular comparison based on whether or not the frequency and file z-scores are valid according to the definition from Section 5.6.1. In the Toolkit, the colors are green, yellow, and red, representing 2, 1 or

0 valid scores, respectively. In the examples here, the darker bar is the green, and the lighter is the yellow.

Given that in this case the sub-corpora are decade-based, the display provides a diachronic view of usage, indicating a dramatic change from a disproportionately high use of the items in the 1960s, to a disproportionately low use in the 1980s and 1990s. Although somewhat less dramatic, this same trend is found with file z-scores, which is seen in Figure 5.14. This lends credibility to the original finding.

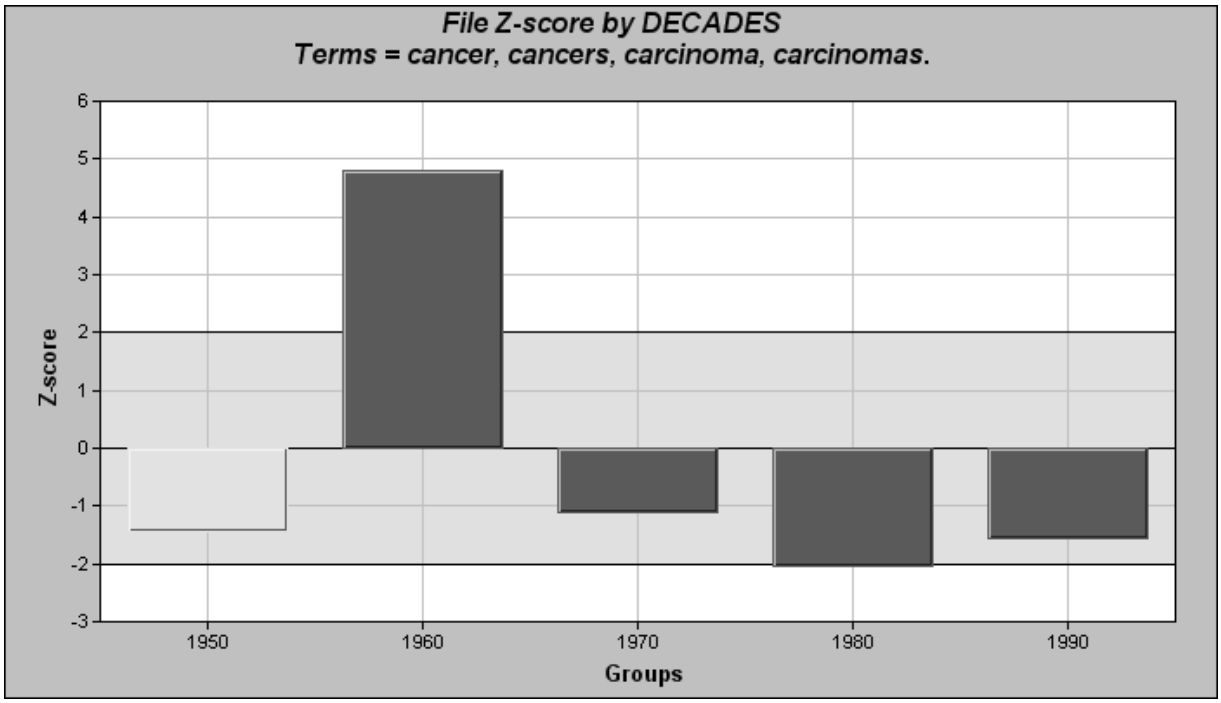


Figure 5.14: PLOT Tool, File Z-score, Cancer Items, Decades

As an interesting counterpart to the above trend, Figure 5.15 is a similar plot of frequency z-scores using the token types *market*, *markets*, and *marketing*. In this case, the trend is nearly the opposite, disproportionally low usage up through the 1970s, but disproportionally high use in the 1980s and 1990s. This too is confirmed by the plot of file z-scores seen in Figure 5.16, although only the 1950s and 1980s show significant variation.

Although certainly informative, the z-score graphs by themselves are lacking in that there is no access to the raw data used to make the comparisons. Although these data are not

strictly necessary for interpretation, they do add considerable depth to any conclusions. As a remedy, the Toolkit results pages provide two additional graphs for each of the z-score graphs, for a total of six. The added graphs are similar to the z-score displays, but contain the actual count data, both as raw counts and as percentages. This is illustrated in Figures 5.17 and 5.18, which are the frequency-data accompaniment to the *market* data in Figure 5.15, and also in Figures 5.19 and 5.20, which are the file-data accompaniment to the *market* data in Figure 5.16. As a final measure of data completion, at the bottom of each results page is a table containing the raw data used to create the graphs. In this manner, the user is provided with the most complete view of the data possible within the confines of the page. An example of this is seen in Figure 5.21. These data were used to construct the market-related graphs in Figures 5.15 through 5.20.

In Chapter 2.4, date-based categories were mentioned as an example of category evolution over the course of the TDC Project. During sampling, decade-based categories were a primary category, so it follows that this would carry over to analysis, which in fact it did, as illustrated in Figures 5.13 through 5.21. However, as the analysis progressed, it was determined that finer divisions of date could provide additional insight into the history of the document set. To this end, but not as a replacement for the decade-based analysis, the Toolkit was reconfigured to provide a view of the data based on sub-corpora constructed from half decades (i.e. 1950–1954, 1955–1959, 1960–1964, et cetera). As an illustration, Figure 5.22 is the plot of *cancer* lemma from Figure 5.13, but using half-decade sub-corpora. Noticeably, the transition period 1960–1979 is much more complex than depicted in Figure 5.13. In fact, the 1970–1974 sub-corpus actually shows a disproportionally low use of the items (although not significant), while the 1970 sub-corpus indicates a disproportionally high use.

As a continuation of the evolution of date-based categories, it was eventually decided that a new tool should be added to the Toolkit to examine the usage of items year-to-year across the full date spectrum of the Quota Sample. This became know as the PEAK tool, as

it allow the discovery of peak usage periods from 1950 through 1999. Quoting from a PEAK output page in the Toolkit,

> The data presented in the graphs. . . were collected using a five-year rolling average to dampen the year-to-year fluctuations and make the data more interpretable. The score for any given year X is calculated by counting the term occurrences in the year span X-2 through X+2 and dividing by 5. For example, the count for a data point representing 1960 is the count of all term occurrences from 1958 through 1962 (59,59,60,61,62) divided by five. This is an average yearly count for the five-year span.

These data were then used to make comparisons to data from the reference corpora which was similarly gathered. As examples, Figure 5.23 is the frequency z-score output from the PEAK tool using the *cancer* lemma from Figure 5.13, and Figure 5.24 is the frequency z-score output from the PEAK tool using the *market* lemma from Figure 5.15. Although these are displays of z-score data, keep in mind that in the Toolkit the PEAK tool output page parallels the output of the PLOT tool. It provides the same six plot types (z-score, raw counts, and percentages for both frequency and file) as well as the table of data used to construct them. In interpreting the graph it is important to remember that the value for each year is based on the average yearly count from a five-year period. In the PLOT tool the given value is based on the sum of the yearly counts for the given period. Thus, while Figure 5.23 is clearly similar to the half-decade PLOT output in Figure 5.22, the peaks and valleys are not expected to match (and in fact they do not). Again, this is simply a different perspective. As a final note on the PEAK displays, notice that together Figure 5.23 and Figure 5.24 add additional support to the trends noted above, the decrease in *cancer* lemma and the increase in *market* lemma.

Although the date-based sub-corpora were selected to illustrate the various counts and analyses provided in this chapter, the Toolkit itself provides access to data from a total of seven groups of sub-corpora. The idea behind this is that by providing multiple views

of the data, the user can gain a more complete understanding of the corpus. Along with the decade and half-decade groups mentioned above, the PLOT tool provides data for the groups 'Shifted Decades,' 'Industry Source,' 'Internal vs. External Audiences,' 'Named vs. Unnamed Audiences,' and 'Bliley and Undated Documents.' As a final illustration of Toolkit analyses, the change in view that can be provided by a different sub-corpus group can be seen in Figures 5.25 and 5.26, which are the frequency and file z-score plots based on industry source for the *cancer* lemma from Figure 5.13. What one immediately realizes with both figures is that cancer-related events are not handled by the individual manufactures, but by the research and publicity groups Council for Tobacco Research (ctr) and the Tobacco Institute (ti).

Admittedly, the fourteen figures provided in this section are a limited representation of the tens-of-thousands of plots that could be created for items with known significant variation in the TDC. The purpose behind these illustrations is to provide a very general overview of the tools and point out some interesting features in order to give the reader an idea of what might be accomplished using Toolkit analyses. In the few minutes that it has taken to read this section, which is comparable to the amount of time it took to generate the plots, evaluate them, and reach the conclusions that were presented, the reader has gained information about trends in the NAAG Snapshot that are yet unknown to most who work with tobacco documents on a daily basis.

Finally, it should be noted that the PLOT and PEAK tools are not the whole of the Toolkit. There are also a number of other features not directly related to quantitative analysis. In particular, all TDC documents are available for download or viewing, sub-corpora of the TDC based on any of the major XML or sampling categories can be created and downloaded, and the TDC can be scanned/reviewed using a Key-Word-In-Context (KWIC) viewer. In combination with the analysis tools these features maximize the useful of the Toolkit as a learning tool by providing a connection to the primary data (i.e. the documents themselves).

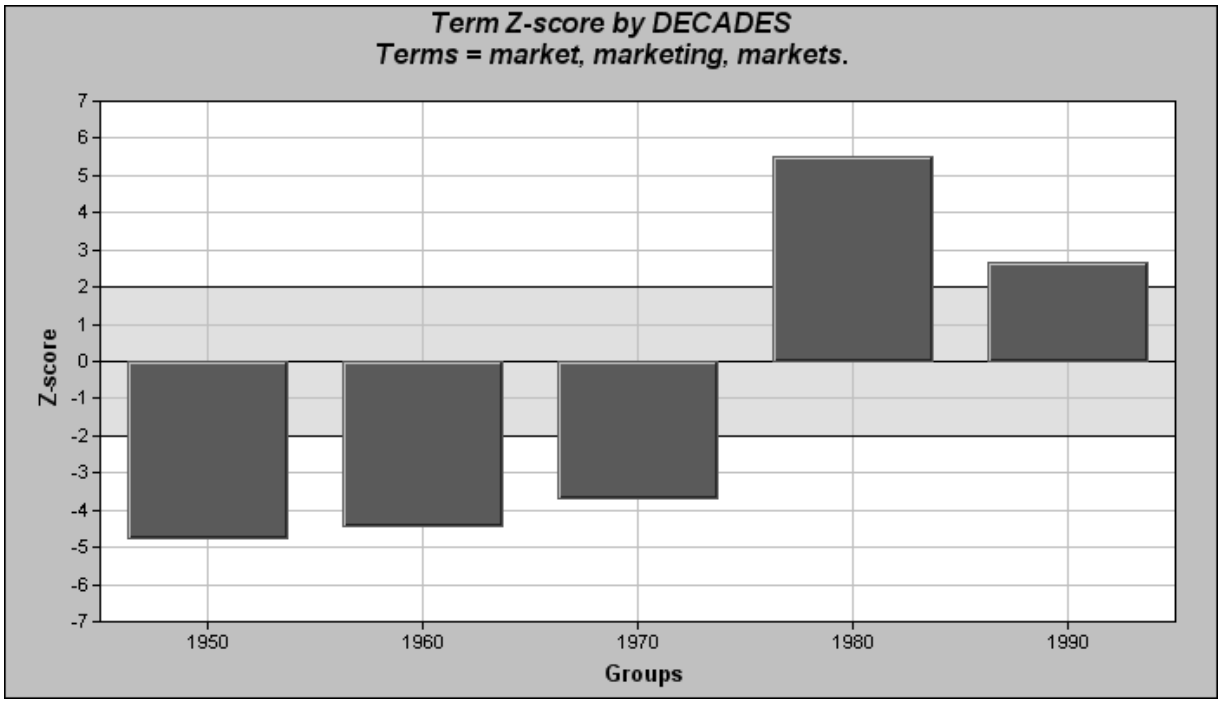


Figure 5.15: PLOT Tool, Frequency Z-score, Market Items, Decades

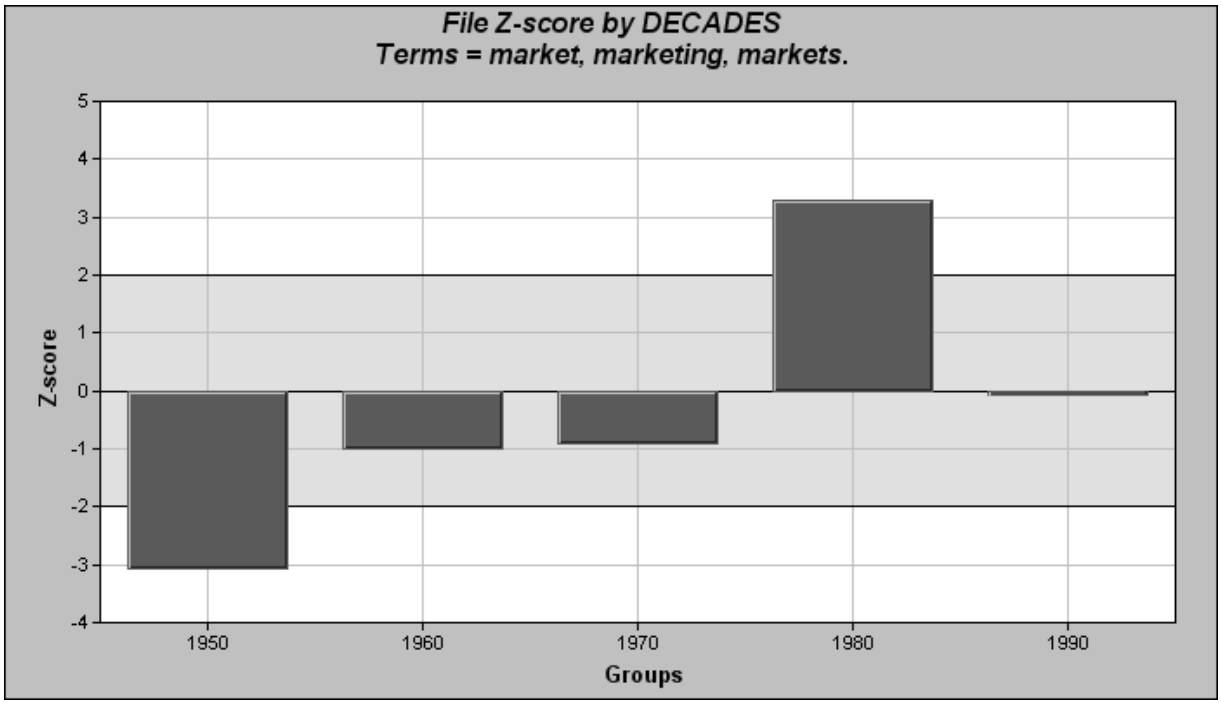


Figure 5.16: PLOT Tool, File Z-score, Market Items, Decades

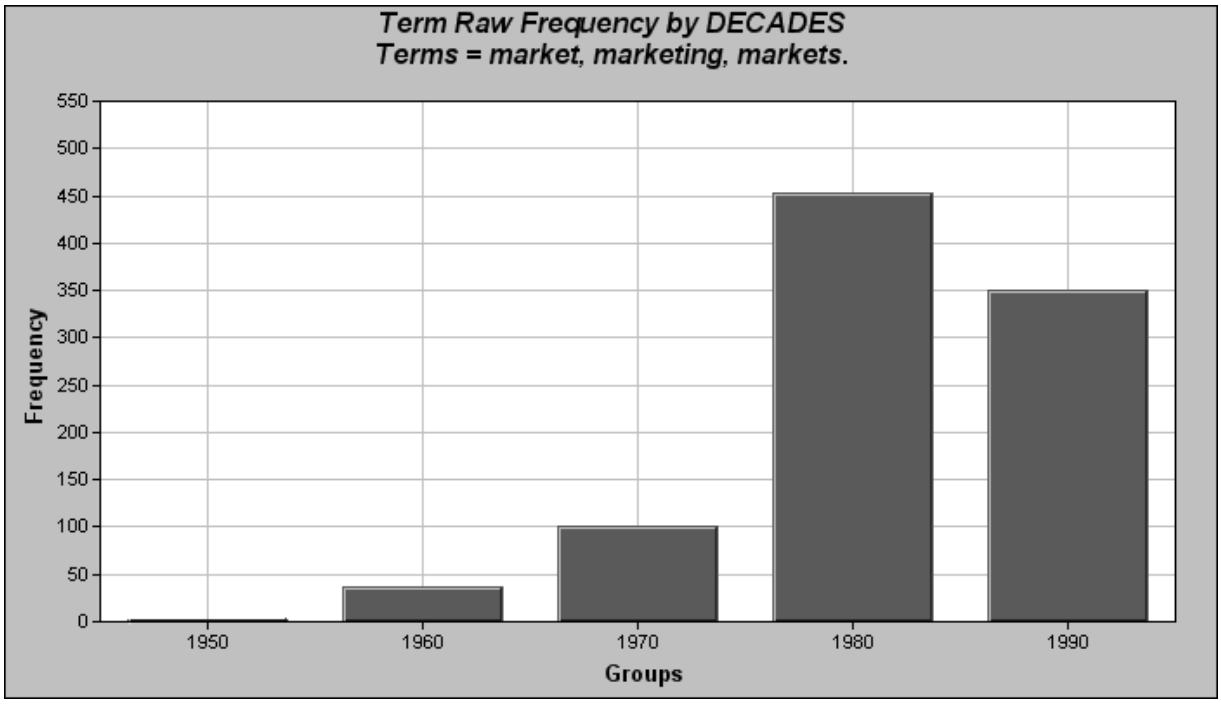


Figure 5.17: PLOT Tool, Frequency Count, Market Items, Decades

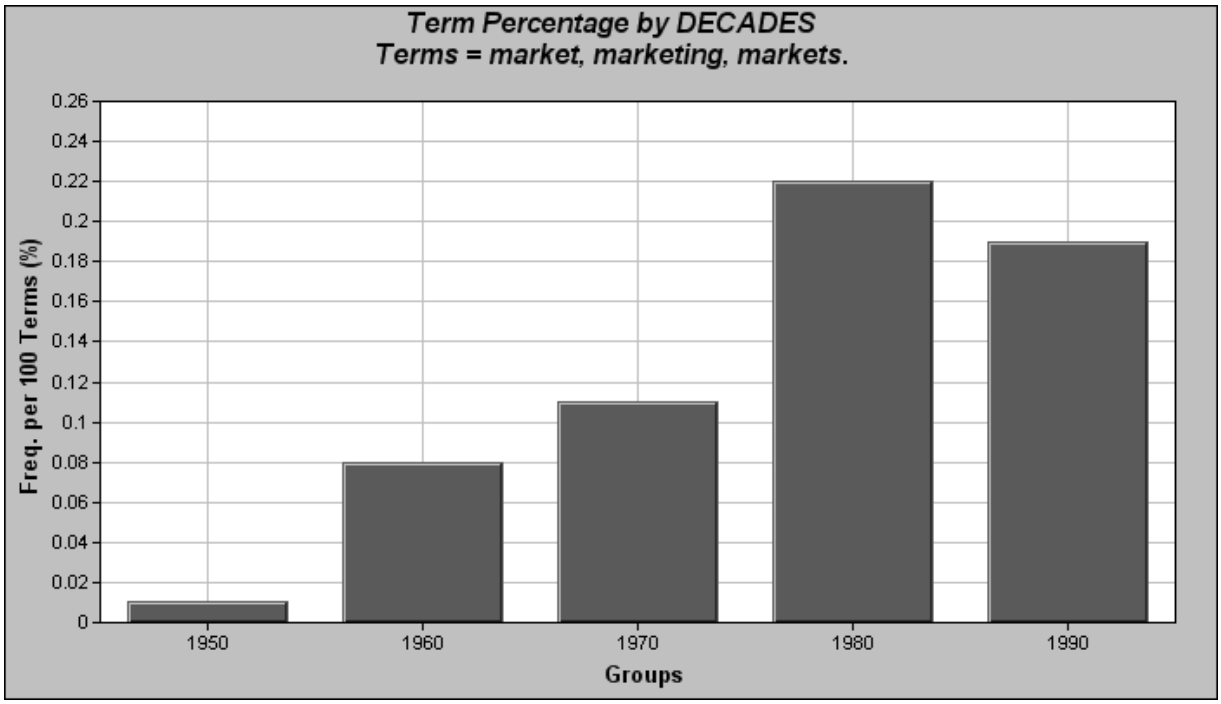


Figure 5.18: PLOT Tool, Frequency Percent, Market Items, Decades

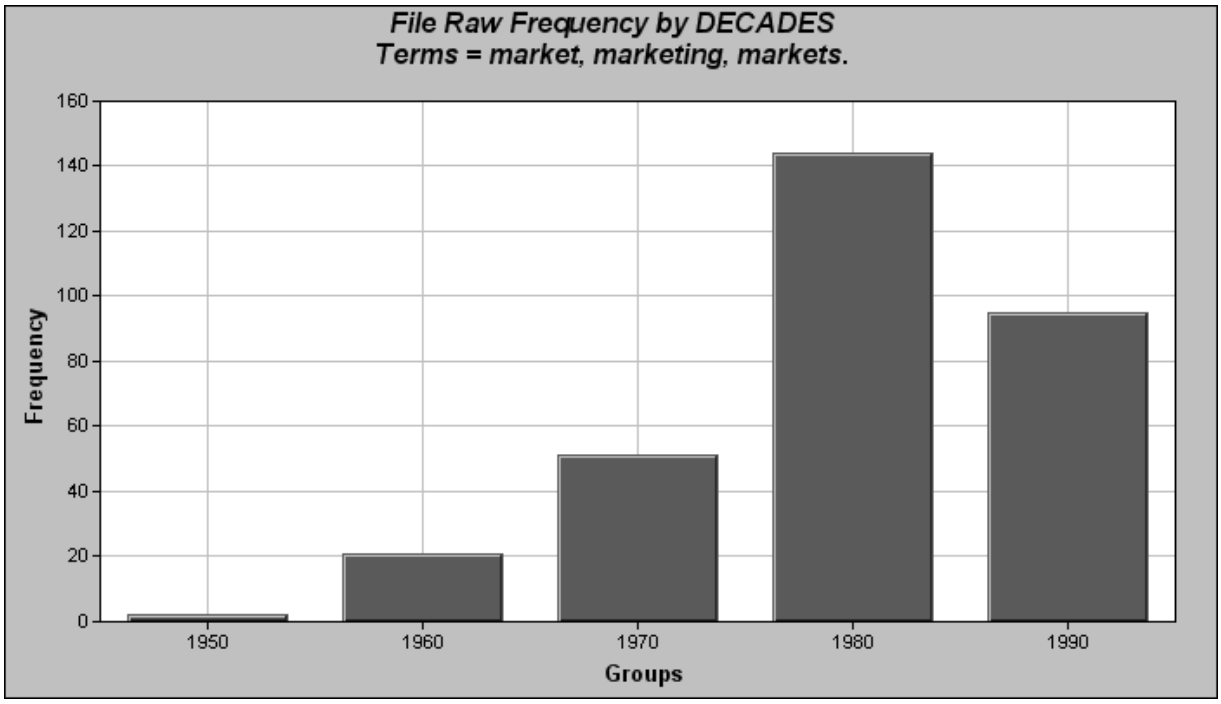


Figure 5.19: PLOT Tool, File Count, Market Items, Decades

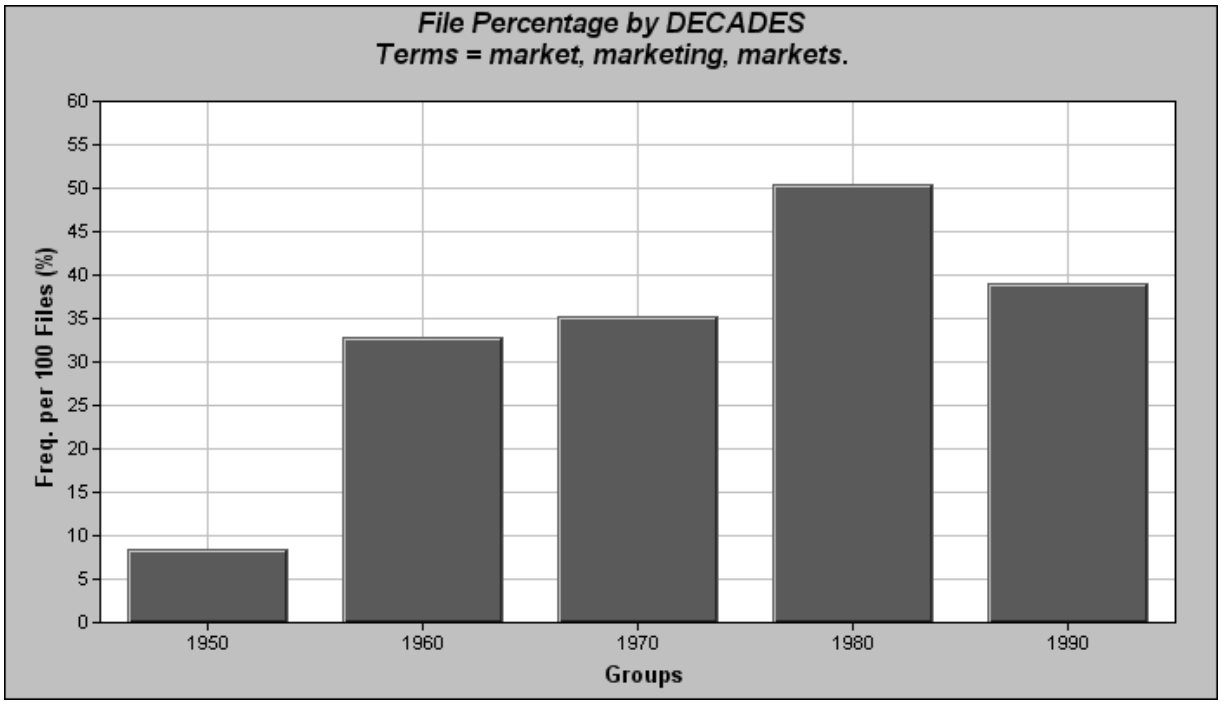


Figure 5.20: PLOT Tool, File Percent, Market Items, Decades

Numerical Data for Terms: market, marketing, markets

Groups	quota	1950	1960	1970	1980	1990
Total Terms	592922	16622	46578	92362	205902	185172
Term Freq.	952	2	36	101	452	351
Term Z-score	n.a.	-4.78	-4.41	-3.7	5.5	2.66
Term Prop.	0.16	0.01	0.08	0.11	0.22	0.19
T. Reliability	ok	ok	ok	ok	ok	ok
Total Files	812	24	64	145	286	244
File Freq.	318	2	21	51	144	95
File Z-score	n.a.	-3.06	-1.0	-0.91	3.3	-0.06
File Prop.	39.16	8.33	32.81	35.17	50.35	38.93
F. Reliability	ok	ok	ok	ok	ok	ok

End of data.

Figure 5.21: PLOT Tool Data, Market Items, Decades

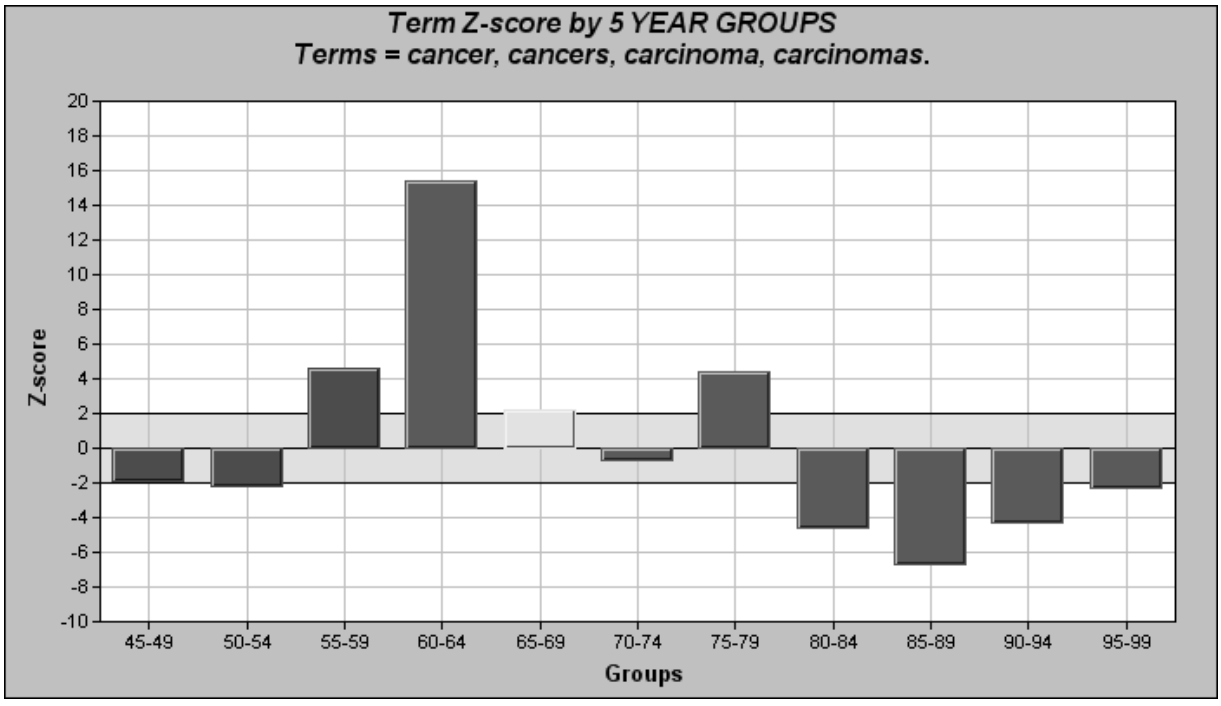


Figure 5.22: PLOT Tool, Frequency Z-score, Cancer Lemma, Half-Decades

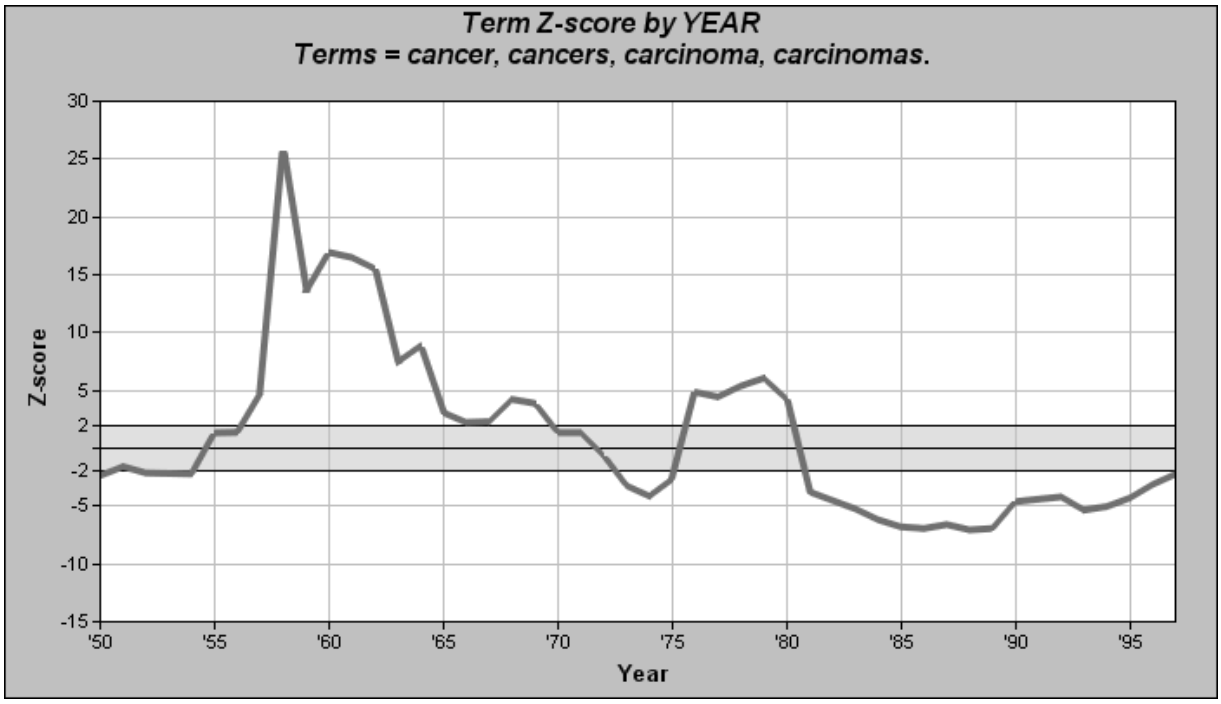


Figure 5.23: PEAK Tool, Frequency Z-score, Cancer Items

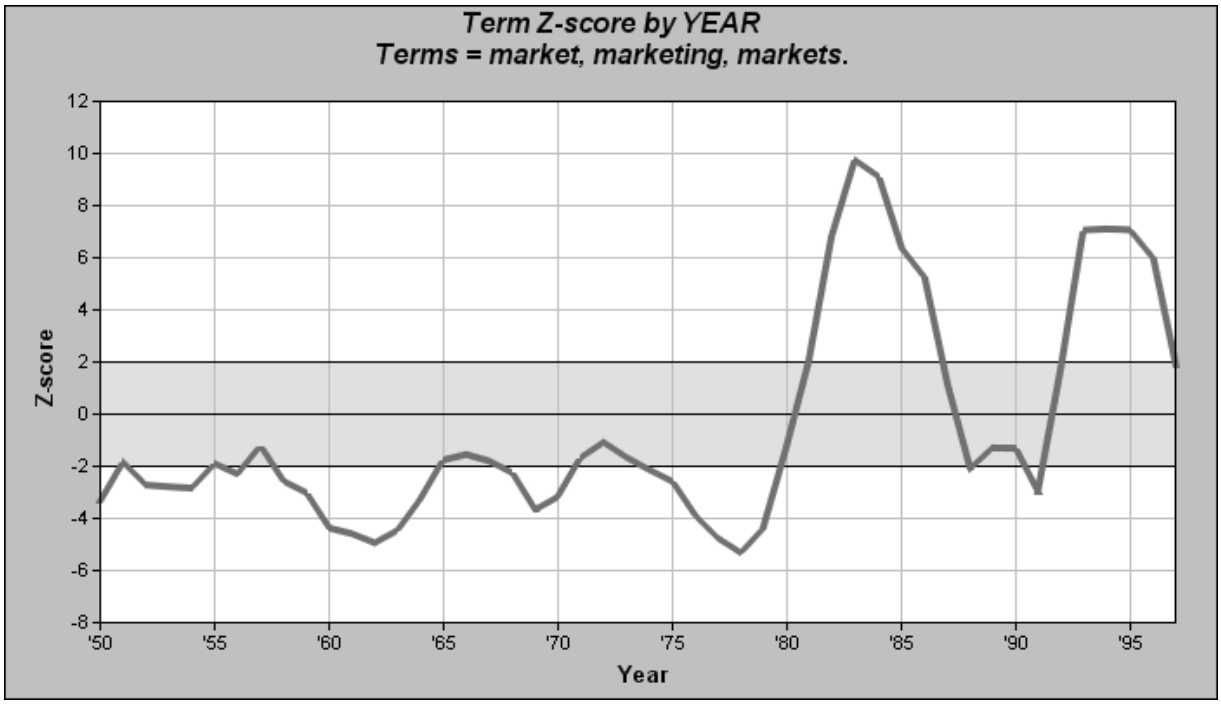


Figure 5.24: PEAK Tool, Frequency Z-score, Market Items

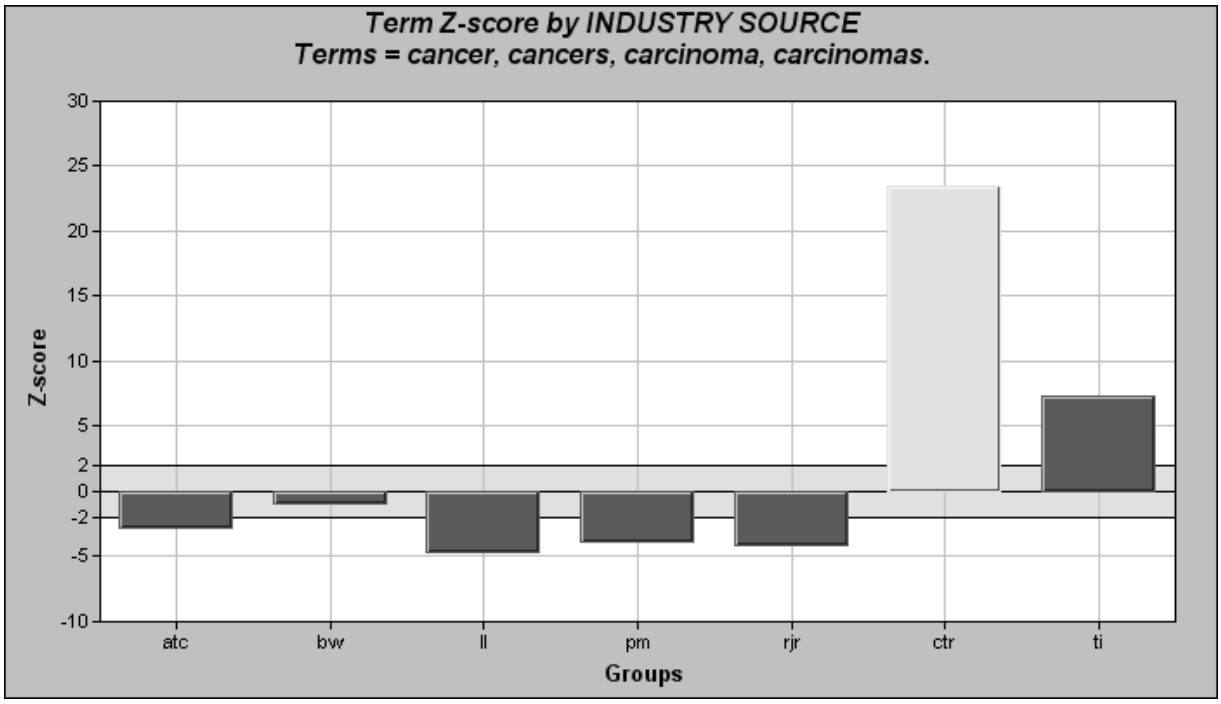


Figure 5.25: PLOT Tool, Frequency Z-score, Cancer Items, Source

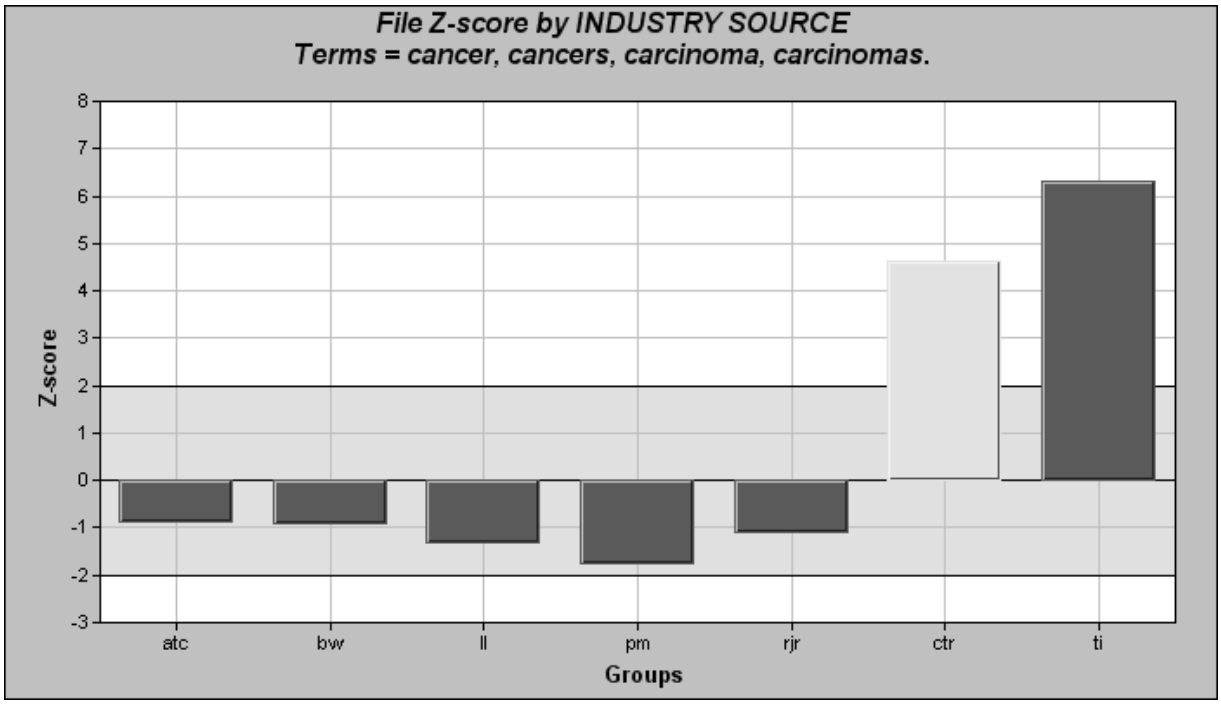


Figure 5.26: PLOT Tool, File Z-score, Cancer Items, Source

Chapter 6

Conclusions

6.1 Introduction

From the beginning my hope in writing this dissertation has been that it might serve as a guide for future studies, both in terms of what was done successfully, and what might be done to improve the procedures. However, not wanting to end with a discussion of what might be improved, I will instead begin with a section on possible improvements and then move to the more general conclusions in Section 6.3 and following.

6.2 Possibilities and Improvements

In presenting the following I have no intention of detracting from the overall success of the TDC. However, as with any practical exercise, knowledge comes with the experience, but it often comes too late to be properly implemented. Thus in fairness to above goal I have reserved this section for a few additional notes concerning things considered but left undone. In this manner, the reader might be prompted to not simply duplicate TDC procedures, but improve them.

6.2.1 Taking Notes

The design decisions, procedure descriptions, and results data presented in this work are based on some forty pages of typed notes and 20 spreadsheets made by the author during the course of the project. This was an attempt to describe and document the procedures and decisions made in the assembly of the Tobacco Documents Corpus over the three-year

period during which I participated. Yet as I sat down to write I was forced to admit that those 14,000 plus words were inadequate. In the preceding chapters there are points, duly noted, where information is lacking, and there are descriptions which seem very simple now that they are on paper that took days, some even weeks, to piece together. As I worked, I could often remember how familiar I had been with the data, but not remember the specific facts. This is, of course, the trap. In the middle of the procedure, when I was intimately acquainted with every detail, it seemed a waste of time to write even the notes that I have. Yet, at this point I could not have written this work without them, and I have frequently regretted that I did not take an extra five minutes each day to make them more thorough.

In short, there is simply no substitution for taking detailed notes. Every decision, every procedure, no matter how seemingly insignificant it seems at the time, must be recorded. In the end, the data produced by a project or study are only useful to the extent that one understands how they were produced. Apart from notes, the details will fade with time, so write then down. And as a final note on taking notes in the computer age, it really does not matter what program one uses to take notes, but it is advisable to export and save them in a text format. Text data can be imported into almost every program and has little risk of becoming obsolete. During the sampling and construction of the TDC, data were kept using spreadsheet software that is now no longer in the mainstream, and recovery of that data was problematic.

6.2.2 Revisiting XML Schema

Given the other schema used for corpora markup, such as the original schema of the Brown Corpus of Standard Written English and the British National Corpus, XML is in my opinion the best option for general corpus archival and markup. Although there are many reasons for this opinion, from the explicitness of the DTD, to the ability to name tags, to the general availability or XML tools, they are all related to the simple fact that data archiving and recovery is very straightforward. However, just as with all categorization exercises (which

is what markup generally is), XML does suffer when there is a lack of a priori knowledge. That is, the DTD must be designed prior to beginning the archival process. In many cases, such as ours, this means prior to any practical knowledge of the structures and data that will be encountered during conversion. And unfortunately, although XML is very open to individual design needs prior to the beginning of archiving, it becomes progressively more fixed as documents are converted, simply because the costs and time involved in returning to the archive files to make corrections and changes becomes greater. Thus what seemed the best method (DTD) for archiving prior to the event may be found to be much less valuable or even detrimental once the process begins, and corrections may require that the process be essentially redone. In other words, if half way through the process it is discovered that rebuttals to certain rhetorically important arguments should have been marked but were not, the only way to do this reliably would be to reread the sample documents, which may be too costly given the available resources.

In the case of the TDC, while it cannot be said that all of the tags used in archiving were necessary, certainly many have not been used as they were intended, I am not aware of any gross errors with the DTD that directly affected text recovery. There were, however, several DTD issues that in terms of processing made recovery of data more complicated. The overall issue is that the DTD was not made as strict as it should have been. The result is that the archive format is less regular than would be optimal, and it is not always clear where certain data will be located. Automated processes require that the location of data be clearly defined, and these leniencies make it more difficult to define the location.

As an example, perhaps the most obvious of errors in this regard is that as a convenience to the archivists the DTD allows the use of empty tags to mark paragraph and non-paragraph breaks (the `<p>` and `<npb>` tags). This was a hold-over from HTML, where the author had previous experience. The common practice for HTML is to mark the beginning of new paragraphs with an empty tag (`<p/>`) in order to alert the browser program to display the paragraph change appropriately. While this is perfectly acceptable for display purposes (the

intent of HTML) where all that matters is the beginning, it is lacking as a means marking text for recovery because it becomes difficult to define the paragraph end. That is, any tag following the text of a paragraph may belong to that paragraph, the subsequent paragraph, or neither. There is really no clear way to decide apart from manually reading the archive. Thus locating paragraphs in an automated manner becomes much less definite. While this is not a huge problem, and certainly does not hinder the ability to recover the text, it is a nagging issue in that it places unnecessary limits on how well the recovered text is defined. That is, one often desires text in more discrete portions rather than the entire text 'blob' from the document (as the database folks would call it).

The cure for this problem, as might be expected, would be to not allow empty tags, but instead require all document data to be within the opening and closing tags of an item-type tag. In this way the `<p>` and `<npb>` tags would become equal siblings with the other item-type tags, such as `<image>` and `<table>` tags, and all unbounded text would be eliminated. This would in turn would remove all the uncertainty and provide a much cleaner, easier-to-manipulate archive. Unfortunately, it was in hindsight that it became clear to us that all data should be bounded, and we were not able to devote the necessary time to make this change in the TDC archives. Given he nature of the problem, there was not a simple automated means of recovery given that what cannot be reliably located, cannot be reliably changed.

The empty-tag problem was by far the most perplexing DTD issue. However, there were a number of other minor issues that, although easily overlooked, add noticeably to the complexity of the archive. For example, a number of times during processing it became apparent that there were more optional items in the DTD than necessary. This was done in order to accommodate the complex structures of the tobacco documents. However, the DTD ended up being too lenient, allowing inconsistencies to creep into the archives. Another example is that the `<p>` tag was applied inconsistently to mark text, allowed by the DTD and required by procedure in the `<text>` tag, but not allowed in the `<pretext>` or `<posttest>`

to mark divisions in text. It was assumed that these data were not analyzable (but they may well be). Finally, several of the tags had overlapping purposes, which aside from confusing the archivists, lead to inconsistencies. The `<note>`, `<description>`, and `<symbol>` tags each allowed unspecified, non-document PCDATA for description; both the `<h>` and `<p>` tags were used to mark text divisions. But again, it was in hindsight that these issues became apparent. By experience it became obvious that more specificity could have added to the consistency of the overall corpus archive and not detracted from the DTD's ability to accommodate complex document formats. I am not certain that more care in the beginning would have overcome our inexperience.

6.2.3 Revisiting the Proportions Test

It was noted earlier in Section 5.6.1 that as sample size increases, for a given proportion z-scores tend to become more extreme. For example, an experimental proportion of 0.49 compared to a normal proportion of 0.51 has a z-score of -0.283 when the population sizes are set to 100. However, when the population sizes are set to 10,000, the z-score becomes -2.83, and with a population of 1,000,000 it becomes -28.3. This is not a fault in the statistic, but a common property of z-scores. As the sample population goes up, it is natural to expect that the probability of error would go down.

In the case of a single comparison, such as comparing a proportion from the 1950-decade corpus to one in the Quota-Sample corpus, this presents no problem. However, when the z-scores from multiple comparisons are displayed together, as with the display of the five decade-based corpora in Chapter 5.6.3, the interpretation is not as straightforward as it was first thought to be. The problem occurs when data from relatively small corpora are compared to data from relatively large corpora, in which case the same proportion may yield noticeably different z-scores. The common mistake is that one misinterprets the z-score as an indication of prevalence, leading to the conclusion that the corpus with the higher z-score has more of whatever it is being examined. Although this would be the case in a

single comparison in which both populations are fixed, or even with multiple comparisons of same-size populations, it is not the case between multiple comparisons with different size populations. It may be that the larger corpus has a smaller proportion, but a higher z-score because of its size.

As an example, Figure 6.1 illustrates the difference in z-scores produced when two hypothetical corpora of different sizes are compared to the same reference corpus using proportionally equal counts. For demonstration, the size of the two experimental corpora differ at a ratio of 1:10 based on total-token count. The smaller of the corpora has a total count of 15,000 tokens, while the larger has a count of 150,000 tokens. These sizes are comparable to those of the 1950-decade and 1990-decade corpora (respectively) used in earlier examples. The size of the reference corpus was set at 500,000 tokens to be similar to that of the Quota-Sample corpus. In terms of type counts, for the reference corpus the count of the hypothetical type T in question was set at 500, which is 0.1 percent of the total token count. This remained constant for all comparisons. For the two experimental corpora, the count of type T ranged from 1 to a value equal to 1.0 percent of the total token count (150 for the smaller corpus and 1,500 for the larger) in 150 equally-spaced steps. The data for the two plots were produced by 150 trials in which the various counts of T in the experimental corpora were compared to the count of T in the reference corpus. This was done using the Python function *prop_compare*$(x1, n1, x2, n2)$ found in Figure 5.9 and in a manner similar to the following:

```
for x in range(1,151,1):
    zv1 = prop_compare(x,   15000, 500,500000) # small corpus
    zv2 = prop_compare(x*10,150000,500,500000) # large corpus
```

In all 150 trials the v-score returned for both comparisons was always 1. Consequently, these data are not accounted for in Figure 6.1.

Interpreting Figure 6.1, the X axis represents the count of T expressed as a percent of the total token counts for the experimental corpora. The Y axis is the z-score returned from the comparison of the given T value to the reference corpus. Notably, the plots of the

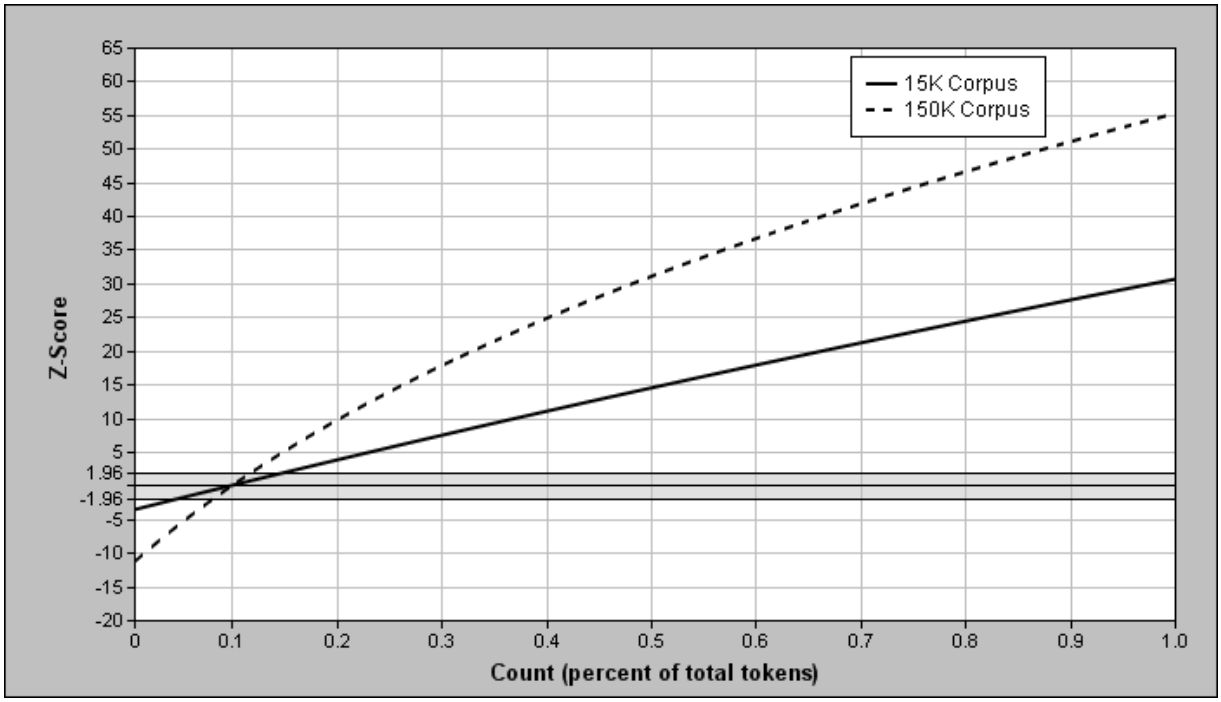


Figure 6.1: Proportions Test: Cross-Corpora Comparisons

two corpora cross where counts are 0.1 percent of the totals, which is the point where the proportions of type T in the experimental corpora are equal to the proportion of type T in the reference corpus (thus a z-score of 0.0). However, moving from this proportion in either direction shows that the z-score of the larger corpus diverges from the zero-line at a higher rate than that of the smaller corpus, even though the proportions remain equal. In order for the z-scores to remain the same, the proportion in the larger corpus would actually have to decrease.

During the early stages of Toolkit development this problem with the juxtaposition of comparison data from corpora of different sizes was noted. As an aid to interpretation, additional plots of raw counts and percentages were added to the Toolkit displays. The hope was that the combination of data would add perspective to the trends seen in the z-score data

and increase the likelihood that the z-score would be interpreted correctly as a confirmation of the observed proportion (given as a percentage).

6.2.4 Revisiting Toolkit Core Data

In writing Chapter 5, I was once again reminded of the utility of scanning well-sorted tables of data, such as those in Appendix E.3, as a means of rapid assimilation. For those of us who study corpora, I would venture to say that the single most formative event in our career was the first time we saw our own data in a KWIC (Key-Word-In-Context) display. As simple as KWIC displays are, they forever alters one's perspective, often providing insight in just a few moments that would take years to develop by more traditional methods. In a similar way, well-constructed data tables provide insight into the text that is difficult to obtain in other ways (if it is even possible). Again I refer to personal pronouns in the Quota-Sample corpus. What is obvious in a data table would be elusive in reading (at least to me).

This being the case, what the TDC Toolkit lacks is the perspective that could be provided from the core data arranged in tables (greatly reduced in size) like the ones in Appendix E.3. Of course, the data exist and are used in the background to generate the comparisons. However, they are not directly available to the Toolkit user. The main reason for this is that at the time data storage was a major consideration. As it stands, the Toolkit has 645 MB of data. Adding the data tables for the 33 sub-corpora used for the PLOT tool could have easily pushed the total to several GB, which was simply too much data to include on a CD, which was originally the primary means for distributing the Toolkit. However, given recent advances in data storage, and the fact that the web-based Toolkit is now the primary means of distribution, the display of core data is worth reconsidering given the value added.

6.3 General Conclusions

When first admitted to candidacy in January of 2003, I had not yet considered the TDC as a topic for my dissertation. In fact, the case was quite the opposite. Although we had made good

progress and had realized a number of smaller successes—the sampling was complete, and conversion and markup procedures had stabilized—at that time I still considered the TDC as a whole to be an experiment. In other words, the TDC was a good idea that had evolved into a rigorous method, but its ultimate success in terms of usefulness was still undetermined. This was compounded by the fact that we were unaware of similar studies/projects that might serve as proof of concept. We believed that the TDC Project would be successful (useful), but were unsure exactly how that success would be manifested. During the remainder of 2003, however, a few key pieces of the puzzle came together and allowed insight into the potential usefulness of the TDC. The major event that year was the completion of document conversion. This moved the TDC from a theoretical entity (a list of Bates Numbers) to a usable product (a set of data that could be manipulated and analyzed). It also allowed us to begin experimenting with XSLT and other data transformations, and to begin evaluating different methods of analysis. By October of 2003 analysis had progressed to the point that we were convinced of the success of the TDC Project. That is, from our experience with the corpus to date we could see that the TDC was producing valuable data and would serve well as a benchmark for further tobacco-document study. Although the evidence was sufficient at this juncture, I believe, to declare the TDC Project a success and therefore complete, our thoughts were that a much more logical end to the project would be making our collective work, both the TDC archive and the associated analyses, accessible to other researchers (not just linguists), not as a statement of finality, but as a means of investigation and an example of what might be done with the TDC. This being the case, from mid October 2003 until the project end in June 2004, the focus shifted from analysis (although we did continue with analysis) to developing the methods of presentation that eventually became the TDC Toolkit.

In many ways, the completion of the Toolkit marked the point which the TDC Project reached maturity. By this time, we had rigorously sampled an overwhelmingly large set of documents to create a representative corpus, which to my knowledge is the most principled

and well-formed representative corpus to date; we had converted the sample to a richly-encoded archive with target data clearly marked to allow straightforward extractions, not relying on machine decisions; we had produced meaningful secondary data using unique applications of proven analysis methods (mass comparisons, rolling averages, cross-corpus comparisons, et cetera); and we had developed a number of methods to make the TDC, both corpus and data, available to other researchers, to include tools for generating user-specific sub-corpora, for producing KWIC displays of target items, and for displaying analysis data (the PLOT and PEAK tools). In this manner, the TDC moved from concept to application, with the end result of the experiment being that a seemingly unapproachable set of documents was made accessible, both in the establishment of a norm, and in providing a means for investigating that norm. Thus the success of the TDC as a project was self-evident in the product.

As our confidence in the TDC as a product began to increase over the final months of the project, I began to realize was that there was also considerable value in the TDC Project as a process, meaning that what was done with the TDC could also serve as a process example for other researchers with similar goals. Of course, in order to serve as a process example, thorough description of the process would be required, to the extent that replication becomes possible. It was at this point that I began to view the TDC process as the logical topic for my dissertation. In other words, it seemed to me that the best conclusion to the TDC Project would be its description. In this way the work done for the TDC would become maximally useful, not ending with our presentation of data, but potentially extending to other similar projects.

Although not apparent at the time, the description of the TDC process actually began in Chapter 2 with the discussion of category. As I began to examine the process itself, it became clear that the overall success of the TDC Project was largely a result of choosing strong categories and insisting on clear definitions, although at the time we were not using Chapter 2 terminology (we simply insisted on rigorous definitions). It also became clear that

the converse was true. Weaknesses in the process, as described in the previous section, were in many respects the result of poor category choices and/or a lack of clear definitions. Even the above section on note taking addresses categories given that a lack of notes is essentially a lack of definition. Thus the common thread that unites the full discussion of the TDC process is category. Any conclusion we make about the data presented in Chapter 5.6 or the Toolkit can (and should) be traced back, category by category, to the earliest decisions made in Chapter 3 related to the sampling domain, while at the same time being conscious of our definition of category itself. In other words, discoveries are pinnacle events, held aloft by a mountain of other categories, which themselves are supported by the category doctrine we embrace. In much the same manner, the core chapters of this work are as much about interdependency of processes as they are about the methods presented.

Ultimately, however, the reader's acceptance of the given premise as a conclusion is based on experiencing discovery while using the described methods and data. If the claim is that a large document set such as the NAAG Snapshot can be managed, described, and manipulated to provide an understanding of document types, content, events and structures apart from extensive review of the individual documents, then the proof of that claim is that this type of discovery has been experienced to the extent that one can say, 'Yes, it works. I've seen it myself.' Although it is admittedly limited, the data presentation in latter half of Chapter 5 does provide the opportunity for discovery, particularly if one makes use of the data provided in Appendix E.3. A well, presentation of data allowed a number of unique adaptations of statistical methods to be demonstrated.

The premise of this work is that there is a straightforward and reliable means for approaching large document sets, that through the integration of methods from Corpus Linguistics, Humanities Computing, and Statistics a general understanding of document types, content, events and structures can be reached apart from extensive manual examination of texts. The foundation for proving the premise, the necessary evidence, is the preceding description of the TDC process (or processes), which I believe is complete to the extent

necessary to validate the premise. That is, the description, examples, and data contained in Chapters 3, 4, and 5 are sufficient to provide the reader with an understanding of the potential these methods have for addressing issues related to large corpora.

6.4 Process Planning

It has been noted a number of times throughout this work that decisions made in the completion of certain sup-processes significantly affected the outcome of later events. For example, in Chapter 5 one can see that analysis depends on count methods, which in turn depend on parsing methods, which depend on tokenization. In actuality, this dependence extends to the very earliest decisions made, such that analysis can be traced directly to decisions concerning the sampling domain. Thus to a great extent the project can be viewed as a series of steps, each of which involves a number of decisions that determine not only the outcome of the current subprocess, but also govern the outcome of subsequent steps. In other words, the subprocesses are like links in a chain from which the outcome of the full process is suspended, such that the weight (or value) of the results is dependent on the strength of each link. This being the case, the overall goal in project design is that each subprocess be as strong as possible. This comes by careful consideration and planning.

As an aid in this regard, the list below provides a number of items, both theoretical and practical, that should be considered in planning a project similar to the TDC. This list was derived from the experience gained over the course of the TDC Project, from initial concept through description, both in what was done and left undone. Although the items are roughly in the order that they would be encountered in the progress of a TDC-style project, they are not intended to represent a protocol or fixed algorithm for success, but instead are given to serve as a summary of the types of decisions necessary for success. As such, they can become the basis for a method, adapted as necessary to the given project. However, the key to the successful use of the following items, because they are interdependent, is that they be

carefully considered during project planning (i.e. prior to the onset of the project) such that all actions (and inaction) are intentional.

1. Project Records - As mentioned above, the usefulness of the data produced over the course of a project such as the TDC Project, particularly in relation to the creation of corpora, is ultimately derived from a description of the process methods such that subsequent researchers can properly interpret and implement those data. This being the case, a detailed record of the full course of the project is a necessity. All decisions, procedures, and resultant data should be permanently recorded. Where the project records are lacking, the project description, often written long after the initial decisions and processes are complete, will also be lacking. Consequently, one of the first tasks of a project manager is to establish a protocol for record keeping sufficient to document the details of not only each steps in the process, but all preliminary decisions and their rationale (to include decisions related to the items below). Refer to Section 6.2.1 above for additional discussion of notes and records.

2. Major Categories - Although the defining and evaluation of categories is the dominant topic throughout all the items listed here, at the onset of any project the major categories of the study should be subjected to particularly rigorous evaluation. It is necessary to not only identify and define the categories that we ultimately want to compare, but also insure that the complements are not being elevated (i.e. being compared to the categories) as this would greatly reduce the value of the project results (if it did not render them completely invalid). As well, the various subprocesses must be designed to insure that sufficient data are collected to allow the desired comparisons. Refer to Chapter 2 for additional discussion of categories and potential errors, and to Chapters 3, 4, and 5 in general for descriptions of the subprocesses necessary for analysis.

3. Sampling Domain - The main objective in considering the sampling domain, apart from defining it clearly, is determining whether or not the domain is bounded in a practical sense in order to plan the sampling procedures. For example, while the domain of both the TDC and the Brown Corpus are defined and theoretically bounded, practically speaking they must be approached differently in terms of sampling. With the TDC, all the documents in the domain are accessible, which makes it bounded in a practical sense and allows a representative corpus to be constructed. However, for the Brown Corpus, which has as its sampling domain all published English documents from the year 1961, the domain is unbounded from a practical standpoint. That is, one does not have access to the full domain, and consequently a representative sample cannot be constructed. In this case, adaptations of the sampling plan must be made to insure a well-rounded (reasonably representative) corpus. Refer to Chapters 2.2 and 3.2 for additional discussion of sampling domain.

4. Sample Size - At a very practical level, in most situations the sample size (and ultimately the corpus size) is limited by resources and time. In other words, we must make a practical compromise between our theoretical desire for large corpora and our ability to complete construction, which includes not just the sampling, but also the resource-intensive tasks of conversion and markup. In making initial estimates of sample size it is better to err toward underestimating resources than overestimating. The reason for this is that underestimation insures completion of the corpus, which can then be augmented by resampling. However, overestimation produces an incomplete corpus which may be difficult to repair such that it remains representative of the sampling domain. An option to consider is a hybrid approach like the one used by the TDC in which multiple smaller (but representative) samples were combined to form the larger representative whole (see Chapter 3.4.4). In this manner, sampling and conversion can continue until resources are exhausted, at which point the larger representative corpus can be assembled from all complete and converted sub-samples. As well, the corpus

can easily be augmented as resources become available. In the case of the TDC, the sampling plan allows the corpus to be augmented in sub-samples of 202 documents. Refer to Appendix A.3 for a discussion of TDC sample size.

5. Preliminary Investigation - The key idea here is that in order to design a sampling procedure that will insure the resultant corpus is representative of the sampling domain, a reasonable understanding of that domain is necessary, both in terms of content and measure (what and how much, for determining quotas), and format and availability (for developing sampling procedures). In most cases, as with the TDC, this cannot be had from casual observation but involves some form of disciplined preliminary investigation. Although the specific method may vary, generally an analysis of a limited random sample of the domain is necessary. Refer to Chapter 3.3 and Appendix A.2 for additional discussion of preliminary investigation and sampling.

6. Sampling Frame - In general, developing a set of sampling quotas (a sampling frame) should be the natural progression of the *information* gained from the preliminary investigation. The key idea is representativeness, that any corpus of documents assembled according to the sampling frame will be a model of the sampling domain (assuming a bounded domain). In practice, there is some expectation that the development of the sampling frame following the preliminary investigation of the sampling domain will cause the domain itself to be redefined. This is simply because the two are explicitly interrelated, the sampling frame actually being a continuation of the definition for the sampling domain. For example, with the TDC the original sampling domain for the reference corpus (later the Quota Sample) was the set of NAAG Snapshot and Bliley documents 'in which manipulation could have occurred.' Following the preliminary investigation (the Core Sample), the 'in which manipulation could have occurred' section of the sampling domain definition became much more specific, namely a narrowly defined set of document types. Refer to Chapter 3, Sections 3.4.1 through 3.4.3, and

Appendix A for additional discussion of the TDC sampling frame and quota development.

7. Sampling Procedures - In most cases, the sampling procedures will evolve directly from the *experience* gained from the preliminary investigation. The key concern is that they be not only well defined, but as intuitive and straightforward as possible, matching both the goals of the project and the sampling environment. This may require modifications to more standard procedures. For example, with the TDC, access to the sampling domain (the documents themselves) was provided by online archives. To facilitate sampling, we adapted more common methods of random selection to accommodate (i.e. work with and through) the search-engine gateway to the document archive. Although the description of the method is tedious, in practice it was much simpler than the alternative, namely gathering metadata for all domain documents and making the selection locally. Refer to Chapter 3.4.4 for additional discussion of general sampling procedures.

8. Sample Validation - Assuming a well designed sampling frame and manageable procedures, the reliability of the sample is determined by how well the quotas are followed, and in particular the accuracy of classifications made in the selection of documents. In most cases, classification will be a qualitative process completed by the individual conducting the sampling. Thus some form of validation, generally a second, independent review of the documents, is necessary to minimize the potential for error (misclassification). As a practical measure, the project manager should plan to validate and finalize the sample (or sub-samples) as early as possible and produce a master record of the sample documents and associated metadata. The record should be clearly formatted and accessible to anyone working on or with the corpus. Although this may seem an odd item to include in this list, a clear and concise record of the sample documents is a valuable tool in working with a corpus, particularly for managing the archivists workload. For the TDC, the primary means for sample validation was a

comparison of sampling metadata with archivist metadata (see Chapter 3.6). While this did provide the opportunity for the independent classification of documents by multiple readers, often that opportunity presented itself late in the overall process and consequently made adjustments to the sample difficult when errors were found (i.e the sampling process was complete prior to the discovery of the error by the archivist). Refer to Chapters 3.4.2 and 3.4.6 for additional discussion of sample validation. Refer to Appendix B for the TDC sampling records.

9. Markup Schema - Assuming that the necessity of a well-defined schema is established, the key consideration in terms of markup schema is whether or not established schema are suited to the needs/goals of the project. Generally speaking, if it is not detrimental to the project goals to use a standard schema (or overly burdensome to the coders), then it is advisable to do so as it will be better understood by secondary researchers. That is, adhering to known standards can allow the resultant work to be distributed and used more easily. However, there is no requirement, and often no advantage, to using existing schema, particularly for non-standard purposes. In these cases the research is better served by a carefully designed schema that captures the target data. Another idea to consider when developing markup schema is that in practice markup is done in a limited number of passes, often only one, simply because markup is very resource-intensive. This being the case, to the extent practical given the available resources, the schema should be augmented to accommodate not just the known data targets, but all potential targets. That is, if one suspects there will be a use for a particular data type, it makes sense to include a tag for those data in the schema and to mark them during the initial coding process. This increases the value of the archive by maximizing the opportunity to mark data during the manual examination of documents (thereby minimizing the need for reexamination), and helps to offset the effects of missing a priori knowledge (i.e. we do not always know all that we want to study when designing markup schema). Refer to Chapter 4.4 and Appendix D for additional discussion of

the TDC markup schema. Also see Section 6.2.2 above for a discussion of errors in the TDC markup schema.

10. Markup Procedures and Training - As mentioned above, practically speaking archiving is often done in a single pass. Thus, once the archivist leaves a document, time and resource constraints generally prevent returning for extensive revision. This being the case, it is necessary to have well defined procedures that match the goals of the project and are as intuitive as possible. As well, plans should be made to thoroughly train the archivists prior to beginning the document conversion. Ultimately, the value of the archives rests in the hands of the archivists, so it should not be assumed that they understand the process apart from training and testing, particularly if the markup schema departs markedly from more common schema, as was the case with the TDC. Refer to Chapter 4.5 for additional discussion of markup procedures.

11. Markup Validation - Although the focus of validation is generally on post-conversion methods, which are certainly necessary to insure consistency, plans should be made to validate coding throughout the course of the conversion process, readjusting procedures and making repairs as needed. This is necessary given that time and resource limits may preclude returning to the archives manually to repair major errors discovered at the end of the conversion process. Specifically, in order to avoid surprises late in the process one might consider pre and concurrent forms of validation such as archivist 'certification' to insure that coders are able to apply the markup procedure accurately (prior to working on the permanent archive), editing software that requires the archivist to follow the project schema, regular sampling and checks of completed documents, computer-assisted consistency checks, and if possible, independent coding of documents by multiple archivists for cross-coder validation. Refer to Chapters 4.5.4 and 4.5.5 for additional discussion of markup validation.

12. Text Production - There are two major questions to consider with text production. First, of all the data available in the archive, how can those data relevant to analysis be separated (extracted) from the irrelevant data. Second, once the text data is extracted, how should it be normalized (tokenized) to allow uniform processing (i.e. consistent according to the researcher's expectation). Of all the subprocesses in corpus studies, I think none are overlooked more than extraction and tokenization. However, as simple as they may be, these are pivotal processes and must be well defined if subsequent processes and data are to be understood. Refer to Chapters 5.2 and 5.3, as well as Appendices E.1 and E.2, for additional discussion of text production.

13. Target Constituents - As with text production, the location of target constituents in the extracted and tokenized text (parsing) and the subsequent counting of those constituents often go undefined in corpus studies. However, given that corpus analysis is heavily quantitative, the counts are important, and consequently the parsing and counting methods need to be well designed (and well defined). In relation to parsing, the key concern is insuring that the methods are defined to the extent that various constituent types can be located consistently according to the researcher's expectations. For counting, a number of decisions must be made in relation to borders. For example, can constituents span sentence, paragraph, or document boundaries, and are counts in relation to documents or corpora. Although easily overlooked, consideration of constituent definition, location, and counting is necessary in that constituents are the connection between the archive and the analysis. Refer to Chapters 5.4 and 5.5 for additional discussion of constituents.

14. Analysis - In a sense, all subprocesses revolve around analysis. All that precede it work to insure that the necessary data are available, and all that follow work to display or describe the resultant data. This being the case, defining the intended method of analysis is key (the first step) to all project planning, even the selection of the sampling domain. Unfortunately, because the appropriate method of analysis will depend on the

overall goals of the project, which may not be the same as with the TDC, there is little to be said at this point other than to insure the appropriateness of the method. Refer to Chapter 5.6 for a discussion of TDC analyses.

15. Data Presentation - The key idea with data presentation is realizing that analyses, however successful they may be, have little value for the larger community of researchers apart from straightforward presentation of the results. The unfortunate reality of past studies is that often huge amounts of valuable data were generated, but the presentation was limited to a few tables of distilled results that could be published in a journal or presented on slides. However, with the widespread availability of server space and high-bandwidth Internet connections this is no longer the case, to the extent that we can now err on the side of presenting too much data. Essentially all potentially valuable data generated over the course of a study/project can now be distributed relatively easily via the Internet. This being the case, it makes sense to plan for the distribution of larger data sets. However, in very practical terms, not all data is valuable, so some additional procedures will be necessary for the selection of display data. In the case of the TDC Project, tens of millions of comparisons were made, but only about a tenth of the resultant data has any tangible value (the rest being both rhetorically and statistically insignificant). This is still a large amount of data, and we were required to develop a number of innovative methods for making it accessible to secondary users. Refer to Chapters 5.6.2 and 5.6.3, Section 6.2.4 above, Appendix E.3, and the TDC Toolkit Glossary for examples and discussion of the methods used for the selection and display of data.

16. Description - Returning to the rationale for the first item, Project Records, the usefulness of the data produced over the course of a project (assuming that they are made available) is ultimately derived from 1) the description of the process methods, and 2) the distribution of that description, such that subsequent researchers can properly

interpret and implement the given data. This being the case, plans should be made to convert the project record to a formal description and make it available.

6.5 Final Thoughts

In closing, my belief is that overall the TDC Project, as both a product and process, has been an overwhelming success and is a contribution to the field of Corpus Linguistics. As products, the TDC and Toolkit are unique in the field and have shown themselves to be incredibly useful tools for the study of the NAAG Snapshot and for the illustration of corpus methods. Although I am no longer directly involved in the study of tobacco documents, I continue to return to the TDC website as an example of how corpus analysis can be made accessible to more general audiences. As a process, I can say without reserve that TDC methods work as intended. They can be used to manage and describe large corpora. Not only have I seen this in the outcome of the TDC Project, but since the project's end I have continued to use similar methods for managing even larger document sets as part of a litigation support and investigation company. What was accomplished over the course of the TDC Project can be applied directly to any study involving large corpora. My hope is that it will be.

Bibliography

Bailey, Charles-James N. 1973. *Variation and linguistic theory*. Arlington, VA: Center for Applied Linguistics.

Berlin, Brent and Paul Kay. 1969, 1991. *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Berlin, Brent. 1967. Categories of eating in Tzeltal and Navaho. *International Journal of American Linguistics* 33:1–6.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2003. Personal conversation.

Bliley, Thomas. US House Commerce Committee. Chairman Tom Bliley releases subpoenaed tobacco documents to the American people. http://www.house.gov/commerce/-TobaccoDocs/documents.html (website no longer available).

Brown, Catherine (Cati) and Donald L. Rubin. 2005. Causal markers in tobacco industry documents: the pragmatics of responsibility. *Journal of Pragmatics* 37.6: 799–811.

Brown, Catherine (Cati). 2004. UGA Tobacco Documents Project: Rhetorical Cases. http://www.tobaccodocs.uga.edu/TDC/rhe/rheinfo.htm (accessed March 21, 2008).

Brown, Catherine G. 2006. Rating tobacco industry documents for corporate deception and public fraud: a corpus linguistic assessment of intent. PhD Dissertation. University of Georgia.

Burnard, Lou, and C. M. Sperberg-McQueen 1995, 2002. *TEI Lite: An introduction to text encoding for interchange*. http://www.tei-c.org/Guidelines/Customization/Lite/-teiu5_en.pdf (accessed March 28, 2008).

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge: MIT Press.

Cohen, William W. 2004. *The Enron email dataset* (March 2, 2004 version). Machine Learning Department, Carnegie Mellon University. Available at http://www.cs.cmu.edu/ enron/ (accessed March 28, 2008).

Covington, Michael A. 1994. *Natural language processing for Prolog programmers*. Englewood Cliffs, N.J.: Prentice Hall.

Covington, Michael, Donald Nute and A. Vellino. 1997. *Prolog programming in depth*. Upper Saddle River, New Jersey: Prentice Hall.

Darwin, Clayton. 2001. Student PARSing Environment (SPARSE II) CGI-Web demonstration version. http://www.ai.uga.edu/mc/sparse/sparsecgi.html (accessed March 28, 2008).

Davis, Lawrence M. 1990. *Statistics in Dialectology*. Tuscaloosa: University of Alabama Press.

Daynard, Richard and Mark Gottlieb. 2003. Influence of legal context on tobacco industry behavior. Paper presented at the Tobacco Industry Documents Research Investigators Meeting, June 17–19, 2003, in Washington, DC.

Derry, Robbin. and Sachin V. Waikar. 2008. Strategic distrust as a legitimation tool in the 50-year battle between public health activists and big tobacco. *Business & Society* 47.1: 102–139.

Fleischer, Richard (director). 1973. *Soylent green*. Film. Prod. Walter Seltzer and Russel Thacher. Written by Harry Harrison (novel) and Stanley R. Greenberg (screenplay). Metro-Goldwyn-Mayer, May 9, 1973.

Francis, W. N. and H. Kucera. 1964 *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers* (revised 1971, 1979). Providence, RI: Department of Linguistics, Brown University.

FRCP: Federal Rules of Civil Procedure. 2006. Available at Legal Information Institute. Cornell Law School. Federal Rules of Civil Procedure. http://www.law.cornell.edu/-rules/frcp (accessed March 21, 2008).

Gilliron, Jules. 1902–10. *Atlas linguistique de France*. Paris: Champion. As conveyed by William Kretzschmar in a graduate course in language variation, University of Georgia, Fall 1998.

Gray, Loretta and Clayton Darwin. 2001. A context-independent search algorithm for phrasal verbs. Paper presented at the North American Symposium on Corpus Linguistics and Language Teaching. March 23–25, 2001, in Boston, MA.

Hirschhorn, Norbert. 2004. Corporate social responsibility and the tobacco industry: hope or hype? *Tobacco Control* 13: 447–453.

Hundt, Marianne, Andrea Sand, and Paul Skandera. 1999. *Manual of information to accompany the Freiburg-Brown corpus of American English*. Freiburg, Germany: Albert-Ludwigs-Universitt.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kennedy, Graeme. 1998. *An Introduction to corpus linguistics*. London: Longman.

Kretzschmar, William A. and Edgar W. Schneider. 1996. *An introduction to quantitative analysis of linguistic survey data: An atlas by the numbers*. Thousand Oaks: Sage Publications.

Kretzschmar, William A., and Lee Pederson. 2000. Graduate course on impressionistic transcription. University of Georgia, Fall 2000.

Kretzschmar, William A., and Susan Tamasi. 2003. Distributional foundations for a theory of language change. *World Englishes* 22.4: 377-401.

Kretzschmar, William A., Jr., Virginia G. McDavid, Theodore K. Lerud, and Ellen Johnson. 1993. *Handbook of the linguistic atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

Kretzschmar, William A. 1998. Analytical procedure and three technical types of dialect. In *From the Gulf States and beyond: The legacy of Lee Pederson and LAGS*, ed. Michael B. Montgomery and Thomas Nunnally, 167–85. Tuscaloosa: University of Alabama Press.

Kretzschmar, William A. 2005. Linguistic atlas projects. Available at http://us.english.uga.edu (accessed March 28, 2008).

Kurath, Hans. 1949. *A word geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Legacy: Legacy Tobacco Documents Library. University of California, San Francisco. http://legacy.library.ucsf.edu (accessed March 21, 2008). See also http://legacy.library.-ucsf.edu/about/about_collections.jsp (accessed March 21, 2008).

Malone, Ruth. 2003. Tobacco industry promotional and public relations strategies. Paper presented at the Tobacco Industry Documents Research Investigators Meeting, June 17–19, 2003, in Washington, DC.

Moore, David S., and George P. McCabe. 1999. *Introduction to the practice of statistics*, 3rd ed. New York: W.H. Freeman.

NAAG: National Association of Attorneys General. 1998. *Multistate settlement with the tobacco industry. A National Association of Attorneys General authored summary of the Master Settlement Agreement*. Available at the Legacy Tobacco Documents Library. University of California, San Francisco. http://www.library.ucsf.edu/tobacco/litigation/-msa.pdf (accessed March 21, 2008). Also see http://legacy.library.ucsf.edu/about/-about_collections.jsp (accessed March 21, 2008).

Pankow, James. 2003. An analysis of tobacco industry documents relating to research on benzene, acrylonitrile, and other carcinogenic volatile organic compounds in tobacco smoke. Paper presented at the Tobacco Industry Documents Research Investigators Meeting, June 17–19, 2003, in Washington, DC.

Patrick, Peter L. 1998. Testing the Creole continuum. Paper presented at the 27th annual meeting on New Ways of Analyzing Variation in English (NWAVE 27), October 1–4, Athens GA. Available at http://privatewww.essex.ac.uk/p̃atrickp/papers/-TestingContinuum.html (accessed March 28, 2008).

Patrick, Peter L. 1999. *Urban Jamaican Creole: Variation in the mesolect*. (*Varieties of English Around the World*, G17.) Amsterdam: Benjamins.

Pederson, Lee. 1988. *Linguistic atlas of the gulf states*. Athens: University of Georgia Press.

Peter H. A. Sneath, and Robert R. Sokol. 1973. *Numerical taxonomy*. San Fransisco: W. H. Freeman.

Rubin, Donald L. and Yuan Hou. 2004. Detecting possible deception in tobacco industry documents: an application of the linguistic inquiry and word count program. Paper presented at the International Conference on Language & Social Psychology (ICLASP), June 30 to July 3, 2004 at Penn State University.

Rubin, Donald. 2001. NIH-NCI Tobacco-Documents Project at The University of Georgia. http://www.tobaccodocs.uga.edu (accessed March 21, 2008).

Saunders, Barbara. 1988. Revisiting basic color terms. Paper presented at the conference on Anthropology and Psychology: The Legacy of the Torres Strait Expedition. August 10–12, 1998 at St. Johns College, Cambridge. Available at http://human-nature.com/science-as-culture/saunders.html (accessed March 29, 2008).

Saussure, Ferdinand de. (1916) 1986. *Course in general linguistics*. Trans by Roy Harris. LaSalle, IL: Open Court.

Sedona Conference, The. 2007. *The Sedona Conference best practices commentary on the use of search and information retrieval methods in e-discovery*. Public Comment Version.http://www.thesedonaconference.org (accessed March 21, 2008).

Shuy, Roger W. 2003. Tobaccospeak: Image repair as a variety of American English. Paper presented at the annual conference of the American Dialect Society, January 2–4, 2003, in Atlanta, GA. Available at http://www.uga.edu/tobaccodocs/papers/tobaccospeak.doc (accessed 30 March, 2008).

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

TDO: Tobacco Documents Online. http://tobaccodocuments.org (accessed March 21, 2008).

USDC: U.S. District Court, District of Columbia. 2002. *U.S. v. Philip Morris INC. (212 FRD 421 DDC 2002)*. Available at the Legacy Tobacco Documents Library. University of

California, San Francisco. http://legacy.library.ucsf.edu/resources/bliley_ref.pdf (accessed March 21, 2008).

van Rossum, Guido. 2008. *Python library reference* (release 2.5.2). Ed. Fred L. Drake, Jr. Python Programming Language–Official Website. http://docs.python.org/lib/lib.html (accessed March 21, 2008).

Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In *Directions for historical linguistics*, ed. Winifred Lehmann and Yakov Malkiel, 95–188. Austin: University of Texas Press.

Wolfram, Walt. 1969. *A linguistic description of Detroit Negro speech*. Washington DC: Center for Applied Linguistics.

Zipf, George. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison Wesley.

Appendix A

Sampling plan for creation of corpora for Tobacco Documents grant

W. Kretzschmar, August 15, 2001.

A.1 Introduction

According to the grant proposal, the task for creation of corpora from the Tobacco Documents (TDs) is twofold:

1. Identify TDs in which rhetorical manipulation may have occurred, and to estimate the extent and prevalence of manipulation. Previous research on manipulation is subject to attack because of highly selective use of data.

2. Analyze any manipulation we find in order to classify it and develop means to identify similar manipulation in other industrial situations. This analysis should be particularly productive in cross-audience and cross-draft comparisons.

To accomplish these tasks, I suggest a three-part strategy for corpus creation which emphasizes rigorous sampling methods. We should first draw a limited sample from the entire body of TDs, so that we can determine the best classification of text types and estimate their proportions within the overall body of texts. We should next create a reference corpus of about 500,000 words from those text types that we consider relevant to (i.e. subject to) rhetorical manipulation; this corpus will be the result of a sample of all relevant TDs, whether or not they are thought to contain any manipulation. Finally, we should compile a corpus which includes all texts which we determine to contain any rhetorical manipulation, along with parallel corpora of earlier drafts of the same texts or versions of the same texts prepared

for other audiences, so that detailed analysis of rhetorical manipulation can be carried out for itself and by comparison with cross-draft and cross-audience TDs. The contents of reference corpus may or may not be reproduced in part in the final set of parallel corpora.

A.2 Part 1: Limited Sample of TDs

The extant set of TDs comprises millions of documents, ranging in length from just a few words to hundreds of pages. It will clearly not be possible to inspect every word of every document. Yet we do need to know what kinds of documents exist in the set of TDs, and more specifically, what kinds of documents relevant to the grant exist in the set of TDs. Further, we need to know the extent of those documents, both the quantity of relevant documents and how long they tend to be. We cannot create a valid sample of relevant documents without this information. We should therefore sample the body of TDs according to a fixed random sampling frame, a procedure that gives every document in the collection an equal chance of selection.

Clayton Darwin inspected documents available online according to a fixed, date-based frame that we had discussed. He found the following:

I looked in the snapshot documents of the tobacco companies, from March 1 - 31, for the 0 years 1930 - 1990. The page numbers are approximated by taking the average pages per document from the first 100 pages of a set.

- 1930 - 13 docs - 21 pages
- 1940 - 87 docs - 388 pages
- 1950 - 169 docs - 1500 pages
- 1960 - 864 docs - 2100 pages
- 1970 - 2509 docs - 3100 pages
- 1980 - 6605 docs - 34000 pages

- 1990 - 11717 docs - 50000 pages (welcome to the age of computing)

So that is roughly 90000 pages at between 25–250 words a page.

If we extrapolate this estimate to the snapshot texts (i.e. multiply by 10 to fill in the decades, and multiply by 12 to fill in the years), we get 2.6 million documents, containing 10.9 million pages. Documents averaged slightly over 4 pages in length, and consisted of an average of about 1000 words (if we assume that a full page contains about 250 words). It is also clear that the average length of the documents varies widely by decade (1.2 pages in 1970 to 5.1 pages in 1980 and 8.9 pages in 1950), which suggests variability in text type.

I propose that we draw a random sample of 200–300 documents from the body of TDs. For each decade in which we have TDs, we should select one year for each decade (put the numbers 0–9 in a hat and draw one; repeat with all numbers in the hat for each decade). For each year selected, we should select one month (put the names of the months in a hat and draw one; repeat with all months in the hat for each year). Because there are few documents from the 1930s, 1940s and 1950s, I would regroup the documents from the selected years/months for those decades into a single set before further selection of documents. Finally, I propose selecting 1% of the documents from each year/month set according to a fixed, sequential process; the percentage could be adjusted according to the actual count of eligible documents so that we selected at least 200 but no more than 300 documents in all. We would require a count of the number of documents in each set so that we knew how many to select, and then we would select every nth document, where n = the last two digits of the year selected for the decade. This sampling frame would have yielded 220 documents from the months that Clayton used above.

I further propose that we classify the documents thus selected according to the following variables which have been identified in the grant proposal as significant for analysis:

- PHS (2): significant for public health, not significant for public health
- AUD (2): internal audience, external audience

- ADR (2): personal addressee (from 1 to 5 individuals), multiple addressees (greater than 5 individuals)

- SRC (3): admin/legal source or author, research source or author, public relations source or author

These three binary variables plus one trinary variable yield a total number of 24 possible text types based on configurations of the variables (2x2x2x3). We need to estimate the proportion of each text type in the body of TDs, and for each text type we need to determine the range in size of the documents. We also need to associate each of these text types with labels for more familiar names for document types, as found in the TDD indexes.

The result of this sampling procedure will be a spreadsheet from which we can make an estimate of the extent of the text types of interest for the grant, and their document sizes, within the body of TDs. The documents will be classified to guide further sampling, but the selected documents will not be keyed. Nonetheless, as the documents are inspected, any that contain evidence of manipulation will be identified for inclusion in the third stage of corpus construction. Moreover, information about the documents observed during the initial limited sampling process can guide the drawing of the reference sample.

A.3 Part 2: Reference Sample/Corpus

The purpose of the Reference Sample/Corpus is to create a control set of TDs from among those in which manipulation could have occurred (but did not necessarily occur), from which we can estimate the general frequency of occurrence of linguistic characteristics of interest in the analysis of rhetorical manipulation. Because many of these characteristics may occur with low frequency, this corpus must be large enough to ensure that the characteristics are represented. On the other hand, the corpus must not be so large that its creation overruns the resources in the grant to create it: a corpus of about 500,000 words appears to be as much as the resources of the grant might handle.

Given the estimate from the initial sampling of what kinds of text types exist in the TDs, and in what proportion, it will be possible to create a quota sample from the set of TDs relevant to the purposes of the grant (NB: not from all TDs). We can decide exactly which text types should be represented in the Reference Sample, and in what numbers. Each of the 24 possible text types will be considered, and either accepted for the Reference Sample/Corpus or rejected. The proportion of each text type in the Limited Sample, and its importance to the purposes of the grant, will determine the quota for each text type in the Reference Sample/Corpus.

Overall, the number of documents in the Reference Sample/Corpus should be about 500. We will accept entire documents of up to 2000 words, and a 2000-word segment of documents larger than that size. It is expected that the inclusion of text types with shorter length (e.g. letters, memos) will balance use of longer text types (e.g. research reports), so that the average length of a document will be about 1000 words.

After establishment of text type quotas, selection of particular documents will be randomized by a fixed, sequential sampling frame. For the Reference Sample/Corpus, the sampling procedure outlined in Part 1 should be repeated; however, for Part 2 only documents which fit the established quotas should be included. For instance, if the nth document cannot be included in the sample, either because it does not fit a quota category or because the quota category that it its is already full, then the investigator will move on to the next nth document in sequence.

Selection of a 2000-word segment from a longer document will also be accomplished by a randomized process. 2000-word segments should to the extent possible consist of coherent sections of text. The investigator will begin the segment at a heading, subheading, or paragraph break in the text, and end the segment at the completion of the paragraph or other text unit closest to but not exceeding the 2000-word limit. The investigator will draw a number from 1 to 4, and then will select the first available heading, subheading, or paragraph of text

in the quartile of the document corresponding to the number drawn, to begin the 2000-word segment.

The result of this procedure will be a corpus of about 500 TDs which will have been compiled according to the current best practices for the creation of balanced corpora. The grant team will be able to apply text markup, tags, and other organizational and analytical tools as necessary for textual storage and analysis.

A.4 Part 3: Parallel Corpora of Manipulated Documents

The purpose of the Parallel Corpora of Manipulated Documents is to enable analysis of TDs across drafts and across audiences, as described in the grant proposal. To this end, all available documents which are determined to show evidence of manipulation will be collected in a corpus; relative scarcity of such documents suggests that sampling will not yield enough exemplars for analysis. Project staff will then attempt to match each document in the corpus of manipulated documents with earlier drafts, and will place the drafts into a second, parallel corpus. project staff will also attempt to locate versions of the document for other audiences, and will locate these documents in a third parallel corpus.

All documents in these parallel corpora will be intentionally selected according to their contents, and so no sampling frame will be used. The parallel corpora will become as large as the staff can make them by identifying manipulated documents and associated cross-audience or cross-draft versions. The larger the size of these corpora, the better modeling of linguistic characteristics will be possible.

All documents in the parallel corpora will be classified according to the criteria identified in part 1, and the quotas established for Part 2, so that they may be compared explicitly to the data gathered in Parts 1 and 2.

Appendix B

Document Metadata

The sections in this chapter contain the metadata for the Core, Quota, and Supplemental Samples collected during the sampling and text conversion processes.

B.1 Description

The columns are labeled according to the following:

1. AMT - the amount of document text recorded in the XML. For short documents, 'ALL' text was recorded. For long documents, as defined in Chapter 4, 'PART' is denoted.

2. BATES-END - the final Bates Number for the document. See Chapter 3 for more details.

3. BATES-START - the starting Bates Number for the document. See Chapter 3 for more details.

4. CDR - the initials of the original coder (archivist) for the document.

5. CLS - the class of document according to the quota-sample specifications as defined in Section 3.4. One of the following: Named-Internal (NI), Named-External (NE), Unnamed-Internal (UI), or Unnamed-External (UE).

6. CV - core sample document selected for classification validation. 1 signifies YES. Additionally, X signifies that there were discrepancies associated with this document during validation. See Section 3.3 for more details.

7. DAT - the date of the document, if known, in the format `yymmdd`.
8. EDT - classified as *edited*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
9. ENG - classified as *English*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
10. FRM - classified as *form*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
11. IA - classified as *internal audience*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
12. IMG - classified as *image*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
13. IS - classified as *internal source*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
14. ITR - the sampling iteration which selected the document, Q1 through Q4.
15. MRG - classified as *marginalia*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
16. NA - classified as *named addressee*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.
17. NUM - a number assigned for ease of reference. It has no significance or value assigned.
18. PGS - the number of pages in the entire document.
19. PH - classified as *Public Health*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.

20. SHT - classified as *short*. 1 signifies YES. 0 signifies NO. See Section 3.3 for more details.

21. SRC - the industry source of the document. One of the following: American Tobacco Company (ATC), Brown and Williamson (BW), Lorillard (LOR), Philip Morris (PM), and R. J. Reynolds (RJR), Council for Tobacco Research (CTR), or and the Tobacco Institute (TI).

22. STRAT - the particular stratum from/for which the document was selected.

23. TKN - the token count for the text portion of the document recorded in the XML, based on the text extracted using the standard counting stylesheet seen in Section D.2.

24. U - usable for deriving quotas (+*English*, −*short*, +*internal source*, +*Public Health*). See Section 3.3 for more details.[1]

25. VER - the initials of the verifying archivist for the document if a document went through verification. Otherwise, NONE.

B.2 Metadata for Core Sample Documents

The following data were derived by examining the sampling and classification archives of the completed Core Sample.

NUM	STRAT	SN	BATES-START	BATES-END	PGS	FRM	IMG	ENG	EDT	MRG	SHT	NA	IA	IS	PH	U	CV
1	1900	1	682329988	682329990	3	0	0	1	0	0	0	0	0	0	1	0	
2	1900	2	88111246	88111248	3	0	0	1	0	0	0	0	1	1	1	1	
3	1900	3	502593497	502593497	1	0	1	1	0	1	0	0	0	1	1	1	
4	1900	4	502595907	502595907	1	0	1	1	0	1	0	0	0	1	1	1	1

[1]The actual number of documents classified as *usable* is 203. By an oversight, this was originally determined to be 202. The count of *usable* documents in the 1990 stratum is 62 rather than the 61 given in Table 3.11. This error reduced the final document total from 812 to 808 (0.50 percent) for the full quota sample, and from 248 to 244 (1.61 percent) in the 1990 stratum. In terms of differences in ratio, the 1990 stratum to full corpus ratio was reduced 1.11 percent. Because this error was not discovered until after the completion of sampling (and given the small varience), 202 is used as a total of *usable* documents in all discussions of quotas. This error is noted to avoid confusion between the Chapter 3 data and the data presented here.

5	1900	5	507142069	507142073	5	0	0	1	0	1	0	0	1	1	1	1	
6	1950	6	2040455378	2040455378	1	0	0	1	0	0	0	1	1	1	1	1	
7	1950	7	503269441	503269445	5	0	0	1	0	1	0	0	0	0	1	0	
8	1950	8	502367598	502367598	1	0	0	1	0	0	1	1	1	1	1	0	
9	1950	9	681841118	681841121	6	0	0	1	0	1	1	0	1	1	1	0	
10	1950	10	1005039213	1005039222	10	0	0	1	0	0	0	1	1	1	1	1	1
11	1960	1	2026344269	2026344271	3	0	0	1	0	1	0	1	1	1	1	1	
12	1960	2	1003541254	1003541257	4	0	0	1	0	0	0	1	1	0	1	0	
13	1960	3	505549425	505549470	46	0	0	1	0	0	0	0	1	1	1	1	
14	1960	4	504559655	504559655	1	1	0	1	1	1	1	0	1	1	1	0	
15	1960	5	1002762912	1002762912	1	0	1	1	0	0	0	0	0	1	1	1	
16	1960	6	1000332691	1000332701	11	0	0	1	0	1	0	1	1	1	1	1	1
17	1960	7	1000316126	1000316126	1	0	0	1	0	0	0	1	0	1	1	1	
18	1960	8	500028420	500028421	2	0	0	1	0	1	0	0	1	1	1	1	
19	1960	9	1000329634	1000329635	2	0	0	1	0	1	0	1	1	1	1	1	1
20	1960	10	1003703089	1003703089	1	0	0	1	0	1	0	1	1	1	1	1	
21	1960	11	1000859411	1000859411	1	0	0	1	0	1	0	1	1	1	1	1	
22	1960	12	1005149916	1005149916	1	0	0	1	0	1	0	0	1	1	1	1	
23	1960	13	88131499	88131499	1	1	0	1	0	0	1	0	1	1	1	0	
24	1960	14	553003553	553003553	1	0	0	1	0	1	0	0	1	1	1	1	
25	1960	15	50018991	50018991	1	0	0	1	0	1	1	1	1	1	1	0	1X
26	1960	16	50008132	50008132	1	0	0	1	0	0	1	1	1	1	1	0	
27	1960	17	11293732	11293732	1	0	0	1	0	0	0	1	1	0	1	0	
28	1960	18	680279589	680279589	1	0	0	1	0	1	0	1	1	1	1	1	
29	1960	19	01200579	01200583	5	0	0	1	0	0	0	1	1	1	1	1	
30	1960	20	2026350535	2026350539	5	0	0	1	1	0	0	0	1	1	1	1	
31	1960	21	1003540996	1003540998	3	0	0	1	0	1	0	0	1	1	1	1	
32	1960	22	501908816	501908821	6	0	0	1	0	0	0	1	1	1	1	1	1
33	1970	1	504369201	504369227	27	0	0	1	0	1	0	0	1	1	1	1	
34	1970	2	2040936649	2040936650	2	0	0	1	0	0	0	0	1	1	1	1	
35	1970	3	2041733937	2041733937	1	0	0	1	0	0	0	1	0	1	1	1	
36	1970	4	2057675287	2057675287	1	0	0	1	0	1	1	0	1	1	1	0	
37	1970	5	1005140250	1005140250	1	0	0	1	0	0	0	0	1	1	1	1	
38	1970	6	1003030465	1003030465	1	0	0	1	0	0	0	0	1	1	1	1	
39	1970	7	2057451687	2057451689	3	1	0	1	0	0	1	0	0	0	0	0	
40	1970	8	2028620932	2028620932	1	0	0	1	0	0	1	0	1	1	1	0	
41	1970	9	2022248873	2022248873	1	0	0	1	1	1	1	1	1	1	1	0	
42	1970	10	500175350	500175350	1	0	0	1	0	1	0	1	1	1	1	1	
43	1970	11	500876387	500876401	15	0	0	1	0	1	0	0	0	0	1	0	
44	1970	12	504853431	504853431	1	0	0	1	0	1	0	0	0	0	1	0	
45	1970	13	504877780	504877781	2	0	0	1	0	1	0	1	1	1	1	1	
46	1970	14	2025997140	2025997140	1	0	0	1	0	0	1	1	1	1	1	0	
47	1970	15	2501015916	2501015968	53	0	0	1	0	1	0	0	0	0	1	0	1
48	1970	16	500927442	500927442	1	0	0	1	0	0	1	1	1	1	1	0	
49	1970	17	502804844	502804850	7	0	0	1	0	1	0	1	1	1	1	1	
50	1970	18	2024469591	2024469591	1	0	0	0	0	1	0	1	1	1	0	0	
51	1970	19	2028839092a	2028839092a	1	1	0	1	0	1	1	1	1	1	1	0	
52	1970	20	0000888497	0000888497	1	1	0	1	0	0	1	0	0	0	0	0	
53	1970	21	502122784	502122787	4	0	0	1	0	0	0	0	1	1	1	1	1
54	1970	22	670654424	670654449	26	0	0	1	0	1	0	0	1	1	1	1	
55	1970	23	670182022	670182025	4	0	0	1	0	0	0	0	1	1	1	1	
56	1970	24	660110697	660110699	3	0	0	1	0	0	1	0	1	1	1	0	1
57	1970	25	660037039	660037041	3	0	0	1	0	1	0	1	1	1	1	1	
58	1970	26	659055253	659055253	1	1	0	1	0	0	1	0	1	1	1	0	1
59	1970	27	657013085	657013086	2	0	0	1	0	1	0	0	1	1	1	1	
60	1970	28	2024265055	2024265055	1	0	0	1	0	0	1	0	1	1	1	0	
61	1970	29	650521702	650521702	1	1	0	1	0	1	1	0	1	1	1	0	
62	1970	30	2025015975	2025015975	1	0	0	1	0	1	1	0	1	1	1	0	
63	1970	31	680029906	680029906	1	0	1	1	0	0	1	0	1	1	1	0	
64	1970	32	508002959	508002959	1	0	0	1	0	1	0	1	1	1	1	1	
65	1970	33	501830043	501830049	7	0	0	1	0	0	0	0	1	1	1	1	1
66	1970	34	500692632	500692651	20	0	0	1	0	1	0	0	1	1	1	1	1
67	1970	35	620044613	620044619	7	0	0	1	0	1	0	0	1	1	1	1	
68	1970	36	1003637041	1003637041	1	0	0	1	0	0	0	1	1	1	1	1	
69	1970	37	1001867542	1001867542	1	0	0	1	0	1	0	1	1	1	1	1	
70	1970	38	500492491	500492492	2	0	0	1	0	0	0	0	1	1	1	1	
71	1970	39	500047098A	500047131	34	0	0	1	0	1	0	0	1	1	1	1	
72	1970	40	621094587	621094590	4	1	0	1	0	0	1	0	1	1	1	0	1X

73	1970	41	1000206950	1000206950	1	0	0	1	0	1	0	0	1	1	1	1	
74	1970	42	621605460	621605463	5	0	0	1	1	1	0	1	1	1	0	0	
75	1970	43	621071059	621071062	4	0	0	1	0	0	0	0	1	1	1	1	
76	1970	44	1003057164	1003057169	6	0	0	1	0	1	0	0	0	0	1	0	1
77	1970	45	1000127820	1000127821	2	1	0	1	0	0	1	0	0	0	1	0	
78	1970	46	1003032493	1003032493	1	0	0	1	0	1	0	0	1	1	1	1	1
79	1970	47	80510036	80510036	1	0	0	1	0	0	0	0	1	1	1	1	
80	1970	48	01073990	01073990	1	0	0	1	0	0	0	1	1	1	1	1	
81	1970	49	1003480966	1003480966	1	0	0	1	1	1	1	1	1	1	1	0	
82	1970	50	1005147696	1005147697	2	0	0	1	0	0	0	1	0	1	1	1	1
83	1970	51	1000773106	1000773114	9	0	0	1	0	0	0	0	1	1	1	1	
84	1970	52	666019391	666019394	4	0	0	1	0	0	0	0	1	1	1	1	
85	1970	53	664037346	664037346	1	0	0	1	0	0	1	1	1	1	1	0	
86	1970	54	SF0610392	SF0610402	11	0	0	1	0	1	0	0	0	0	1	0	
87	1970	55	2001205386	2001205386	1	0	0	0	0	1	0	1	0	0	1	0	
88	1970	56	2024928709	2024928709	1	0	0	1	0	1	0	1	1	0	1	0	
89	1970	57	570216725	570216751	27	0	0	1	0	0	0	1	1	1	1	1	
90	1970	58	CTRSPFILES01319731	CTRSPFILES01319731	1	0	0	1	0	1	0	1	1	1	1	1	
91	1970	59	85705173	85705175	3	0	0	1	0	1	0	1	1	1	1	1	
92	1970	60	502134637	502134639	3	0	0	1	0	0	0	0	1	1	1	1	
93	1970	61	2024976529	2024976529	1	0	0	1	0	1	0	0	1	1	1	1	
94	1970	62	2028355706	2028355706	1	0	0	0	0	0	0	1	0	1	1	0	
95	1970	63	04302510	04302515	6	0	0	1	0	0	0	1	1	1	1	1	
96	1970	64	CTRLRD001741	CTRLRD001742	2	1	0	1	0	0	1	0	0	0	0	0	
97	1970	65	501381670	501381671	2	0	0	1	0	1	0	1	1	1	1	1	
98	1970	66	500359355	500359355	1	0	0	1	0	1	1	1	1	1	1	0	
99	1980	1	500879570	500879594	25	0	0	1	0	1	0	0	0	0	1	0	
100	1980	2	2026257802	2026257802	1	0	0	1	0	1	0	0	0	0	1	0	
101	1980	3	2040756966	2040756966	1	1	0	1	0	0	0	1	1	1	1	1	
102	1980	4	2040447205	2040447205	1	0	0	1	0	0	1	0	1	1	1	0	1
103	1980	5	2025662575	2025662575	1	0	0	1	0	1	1	1	1	1	1	0	
104	1980	6	2055254978	2055254983	6	0	0	1	0	0	0	0	1	1	1	1	
105	1980	7	2056121564A	2056121564A	1	0	0	1	0	0	1	0	1	1	0	0	
106	1980	8	2500126669	2500126669	1	0	0	1	0	1	0	0	0	0	1	0	
107	1980	9	2501029902	2501029918	17	0	0	1	0	0	0	0	1	1	1	1	
108	1980	10	1005114772	1005114773	2	0	0	1	0	0	0	0	0	0	1	0	1
109	1980	11	2054389866	2054389867	2	1	0	1	0	1	1	0	0	0	0	0	
110	1980	12	2028793879	2028793879	1	1	0	1	0	0	1	0	1	1	1	0	
111	1980	13	2021576101	2021576101	1	0	0	1	0	1	1	1	1	1	1	0	
112	1980	14	500158408	500158409	2	0	0	1	0	1	0	1	1	1	1	1	1
113	1980	15	513245257	513245257	1	0	0	1	0	0	1	1	1	1	1	0	
114	1980	16	2028846475	2028846475	1	1	0	0	0	0	1	0	1	1	1	0	
115	1980	17	504922505	504922505	1	0	0	1	0	1	0	1	1	1	1	1	1
116	1980	18	2501634965	2501634991	27	0	0	1	0	0	1	0	1	1	1	0	
117	1980	19	2500044817	2500044817	1	0	0	1	0	1	0	0	1	1	1	1	
118	1980	20	505081328	505081328	1	0	0	1	0	1	1	1	1	1	1	0	
119	1980	21	503274590	503274591	2	0	0	1	0	0	0	0	1	1	1	1	
120	1980	22	504482608	504482615	8	0	0	1	0	0	0	0	1	1	1	1	
121	1980	23	512504816	512504816	1	0	0	1	0	0	0	1	1	1	1	1	
122	1980	24	502032476	502032480	5	0	0	1	0	1	0	1	1	1	1	1	
123	1980	25	517003323	517003324	2	0	0	1	0	1	0	0	0	0	1	0	
124	1980	26	503688315	503688324	10	0	0	1	0	1	0	0	0	0	1	0	
125	1980	27	521031660	521031660	1	0	0	1	0	1	0	1	1	1	1	1	1
126	1980	28	510994920	510994920	1	0	0	1	0	1	0	0	1	1	1	1	1
127	1980	29	670164309	670164317	9	0	0	1	0	0	0	0	1	1	1	1	
128	1980	30	660113868	660113869	2	0	0	1	0	1	0	1	1	1	1	1	
129	1980	31	660064268	660064268	1	0	0	1	0	0	0	1	1	1	1	1	
130	1980	32	504879216	504879216	1	0	0	1	0	1	0	0	0	0	1	0	
131	1980	33	2026442432	2026442432	1	0	0	1	0	0	1	1	1	1	1	0	
132	1980	34	655006218	655006223	6	0	0	1	0	0	0	0	1	1	1	1	
133	1980	35	650537032	650537032	1	0	0	1	0	1	0	1	1	1	1	1	
134	1980	36	650332587	650332623	37	0	0	1	0	0	0	1	1	1	1	1	
135	1980	37	503452481	503452482	2	1	0	1	0	0	0	0	1	1	1	1	
136	1980	38	542006663	542006664	2	0	0	1	0	0	1	0	1	1	1	0	
137	1980	39	501542120	501542123	4	0	0	1	0	0	0	1	0	1	1	1	
138	1980	40	502248782	502248783	2	0	0	1	0	0	0	0	1	1	1	1	
139	1980	41	2022191558	2022191558	1	1	0	1	0	1	1	1	1	1	0	0	
140	1980	42	1005071861	1005071862	2	0	0	1	0	1	0	1	1	1	1	1	

141	1980	43	1003656885	1003656885	1	0	0	1	0	0	0	0	1	1	1	1	1
142	1980	44	1001869408	1001869409	2	1	0	1	0	1	1	1	1	1	1	0	1
143	1980	45	661046550	661046551	3	0	0	1	0	0	0	1	1	1	1	1	1
144	1980	46	620232135	620232137	3	1	0	1	0	0	1	0	1	1	1	0	
145	1980	47	660909802	660909866	65	0	0	1	0	1	0	0	1	1	1	1	
146	1980	48	1000006620	1000006622	3	0	0	1	0	0	0	0	0	0	1	0	
147	1980	49	03028493	03028494	2	0	0	1	0	0	0	0	1	1	1	1	
148	1980	50	1003155118	1003155118	1	0	0	1	0	1	0	1	1	1	0	0	
149	1980	51	465825764	465825790	27	0	0	1	0	1	0	1	1	1	1	1	
150	1980	52	85628850	85628850	1	0	0	1	0	1	1	1	1	1	1	0	
151	1980	53	1000089004	1000089004	1	0	0	1	0	1	0	0	1	1	1	1	
152	1980	54	88698582	88698590	9	1	0	1	0	1	0	1	1	0	1	0	
153	1980	55	509074281	509074281	1	0	0	1	1	1	1	0	1	1	1	0	
154	1980	56	504006142	504006155	14	0	0	1	0	1	0	0	1	1	1	1	
155	1980	57	1002646644	1002646691	48	0	0	1	0	0	0	0	1	0	1	0	
156	1980	58	1003635395	1003635405	11	0	0	1	0	0	0	1	1	1	1	1	
157	1980	59	1003481534	1003481534	1	0	0	1	0	1	0	1	1	1	1	1	
158	1980	60	89467002	89467028	27	0	0	1	0	1	0	1	1	1	1	1	1
159	1980	61	1000799370	1000799374	5	0	0	1	0	1	0	1	1	1	1	1	
160	1980	62	508897150	508897152	3	0	0	1	0	0	0	1	1	1	1	1	
161	1980	63	621604039	621604040	2	0	0	1	0	1	0	1	1	1	1	1	
162	1980	64	HT0026228	HT0026228	1	0	0	1	0	0	1	0	1	1	1	0	
163	1980	65	620120474	620120480	7	0	0	1	0	1	0	1	1	1	1	1	
164	1980	66	2020277095	2020277095	1	0	0	1	0	0	0	0	1	1	0	0	1
165	1980	67	87103823	87103823	1	0	0	1	0	1	1	1	1	1	1	0	
166	1980	68	2023269446	2023269446	1	0	0	1	0	1	0	1	1	1	1	1	
167	1980	69	504317190	504317190	1	1	0	1	0	0	1	0	1	1	1	0	
168	1980	70	690133003	690133018	16	0	0	1	0	1	0	0	1	1	1	1	
169	1980	71	84433745	84433746	2	0	0	1	0	1	0	1	1	1	1	1	
170	1980	72	04236214	04236214	1	0	0	1	0	1	1	1	1	1	1	0	
171	1980	73	501526381	501526381	1	0	0	1	0	1	0	1	1	1	1	1	
172	1980	74	03014027	03014027	1	0	0	1	0	0	0	0	0	0	1	0	
173	1980	75	503747210	503747210	1	0	0	1	0	1	0	1	1	1	1	1	
174	1980	76	502145104	502145104	1	0	0	1	0	0	0	1	1	1	1	1	
175	1980	77	683026538	683026540	3	0	0	1	0	0	0	0	0	0	1	0	
176	1980	78	03673206	03673206	1	0	0	1	0	1	1	1	1	1	1	0	
177	1980	79	03615017	03615018	2	0	0	1	0	1	0	0	0	0	1	0	
178	1980	80	TIMN0154697	TIMN0154706	10	0	0	1	0	1	0	0	0	0	1	0	
179	1980	81	2025998796	2025998800	5	1	0	1	0	0	1	1	1	1	1	0	
180	1980	82	655026138	655026142	5	1	0	1	0	1	1	0	1	1	1	0	
181	1980	83	TIMN0096786	TIMN0096787	2	0	0	1	0	0	0	0	0	0	1	0	
182	1980	84	TIMN0086661	TIMN0086664	4	0	0	1	0	0	0	0	0	0	1	0	
183	1980	85	502785505	502785506	2	0	0	1	0	1	0	0	1	1	1	1	
184	1980	86	TIMN0063592	TIMN0063592	1	0	0	1	0	1	0	1	1	1	1	1	
185	1980	87	00921442	00921442	1	0	0	1	0	0	1	1	1	1	1	0	
186	1980	88	516003511	516003511	1	0	0	1	0	1	0	0	1	1	1	1	
187	1980	89	00833182	00833192	11	0	0	1	0	1	0	0	1	1	1	1	
188	1980	90	500148640	500148643	4	0	0	1	0	1	0	0	1	1	1	1	
189	1980	91	2026257800	2026257801	2	0	0	1	0	1	0	0	0	0	1	0	
190	1980	92	2040756965	2040756965	1	1	0	1	0	1	0	1	1	1	1	1	
191	1980	93	2040447134	2040447204	71	0	0	1	0	1	0	0	1	1	1	1	1
192	1980	94	2025662583	2025662584	2	0	0	1	0	1	0	1	1	1	1	1	
193	1980	95	2055254970	2055254977	8	0	0	1	0	0	0	0	1	1	1	1	
194	1980	96	2056121547	2056121564	18	0	0	1	0	1	0	0	1	1	1	1	
195	1980	97	1000017145	1000017152	8	0	0	1	0	1	0	1	1	1	1	1	
196	1980	98	2501029891	2501029901	11	0	0	1	0	0	0	0	1	1	1	1	
197	1980	99	1005114768	1005114771	4	0	0	1	0	0	0	0	0	0	1	0	1
198	1980	100	2053578305	2053578306	2	1	0	1	0	0	1	1	0	0	0	0	
199	1980	101	2028793877	2028793877	1	1	0	1	0	0	1	0	1	1	1	0	
200	1980	102	2015027327	2015027327	1	0	0	1	0	1	0	1	1	1	1	1	
201	1980	103	500158380	500158380	1	0	0	1	0	1	0	1	1	1	1	1	1
202	1980	104	513245232	513245232	1	0	0	1	0	0	1	1	1	1	1	0	
203	1980	105	2028846472	2028846473	2	1	0	0	0	1	1	0	1	1	1	0	1
204	1980	106	504800635	504800638	4	0	0	1	0	1	0	1	1	1	1	1	
205	1980	107	2501634963	2501634964	2	1	0	1	0	0	1	0	1	1	1	0	
206	1980	108	502499805	502499807	3	0	0	1	0	0	0	1	1	1	1	1	
207	1980	109	505081315	505081315	1	0	0	1	0	1	1	1	1	1	1	0	1
208	1980	110	503397266	503397267	2	0	0	1	0	0	0	0	1	1	1	1	

209	1980	111	504418496	504418498	3	0	0	1	0	0	0	0	1	1	1	1	
210	1980	112	2022262932	2022262943	12	0	0	1	1	0	0	0	0	0	1	0	
211	1980	113	500133169	500133170	2	0	0	1	0	1	0	1	1	1	1	1	
212	1980	114	680594171	680594175	5	0	0	1	1	0	0	0	1	1	1	1	
213	1980	115	503254348	503254352	5	0	0	1	0	1	0	0	0	0	1	0	1
214	1980	116	680147228	680147230	3	0	0	1	0	1	0	0	0	0	1	0	
215	1980	117	502741860	502741860	1	0	0	1	0	1	1	1	1	1	1	0	
216	1980	118	2040698944	2040698944	1	0	0	1	0	0	0	1	1	1	1	1	
217	1980	119	660113863	660113866	4	0	0	1	0	1	1	1	1	1	1	0	1X
218	1980	120	660064260	660064266	7	0	0	1	0	0	0	0	1	1	1	1	
219	1980	121	504879019	504879019	1	0	0	1	0	1	0	0	0	0	1	0	
220	1980	122	504867962	504867962	1	0	0	1	0	1	0	0	1	1	1	1	
221	1980	123	655006208	655006216	9	0	0	1	0	0	0	0	1	1	1	1	
222	1980	124	650537031	650537031	1	0	0	1	0	1	0	1	1	1	1	1	
223	1980	125	650332073	650332103	32	0	0	1	0	1	0	1	1	1	1	1	
224	1980	126	675210102	675210103	2	0	0	1	0	1	1	1	1	1	1	0	
225	1980	127	503089604	503089716	113	0	0	1	0	0	0	0	1	1	1	1	
226	1980	128	2024920272	2024920272	1	0	0	1	0	1	1	1	1	1	1	0	
227	1980	129	502248747	502248750	4	0	0	1	0	0	0	1	1	1	1	1	1
228	1980	130	2022191551	2022191557	7	0	0	1	0	1	0	0	1	1	1	1	
229	1980	131	682172843	682172843	1	0	0	1	0	0	1	1	1	1	1	0	
230	1980	132	1003639962	1003639962	1	1	0	1	0	0	1	0	1	1	1	0	1
231	1990	1	518202037	518202070	34	0	0	1	0	1	0	0	1	1	1	1	1
232	1990	2	2063118422	2063118438	17	1	0	1	0	1	1	0	0	1	1	0	
233	1990	3	2063657859	2063657859	1	0	0	1	0	1	0	1	1	1	1	1	
234	1990	4	2063590781	2063590781	1	1	0	1	0	0	1	0	1	1	1	0	
235	1990	5	2063053037	2063053037	1	1	0	1	0	0	1	0	1	1	1	0	1
236	1990	6	518643744	518643746	3	0	0	1	0	1	1	1	1	1	1	0	
237	1990	7	517404241	517404250	10	0	0	1	0	0	0	0	0	0	1	0	
238	1990	8	518191558	518191558	1	1	0	1	0	1	1	0	1	1	1	0	
239	1990	9	518050808	518050808	1	0	0	1	0	0	0	1	1	1	1	1	
240	1990	10	518210211	518210215	5	0	0	1	0	1	0	0	1	1	1	1	
241	1990	11	2062391348	2062391348	1	0	0	1	0	0	1	0	1	1	1	0	
242	1990	12	621967784	621967785	2	0	0	1	0	0	1	0	1	1	1	0	
243	1990	13	94573526	94573530	5	0	0	1	0	0	0	0	1	1	1	1	
244	1990	14	94547591	94547595	5	0	0	1	0	1	0	1	1	1	1	1	
245	1990	15	94525903	94525903	1	0	0	1	0	1	0	0	0	0	1	0	
246	1990	16	94403084	94403085	2	1	0	1	0	0	0	0	1	1	1	1	
247	1990	17	60015577B	60015577B	1	0	0	1	0	0	0	0	1	1	1	1	
248	1990	18	60015554	60015554	1	0	0	1	0	0	0	0	1	1	1	1	
249	1990	19	518650349	518650351	3	0	0	1	0	0	0	1	1	1	1	1	
250	1990	20	2063118421	2063118421	1	0	0	1	0	0	1	1	0	1	1	0	
251	1990	21	2063657858	2063657858	1	0	0	1	0	1	1	1	1	1	1	0	
252	1990	22	2063590122	2063590133	12	0	0	1	0	1	0	0	1	1	1	1	
253	1990	23	2063008025	2063008025	1	0	0	1	0	0	0	1	1	1	1	1	1
254	1990	24	518643736	518643740	5	0	0	1	0	1	0	1	1	1	1	1	
255	1990	25	517404239	517404240	2	0	0	1	0	0	0	0	0	0	1	0	
256	1990	26	518191557	518191557	1	0	0	1	0	1	0	0	1	1	1	1	
257	1990	27	518050807	518050807	1	0	0	1	0	0	0	1	1	1	1	1	
258	1990	28	518210210	518210210	1	0	0	1	0	1	0	0	1	1	1	1	
259	1990	29	2062391333	2062391333	1	0	0	1	0	0	1	0	1	1	1	0	
260	1990	30	621967762	621967762	1	0	0	1	0	0	0	1	1	1	1	1	
261	1990	31	776221554	776221555	2	0	0	1	0	0	0	0	1	1	1	1	
262	1990	32	94547582	94547590	9	0	0	1	0	1	0	1	1	1	1	1	
263	1990	33	94525487	94525488	2	1	0	1	0	0	1	0	1	1	1	0	
264	1990	34	94403082	94403083	2	1	0	1	0	0	0	0	1	1	1	1	1
265	1990	35	60015577A	60015577A	1	0	0	1	0	0	0	0	1	1	1	1	
266	1990	36	60015553	60015553	1	0	0	1	0	0	0	0	1	1	1	1	
267	1990	37	518574535	518574538	4	0	0	1	0	1	0	1	1	1	1	1	
268	1990	38	2063118371	2063118372	2	0	0	1	0	1	1	1	0	1	1	0	1
269	1990	39	2063656096	2063656117	22	0	0	1	0	0	0	0	1	1	1	1	
270	1990	40	2063588303	2063588303	1	1	0	1	1	0	1	1	1	1	0	0	1X
271	1990	41	2054445377A	2054445377A	1	0	0	1	0	0	1	0	1	1	1	0	
272	1990	42	518643618	518643621	4	0	0	1	0	1	1	1	1	1	1	0	
273	1990	43	517743372	517743406	35	0	0	1	1	1	0	0	1	1	1	1	
274	1990	44	518191554	518191556	3	0	0	1	0	1	0	0	1	1	1	1	
275	1990	45	518050804	518050805	2	0	0	1	0	0	0	0	1	1	1	1	
276	1990	46	518210208	518210209	2	0	0	1	0	1	0	0	1	1	1	1	

277	1990	47	2062391328	2062391328	1	0	0	1	0	0	1	0	1	1	1	0	
278	1990	48	621967682	621967682	1	0	0	1	0	0	0	0	1	1	1	1	
279	1990	49	776221552	776221553	2	0	0	1	0	0	0	0	1	1	1	1	
280	1990	50	94547575	94547581	7	0	0	1	0	1	0	1	1	1	1	1	
281	1990	51	94525144	94525144	1	0	0	1	0	0	0	0	1	1	1	1	1
282	1990	52	94403080	94403081	2	1	0	1	0	0	0	0	1	1	1	1	
283	1990	53	60015576B	60015576B	1	0	0	1	0	0	0	0	1	1	1	1	
284	1990	54	60015552	60015552	1	0	0	1	0	0	0	0	1	1	1	1	
285	1990	55	518191528	518191531	4	1	0	1	0	1	1	0	1	1	1	0	
286	1990	56	2063117627	2063117628	2	1	0	1	1	0	1	0	0	1	1	0	
287	1990	57	2063656086	2063656086	1	1	0	1	0	0	1	0	1	1	1	0	
288	1990	58	2063587022	2063587022	1	0	0	1	0	1	0	1	1	1	1	1	
289	1990	59	2054445119A	2054445119A	1	0	0	1	0	0	1	0	1	1	0	0	
290	1990	60	518643615	518643617	3	0	0	1	0	1	0	1	1	1	1	1	
291	1990	61	517743100	517743133	34	0	0	1	0	0	0	0	1	1	1	1	
292	1990	62	518191550	518191553	4	1	0	1	1	1	1	0	1	1	1	0	
293	1990	63	518050799	518050799	1	0	0	1	0	0	0	1	1	1	1	1	
294	1990	64	518210202	518210207	6	0	0	1	0	0	0	0	1	1	1	1	
295	1990	65	2062274709	2062274709	1	0	0	1	0	1	0	1	1	1	1	1	
296	1990	66	621967415	621967415	1	0	0	1	0	1	0	0	1	1	1	1	1
297	1990	67	776221550	776221551	2	0	0	1	0	0	0	0	1	1	1	1	
298	1990	68	94547570	94547574	5	0	0	1	0	1	0	1	1	1	1	1	
299	1990	69	94525142	94525142	1	0	0	1	0	0	0	0	1	1	1	1	
300	1990	70	94403078	94403079	2	1	0	1	0	0	0	0	1	1	1	1	
301	1990	71	60015575B	60015575B	1	0	0	1	0	0	0	0	1	1	1	1	
302	1990	72	60015551	60015551	1	0	0	1	0	0	0	0	1	1	1	1	
303	1990	73	518520617	518520618	2	0	0	1	0	0	0	1	1	1	1	1	
304	1990	74	2063117616	2063117625	10	1	0	1	0	1	1	0	0	1	1	0	
305	1990	75	2063656000	2063656000	1	0	0	1	0	1	1	0	1	1	1	0	
306	1990	76	2063587020	2063587020	1	0	0	1	0	1	1	1	1	1	1	0	
307	1990	77	2061878710	2061878711	2	0	0	1	0	0	0	1	0	0	1	0	
308	1990	78	518643469	518643472	4	0	0	1	0	1	0	1	1	1	1	1	
309	1990	79	518207517	518207518	2	1	0	1	0	1	1	0	1	1	1	0	
310	1990	80	518191548	518191549	2	0	0	1	0	0	0	0	1	1	1	1	1
311	1990	81	518050788	518050788	1	1	0	1	0	0	1	1	1	1	1	0	
312	1990	82	518210193	518210194	2	0	0	1	0	0	0	0	1	1	1	1	
313	1990	83	2062274706	2062274706	1	0	0	1	0	1	0	1	1	1	1	1	
314	1990	84	621967091	621967091	1	0	0	1	0	1	0	1	1	1	1	1	
315	1990	85	94571741	94571742	2	0	0	1	0	0	0	0	1	1	1	1	
316	1990	86	94547497	94547497	1	1	0	1	0	0	1	0	1	1	1	0	1
317	1990	87	94525133	94525133	1	0	0	1	0	0	0	0	1	1	1	1	1
318	1990	88	94403077	94403077	1	0	0	1	0	0	1	0	1	1	0	0	
319	1990	89	60015575A	60015575A	1	0	0	1	0	0	0	0	1	1	1	1	
320	1990	90	60015557	60015557	1	0	0	1	0	0	0	0	1	1	1	1	
321	1990	91	518520616	518520616	1	0	0	1	0	0	0	1	1	1	1	1	
322	1990	92	2063116805	2063116805	1	0	0	1	0	0	1	0	0	0	1	0	
323	1990	93	2063653357	2063653357	1	0	0	1	0	1	1	0	1	1	1	0	
324	1990	94	2063580142	2063580147	6	0	0	1	0	0	0	0	1	1	1	1	
325	1990	95	2061878709	2061878709	1	0	0	1	0	0	0	1	1	1	1	1	1X
326	1990	96	518643437	518643439	3	0	0	1	1	1	1	1	1	1	1	0	
327	1990	97	518207291	518207296	6	1	0	1	0	1	0	0	1	1	1	1	
328	1990	98	518191547	518191547	1	0	0	1	0	0	1	0	1	1	1	0	
329	1990	99	518050747	518050753	7	0	0	1	0	1	1	1	1	1	1	0	1
330	19xx	1	503848104	503848142	39	0	0	1	0	0	0	0	1	1	1	1	
331	19xx	2	504564495	504564495	1	1	0	1	0	0	1	0	1	1	1	0	
332	19xx	3	504552019	504552029	11	0	0	1	0	1	0	0	1	1	1	1	
333	19xx	4	504852919	504852919	1	0	0	1	0	1	0	0	1	1	1	1	
334	19xx	5	502271326	502271329	4	0	0	1	1	1	0	0	1	1	1	1	1
335	19xx	6	620336290	620336294	5	0	0	1	1	0	0	0	1	1	1	1	
336	19xx	7	681720801	681720804	4	0	0	1	0	0	0	0	0	0	0	0	
337	19xx	8	681642939	681642940	2	0	0	1	0	0	0	0	0	0	1	0	
338	19xx	9	465694114	465694114	1	1	0	1	0	0	1	0	1	1	1	0	
339	19xx	10	682234339	682234339	1	0	0	1	1	1	1	0	1	1	1	0	
340	Bliley	1	TIMN0012764	TIMN0012774	11	0	0	1	0	1	0	0	1	1	1	1	
341	Bliley	2	1005045350	1005045350	1	0	0	1	0	1	1	1	1	1	1	0	
342	Bliley	3	1005154385	1005154388	4	0	0	1	0	1	0	1	1	1	1	1	
343	Bliley	4	2023205580	2023205580	1	0	0	1	0	1	0	1	1	1	1	1	
344	Bliley	5	2028379598	2028379602	5	0	0	0	1	1	0	0	0	0	1	0	

345	Bliley	6	2501024514	2501024514	1	0	0	1	0	1	0	1	1	1	1	1
346	Bliley	7	03746244	03746244	1	0	0	1	0	0	1	0	1	1	1	0
347	Bliley	8	521031920	521031921	2	0	0	1	0	1	0	1	1	1	1	1
348	Bliley	9	682040630	682040633	4	0	0	1	0	1	0	1	1	1	1	1
349	Bliley	10	536510392	536510393	2	0	0	1	0	1	0	1	1	1	1	1
				TOTALS:	1818	49	4	343	17	176	101	145	294	302	334	203	50

B.3 Metadata for Quota Sample Documents

The following data were derived by examining the XML archives of the completed Quota Sample.

NUM	STRAT	ITR	BATES-START	BATES-END	SRC	CLS	DAT	PGS	AMT	TKN	CDR	VER
1	1950	Q1	04350077	04350077	LL	NI	19281024	1	ALL	126	AW	NONE
2	1950	Q1	500418376	500418377	RJR	UE	19520314	2	ALL	230	AW	NONE
3	1950	Q1	514564391	514564400	RJR	UE	19390207	10	ALL	1574	ET	NONE
4	1950	Q1	514597415	514597435	RJR	UI	19480626	21	ALL	2134	AW	NONE
5	1950	Q1	93219094	93219094	LL	NI	19560309	1	ALL	82	AW	NONE
6	1950	Q1	MNAT00374451	MNAT00374458	ATC	UI	19580912	8	ALL	208	AW	NONE
7	1950	Q2	2022216785	2022216786	PM	UI	19540602	2	ALL	389	AW	NONE
8	1950	Q2	2048017646	2048017646	PM	UE	19531230	1	ALL	256	AW	NONE
9	1950	Q2	514538372	514538376	RJR	UE	19480330	5	ALL	916	AW	NONE
10	1950	Q2	ATX040782766	ATX040782766	ATC	NI	19550304	1	ALL	116	AW	NONE
11	1950	Q2	MNAT00115479	MNAT00115481	ATC	UI	19480305	3	ALL	746	AW	NONE
12	1950	Q2	MNAT00606104	MNAT00606108	ATC	NI	19341105	5	ALL	1535	SM	NONE
13	1950	Q3	1001534667	1001534686	PM	UE	19521229	20	PART	2193	HW	NONE
14	1950	Q3	502470877	502470877	RJR	UE	19470418	1	ALL	166	HW	NONE
15	1950	Q3	ATX040194304	ATX040194306	ATC	UI	19521023	3	ALL	439	HW	AW
16	1950	Q3	ATX05_0170990	ATX05_0170990	ATC	UI	19270829	2	ALL	156	HW	NONE
17	1950	Q3	ATX05_0227117	ATX05_0227131	ATC	NI	19401115	15	ALL	261	HW	NONE
18	1950	Q3	HK2389264	HK2389264	CTR	NI	19551213	1	ALL	63	HW	NONE
19	1950	Q4	1002763364	1002763364	PM	UE	19350301	1	ALL	252	SM	NONE
20	1950	Q4	1003077754	1003077756	PM	NI	19480514	3	ALL	228	SM	NONE
21	1950	Q4	514564158	514564173	RJR	UI	19380913	16	ALL	1380	SM	NONE
22	1950	Q4	620083400	620083402	BW	NI	19590105	3	ALL	105	ET	NONE
23	1950	Q4	680255457	680255560	BW	UE	19540907	104	PART	2081	ET	NONE
24	1950	Q4	MNAT00467525	MNAT00467525	ATC	UI	19570516	1	ALL	136	ET	NONE
25	1960	Q1	01137715	01137715	LL	NI	19650112	1	ALL	159	AW	NONE
26	1960	Q1	01195752	01195752	LL	NI	19630805	1	ALL	161	AW	NONE
27	1960	Q1	1000862060	1000862061	PM	NI	19611205	2	ALL	486	ET	NONE
28	1960	Q1	1001901195	1001901199	PM	UI	19640000	5	ALL	1152	AW	NONE
29	1960	Q1	1002968498	1002968546	PM	UI	19690000	49	ALL	2190	AW	NONE
30	1960	Q1	11318888	11318888	CTR	NI	19631213	1	ALL	43	AW	NONE
31	1960	Q1	11319980	11319980	CTR	NE	19620813	1	ALL	178	AW	NONE
32	1960	Q1	2026367091	2026367092	PM	UI	19671112	2	ALL	795	AW	NONE
33	1960	Q1	2048925980	2048926014	PM	UI	19641000	35	PART	2355	AW	NONE
34	1960	Q1	2056159295	2056159295	PM	NI	19620813	1	ALL	179	AW	NONE
35	1960	Q1	500325578	500325614	ATC	UE	19610214	37	ALL	2078	AW	NONE
36	1960	Q1	681841016	681841016	BW	NI	19640826	1	ALL	103	AW	NONE
37	1960	Q1	ATX080023249	ATX080023250	ATC	NI	19640824	2	ALL	166	AW	NONE
38	1960	Q1	MNAT00558527	MNAT00558532	ATC	UI	19660615	6	ALL	606	AW	NONE
39	1960	Q1	MNAT00594069	MNAT00594069	ATC	UI	19670220	1	ALL	73	AW	NONE
40	1960	Q1	MNAT00820755	MNAT00820755	ATC	NI	19620219	1	ALL	53	AW	NONE
41	1960	Q2	1001604577	1001604577	PM	NI	19650305	1	ALL	93	AW	NONE
42	1960	Q2	1005110378	1005110381	TI	NE	19690000	4	ALL	845	AW	NONE
43	1960	Q2	500170004	500170004	RJR	NI	19667021	1	ALL	111	AW	NONE
44	1960	Q2	500612486	500612487	RJR	UI	19630403	2	ALL	140	AW	NONE
45	1960	Q2	502059858	502059858	RJR	UI	19641000	1	ALL	524	AW	NONE
46	1960	Q2	650203481	650203481	BW	NI	19600816	1	ALL	135	AW	NONE

47	1960	Q2	85669259	85669261	LL	NI	19620831	3	ALL	362	AW	NONE
48	1960	Q2	ATX03_0044805	ATX03_0044814	ATC	UI	19631007	10	ALL	893	AW	NONE
49	1960	Q2	ATX05_0022017	ATX05_0022030	ATC	UI	19671102	14	ALL	1246	AW	NONE
50	1960	Q2	ATX05_0113396	ATX05_0113396	ATC	NI	19610303	1	ALL	269	AW	NONE
51	1960	Q2	ATX080007098	ATX080007098	ATC	NI	19630813	1	ALL	142	AW	NONE
52	1960	Q2	ATX080010894	ATX080010895	ATC	NI	19640514	2	ALL	158	AW	NONE
53	1960	Q2	ATX_795162_0060	ATX_795162_0061	ATC	UE	19670623	2	ALL	64	AW	NONE
54	1960	Q2	HK0982139	HK0982149	CTR	UI	19601104	11	PART	2253	AW	NONE
55	1960	Q2	MNAT00778295	MNAT00778295	ATC	NI	19670419	1	ALL	149	AW	NONE
56	1960	Q2	TIMN0127604	TIMN0127605	TI	UI	19630415	2	ALL	355	ET	NONE
57	1960	Q3	1000249727	1000249728	PM	UI	19621206	2	ALL	590	AW	NONE
58	1960	Q3	1000322366	1000322367	PM	NI	19660315	2	ALL	179	HW	NONE
59	1960	Q3	1000849881	1000849882	PM	NI	19640624	2	ALL	92	AW	NONE
60	1960	Q3	1002430081	1002430125	PM	NI	19691122	45	ALL	2272	AW	NONE
61	1960	Q3	1005038370	1005038376	PM	UI	19621229	7	ALL	2121	AW	NONE
62	1960	Q3	2022241482	22022241483	PM	NI	19620531	2	ALL	208	HW	NONE
63	1960	Q3	2049387255	2049387255	PM	NI	19691229	1	ALL	187	HW	NONE
64	1960	Q3	50041884	50041884	CTR	NI	19670525	1	ALL	130	HW	NONE
65	1960	Q3	50044080	50044082	CTR	UI	19661001	3	ALL	869	HW	NONE
66	1960	Q3	50047883	50047884	CTR	NI	19660817	2	ALL	427	HW	NONE
67	1960	Q3	501772729	501772741	RJR	UI	19620000	13	ALL	2059	HW	NONE
68	1960	Q3	502218769	502218769	RJR	UE	19610320	1	ALL	93	HW	NONE
69	1960	Q3	502794415	502794426	RJR	UI	19610920	12	ALL	2620	HW	NONE
70	1960	Q3	MNAT00290815	MNAT00290815	ATC	NI	19631220	1	ALL	173	HW	AW
71	1960	Q3	MNAT00528154	MNAT00528157	ATC	UI	19650402	4	ALL	739	AW	NONE
72	1960	Q3	MNAT00887147	MNAT00887147	ATC	NE	19610421	1	ALL	95	SM	NONE
73	1960	Q4	01148569	01148569	LL	NI	19600514	1	ALL	141	SM	NONE
74	1960	Q4	1000292034	1000292034	PM	NI	19690203	1	ALL	109	SM	NONE
75	1960	Q4	2015033720	2015033720	PM	UI	19660000	1	ALL	172	SM	NONE
76	1960	Q4	2022177237	2022177239	PM	NI	19640422	3	ALL	475	SM	AW
77	1960	Q4	2022241016	2022241030	PM	UI	19610000	15	ALL	1999	SM	NONE
78	1960	Q4	2026452194	2026452195	PM	UI	19660000	2	ALL	200	SM	NONE
79	1960	Q4	501889349	501889349	RJR	NE	19670202	1	ALL	138	SM	AW
80	1960	Q4	570394164	570394183	BW	UI	19630226	20	ALL	2073	AW	NONE
81	1960	Q4	680258078	680258078	BW	NI	19660917	1	ALL	238	SM	NONE
82	1960	Q4	680279579	680279581	BW	UI	19600505	3	ALL	566	SM	NONE
83	1960	Q4	80630953	80630953	LL	NI	19670622	1	ALL	76	SM	NONE
84	1960	Q4	ATX040060125	ATX040060125	ATC	NI	19630418	1	ALL	68	SM	NONE
85	1960	Q4	ATX040310704	ATX040310726	ATC	UE	19660406	23	PART	2138	SM	NONE
86	1960	Q4	ATX040843121	ATX040843129	ATC	NI	19621212	9	ALL	1374	SM	NONE
87	1960	Q4	AXT020012063	AXT020012082	ATC	UI	19651210	20	PART	2056	SM	NONE
88	1960	Q4	MNAT00616603	MNAT00616604	ATC	NI	19600814	2	ALL	257	SM	AW
89	1970	Q1	03722346	03722346	LL	UI	19720000	1	ALL	97	AW	NONE
90	1970	Q1	03750675	03750675	LL	NI	19781023	1	ALL	170	AW	NONE
91	1970	Q1	1000027093	1000027094	PM	UI	19710708	2	ALL	359	HW	NONE
92	1970	Q1	1000147911	1000147911	PM	UI	19781130	1	ALL	205	AW	NONE
93	1970	Q1	1000292993	1000292993	PM	NI	19710806	1	ALL	312	AW	NONE
94	1970	Q1	1000364896	1000364906	PM	UI	19770323	11	ALL	271	AW	NONE
95	1970	Q1	1000736375	1000736379	PM	UI	19750100	5	ALL	1006	AW	NONE
96	1970	Q1	1001511601	1001511604	PM	UI	19760927	4	ALL	898	HW	NONE
97	1970	Q1	1005064360	1005064361	PM	NI	19750512	2	ALL	413	AW	NONE
98	1970	Q1	1005133325	1005133325	TI	UI	19710000	1	ALL	215	AW	NONE
99	1970	Q1	2000761001	2000761002	PM	UI	19700312	2	ALL	522	AW	NONE
100	1970	Q1	2001208818	2001208818	PM	NE	19760422	1	ALL	70	AW	NONE
101	1970	Q1	2028626055	2028626057	PM	UI	19750605	3	ALL	747	AW	NONE
102	1970	Q1	2040260786	2040260796	PM	UI	19790725	11	ALL	913	AW	NONE
103	1970	Q1	2050941296	2050941328	PM	UI	19750219	33	ALL	1355	AW	NONE
104	1970	Q1	500008576	500008578	BW	NI	19771201	3	ALL	500	AW	NONE
105	1970	Q1	500531640	500531649	RJR	UI	19731227	10	ALL	2124	AW	NONE
106	1970	Q1	500872831	500872832	RJR	NI	19790928	2	ALL	570	AW	NONE
107	1970	Q1	501791319	501791319	RJR	NI	19710811	1	ALL	107	AW	NONE
108	1970	Q1	502090824	502090824	RJR	NI	19790230	1	ALL	240	AW	NONE
109	1970	Q1	503257271	503257271	RJR	NI	19700603	1	ALL	244	AW	NONE
110	1970	Q1	570320909	570320950	BW	UI	19740430	42	ALL	2000	HW	NONE
111	1970	Q1	60172127	60172128	ATC	NI	19780518	2	ALL	118	AW	NONE
112	1970	Q1	60198802	60198809	ATC	NI	19791102	8	ALL	75	AW	NONE
113	1970	Q1	650511474	650511474	BW	UI	19780413	1	ALL	123	AW	NONE
114	1970	Q1	675023264	675023264	BW	NI	19720524	1	ALL	59	ET	NONE

115	1970	Q1	680106368	680106370	BW	NI	19741008	3	ALL	727	AW	NONE
116	1970	Q1	686056400	686056400	BW	UI	19750905	1	ALL	102	AW	NONE
117	1970	Q1	91046533	91046536	LL	UI	19750600	4	ALL	986	AW	NONE
118	1970	Q1	ATX040209930	ATX040209937	TI	UI	19731112	8	ALL	2127	AW	NONE
119	1970	Q1	ATX040307221	ATX040307224	ATC	UI	19790609	4	ALL	1028	AW	NONE
120	1970	Q1	ATX040665728	ATX040665728	ATC	UI	19770210	1	ALL	359	AW	NONE
121	1970	Q1	ATX040827617	ATX040827619	ATC	NI	19730224	3	ALL	168	AW	NONE
122	1970	Q1	ATX300001016	ATX300001016	ATC	UI	19770912	1	ALL	120	AW	NONE
123	1970	Q1	CTRCONTRACTS022990	CTRCONTRACTS022990	CTR	NE	19750602	1	ALL	69	AW	NONE
124	1970	Q1	MNAT00718324	MNAT00718325	ATC	UI	19721025	2	ALL	367	AW	NONE
125	1970	Q2	03593594	03593594	LL	NI	19700522	1	ALL	162	AW	NONE
126	1970	Q2	03671878	03671879	LL	NI	19760203	2	ALL	120	AW	NONE
127	1970	Q2	03732265	03732270	LL	NI	19790416	6	ALL	1983	AW	NONE
128	1970	Q2	1001517814	1001517815	PM	NI	19720602	2	ALL	274	AW	NONE
129	1970	Q2	1001853457	1001853457	PM	UI	19750506	1	ALL	146	AW	NONE
130	1970	Q2	10408940	10408945	CTR	UI	19780600	6	ALL	2601	AW	NONE
131	1970	Q2	2001216666	2001216669	PM	UI	19700808	4	ALL	577	AW	NONE
132	1970	Q2	2025027475	2025027476	PM	NI	19750105	2	ALL	329	AW	NONE
133	1970	Q2	2047405735	2047405735	PM	UI	19750000	1	ALL	163	AW	NONE
134	1970	Q2	2051011027	2051011028	PM	UI	19780000	2	ALL	280	AW	NONE
135	1970	Q2	2501241846	2501241846	PM	UI	19721223	1	ALL	224	AW	NONE
136	1970	Q2	464401000	464401044	BW	UI	19790700	45	PART	2081	AW	NONE
137	1970	Q2	500006704	500006705	BW	NI	19730131	2	ALL	433	AW	NONE
138	1970	Q2	500193626	500193627	RJR	UI	19780115	2	ALL	443	AW	NONE
139	1970	Q2	501368300	501368302	RJR	UI	19780123	3	ALL	238	AW	NONE
140	1970	Q2	503247364	503247365	RJR	NI	19791010	2	ALL	225	AW	NONE
141	1970	Q2	507877294	507877299	RJR	UI	19790000	6	ALL	1369	AW	NONE
142	1970	Q2	570200037	570200046	BW	UI	19750307	10	ALL	726	AW	NONE
143	1970	Q2	670300714	670300714	BW	NI	19741017	1	ALL	51	AW	NONE
144	1970	Q2	674149405	674149406	BW	NI	19770824	2	ALL	501	AW	NONE
145	1970	Q2	680038272	680038272	BW	NI	19740509	1	ALL	94	AW	NONE
146	1970	Q2	680065763	680065768	BW	UI	19750607	6	ALL	716	AW	NONE
147	1970	Q2	680233000	680233001	BW	UI	19760913	2	ALL	496	AW	NONE
148	1970	Q2	684033262	684033262	BW	NI	19780330	1	ALL	71	AW	NONE
149	1970	Q2	686017111	686017113	BW	NE	19771012	3	ALL	747	AW	NONE
150	1970	Q2	776089304	776089314	BW	NI	19760310	11	ALL	2173	AW	NONE
151	1970	Q2	777088699	777088699	BW	UI	19760127	1	ALL	737	ET	NONE
152	1970	Q2	85173191	85173191	LL	UI	19760619	1	ALL	157	AW	NONE
153	1970	Q2	89301401	89301424	LL	UI	19731031	24	ALL	2061	AW	NONE
154	1970	Q2	ATX040085814	ATX040085819	ATC	UI	19720000	6	ALL	467	AW	NONE
155	1970	Q2	ATX05_0006533	ATX05_0006534	ATC	NI	19760205	2	ALL	125	AW	NONE
156	1970	Q2	ATX080011761	ATX080011761	ATC	UI	19720600	1	ALL	86	AW	NONE
157	1970	Q2	HK0042150	HK0042150	CTR	UI	19730000	1	ALL	87	AW	NONE
158	1970	Q2	SF0823867	SF0823892	CTR	UI	19760206	26	ALL	2231	AW	NONE
159	1970	Q2	TIMN0109658	TIMN0109658	TI	NE	19700221	1	ALL	190	AW	NONE
160	1970	Q2	ZN17256	ZN17291	CTR	UI	19751005	36	PART	2208	AW	NONE
161	1970	Q3	00103870	00103870	LL	UI	19710413	1	ALL	53	AW	NONE
162	1970	Q3	1000016672	1000016676	PM	UI	19750829	5	ALL	1047	AW	NONE
163	1970	Q3	1000212367	1000212368	PM	NI	19770630	2	ALL	448	AW	NONE
164	1970	Q3	1000311074	1000311074	RJR	NI	19710505	1	ALL	253	AW	NONE
165	1970	Q3	1003290426	1003290427	PM	NE	19751108	2	ALL	520	AW	NONE
166	1970	Q3	1003702983	1003702983	PM	NI	19740102	1	ALL	250	AW	NONE
167	1970	Q3	1005049184	1005049185	PM	NI	1970000	2	ALL	328	AW	NONE
168	1970	Q3	1005075578	1005075578	PM	NI	19740000	1	ALL	268	AW	NONE
169	1970	Q3	1005200215	1005200247	PM	UI	19750106	33	PART	2086	AW	NONE
170	1970	Q3	11294502	11294502	CTR	NI	19720829	1	ALL	84	AW	NONE
171	1970	Q3	2020212318	2020212319	PM	UI	19720209	2	ALL	367	AW	NONE
172	1970	Q3	2022143088	2022143089	PM	UI	19781107	2	ALL	283	AW	NONE
173	1970	Q3	2024975552	2024975552	PM	NI	19780308	1	ALL	171	AW	NONE
174	1970	Q3	2026500735	2026500735	PM	UI	19750000	1	ALL	160	AW	NONE
175	1970	Q3	500810054	500810056	RJR	UI	19700909	3	ALL	1286	AW	NONE
176	1970	Q3	500877719	500877720	RJR	NE	19791102	2	ALL	129	AW	NONE
177	1970	Q3	501002809	501002822	RJR	NI	19721106	14	PART	2080	AW	NONE
178	1970	Q3	501018759	501018760	RJR	NI	19710126	2	ALL	444	AW	NONE
179	1970	Q3	501239203	501239203	RJR	NI	19760924	1	ALL	356	AW	NONE
180	1970	Q3	501477276	501477280	RJR	NI	19770716	5	PART	2132	AW	NONE
181	1970	Q3	501477475	501477479	RJR	UI	19730228	5	ALL	766	AW	NONE
182	1970	Q3	501497202	501497202	RJR	UI	19730416	1	ALL	234	AW	NONE

183	1970	Q3	508370241	508370241	RJR	UI	19720526	1	ALL	159	AW	NONE
184	1970	Q3	511922328	511922330	RJR	UI	19711005	3	ALL	652	AW	NONE
185	1970	Q3	664063379	664063381	BW	NI	19790807	3	ALL	528	ET	NONE
186	1970	Q3	676146461	676146463	BW	UI	19720703	3	ALL	856	AW	NONE
187	1970	Q3	677191064	677191067	BW	UI	19740513	4	ALL	784	AW	NONE
188	1970	Q3	680092883	680092885	BW	UI	19730303	3	ALL	964	AW	NONE
189	1970	Q3	680112152	680112152	BW	UI	19760504	1	ALL	133	AW	NONE
190	1970	Q3	ATX05_0006007	ATX05_0006007	ATC	UI	19740213	1	ALL	129	AW	NONE
191	1970	Q3	ATX05_0171658	ATX05_0171658	ATC	NI	19720912	1	ALL	95	BH	NONE
192	1970	Q3	HK0668029	HK0668029	CTR	UI	19760511	1	ALL	122	AW	NONE
193	1970	Q3	HK1691097	HK1691097	CTR	UI	19740819	1	ALL	193	AW	NONE
194	1970	Q3	MNAT00752356	MNAT00752356	ATC	UI	19730822	1	ALL	95	AW	NONE
195	1970	Q3	TIMN0064609	TIMN0064609	TI	UI	19731125	1	ALL	100	AW	NONE
196	1970	Q3	TITX0011710	TITX0011710	TI	UI	19730119	1	ALL	486	AW	NONE
197	1970	Q4	03712383	03712384	LL	NE	19760511	2	ALL	305	SM	NONE
198	1970	Q4	03724617	03724617	LL	NI	19730223	1	ALL	81	SM	NONE
199	1970	Q4	03730856	03730856	LL	NI	19720519	1	ALL	58	SM	NONE
200	1970	Q4	04233241	04233250	LL	UI	19790000	10	ALL	1507	SM	NONE
201	1970	Q4	1000774597	1000774599	PM	NI	19790226	3	ALL	838	SM	NONE
202	1970	Q4	1000845352	1000845352	PM	NI	19740810	1	ALL	352	SM	NONE
203	1970	Q4	1001507531	1001507531	PM	NI	19790908	1	ALL	155	SM	NONE
204	1970	Q4	1001863667	1001863667	PM	UI	19780622	1	ALL	168	SM	NONE
205	1970	Q4	2022147756	2022147758	PM	UI	19710809	3	ALL	593	SM	NONE
206	1970	Q4	2022224303	2022224303	PM	NI	19760711	1	ALL	133	SM	NONE
207	1970	Q4	2026341218	2026341225	PM	UI	19720712	8	ALL	1643	SM	NONE
208	1970	Q4	2026514690	2026514690	PM	UI	19731122	1	ALL	185	SM	AW
209	1970	Q4	500208804	500208804	RJR	NE	19761101	1	ALL	82	SM	NONE
210	1970	Q4	50060867	50060872	CTR	UI	19711026	6	ALL	1832	SM	NONE
211	1970	Q4	50136358	50136358	CTR	NI	19790718	1	ALL	113	SM	NONE
212	1970	Q4	501556925	501556925	RJR	NI	19711111	1	ALL	62	SM	NONE
213	1970	Q4	501720844	501720846	RJR	UI	19750409	3	ALL	498	SM	NONE
214	1970	Q4	502878463	502878515	RJR	UI	19720127	53	ALL	1588	SM	NONE
215	1970	Q4	503681521	503681522	RJR	UI	19781111	2	ALL	182	SM	NONE
216	1970	Q4	503821158	503821158	RJR	NI	19771012	1	ALL	136	SM	NONE
217	1970	Q4	654072906	654072908	BW	UI	19770708	3	ALL	440	SM	NONE
218	1970	Q4	777115829	777115829	BW	UI	19750702	1	ALL	148	SM	NONE
219	1970	Q4	81331180	81331180	LL	NI	19790319	1	ALL	110	SM	NONE
220	1970	Q4	ATX040120514	ATX040120532	ATC	UI	19740803	19	ALL	1991	SM	NONE
221	1970	Q4	ATX040154856	ATX040154858	ATC	UI	19750000	3	ALL	159	SM	NONE
222	1970	Q4	ATX080002785	ATX080002788	ATC	UI	19760618	4	ALL	978	SM	NONE
223	1970	Q4	CTRSP_FILES026021	CTRSP_FILES026021	CTR	NI	19710404	1	ALL	99	SM	NONE
224	1970	Q4	DM0140020	DM0140020	CTR	UI	19751210	1	ALL	80	SM	NONE
225	1970	Q4	HK1871057	HK1871068	CTR	UI	19700630	12	ALL	745	SM	NONE
226	1970	Q4	HK2186086	HK2186086	CTR	UI	19720000	1	ALL	124	SM	NONE
227	1970	Q4	HT0143032	HT0143032	CTR	UI	19760701	1	ALL	221	SM	NONE
228	1970	Q4	MNAT00275767	MNAT00275767	ATC	NI	19760720	1	ALL	286	SM	NONE
229	1970	Q4	MNAT00290341	MNAT00290343	ATC	NI	19740222	3	ALL	508	SM	NONE
230	1970	Q4	MNAT00757792	MNAT00757793	ATC	UI	19741228	2	ALL	138	SM	AW
231	1970	Q4	MNAT00757808	MNAT00757808	ATC	UI	19730603	1	ALL	85	SM	NONE
232	1970	Q4	MNAT00777199	MNAT00777199	ATC	UI	19710927	1	ALL	136	SM	NONE
233	1980	Q1	03753086	03753088	LL	NI	19820120	3	ALL	180	AW	NONE
234	1980	Q1	1003195215	1003195215	PM	NI	19831104	1	ALL	67	AW	NONE
235	1980	Q1	2000596903	2000596914	PM	UI	19870718	12	ALL	2417	AW	NONE
236	1980	Q1	2001115061	2001115148	PM	UI	19840630	88	PART	2203	AW	NONE
237	1980	Q1	2021651791	2021651794	PM	UI	19820000	4	ALL	408	AW	NONE
238	1980	Q1	2022946920	2022946920	PM	NI	19880909	1	ALL	256	AW	NONE
239	1980	Q1	2023103364	2023103364	PM	NI	19830522	1	ALL	145	AW	NONE
240	1980	Q1	2026009338	2026009521	PM	UI	19890325	184	PART	2132	AW	NONE
241	1980	Q1	2028456398	2028456398	PM	UI	19860611	1	ALL	57	AW	NONE
242	1980	Q1	2029260117	2029260118	PM	NI	19880530	2	ALL	220	AW	NONE
243	1980	Q1	2031426348	2031426348	PM	UI	19860410	1	ALL	329	AW	NONE
244	1980	Q1	2040797915	2040797915	PM	NI	19800815	1	ALL	62	AW	NONE
245	1980	Q1	2043375178	2043375197	PM	NI	19821031	20	PART	2084	AW	NONE
246	1980	Q1	2049347698	2049347698	PM	UI	19860300	1	ALL	295	AW	NONE
247	1980	Q1	2051197420	2051197420	PM	UI	19810119	1	ALL	128	AW	NONE
248	1980	Q1	2501662150	2501662150	PM	NI	19870311	1	ALL	304	AW	NONE
249	1980	Q1	464002536	464002592	BW	NI	19830527	57	PART	2462	ET	NONE
250	1980	Q1	500627305	500627307	RJR	UI	19840300	3	ALL	429	AW	NONE

251	1980	Q1	502012325	502012326	RJR	NI	19830825	2	ALL	246	AW	NONE
252	1980	Q1	503631680	503631682	RJR	NI	19850000	3	ALL	312	AW	NONE
253	1980	Q1	504443169	50443171	RJR	UI	19831012	3	ALL	491	AW	NONE
254	1980	Q1	504589559	504589561	RJR	UI	19850500	3	ALL	987	AW	NONE
255	1980	Q1	504740775	504740777	RJR	UI	19850614	3	ALL	514	AW	NONE
256	1980	Q1	504744312	504744312	RJR	NI	19850208	1	ALL	199	AW	NONE
257	1980	Q1	504862231	504862231	LL	NI	19810507	1	ALL	450	AW	NONE
258	1980	Q1	504905532	504905535	RJR	NI	19860300	4	ALL	1341	AW	NONE
259	1980	Q1	505353277	50353280	RJR	NI	19860708	4	ALL	1408	HW	NONE
260	1980	Q1	505746853	505746853	RJR	UI	19841229	1	ALL	364	AW	NONE
261	1980	Q1	505876974	505877022	RJR	NI	19850311	49	PART	2204	AW	NONE
262	1980	Q1	506044286	506044300	PM	UI	19840823	15	ALL	2185	AW	NONE
263	1980	Q1	506371345	506371346	RJR	UI	19880201	2	ALL	222	AW	NONE
264	1980	Q1	506428080	506428282	RJR	UI	19870909	203	ALL	637	AW	NONE
265	1980	Q1	506535622	506535622	RJR	NI	19880213	1	ALL	207	AW	NONE
266	1980	Q1	506808914	506808920	RJR	UI	19890110	7	ALL	1967	HW	NONE
267	1980	Q1	507066381	507066386	RJR	UI	19890711	6	ALL	348	HW	NONE
268	1980	Q1	507137069	507137070	RJR	NI	19890000	2	ALL	391	AW	NONE
269	1980	Q1	507937482	507937482	RJR	UI	19890822	1	ALL	73	AW	NONE
270	1980	Q1	508065034	508065042	RJR	UI	19890630	9	ALL	866	AW	NONE
271	1980	Q1	509859136	509859136	RJR	NI	19890704	1	ALL	80	AW	NONE
272	1980	Q1	510434143	510434146	RJR	NI	19890108	4	ALL	716	AW	NONE
273	1980	Q1	510954816	510954818	RJR	UI	19830605	3	ALL	163	HW	NONE
274	1980	Q1	511580777	511580779	RJR	NI	19880813	3	ALL	151	AW	NONE
275	1980	Q1	512816109	512816109	RJR	NI	19891212	1	ALL	164	AW	NONE
276	1980	Q1	514919407	514919408	RJR	NI	19880301	2	ALL	111	AW	NONE
277	1980	Q1	521045421	521045421	BW	NE	19850413	1	ALL	172	AW	NONE
278	1980	Q1	620464175	6240464201	BW	UI	19870801	27	ALL	2014	AW	NONE
279	1980	Q1	620682024	620682029	BW	UI	19800910	6	ALL	1631	AW	NONE
280	1980	Q1	620802637	620802654	BW	NI	19800424	18	ALL	1704	AW	NONE
281	1980	Q1	621015665	621015666	BW	NI	19861028	2	ALL	172	AW	NONE
282	1980	Q1	635417360	635417368	BW	NI	19820224	9	ALL	816	AW	NONE
283	1980	Q1	656024403	656024405	RJR	NI	19850820	3	ALL	66	AW	NONE
284	1980	Q1	659032972	659032973	BW	NI	19810920	2	ALL	60	AW	NONE
285	1980	Q1	660059719	660059722	BW	NI	19810424	4	ALL	943	AW	NONE
286	1980	Q1	660064385	660064385	BW	NI	19811007	1	ALL	169	AW	NONE
287	1980	Q1	670576157	670576286	BW	UI	19821010	130	PART	2082	AW	NONE
288	1980	Q1	675109665	675109676	BW	NI	19830603	12	ALL	799	AW	NONE
289	1980	Q1	680580450	680580450	BW	NI	19850427	1	ALL	302	AW	NONE
290	1980	Q1	680591270	680591273	BW	UI	19830616	4	ALL	466	AW	NONE
291	1980	Q1	682341727	682341740	BW	UI	19831020	14	PART	2143	AW	NONE
292	1980	Q1	690149447	690149470	BW	UI	19850000	24	PART	2230	AW	NONE
293	1980	Q1	89397458	89397488	LL	UI	19860227	31	PART	1992	AW	NONE
294	1980	Q1	93455341	93455344	LL	UI	19860731	4	ALL	261	AW	NONE
295	1980	Q1	ATX03_0025429	ATX03_0025431	ATC	NI	19820719	3	ALL	176	AW	NONE
296	1980	Q1	ATX040102056	ATX040102056	ATC	UI	19880309	1	ALL	148	AW	NONE
297	1980	Q1	ATX040421892	ATX040421892	ATC	NI	19820111	1	ALL	187	AW	NONE
298	1980	Q1	ATX040980579	ATX040980580	ATC	NI	19880000	2	ALL	222	AW	NONE
299	1980	Q1	CTRSP_FILES023888	CTRSP_FILES023888	CTR	UI	19841206	1	ALL	56	AW	NONE
300	1980	Q1	MNAT00559787	MNAT00559787	ATC	UI	19880211	1	ALL	180	AW	NONE
301	1980	Q1	NYO_625_(6_69)	NYO_625_(6_69)	ATC	NI	19850411	2	ALL	115	AW	NONE
302	1980	Q1	TIDN0004236	TIDN0004238	TI	NI	19891025	3	ALL	629	AW	NONE
303	1980	Q1	TIMN0167899	TIMN0167899	TI	NI	19860108	1	ALL	110	AW	NONE
304	1980	Q2	04225201	04225202	LL	NI	19800313	2	ALL	394	AW	NONE
305	1980	Q2	1000779683	1000779688	PM	UI	19800423	6	ALL	1042	AW	NONE
306	1980	Q2	1005063369	1005063369	PM	NI	19830216	1	ALL	53	AW	NONE
307	1980	Q2	11299725	11299726	CTR	NI	19800812	2	ALL	384	AW	NONE
308	1980	Q2	2001300163	2001300163	PM	UI	19860000	1	ALL	221	AW	NONE
309	1980	Q2	2021610427	2021610427	PM	NI	19870318	1	ALL	138	AW	NONE
310	1980	Q2	2023272007	2023272007	PM	UI	19840524	1	ALL	81	AW	NONE
311	1980	Q2	2024058560	2024058560	PM	UI	19840000	1	ALL	171	AW	NONE
312	1980	Q2	2024481688	2024481692	PM	NI	19820517	5	ALL	381	AW	NONE
313	1980	Q2	2028623219	2028623220	PM	UI	19810821	2	ALL	341	AW	NONE
314	1980	Q2	2028624748	2028624749	PM	UI	19820204	2	ALL	507	AW	NONE
315	1980	Q2	2029137905	2029137905	PM	NI	19831021	1	ALL	214	AW	NONE
316	1980	Q2	2029252048	2029252048	PM	UI	19860000	1	ALL	226	AW	NONE
317	1980	Q2	2029271508	2029271513	PM	UI	19830000	6	ALL	2148	AW	NONE
318	1980	Q2	2031402953	2031402953	PM	NI	19851022	1	ALL	52	AW	NONE

319	1980	Q2	2040204289	2040204291	PM	UI	19800329	3	ALL	1127	AW	NONE
320	1980	Q2	2040918451	2040918451	PM	NI	19850821	1	ALL	148	AW	NONE
321	1980	Q2	2043919556	2043919571	PM	UI	19870620	16	PART	2188	AW	NONE
322	1980	Q2	2044223026	2044223034	PM	UI	19840800	9	ALL	1460	AW	NONE
323	1980	Q2	2044290615	2044290616	PM	NI	19870609	2	ALL	362	AW	NONE
324	1980	Q2	2044927694	2044927694	PM	UI	19870916	1	ALL	108	ET	NONE
325	1980	Q2	2044983497	2044983497	PM	UI	19890727	1	ALL	61	AW	NONE
326	1980	Q2	2045083228	2045083228	PM	NI	19850312	1	ALL	98	AW	NONE
327	1980	Q2	2047938973	2047938973	PM	UI	19891101	1	ALL	208	AW	NONE
328	1980	Q2	2048526822	2048526822	PM	NI	19890222	1	ALL	128	AW	NONE
329	1980	Q2	2051989496	2051989496	PM	NI	19820826	1	ALL	133	AW	NONE
330	1980	Q2	2056165060A	2056165060A	PM	NI	19850220	1	ALL	58	AW	NONE
331	1980	Q2	2061643106	2061643106	PM	NI	19830715	1	ALL	96	AW	NONE
332	1980	Q2	2061678775	2061678776	PM	NI	19871202	2	ALL	188	AW	NONE
333	1980	Q2	2062555037	2062555117	PM	UI	19830100	81	ALL	1965	AW	NONE
334	1980	Q2	2501000064	2501000069	PM	UI	19880811	6	ALL	2013	AW	NONE
335	1980	Q2	464001789	464001791	PM	NI	19830831	3	ALL	628	AW	NONE
336	1980	Q2	500905619	500905627	RJR	UI	19840618	9	ALL	824	AW	NONE
337	1980	Q2	500928158	500928159	RJR	NI	19800917	2	ALL	269	AW	NONE
338	1980	Q2	501010054	501010057	RJR	UI	19831031	4	ALL	1171	AW	NONE
339	1980	Q2	501625859	501625861	RJR	UI	19800901	3	ALL	448	AW	NONE
340	1980	Q2	502203246	502203247	RJR	UI	19841024	2	ALL	222	ET	NONE
341	1980	Q2	503518982	503518984	RJR	UI	19830318	3	ALL	625	AW	NONE
342	1980	Q2	503905326	503905326	RJR	NI	19811121	1	ALL	195	AW	NONE
343	1980	Q2	503905718	503905719	RJR	NI	19820826	2	ALL	398	AW	NONE
344	1980	Q2	503948629	503948635	RJR	NI	19841214	7	ALL	1908	AW	NONE
345	1980	Q2	504337601	504337603	RJR	NI	19850123	3	ALL	737	AW	NONE
346	1980	Q2	504585128	504585129	RJR	UI	19850415	2	ALL	402	AW	NONE
347	1980	Q2	505868965	505868968	RJR	NI	19860820	4	ALL	342	AW	NONE
348	1980	Q2	506282427	506282454	RJR	UI	19870506	28	ALL	384	AW	NONE
349	1980	Q2	506423529	506423530	RJR	UI	19870916	2	ALL	352	AW	NONE
350	1980	Q2	506559078	506559084	RJR	NI	19870127	7	ALL	2082	AW	NONE
351	1980	Q2	506686472	506686473	RJR	NI	19871029	2	ALL	370	AW	NONE
352	1980	Q2	509808181	509808187	RJR	UI	19870719	7	ALL	1919	AW	NONE
353	1980	Q2	512180224	512180226	RJR	UI	19841014	3	ALL	1286	AW	NONE
354	1980	Q2	512473163	512473181	RJR	NI	19890614	19	ALL	868	AW	NONE
355	1980	Q2	516712072	516712073	RJR	NI	19880102	2	ALL	412	AW	NONE
356	1980	Q2	521043633	521043634	BW	NI	19810828	2	ALL	345	AW	NONE
357	1980	Q2	620147516	620147525	BW	NI	19880301	10	ALL	276	AW	NONE
358	1980	Q2	620150885	620150888	BW	NI	19870612	4	ALL	544	AW	NONE
359	1980	Q2	621115496	621115499	BW	NI	19841005	4	ALL	689	AW	NONE
360	1980	Q2	623085522	623085524	BW	UI	19890520	3	ALL	260	AW	NONE
361	1980	Q2	656048223	656048224	BW	NI	19880308	2	ALL	86	AW	NONE
362	1980	Q2	675094380	675094403	BW	UI	19821216	24	ALL	787	AW	NONE
363	1980	Q2	689478896	689478897	BW	NI	19850812	2	ALL	384	AW	NONE
364	1980	Q2	690106979	690106980	BW	NI	19801018	2	ALL	270	AW	NONE
365	1980	Q2	80830699	80830700	LL	NI	19860502	2	ALL	503	AW	NONE
366	1980	Q2	85384790	85384816	LL	NI	19850520	27	ALL	1287	AW	NONE
367	1980	Q2	85682703	85682704	LL	NI	19820311	2	ALL	369	AW	NONE
368	1980	Q2	85709714	85709718	LL	UI	19840816	5	ALL	924	AW	NONE
369	1980	Q2	88208744	88208770	LL	UI	19870603	27	ALL	2652	AW	NONE
370	1980	Q2	91539011	91539011	LL	NE	19830619	1	ALL	119	AW	NONE
371	1980	Q2	91795716	91795721	LL	UI	19820303	6	ALL	1375	AW	NONE
372	1980	Q2	ATX05_0069812	ATX05_0069812	ATC	UI	19821229	1	ALL	257	AW	NONE
373	1980	Q2	CTRSP_FILES026242	CTRSP_FILES026242	CTR	NI	19880516	1	ALL	144	AW	NONE
374	1980	Q2	MNAT00788671	MNAT00788673	ATC	NI	19860725	3	ALL	377	AW	NONE
375	1980	Q3	03023603	03023603	LL	NI	19830426	1	ALL	82	AW	NONE
376	1980	Q3	03730267	03730267	LL	NI	19840125	1	ALL	175	AW	NONE
377	1980	Q3	03751359	03751359	LL	NI	19800306	1	ALL	101	AW	NONE
378	1980	Q3	1000081665	1000081665	PM	NI	19820823	1	ALL	230	AW	NONE
379	1980	Q3	1000134314	1000134332	PM	UI	19831122	19	ALL	2024	AW	NONE
380	1980	Q3	1000143124	1000143124	PM	UI	19810108	1	ALL	162	AW	NONE
381	1980	Q3	1000143680	1000143683	PM	UI	19800304	4	ALL	766	ET	NONE
382	1980	Q3	1003390563	1003390563	PM	NI	19800112	1	ALL	341	AW	NONE
383	1980	Q3	2021271110	2021271117	PM	UI	19850117	8	ALL	1871	AW	NONE
384	1980	Q3	2021310202	2021310205	PM	NI	19870604	4	ALL	562	AW	NONE
385	1980	Q3	2022164468	2022164469	PM	NI	19800812	2	ALL	517	AW	NONE
386	1980	Q3	2025688065	2025688065	PM	UI	19860301	1	ALL	239	AW	NONE

387	1980	Q3	2028456453	2028456453	PM	NI	19860918	1	ALL	129	AW	NONE
388	1980	Q3	2031403574	2031403574	PM	NI	19860510	1	ALL	131	AW	NONE
389	1980	Q3	2040309238	2040309239	PM	NI	19860131	2	ALL	463	AW	NONE
390	1980	Q3	2042821310	2042821311	PM	UI	19840215	2	ALL	319	AW	NONE
391	1980	Q3	2043919694	2043919701	PM	UI	19870428	8	ALL	2012	AW	NONE
392	1980	Q3	2044369823	2044369831	PM	UI	19841112	9	ALL	1723	AW	NONE
393	1980	Q3	2057089982	2057089984	PM	NI	19871025	3	ALL	338	AW	NONE
394	1980	Q3	2059264000	2059264000	PM	NI	19800305	1	ALL	147	AW	NONE
395	1980	Q3	2501205056	2501205058	PM	UI	19890624	3	ALL	1208	AW	NONE
396	1980	Q3	2501324157	2501324159	PM	UI	19830311	3	ALL	637	BH	NONE
397	1980	Q3	500627230	500627232	RJR	NI	19840710	3	ALL	328	AW	NONE
398	1980	Q3	500865785	500865785	RJR	NI	19801123	1	ALL	331	AW	NONE
399	1980	Q3	500928452	500928458	RJR	UI	19800331	7	ALL	1557	AW	NONE
400	1980	Q3	502455761	502455763	RJR	NI	19820701	3	ALL	467	AW	NONE
401	1980	Q3	503025412	503025448	RJR	UI	19820624	37	ALL	2121	AW	NONE
402	1980	Q3	503251007	503251008	RJR	UI	19840203	2	ALL	264	AW	NONE
403	1980	Q3	503495035	503495037	RJR	NI	19841220	3	ALL	750	AW	NONE
404	1980	Q3	503607592	503607593	RJR	NI	19810914	2	ALL	90	AW	NONE
405	1980	Q3	504221699	504221702	RJR	NE	19850626	4	ALL	847	AW	NONE
406	1980	Q3	505158916	505158924	RJR	UI	19850221	9	ALL	1287	AW	NONE
407	1980	Q3	505482542	505482547	TI	UI	19850716	6	PART	1834		NONE
408	1980	Q3	506205771	506205774	RJR	UI	19850503	4	ALL	763	AW	NONE
409	1980	Q3	506384359	506384360	RJR	NI	19880308	2	ALL	475	AW	NONE
410	1980	Q3	506464546	506464547	RJR	NI	19871029	2	ALL	405	AW	NONE
411	1980	Q3	506559844	506559863	RJR	UI	19880824	20	ALL	1093	ET	NONE
412	1980	Q3	506772178	506772179	RJR	NI	19870823	2	ALL	282	AW	NONE
413	1980	Q3	508074644	508074648	RJR	UI	19821029	5	ALL	1005	AW	NONE
414	1980	Q3	509833606	509833619	RJR	UI	19810000	14	ALL	1677	AW	NONE
415	1980	Q3	510627850	510627850	RJR	NI	19841214	1	ALL	126	AW	NONE
416	1980	Q3	511572487	511572487	RJR	NI	19830318	1	ALL	146	AW	NONE
417	1980	Q3	512214381	512214396	RJR	NI	19880426	16	ALL	880	AW	NONE
418	1980	Q3	514927939	514927969	RJR	UI	19870429	31	ALL	2246	AW	NONE
419	1980	Q3	521006354	521006360	BW	NI	19810107	7	ALL	86	AW	NONE
420	1980	Q3	539005241	539005243	BW	UI	19810800	3	ALL	365	AW	NONE
421	1980	Q3	60007772	60007772	CTR	NI	19870922	1	ALL	242	AW	NONE
422	1980	Q3	60032351	60032351	CTR	UI	19831222	1	ALL	101	BH	AW
423	1980	Q3	620759686	620759713	BW	NI	19880721	28	ALL	121	AW	NONE
424	1980	Q3	621616295	621616295	BW	NI	19800129	1	ALL	53	AW	NONE
425	1980	Q3	655059702	655059764	BW	NI	19811202	63	ALL	187	AW	NONE
426	1980	Q3	660916357	660916361	BW	UI	19820000	5	ALL	778	AW	NONE
427	1980	Q3	660929284	660929284	BW	NI	19830212	1	ALL	101	BH	NONE
428	1980	Q3	670901238	670901243	BW	NI	19890328	6	ALL	719	AW	NONE
429	1980	Q3	677180934	677180939	BW	NI	19830506	6	ALL	508	AW	NONE
430	1980	Q3	680557538	680557539	BW	NI	19800702	2	ALL	227	AW	NONE
431	1980	Q3	682817176	682817179	RJR	UI	19881121	4	ALL	1106	AW	NONE
432	1980	Q3	689307748	689307750	BW	UI	19891129	3	ALL	266	AW	NONE
433	1980	Q3	690126195	690126195	BW	UI	19851011	1	ALL	106	AW	NONE
434	1980	Q3	690838151	690838151	BW	NI	19841011	1	ALL	161	AW	NONE
435	1980	Q3	87117710	87117710	LL	NI	19870809	1	ALL	59	AW	NONE
436	1980	Q3	87789235	87789235	LL	NI	19860310	1	ALL	148	AW	NONE
437	1980	Q3	88197622	88197640A	LL	UI	19851016	20	ALL	1008	AW	NONE
438	1980	Q3	88765193	88765195	LL	NI	19880115	3	ALL	66	AW	NONE
439	1980	Q3	89355946	89355989	BW	UI	19860429	44	ALL	1087	AW	NONE
440	1980	Q3	89434204	89434217	LL	UI	19810410	14	ALL	2851	SM	NONE
441	1980	Q3	ATX02_0151102	ATX02_0151105	ATC	UI	19891104	4	ALL	862	AW	NONE
442	1980	Q3	ATX040895180	ATX040895180	ATC	NI	19860403	1	ALL	103	AW	NONE
443	1980	Q3	B00747447	B00747455	BW	NI	19851004	9	ALL	1387	AW	NONE
444	1980	Q3	TIDN0004239	TIDN0004248	TI	UI	19891025	10	PART	2154	AW	NONE
445	1980	Q3	TIMN0350521	TIMN0350522	TI	UI	19840923	2	ALL	470	AW	NONE
446	1980	Q4	03763064	03763064	LL	UI	19800000	1	ALL	153	SM	NONE
447	1980	Q4	03923293	03923319	LL	NI	19830906	27	ALL	1022	SM	NONE
448	1980	Q4	1000039109	1000039110	RJR	UI	19800710	2	ALL	436	SM	NONE
449	1980	Q4	2001205111	2001205112	PM	NI	19800619	2	ALL	113	SM	NONE
450	1980	Q4	2012602322	2012602322	PM	UI	19821029	1	ALL	155	SM	NONE
451	1980	Q4	2022141986	2022141988	PM	NI	19880525	3	ALL	734	SM	NONE
452	1980	Q4	2022196159	2022196175	PM	NI	19860307	17	ALL	1397	SM	AW
453	1980	Q4	2022826578	2022826586	PM	UI	19840000	9	ALL	87	SM	NONE
454	1980	Q4	2024491993	2024491993	PM	NI	19891005	1	ALL	81	SM	NONE

455	1980	Q4	2024964051	2024964053	PM	NI	19810722	3	ALL	1205	SM	NONE
456	1980	Q4	2028390412	2028390426	RJR	UI	19880226	15	ALL	1433	SM	NONE
457	1980	Q4	2029024810	2029024811	PM	UI	19870126	2	ALL	260	SM	NONE
458	1980	Q4	2037005280	2037005281	PM	UI	19890310	2	ALL	288	SM	AW
459	1980	Q4	2040940418	2040940418	PM	NI	19820416	1	ALL	76	SM	NONE
460	1980	Q4	2047422826	2047422832	PM	UI	19850618	7	ALL	356	SM	NONE
461	1980	Q4	2050702876	2050702880	PM	UI	19890806	5	ALL	179	SM	NONE
462	1980	Q4	2057475579	2057475582	PM	UI	19830212	4	ALL	834	SM	NONE
463	1980	Q4	2501659008	2501659008	PM	UI	19851122	1	ALL	424	SM	NONE
464	1980	Q4	467011629	467011694	BW	NI	19800827	66	PART	1995	SM	NONE
465	1980	Q4	500952817	500952817	RJR	NI	19810730	1	ALL	157	SM	NONE
466	1980	Q4	501254820	501254850	RJR	NI	19840509	31	ALL	1945	SM	NONE
467	1980	Q4	502091860	502091862	RJR	UI	19800325	3	ALL	596	SM	NONE
468	1980	Q4	502131635	502131636	RJR	UI	19800917	2	ALL	388	SM	NONE
469	1980	Q4	503411724	503411726	RJR	NI	19810813	3	ALL	533	SM	AW
470	1980	Q4	503906090	503906090	RJR	NI	19821129	1	ALL	191	SM	NONE
471	1980	Q4	503907156	503907157	RJR	UI	19830525	2	ALL	655	SM	AW
472	1980	Q4	503983481	503983482	RJR	UI	19850700	2	ALL	541	SM	NONE
473	1980	Q4	504180999	504181049	RJR	UI	19800113	51	ALL	431	SM	NONE
474	1980	Q4	504221868	504221871	RJR	NI	19850603	4	ALL	664	SM	NONE
475	1980	Q4	504476087	504476090	RJR	NI	19830822	4	ALL	558	AW	NONE
476	1980	Q4	504845812	504845813	RJR	UI	19801003	2	ALL	1017	SM	NONE
477	1980	Q4	505047412	505047414	RJR	UI	19860303	3	ALL	441	SM	AW
478	1980	Q4	505406114	505406126	RJR	UI	19860000	13	ALL	929	SM	NONE
479	1980	Q4	505704256	505704263	RJR	UI	19870602	8	ALL	556	SM	NONE
480	1980	Q4	505870192	505870194	RJR	UI	19870707	3	ALL	360	SM	NONE
481	1980	Q4	505955188	505955190	RJR	NI	19841009	3	ALL	549	SM	NONE
482	1980	Q4	506674675	506674689	RJR	UI	19880405	15	ALL	907	SM	NONE
483	1980	Q4	507001699	507001708	RJR	NI	19880718	10	PART	2143	SM	NONE
484	1980	Q4	508711412	508711412	PM	NI	19890616	1	ALL	193	SM	NONE
485	1980	Q4	511025094	511025095	RJR	NI	19881222	2	ALL	211	SM	AW
486	1980	Q4	513232640	513232643	RJR	NI	19880330	4	ALL	216	SM	NONE
487	1980	Q4	514421368	514421369	RJR	NE	19860714	2	ALL	485	SM	NONE
488	1980	Q4	521018677	521018677	RJR	NI	19810213	1	ALL	167		NONE
489	1980	Q4	576100892	576100897	BW	NI	19890705	6	ALL	189	SM	NONE
490	1980	Q4	620752148	620752148	BW	NI	19831121	1	ALL	59	SM	NONE
491	1980	Q4	620848485	620848490	BW	NI	19890703	6	ALL	1674	SM	NONE
492	1980	Q4	634341854	634341855	BW	NI	19890816	2	ALL	88	SM	NONE
493	1980	Q4	634341957	634341960	BW	UI	19890809	4	ALL	394	SM	NONE
494	1980	Q4	635608538	635608541	BW	NI	19870219	4	ALL	741	SM	NONE
495	1980	Q4	650565527	650565527	BW	NI	19840328	1	ALL	92	SM	NONE
496	1980	Q4	670224756	670224756	BW	NI	19840120	1	ALL	121	SM	NONE
497	1980	Q4	670621869	670622002	BW	NI	19820215	134	PART	1510	SM	NONE
498	1980	Q4	680593531	680593533	BW	NI	19811102	3	ALL	987	SM	NONE
499	1980	Q4	690115044	690115050	BW	UI	19850515	7	ALL	2092	SM	NONE
500	1980	Q4	81187645	81187646	ATC	UI	19800000	2	ALL	282	SM	NONE
501	1980	Q4	85703560	85703567	LL	UI	19800900	8	ALL	3149	SM	NONE
502	1980	Q4	87394887	87394888	LL	NI	19880819	2	ALL	129	SM	NONE
503	1980	Q4	87651678	87651678	LL	UI	19890126	1	ALL	168	SM	NONE
504	1980	Q4	88762722	88762723	LL	NI	19841116	2	ALL	83	SM	AW
505	1980	Q4	88984366	88984369	LL	UI	19881119	4	ALL	1037	SM	AW
506	1980	Q4	91719524	91719529	LL	UI	19850000	6	ALL	831	SM	NONE
507	1980	Q4	ATX040221325	ATX040221325	ATC	NI	19860220	1	ALL	142	SM	NONE
508	1980	Q4	ATX040440614	ATX040440614	ATC	NI	19800304	1	ALL	181		AW
509	1980	Q4	MNAT00690879	MNAT00690880	ATC	NI	19890413	2	ALL	239	SM	AW
510	1980	Q4	MNAT00805749	MNAT00805750	ATC	NI	19810000	2	ALL	571	SM	AW
511	1980	Q4	TIDN0018279	TIDN0018279	TI	NI	19880123	1	ALL	116	SM	NONE
512	1980	Q4	TIFL0525271	TIFL0525271	TI	UI	19800718	1	ALL	105	SM	NONE
513	1980	Q4	TIFL0525572	TIFL0525573	TI	NI	19820000	2	ALL	273	SM	NONE
514	1980	Q4	TIMN0256410	TIMN0256410	BW	NI	19830112	1	ALL	129	SM	NONE
515	1980	Q4	TIMN0342280	TIMN0342280	TI	UI	19840000	1	ALL	60	SM	NONE
516	1980	Q4	TIMN0350315	TIMN0350316	TI	UI	19850000	2	ALL	525	SM	NONE
517	1990	Q1	2022913619	2022913620	PM	UI	19930607	2	ALL	257	AW	NONE
518	1990	Q1	2028396112	2028396116	PM	UI	19940000	5	ALL	951	AW	NONE
519	1990	Q1	2029152467	2029152467	PM	UI	19940000	1	ALL	95	AW	NONE
520	1990	Q1	2029182650	2029182651	PM	UI	19950503	2	ALL	528	AW	NONE
521	1990	Q1	2031644014	2031644014	PM	UI	19950626	1	ALL	147	AW	NONE
522	1990	Q1	2043213565	2043213565	PM	NI	19920630	1	ALL	49	AW	NONE

```
523 1990   Q1  2044938392A       2044938393        PM  NI  19911211 2   ALL  462  AW  NONE
524 1990   Q1  2046105036        2046105037        PM  UI  19940700 2   ALL  655  AW  NONE
525 1990   Q1  2047319403        2047319403        PM  UI  19940228 1   ALL  178  AW  NONE
526 1990   Q1  2047715163        2047715164        PM  UI  19940301 2   ALL  245  AW  NONE
527 1990   Q1  2048815665        2048815724        PM  UI  19940000 60  ALL  2086 AW  NONE
528 1990   Q1  2057068691C       2057068691C       PM  NI  19960204 1   ALL  161  AW  NONE
529 1990   Q1  2058097498        2058097498        PM  UI  19950430 1   ALL  196  AW  NONE
530 1990   Q1  2060549643        2060549644        PM  UI  19970304 2   ALL  501  AW  NONE
531 1990   Q1  2061512715        2061512716        PM  NI  19940624 2   ALL  289  AW  NONE
532 1990   Q1  2062996579        2062996579        PM  NI  19960408 1   ALL  68   AW  NONE
533 1990   Q1  2063593224        2063593224        PM  UI  19981000 1   ALL  373  AW  NONE
534 1990   Q1  2063593895        2063593895        PM  NI  19980318 1   ALL  98   AW  NONE
535 1990   Q1  2063608777        2063608778        PM  UI  19980206 2   ALL  648  AW  NONE
536 1990   Q1  2063621826        2063621835        PM  NI  19980128 10  ALL  489  AW  NONE
537 1990   Q1  2063654332        2063654333        PM  UI  19950000 2   ALL  248  AW  NONE
538 1990   Q1  2063658624        2063658624        PM  UI  19971222 1   ALL  177  AW  NONE
539 1990   Q1  2501196964        2501196986        PM  UI  19930429 23  ALL  2183 AW  NONE
540 1990   Q1  2501241108        2501241122        PM  NI  19901204 15  PART 1540 ET  NONE
541 1990   Q1  2501421504        2501421508        PM  UI  19920221 5   ALL  1105 AW  NONE
542 1990   Q1  450260020         450260020         BW  UI  19961030 1   ALL  78   AW  NONE
543 1990   Q1  507417689         5407417690        RJR NI  19900323 2   ALL  182  AW  NONE
544 1990   Q1  507543286         507543339         RJR UI  19910100 54  ALL  1915 AW  NONE
545 1990   Q1  507907303         507907306         RJR UI  19910912 4   ALL  1357 AW  NONE
546 1990   Q1  508676946         508676959         RJR UI  19901230 14  ALL  2162 AW  NONE
547 1990   Q1  508688676         508688688         RJR UI  19930717 13  ALL  1498 AW  NONE
548 1990   Q1  508693878         508693897         RJR UI  19911115 20  ALL  2116 AW  NONE
549 1990   Q1  508832248         508832248         RJR UI  19931203 1   ALL  172  AW  NONE
550 1990   Q1  509475548         509475548         RJR NI  19910306 1   ALL  141  AW  NONE
551 1990   Q1  511466579         511466579         RJR NI  19900626 1   ALL  317  AW  NONE
552 1990   Q1  512117418         512117424         RJR UI  19930125 7   ALL  494  AW  NONE
553 1990   Q1  512694192         512694194         RJR UI  19940124 3   ALL  475  AW  NONE
554 1990   Q1  513201516         513201518         RJR UI  19931124 3   ALL  709  AW  NONE
555 1990   Q1  513487419         513487427         RJR NI  19920914 9   ALL  2167 AW  NONE
556 1990   Q1  514854756         514854785         RJR UI  19950000 30  PART 2830 AW  NONE
557 1990   Q1  516807622         516807622         RJR NI  19961025 1   ALL  190  AW  NONE
558 1990   Q1  516927839         516927839         RJR NI  19971115 1   ALL  58   AW  NONE
559 1990   Q1  517127602         517127603         RJR NI  19930713 2   ALL  100  AW  NONE
560 1990   Q1  517135523         517135534         RJR UI  19960000 12  PART 2100 AW  NONE
561 1990   Q1  517904260         517904308         RJR UI  19950905 49  ALL  1016 AW  NONE
562 1990   Q1  518035912         518035912         RJR UI  19950000 1   ALL  196  AW  NONE
563 1990   Q1  518088189         518088193         RJR UI  19951206 5   ALL  1976 AW  NONE
564 1990   Q1  518089551         518089551         RJR NI  19960902 1   ALL  190  AW  NONE
565 1990   Q1  518204342         518204347         RJR UI  19960000 6   ALL  249  AW  NONE
566 1990   Q1  518207349         518207349         RJR UI  19940916 1   ALL  60   AW  NONE
567 1990   Q1  518477856         518477857         RJR NI  19980520 2   ALL  436  AW  NONE
568 1990   Q1  518584539         518584574         RJR UI  19960000 36  PART 2136 AW  NONE
569 1990   Q1  518680270         518680272         RJR NI  19970103 3   ALL  113  AW  NONE
570 1990   Q1  518698215         518698215         RJR UI  19980225 1   ALL  115  ET  AW
571 1990   Q1  525428160         525428162         RJR UI  19980426 3   ALL  1003 AW  NONE
572 1990   Q1  618000535         618000535         BW  NI  19930605 1   ALL  213  AW  NONE
573 1990   Q1  88523806          88523806          LL  UI  19930700 1   ALL  341  AW  NONE
574 1990   Q1  94551590          94551590          LL  UI  19970219 1   ALL  153  AW  NONE
575 1990   Q1  94561260          94561260          LL  UI  19950605 1   ALL  93   AW  NONE
576 1990   Q1  ATX040310269      ATX040310270      ATC NI  19970712 2   ALL  409  AW  NONE
577 1990   Q1  TIFL0047260       TIFL0047260       TI  NI  19950502 1   ALL  284  AW  NONE
578 1990   Q2  2020256223        2020256223        PM  NI  19911024 1   ALL  75   AW  NONE
579 1990   Q2  2023004370        2023004370        PM  NI  19920710 1   ALL  184  AW  NONE
580 1990   Q2  2023141738        2023141773        PM  UI  19920301 36  ALL  975  AW  NONE
581 1990   Q2  2023894116        2023894118        PM  UI  19940503 3   ALL  66   AW  NONE
582 1990   Q2  2024708863        2024708865        PM  UI  19930313 3   ALL  1091 AW  NONE
583 1990   Q2  2026363725        2026363749        PM  UI  19920614 25  ALL  2052 AW  NONE
584 1990   Q2  2030033505        2030033505        PM  UI  19910104 1   ALL  134  AW  NONE
585 1990   Q2  2041555731        2041555732        PM  NI  19941221 2   ALL  70   AW  NONE
586 1990   Q2  2044901862        2044901888        PM  UI  19911231 27  ALL  1033 AW  NONE
587 1990   Q2  2047028764C       2047028764C       PM  NI  19951026 1   ALL  139  AW  NONE
588 1990   Q2  2047035955        2047035957        PM  UI  19950602 3   ALL  689  AW  NONE
589 1990   Q2  2047036052        2047036052        PM  UI  19940327 1   ALL  206  AW  NONE
590 1990   Q2  2047073435        2047073435        PM  UI  19960220 1   ALL  97   AW  NONE
```

```
591 1990   Q2  2047263020        2047263020        PM  NI  19930901 1   ALL  1593 AW  NONE
592 1990   Q2  2047875511A       2047875511A       PM  NI  19940204 1   ALL  159  AW  NONE
593 1990   Q2  2048489771        2048489774        PM  UI  19900421 4   ALL  457  AW  NONE
594 1990   Q2  2050488068        2050488069        PM  NI  19910326 2   ALL  378  AW  NONE
595 1990   Q2  2050803579        2050803579        PM  UI  19910326 1   ALL  262  AW  NONE
596 1990   Q2  2050852733        2050852746        PM  UI  19920724 14  ALL  898  AW  NONE
597 1990   Q2  2050931138        2050931141        PM  UI  19950704 4   ALL  312  ET  NONE
598 1990   Q2  2051334885        2051334912        PM  UI  19920618 28  PART 2025 ET  NONE
599 1990   Q2  2054515613        2054515614        PM  UI  19910414 2   ALL  154  AW  NONE
600 1990   Q2  2057063728        20570633738       PM  UI  19951223 11  ALL  2120 AW  NONE
601 1990   Q2  2057992469        2057992469        PM  NI  19951926 1   ALL  112  ET  NONE
602 1990   Q2  2060544471        2060544475        PM  UI  19970505 5   ALL  1340 AW  NONE
603 1990   Q2  2060578937        2060578947        PM  UI  19980318 11  ALL  2020 AW  NONE
604 1990   Q2  2062098111        2062098112        PM  NI  19900822 2   ALL  144  AW  NONE
605 1990   Q2  2062896705        2062896705        PM  UI  19961028 1   ALL  240  ET  NONE
606 1990   Q2  2063600621        2063600630        PM  UI  19970000 10  PART 2076 AW  NONE
607 1990   Q2  2063610088        2063610088        PM  UI  19970605 1   ALL  78   AW  NONE
608 1990   Q2  2063653956        2063653957        PM  UI  19970319 2   ALL  287  AW  NONE
609 1990   Q2  2065526277        2065526277        PM  NI  19990402 1   ALL  125  BH  AW
610 1990   Q2  2070315456        2070315457        PM  UI  19980700 2   ALL  448  ET  NONE
611 1990   Q2  2072497196        2072497196        PM  UI  19970711 1   ALL  233  BH  NONE
612 1990   Q2  2074516331        2074516331        PM  UI  19990506 1   ALL  332  AW  NONE
613 1990   Q2  344000805         344000810         BW  UI  19950802 6   ALL  1129 AW  NONE
614 1990   Q2  509444694         509444696         RJR UI  19911025 3   ALL  645  BH  NONE
615 1990   Q2  509495462         509495463         RJR NI  19921109 2   ALL  235  AW  NONE
616 1990   Q2  509910061         509910074         RJR UI  19921121 14  ALL  867  AW  NONE
617 1990   Q2  512542259         512542261         RJR UI  19930421 3   ALL  920  AW  NONE
618 1990   Q2  512779350         512779351         RJR UI  19910913 2   ALL  153  AW  NONE
619 1990   Q2  513157996         513158022         RJR UI  19900204 27  ALL  1265 AW  NONE
620 1990   Q2  514853229         514853229         RJR NI  19961126 1   ALL  149  AW  NONE
621 1990   Q2  515923306         515923306         RJR UI  19920302 1   ALL  125  AW  NONE
622 1990   Q2  516924574         516924577         RJR UI  19980609 4   ALL  1485 AW  NONE
623 1990   Q2  516924782         516924783         RJR UI  19980611 2   ALL  137  AW  NONE
624 1990   Q2  516924787         516924789         RJR UI  19980424 3   ALL  380  AW  NONE
625 1990   Q2  516925163         516925163         RJR NI  19980601 1   ALL  91   AW  NONE
626 1990   Q2  518050238         518050241         RJR NI  19980302 4   ALL  137  AW  NONE
627 1990   Q2  518805178         518805178         RJR NI  19970303 1   ALL  569  AW  NONE
628 1990   Q2  518830555         518830556         RJR NI  19960216 2   ALL  146  AW  NONE
629 1990   Q2  621967092         621967092         BW  UI  19960719 1   ALL  255  AW  NONE
630 1990   Q2  70005392          70005408          CTR UI  19981210 17  PART 2553 AW  NONE
631 1990   Q2  86000341          86000343          CTR UI  19910703 3   ALL  675  AW  NONE
632 1990   Q2  88355846          88355856          LL  NI  19951026 11  ALL  1610 AW  NONE
633 1990   Q2  89237599          89237600          LL  NI  19931209 2   ALL  712  AW  NONE
634 1990   Q2  94375156          94375156          LL  NI  19960528 1   ALL  220  AW  NONE
635 1990   Q2  94543716          94543716          LL  NI  19980119 1   ALL  99   AW  NONE
636 1990   Q2  ATX040246229      ATX040246251      ATC UI  19940900 23  PART 3460 AW  NONE
637 1990   Q2  MNAT00751820      MNAT00751820      ATC UI  19900000 1   ALL  667  AW  NONE
638 1990   Q2  TIOK0011637       TIOK0011637       TI  NI  19910109 1   ALL  144  AW  NONE
639 1990   Q3  2020135926        2020135927        PM  UI  19911101 2   ALL  367  AW  NONE
640 1990   Q3  2021310739        2021310740        PM  UI  19921015 2   ALL  308  AW  NONE
641 1990   Q3  2023958318        2023958325        PM  UI  19920706 8   ALL  1116 AW  NONE
642 1990   Q3  2026387021        2026387021        PM  NI  19920324 1   ALL  77   AW  NONE
643 1990   Q3  2028540117        2028540117        PM  NI  19921026 1   ALL  112  AW  NONE
644 1990   Q3  2029239018        2029239057        PM  UI  19950000 40  ALL  1306 AW  NONE
645 1990   Q3  2040583534        2040583558        PM  UI  19950415 25  PART 2355 ET  NONE
646 1990   Q3  2041020892        2041020892        PM  UI  19930000 1   ALL  167  AW  NONE
647 1990   Q3  2041755683        2041755732        PM  UI  19930000 50  ALL  1488 AW  NONE
648 1990   Q3  2045418043        2045418044        PM  NI  19930415 2   ALL  485  AW  NONE
649 1990   Q3  2046553039        2046553039        PM  UI  19950526 1   ALL  232  AW  NONE
650 1990   Q3  2054935246        2054935247        PM  UI  19950615 2   ALL  402  AW  NONE
651 1990   Q3  2057068992B       2057068992B       PM  NI  19951031 1   ALL  66   AW  NONE
652 1990   Q3  2060171077        2060171079        PM  UI  19930205 3   ALL  395  BH  NONE
653 1990   Q3  2060532735        2060532735        PM  UI  19971023 1   ALL  226  AW  NONE
654 1990   Q3  2060547122        2060547122        PM  NI  19970802 1   ALL  51   AW  NONE
655 1990   Q3  2060548894        2060548894        PM  NI  19970324 1   ALL  139  AW  NONE
656 1990   Q3  2060553932        2060553933        PM  UI  19960429 2   ALL  622  AW  NONE
657 1990   Q3  2062101019        2062101024        PM  UI  19901220 6   ALL  1099 AW  NONE
658 1990   Q3  2062169263        2062169263        PM  NI  19960604 1   ALL  180  AW  NONE
```

659	1990	Q3	2062403055	2062403056	PM	UI	19951227	2	ALL	828	AW	NONE
660	1990	Q3	2063607687	2063607687	PM	NI	19980209	1	ALL	111	AW	NONE
661	1990	Q3	2063608784	2063608786	PM	NI	19971024	3	ALL	881	AW	NONE
662	1990	Q3	2078641670	2078641672	PM	UI	19970519	3	ALL	967	BH	NONE
663	1990	Q3	2501296750	2501296753	PM	NI	19931026	4	ALL	94	AW	NONE
664	1990	Q3	498101000	498101002	BW	UI	19950704	3	ALL	820	BH	NONE
665	1990	Q3	507402666	507402667	RJR	UI	19900000	2	ALL	320	BH	NONE
666	1990	Q3	507625024	507625025	RJR	NI	19901222	2	ALL	328	AW	NONE
667	1990	Q3	508739487	508739497	RJR	UI	19910718	11	ALL	1178	AW	NONE
668	1990	Q3	508950387	508950396	RJR	UI	19921020	10	ALL	1917	AW	NONE
669	1990	Q3	509226990	509926991	RJR	NI	19911207	2	ALL	600	BH	NONE
670	1990	Q3	509735577	509735577	RJR	UI	19930526	1	ALL	102	AW	NONE
671	1990	Q3	509744465	509744467	RJR	NI	19910120	3	ALL	766	BH	NONE
672	1990	Q3	509785500	509785504	RJR	UI	19910628	5	ALL	924	AW	NONE
673	1990	Q3	511920268	511920269	RJR	UI	19910405	2	ALL	488	AW	NONE
674	1990	Q3	515089782	515089785	RJR	UI	19931027	4	ALL	530	BH	NONE
675	1990	Q3	515502497	515502497	RJR	UI	19931222	1	ALL	153	AW	NONE
676	1990	Q3	516801931	516801932	RJR	NI	19961112	2	ALL	472	AW	NONE
677	1990	Q3	517106095	517107083	RJR	UI	19970919	989	PART	8789	AW	NONE
678	1990	Q3	517262932	517262976	RJR	UI	19920810	45	PART	2065	AW	NONE
679	1990	Q3	518026662	518026662	RJR	UI	19971205	1	ALL	173	AW	NONE
680	1990	Q3	518653607	518653608	RJR	NI	19961101	2	ALL	288	AW	NONE
681	1990	Q3	518654464	518654464	RJR	UI	19950701	1	ALL	114	AW	NONE
682	1990	Q3	60016287A	60016287A	CTR	UI	19950628	1	ALL	578	BH	NONE
683	1990	Q3	87603677	87603677	LL	UI	19930211	1	ALL	295	AW	NONE
684	1990	Q3	89113004	89113005	LL	UI	19930406	2	ALL	161	BH	NONE
685	1990	Q3	92043638	92043645	LL	UI	19920100	8	ALL	368	BH	NONE
686	1990	Q3	95570940	95570940	LL	NI	19960325	1	ALL	141	BH	NONE
687	1990	Q3	ATX02_0075226	ATX02_0075229	ATC	UI	19941119	4	ALL	1183	SM	NONE
688	1990	Q3	ATX02_0075607	ATX02_0075608	ATC	UI	19941008	2	ALL	433	BH	NONE
689	1990	Q3	B00717879	B00717890	BW	UI	19941003	12	ALL	497	AW	NONE
690	1990	Q3	TIIL0004756	TIIL0004761	TI	UI	19961218	6	ALL	105	AW	NONE
691	1990	Q3	TIILBC0006256	TIILBC0006256	TI	NI	19970909	1	ALL	77	AW	NONE
692	1990	Q3	TIILBC0021181	TIILBC0021181	TI	NI	19900115	1	ALL	136	AW	NONE
693	1990	Q3	TIILBC0024578	TIILBC0024578	TI	NI	19900801	1	ALL	4	AW	NONE
694	1990	Q3	TIMN0150365	TIMN0150366	TI	NI	19900209	2	ALL	498	AW	NONE
695	1990	Q3	TIMN0219994	TIMN0220004	TI	UI	19910730	11	PART	2037	AW	NONE
696	1990	Q3	TIMN0378880	TIMN0378881	TI	UI	19920806	2	ALL	281	AW	NONE
697	1990	Q3	TIMN0427229	TIMN0427229	TI	UI	19930514	1	ALL	254	AW	NONE
698	1990	Q3	TIMS0008642	TIMS0008642	TI	NI	19920215	1	ALL	168	AW	NONE
699	1990	Q3	TINY0003025	TINY0003042	TI	UI	19980409	18	ALL	2079	AW	NONE
700	1990	Q4	11320767	11320767	CTR	NI	19921017	1	ALL	98	SM	NONE
701	1990	Q4	2023295614	2023295616	PM	NI	19940225	3	ALL	828	SM	NONE
702	1990	Q4	2023637119	2023637119	PM	UI	19930226	1	ALL	180	SM	NONE
703	1990	Q4	2023769347	2023769347	PM	UI	19931013	1	ALL	282	SM	NONE
704	1990	Q4	2024169969	2024169971	PM	UI	19930302	3	ALL	841	SM	AW
705	1990	Q4	2025362838	2025362838	PM	NI	19900517	1	ALL	232	SM	NONE
706	1990	Q4	2028354687	2028354729	PM	UI	19920000	43	PART	2126	SM	NONE
707	1990	Q4	2028389199	2028389221	PM	UI	19930906	23	PART	2109	SM	NONE
708	1990	Q4	2031579110	2031579111	PM	UI	19941108	2	ALL	648	SM	NONE
709	1990	Q4	2041839296	2041839296	PM	NI	19910807	1	ALL	75	SM	NONE
710	1990	Q4	2043025719	2043025723	PM	UI	19940731	5	ALL	218	SM	NONE
711	1990	Q4	2043177229	2043177229	PM	NI	19920610	1	ALL	101	SM	NONE
712	1990	Q4	2044761311	2044761314	PM	NI	19901219	4	ALL	1069	SM	AW
713	1990	Q4	2045857649	2045857833	PM	UI	19950327	185	PART	1767	SM	AW
714	1990	Q4	2047724038	2047724041	PM	NI	19951215	4	ALL	1186	SM	NONE
715	1990	Q4	2048299420	2048299421	PM	NI	19901219	2	ALL	521	SM	NONE
716	1990	Q4	2051334401	2051334403	PM	NI	19920602	3	ALL	882	SM	NONE
717	1990	Q4	2054935252	2054935252	PM	UI	19950804	1	ALL	60	SM	NONE
718	1990	Q4	2055151869	2055151870	PM	UI	19940208	2	ALL	962	SM	NONE
719	1990	Q4	2058105033	2058105033	PM	NI	19930118	1	ALL	97	SM	NONE
720	1990	Q4	2060547154	2060547154	PM	NI	19970527	1	ALL	58	SM	NONE
721	1990	Q4	2060552049	2060552049	PM	NI	19980727	1	ALL	55	SM	NONE
722	1990	Q4	2060569963	2060569964	PM	UI	19980720	2	ALL	372	SM	NONE
723	1990	Q4	2062904484	2062904486	PM	UI	19960523	3	ALL	629	SM	NONE
724	1990	Q4	2063018192	2063018202	PM	UI	19980417	11	PART	2228	SM	NONE
725	1990	Q4	2063593226	2063593235	PM	UI	19981000	10	PART	2211	SM	NONE
726	1990	Q4	2063595051	2063595072	PM	UI	19980900	22	PART	2090	SM	AW

727	1990	Q4	2501383695	2501383701	PM	UI	19920000	7	ALL	682	SM	NONE
728	1990	Q4	505301582	505301583	BW	NI	19920302	2	ALL	181	SM	AW
729	1990	Q4	507593679	507593680	RJR	UI	19910326	2	ALL	614	SM	NONE
730	1990	Q4	507961859	507961862	RJR	UI	19900110	4	ALL	1077	SM	AW
731	1990	Q4	509048419	509048422	RJR	UI	19920724	4	ALL	921	SM	NONE
732	1990	Q4	510325833	510325845	RJR	UI	19920104	13	ALL	363	SM	NONE
733	1990	Q4	511474980	511474980	RJR	NI	19910501	1	ALL	157	SM	NONE
734	1990	Q4	511989625	511989671	RJR	UI	19930807	47	ALL	1533	SM	NONE
735	1990	Q4	516925014	516925017	RJR	UI	19980420	4	ALL	134	SM	NONE
736	1990	Q4	517095434	517095435	RJR	NI	19960911	2	ALL	499	SM	NONE
737	1990	Q4	517974873	517974873	RJR	UI	19970215	1	ALL	200	SM	NONE
738	1990	Q4	518025044	518025044	RJR	UI	19970820	1	ALL	156	SM	NONE
739	1990	Q4	518049113	518049113	RJR	UI	19941014	1	ALL	286	SM	NONE
740	1990	Q4	518199214	518199214	RJR	UI	19960822	1	ALL	69	SM	AW
741	1990	Q4	518643051	518643051	RJR	UI	19931123	1	ALL	161	SM	NONE
742	1990	Q4	591003313	591003315	BW	UI	19931213	3	ALL	190	SM	NONE
743	1990	Q4	605130239	605130250	BW	NI	19951207	12	PART	2188	SM	NONE
744	1990	Q4	620102421	620102422	BW	NI	19900604	2	ALL	571	SM	NONE
745	1990	Q4	620944586	620944676	BW	UI	19900000	91	ALL	1663	SM	NONE
746	1990	Q4	621969343	621969343	BW	UI	19971024	1	ALL	162	SM	NONE
747	1990	Q4	88026355	88026359	LL	UI	19950503	5	ALL	1636	SM	NONE
748	1990	Q4	89450916	89450931	LL	UI	19950509	16	ALL	2500	SM	NONE
749	1990	Q4	89946952	89946954	LL	NI	19910909	3	ALL	590	SM	AW
750	1990	Q4	91782852	91782852	LL	UI	19920316	1	ALL	78	SM	NONE
751	1990	Q4	93777832	93777832	LL	UI	19940428	1	ALL	492	SM	NONE
752	1990	Q4	94412382	94412392	LL	NI	19960619	11	ALL	107	SM	NONE
753	1990	Q4	94530509	94530509	LL	UI	19951102	1	ALL	83	SM	NONE
754	1990	Q4	ATX040055394	ATX040055399	ATC	UI	19910308	6	ALL	574	SM	AW
755	1990	Q4	ATX040880021	ATX040880023	ATC	NI	19920716	3	ALL	309	SM	AW
756	1990	Q4	B01281539	B01281545	BW	UI	19910924	7	ALL	896	SM	NONE
757	1990	Q4	MNAT00380663	MNAT00380664	ATC	UI	19930519	2	ALL	471	SM	AW
758	1990	Q4	MNAT00889212	MNAT00889213	ATC	NI	19900709	2	ALL	523	SM	NONE
759	1990	Q4	TICT0005854	TICT0005855	TI	UI	19961001	2	ALL	481	SM	NONE
760	1990	Q4	TICT0006113	TICT0006113	TI	UI	19951011	1	ALL	180	SM	NONE
761	19XX	Q1	TIMN0018869	TIMN0018921	TI	UI	19000000	53	PART	2194	AW	NONE
762	19XX	Q1	TIMN0065827	TIMN0065829	TI	UI	19000000	3	ALL	224	AW	NONE
763	19XX	Q1	TIMN0067774	TIMN0067776	TI	UI	19000000	3	ALL	642	AW	NONE
764	19XX	Q1	TIMN0069106	TIMN0069108	TI	UI	19000000	3	ALL	454	AW	NONE
765	19XX	Q1	TIMN0119282	TIMN0119319	TI	UI	19000000	38	ALL	2102	AW	NONE
766	19XX	Q2	TIMN0063797	TIMN0063798	TI	UI	19000000	2	ALL	247	AW	NONE
767	19XX	Q2	TIMN0068393	TIMN0068393	TI	UI	19000000	1	ALL	182	AW	NONE
768	19XX	Q2	TIMN0071692	TIMN0071693	TI	UI	19000000	2	ALL	417	AW	NONE
769	19XX	Q2	TIMN0100202	TIMN0100204	TI	UI	19000000	3	ALL	1207	ET	NONE
770	19XX	Q2	TIMN0121859	TIMN0121878	TI	UI	19000000	20	ALL	2326	AW	NONE
771	19XX	Q3	501007677	501007693	BW	UI	19000000	17	PART	2273	ET	AW
772	19XX	Q3	TIFL0511054	TIFL0511054	TI	UI	19000000	1	ALL	682	HW	NONE
773	19XX	Q3	TIMN0064660	TIMN0064660	TI	UI	19000000	1	ALL	79	HW	NONE
774	19XX	Q3	TIMN0073250	TIMN0073258	TI	UI	19000000	9	ALL	2065	HW	NONE
775	19XX	Q3	TIMN0121917	TIMN0121928	TI	UI	19000000	12	ALL	2254	HW	NONE
776	19XX	Q4	TIMN0014092	TIMN0014096	TI	UI	19000000	5	ALL	513	SM	NONE
777	19XX	Q4	TIMN0016198	TIMN0016209	TI	UI	19000000	12	ALL	2033	SM	NONE
778	19XX	Q4	TIMN0065358	TIMN0065359	TI	UI	19000000	2	ALL	553	SM	NONE
779	19XX	Q4	TIMN0074731	TIMN0074732	TI	UI	19000000	2	ALL	915	SM	NONE
780	19XX	Q4	TIMN0088964	TIMN0089043	TI	UI	19000000	80	PART	2142	SM	NONE
781	BLILEY	Q1	00002640	00002640	TI	NI	19840207	1	ALL	93	AW	NONE
782	BLILEY	Q1	1005112517	1005112518	PM	NI	19680918	2	ALL	225	AW	NONE
783	BLILEY	Q1	500534116	500534118	RJR	NI	19820317	3	ALL	855	AW	NONE
784	BLILEY	Q1	500872346	500872346	RJR	NI	19790830	1	ALL	270	AW	NONE
785	BLILEY	Q1	501555286	501555286	RJR	NI	19711209	1	ALL	91	AW	NONE
786	BLILEY	Q1	501869639	501869657	RJR	UI	19690901	19	ALL	2038	SM	NONE
787	BLILEY	Q1	515623581	515623583	RJR	NI	19910531	3	ALL	727	AW	NONE
788	BLILEY	Q2	01335792	01335793	LL	NI	19800403	2	ALL	483	AW	NONE
789	BLILEY	Q2	1000134454	1000134457	PM	NI	19830720	4	ALL	593	AW	NONE
790	BLILEY	Q2	1005149330	1005149330	PM	UI	19621011	1	ALL	318	AW	NONE
791	BLILEY	Q2	2023246959	2023246959	PM	NI	19911126	1	ALL	84	AW	NONE
792	BLILEY	Q2	501624207	501624207	RJR	NI	19790301	1	ALL	350	AW	NONE
793	BLILEY	Q2	502850617	502850627	RJR	NI	19780822	11	ALL	1907	SM	NONE
794	BLILEY	Q2	507731404	507731405	RJR	NI	19900831	2	ALL	345	AW	NONE

```
795 BLILEY Q3  1000320737         1000320737         PM  NI  19700624 1   ALL  237  AW  NONE
796 BLILEY Q3  22986              22986              PM  NI  19850521 1   ALL  53   AW  NONE
797 BLILEY Q3  501626400          501626400          RJR NI  19810114 1   ALL  219  AW  NONE
798 BLILEY Q3  502851703          502851706          RJR NI  19680523 4   ALL  1206 AW  NONE
799 BLILEY Q3  503647461          503647464          RJR NI  19770509 4   ALL  1274 AW  NONE
800 BLILEY Q3  504220960          504220961          RJR NI  19841120 2   ALL  197  AW  NONE
801 BLILEY Q3  509397100          509397117          RJR UI  19910530 18  ALL  2228 AW  NONE
802 BLILEY Q4  2023264674         2023264674         PM  NI  19860324 1   ALL  116  SM  NONE
803 BLILEY Q4  20485              20485              TI  NI  19810630 1   ALL  91   SM  NONE
804 BLILEY Q4  20876              20876              PM  NI  19760116 1   ALL  119  SM  NONE
805 BLILEY Q4  26425              26425              PM  NI  19890913 1   ALL  200  SM  NONE
806 BLILEY Q4  500016109          500016109          RJR NI  19760429 1   ALL  285  SM  NONE
807 BLILEY Q4  500875427          500875428          RJR NI  19810819 2   ALL  726  SM  NONE
808 BLILEY Q4  504350137          504350142          RJR UI  19750000 6   ALL  2195 SM  NONE
```

B.4 Metadata for Supplemental Sample Documents

The following data[2] were derived by examining the XML archives of the completed Supplemental Sample.

```
NUM STRAT  ITR BATES-START       BATES-END         SRC CLS DAT      PGS AMT  TKN  CDR VER
-------------------------------------------------------------------------------------------
1   1950   S1  502218497         502218497         RJR UE  19580507 1   ALL  128  SM  NONE
2   1950   S1  502598657         502598657         RJR UE  19530919 1   ALL  198  SM  NONE
3   1950   S1  620082520         620082520         BW  NE  19590429 1   ALL  61   SM  NONE
4   1950   S1  MNAT00605940      MNAT00605940      ATC NE  19580811 1   ALL  196  SM  NONE
5   1960   S1  1005098873        1005098879        PM  UE  19690925 7   ALL  1609 SM  NONE
6   1960   S1  2026438545        2026438550        PM  NE  19671211 6   ALL  1042 SM  NONE
7   1960   S1  2058501072        2058501072        PM  UE  19680908 1   ALL  162  SM  NONE
8   1960   S1  2061001687        2061001687        PM  UE  19600306 1   ALL  97   SM  NONE
9   1960   S1  500599608         500599609         RJR NE  19690218 2   ALL  412  SM  NONE
10  1960   S1  501934402         501934402         RJR NE  19640625 1   ALL  53   SM  NONE
11  1970   S1  00498006          00498007          LL  NE  19780308 2   ALL  80   AW  NONE
12  1970   S1  01328445          01328446          PM  NE  19730507 2   ALL  176  AW  NONE
13  1970   S1  03602371          03602372          LL  UE  19760200 2   ALL  493  AW  NONE
14  1970   S1  1003290099        1003290099        PM  NE  19790116 1   ALL  200  AW  NONE
15  1970   S1  1003290546        1003290546        PM  NE  19740606 1   ALL  92   AW  NONE
16  1970   S1  1003639567        1003639567        PM  NE  19751028 1   ALL  149  AW  NONE
17  1970   S1  10398864          10398864          CTR UE  19740104 1   ALL  204  AW  NONE
18  1970   S1  11277820          11277821          CTR UE  19790824 2   ALL  770  AW  NONE
19  1970   S1  2000513684        200513684         PM  NE  19790925 1   ALL  70   AW  NONE
20  1970   S1  501513407         501513407         RJR NE  19760517 1   ALL  114  AW  NONE
21  1970   S1  660003890         660003891         BW  NE  19760116 2   ALL  353  AW  NONE
22  1970   S1  660045117         660045124         ATC UE  19770818 8   ALL  786  AW  NONE
23  1970   S1  675018704         675018705         BW  UE  19780517 2   ALL  546  AW  NONE
24  1970   S1  676152358         676152358         BW  NE  19710607 1   ALL  58   AW  NONE
25  1970   S1  680003504         680003509         BW  UE  19740402 6   PART 2199 AW  NONE
26  1970   S1  775008770         775008770         BW  UE  19761129 1   ALL  165  AW  NONE
27  1970   S1  89789777          89789778          LL  UE  19790528 2   ALL  97   AW  NONE
28  1970   S1  91828704          91828710          LL  UE  19741120 7   ALL  1123 AW  NONE
29  1970   S1  CTRSP_FILES026010 CTRSP_FILES026010 CTR NE  19710616 1   ALL  238  AW  NONE
30  1970   S1  TIMN0249833       TIMN0249854       TI  UE  19720616 22  ALL  2739 AW  NONE
31  1980   S1  01331184          01331184          LL  UE  19810326 1   ALL  108  AW  NONE
32  1980   S1  2024271757        2024271758        PM  NE  19871119 2   ALL  326  AW  NONE
33  1980   S1  2024275212        2024275212        PM  NE  19860117 1   ALL  147  AW  NONE
```

[2]Careful scrutiny of following data will reveal that the data given do not match the Table 3.16 data. The 1980 decade stratum contains 36 documents rather than the prescribed quota of 38, and the 1990 decade stratum contains 32 documents rather than 30. I have no explanation for this.

34	1980	S1	2024326901	2024326901	PM	NE	19870522	1	ALL	78	AW	NONE
35	1980	S1	2025849172	2025849181	PM	UE	19870830	10	ALL	413	AW	NONE
36	1980	S1	2025860759	2025860759	TI	NE	19880202	1	ALL	162	AW	NONE
37	1980	S1	2040722635	2040722641	PM	UE	19880502	7	ALL	380	AW	NONE
38	1980	S1	2043531285	2043531296	BW	UE	19880915	12	ALL	663	SM	NONE
39	1980	S1	2044384374	2044384374	PM	NE	19870321	1	ALL	136	AW	NONE
40	1980	S1	2049005450	2049005450	PM	UE	19871100	1	ALL	131	AW	NONE
41	1980	S1	501439328	501439455	RJR	UE	19830411	128	ALL	4095	AW	NONE
42	1980	S1	501977801	501977802	RJR	UE	19840509	2	ALL	410	AW	NONE
43	1980	S1	502635254	502635272	RJR	UE	19840516	19	PART	1933	ET	NONE
44	1980	S1	504306626	504306627	RJR	UE	19840101	2	ALL	1526	AW	NONE
45	1980	S1	505438020	505438020	RJR	NE	19851029	1	ALL	180	AW	NONE
46	1980	S1	506812514	506812514	RJR	UE	19880401	1	ALL	86	AW	NONE
47	1980	S1	509838376	509838378	RJR	NE	19860213	3	ALL	687	AW	NONE
48	1980	S1	521004563	521004563	BW	NE	19811222	1	ALL	192	AW	NONE
49	1980	S1	60017333	60017333	CTR	NE	19870326	1	ALL	314	AW	NONE
50	1980	S1	60018630	60018630	CTR	NE	19880510	1	ALL	313	AW	NONE
51	1980	S1	634332676	634332677	BW	UE	19850625	2	ALL	395	ET	NONE
52	1980	S1	655004405	655004407	BW	NE	19860210	3	ALL	249	AW	NONE
53	1980	S1	682207578	682207578	BW	UE	19800829	1	ALL	1005	AW	NONE
54	1980	S1	690126193	690126193	BW	UE	19851001	1	ALL	115	AW	NONE
55	1980	S1	87113873	87113873	LL	NE	19871230	1	ALL	115	AW	NONE
56	1980	S1	89671712	89671713	LL	UE	19810620	2	ALL	215	SM	NONE
57	1980	S1	93771584	93771585	LL	UE	19890210	2	ALL	334	AW	NONE
58	1980	S1	ATX040785436	ATX040785437	ATC	UE	19860900	2	ALL	307	AW	NONE
59	1980	S1	HK1360117	HK1360117	CTR	UE	19810723	1	ALL	208	AW	NONE
60	1980	S1	TIFL0040165	TIFL0040165	TI	NE	19831029	1	ALL	416	AW	NONE
61	1980	S1	TIMN0054398	TIMN0054398	TI	NE	19850626	1	ALL	151	AW	NONE
62	1980	S1	TIMN0315418	TIMN0315420	TI	NE	19880217	3	ALL	869	AW	NONE
63	1980	S1	TIMN0457513	TIMN0457513	TI	UE	19881119	1	ALL	287	AW	NONE
64	1980	S1	TINY0011140	TINY0011146	TI	NE	19830701	7	ALL	72	AW	NONE
65	1980	S1	TINY0011245	TINY0011245	TI	NE	19841121	1	ALL	154	AW	NONE
66	1980	S1	TNWL0032462	TNWL0032462	TI	NE	19860411	1	ALL	413	AW	NONE
67	1990	S1	2023343220	2023343221	PM	UE	19931028	2	ALL	712	AW	NONE
68	1990	S1	2024254075	2024254075	PM	NE	19910125	1	ALL	163	ET	NONE
69	1990	S1	2024298412	2024298414	PM	UE	19940115	3	ALL	526	ET	AW
70	1990	S1	2041772350	2041772351	PM	NE	19940816	2	ALL	465	SM	NONE
71	1990	S1	2043934309	2043934311	BW	UE	19920801	3	ALL	1161	ET	NONE
72	1990	S1	2047319752	2047319752	PM	UE	19960215	1	ALL	192	CB	NONE
73	1990	S1	2047766713	2047766713	PM	UE	19960623	1	ALL	179	SM	NONE
74	1990	S1	2062542383	2062542384	PM	NE	19920102	2	ALL	307	AW	NONE
75	1990	S1	2064700804	2064700805	PM	UE	19970100	2	ALL	2206	AW	NONE
76	1990	S1	2065215083	2065215083	PM	UE	19980100	1	ALL	278	SM	NONE
77	1990	S1	2069563493	2069563494	PM	NE	19981130	2	ALL	758	SM	NONE
78	1990	S1	2070039086	2070039086	PM	NE	19950216	1	ALL	132	ET	NONE
79	1990	S1	2070915802	2070915802	PM	UE	19960808	1	ALL	197	SM	NONE
80	1990	S1	2072197631	2072197631	PM	UE	19980106	1	ALL	151	SM	NONE
81	1990	S1	2072556372	2072556372	PM	UE	19961127	1	ALL	569	SM	NONE
82	1990	S1	2501343358	2501343361	PM	UE	19900923	4	ALL	357	ET	NONE
83	1990	S1	308001479	308001481	BW	UE	19950930	3	ALL	181	ET	NONE
84	1990	S1	50399835	50399835	CTR	NE	19961030	1	ALL	95	SM	NONE
85	1990	S1	50494256	50494256	CTR	NE	19931004	1	ALL	95	ET	NONE
86	1990	S1	507724242	507724242	RJR	UE	19910301	1	ALL	304	AW	NONE
87	1990	S1	511415571	511415572	RJR	NE	19920803	2	ALL	432	SM	NONE
88	1990	S1	511415865	511415865	RJR	NE	19920206	1	ALL	70	AW	NONE
89	1990	S1	515616123	515616129	RJR	NE	19971211	7	ALL	515	SM	NONE
90	1990	S1	515762055	515762059	RJR	NE	19960124	5	ALL	201	SM	NONE
91	1990	S1	515957783	515957784	RJR	NE	19970529	2	ALL	127	ET	NONE
92	1990	S1	518008339	518008341	RJR	UE	19971116	3	ALL	338	ET	NONE
93	1990	S1	522479450	522479450	RJR	UE	19970425	1	ALL	214	SM	NONE
94	1990	S1	522739495	522739500	RJR	NE	19991215	6	ALL	899	SM	NONE
95	1990	S1	83295344	83295345	LL	NE	19991108	2	ALL	170	SM	NONE
96	1990	S1	88017030	88017031	LL	UE	19951200	2	ALL	903	SM	AW
97	1990	S1	91819326	91819327	RJR	NE	19940228	2	ALL	392	ET	NONE
98	1990	S1	TIMN0435678	TIMN0435679	TI	NE	19910612	2	ALL	561	ET	NONE
99	19XX	S1	TIMN0067458	TIMN0067462	TI	NE	19000000	5	ALL	934	SM	NONE
100	BLILEY	S1	TIMN_256918	TIMN_256940	TI	UE	19790101	23	PART	1937	SM	NONE

Appendix C

OCR Test Data

The following data relates to the Optical Character Recognition (OCR) test. Refer to Section 4.3 for further information.

C.1 OCR Sample Documents

The following twenty documents below were randomly selected from the Quota Sample for the OCR test.

NUM	STRAT	ITR	BATES-START	BATES-END	SRC	CLS	DAT	PGS	AMT	TKN	CDR	VER	
67	1960	Q3	501772729	501772741	RJR	UI	19620000	13	ALL	2059	HW	NONE	
73	1960	Q4	01148569	01148569	LL	NI	19600514	1	ALL	141	SM	NONE	
81	1960	Q4	680258078	680258078	BW	NI	19660917	1	ALL	238	SM	NONE	
82	1960	Q4	680279579	680279581	BW	UI	19600505	3	ALL	566	SM	NONE	no usable OCR text
121	1970	Q1	ATX040827617	ATX040827619	ATC	NI	19730224	3	ALL	168	AW	NONE	OCR Error
127	1970	Q2	03732265	03732270	LL	NI	19790416	6	ALL	1983	AW	NONE	
153	1970	Q2	89301401	89301424	LL	UI	19731031	24	ALL	2061	AW	NONE	
284	1980	Q1	659032972	659032973	BW	NI	19810920	2	ALL	60	AW	NONE	no usable OCR text
295	1980	Q1	ATX03_0025429	ATX03_0025431	ATC	NI	19820719	3	ALL	176	AW	NONE	OCR-Error
299	1980	Q1	CTRSP_FILES023888	CTRSP_FILES023888	CTR	UI	19841206	1	ALL	56	AW	NONE	
326	1980	Q2	2045083228	2045083228	PM	NI	19850312	1	ALL	98	AW	NONE	
345	1980	Q2	504337601	504337603	RJR	NI	19850123	3	ALL	737	AW	NONE	
353	1980	Q2	512180224	512180226	RJR	UI	19841014	3	ALL	1286	AW	NONE	
369	1980	Q2	88208744	88208770	LL	UI	19870603	27	ALL	2652	AW	NONE	
442	1980	Q3	ATX040895180	ATX040895180	ATC	NI	19860403	1	ALL	103	AW	NONE	OCR-Error
459	1980	Q4	2040940418	2040940418	PM	NI	19820416	1	ALL	76	SM	NONE	
551	1990	Q1	511466579	511466579	RJR	NI	19900626	1	ALL	317	AW	NONE	
558	1990	Q1	516927839	516927839	RJR	NI	19971115	1	ALL	58	AW	NONE	
700	1990	Q4	11320767	11320767	CTR	NI	19921017	1	ALL	98	SM	NONE	no usable OCR text
786	BLILEY	Q1	501869639	501869657	RJR	UI	19690901	19	ALL	2038	SM	NONE	

C.2 OCR XSL Stylesheet

The XSL stylesheet used for extracting the typed text in the OCR study can be viewed in Appendix D.3.

C.3 TYPED Text Data

The following block of text lists the 540 word types found in the TYPED text data but not in the OCR text data. Each word is followed by its count in the TYPED text data.

accept (2), accomplished (1), actuarially (1), addendum (1), adding (1), advertise (1), advice (1), advising (1), agents (1), agriculture (1), aim (1), algorithms (1), almost (1), alternative (1), amadori (1), amendment (10), ames (1), amongst (3), announcement (1), answer (1), anti (2), antitrust (1), anyone (2), anything (2), anyway (1), appealed (1), applying (1), appointed (1), appointment (1), appreciably (1), approximates (1), apr (1), architecture (1), arises (1), arriving (1), artek (1), arthur (1), ask (1), assembled (1), assessment (1), attendance (1), attitude (1), author (1), autopsy (1), avenues (1), aware (1), awh (1), barbara (1), bars (1), bat (1), bear (1), becomes (1), benzo (1), bethseda (1), bm (1), boeing (1), bolivia (1), bounds (1), breakdown (1), brown (1), bulk (1), bureaus (1), calculated (1), calculation (2), can't (1), carcinogenic (1), cardinal (1), cardio (1), caring (1), categories (1), causative (1), cautions (1), cellulose (1), cent (1), center's (2), centrifuged (1), characteristic (1), charge (1), children's (2), chosen (1), chromatogram (1), chromatograph (2), churchill (1), circles (1), closely (1), clubs (1), codes (1), coincide (1), collected (1), colleges (1), column (2), commission (1), committee's (1), commons (4), comparability (1), compares (1), compliments (1), compressed (1), compute (1), conceiving (1), concepts (1), conduct (1), conflicting (1), confronted (1), conscience (1), considering (1), considers (1), constituted (1), constitutes (1), constitutions (1), contingencies (1), contribution (2), contributions (1), contributory (1), conversely (1), converted (1), cooperation (1), correlation (1), correlations (1), corresponded (1), corresponding (1), counting (2), covers (1), create (2), creation (1), cross (2), crude (3), crume (1), crushingly (1), ctrsp (1), cue (1), curtis (1), cytoplasm (1), dakota (1), dangers (1), deadwyler (1), decline (1), deficiency (1), demanding (1), depends (1), determinable (1), deterrant (1), diagnostic (1), difficulties (1), difficulty (1), dimethylbenzantracene (1), dimethylsulfoxide (1), dioramas (1), direct (6), distillation (2), distributed (1), doll's (1), donnie (1), doubt (1), dramatically (1), drawn (1), durvasula (1), dust (2), dying (1), economy (1), ecusta (1), educational (1), effectively (1), ehtyl (1), electrostatistically (1), eliminated (1), emphysema (1), empirical (1), employed (1), employer (1), enable (1), enclosed (1), enclosures (1), entertainment (1), entire (1), epidemologic (1), equalize (1), erroneous (1), especially (2), est (1), estimates (1), eventual (1), exchanged (1), exert (1), exhibit (11), exhibitions (1), exhibits (3), existence (1), expanse (1), explore (1), expression (1), factual (1), fair (3), fairfield (1), fairs (5), felt (2), fields (1), filamatic (1), file (2), files (1), film (2), finds (1), firms (1), forest (1), formulate (1), frank (1), frederick (1), frequencies (1), frequency (2), functions (1), furnished (1), gail (1), gathered (1), gavitt (1), generated (1), gercken (1), giving (1), glutamic (2), gone (1), greatest (1), greatly (1), griffith (1), grose (1), gsg (1), haepszel (1), hamburg's (1), hamster (1), hardly (2), haven't (1), hayes (1), headquarters (1), heating (1), heptaocytes (1), heterocycles (1), heterogeneous (1), himpang (1), historical (1), home (1), homogeneous (1), hope (1), housewill (1), howard (1), humidity (1), hygiene (1), hypothalmic (1), ibm (1), identical (1), imminent (1), imperial (2), importance (1), inception (1), incidentally (1), incontrovertible (1), indicator (1), indirect (2), indirectly (1), individual (1), inevitable (1), inferences (2), influence (1), inquire (1), inr (1), instance (1), instrument (1), intelligence (1), intend (1), intensified (1), interesting (3), interpreted (1), intervals (1), interview (1), investigate (1), investigates (1), investigating (1), investigations (1), investigator (4), investigators (1), involved (1), irregular (1), irritating (1), james (1), jane (1), jcs (1), joseph (1), jun (1), justifiably (1), katherine (1), kathleen (1), kept (1), ketones (1), latter (2), launched (1), learned (1), leave (1), legislative (1), lessees (1), list (1), listing (1), location (5), longshore (1), lorraine (1), lotus (1), lt (1), lungs (1), magnetic (1), mailed (1), maintain (1), making (1), males (2), malignant (1), mandatory (1), manliness (1), manning (1), mantel (1), margaret (1), marquees (1), materially (1), mc (1), mckennal (1), meaningless (1), measures (1), medicine (1), membranes (1), mercury (1), micromules (1), migration (2), minorities (1), misleading (2), mission (1), mixtures (1), mobile (1), monkhouse (1), moreover (1), moretem (1), multiply (1), multiplying (2), myerson (2), nagasundarl (1), namely (3), nation's (1), neglect (1), neither (1), neoplasms (1), news (1), nice (1), nitriles (1), nodify (1), nomenclature (1), objectionable (1), objective (1), objectives (1), obligation (1), obscure (1), occurring (1), older (1), olin (1), omission (1), operative (1), original (1), ourselves (1), overlapping (1), owners (1), pantotherm (1), parnele (1), payments (1), peculiarities (1), people's (1), peptic (1), perfetti (1), perforated (1), performing (1), perfume (1), permits (1), photoinactivation (1), physician's (1), pisgah (1), planning (1), pleasing (1), pleasure (1), pleasures (1), polyphenols (1), populations (1), porosity (1), posner (1), posses (1), possibility (1), posters (1), postmaster (1), postponed (1), powder (1), powerful (1), practical (1), predecessor (1), premise (1), presenting (1), principles (1), printouts (1), probability (2), probable (1), producing (2), programmes (2), prudential (1), publications (1), pulmonary (1), putting (1), qualitatively (2), quarantine (1), quickly (1), railroad (1), readily (1), ready (1), reasonable (1), reassayed (1), rec'd (1), recognised (1), referenced (1), referred (2), refers (1), relevant (2), remind (1), repiratory (1), requested (1), rescission (1),

restaurant (1), restrict (1), restricting (1), revealed (2), risks (1), rojner (1), roy (1), satisfactory (1), schedules (2), schools (1), scoring (1), season (1), seeding (1), seemed (1), seemingly (1), send (1), separately (1), serious (1), setting (2), shape (1), shouldn't (1), simultaneous (1), slowly (1), social (2), society (1), somewhat (1), sort (2), span (4), specialized (1), specializing (1), split (1), sprigett (1), static (1), stations (1), statistically (2), statisticians (1), steady (2), steinle (1), stevens (1), strenthened (1), stressed (1), studied (1), subcommittee (6), subm (2), submission (1), successes (1), suggesting (1), sum (1), summarised (1), summer (1), supernatant (1), supervisor (1), supposed (1), supraoptic (1), taking (2), talked (1), technical (1), technology (1), telephone (1), tend (1), tends (1), terminals (1), thank (2), themes (1), thiocynates (1), thrombosis (1), throughout (2), tipping (1), topic (1), toured (1), toxicities (1), traders (1), traffic (1), transportation (1), travel (1), traveling (8), tropical (1), trouble (1), twice (1), ulcer (1), um (1), unable (1), unaffected (1), unanimous (1), uncomplicated (1), understand (1), undertaking (1), undoubtedly (1), unites (1), unknown (2), unnecessary (1), unrelated (2), unusually (1), useless (1), utilize (1), utilized (1), vested (1), via (1), viewers (1), viewing (1), views (2), vital (3), viux (1), voluntary (1), walking (1), wall (1), warning (1), williamson (1), wire (1), wish (1), wonderful (2), worthwhile (1), write (1), writing (1), wug (1), yeaman (1), yellow (1)

C.4 OCR Text Data

The following block of text lists the 3,146 word types found in the OCR text data but not in the TYPED text data. Each word is followed by its count in the OCR text data.

a'nother (1), aa (1), aaaee (1), abiilities (1), abitual (1), abro (1), abvertigin (1), ac (1), acce (2), acceptance (2), accommodate (1), accommodations (2), accompanyinglstatement (1), accord (2), accou'ntspayable (1), accoun (1), accounted (1), accountingl (1), accounts (3), accumu (1), accumulated (1), acd (1), acent (1), acetyl (1), ach (1), achievement (1), aclds (1), acoo (1), acoountants (1), acq (1), acqluisition (1), acqui (1), acquired (5), acre (3), act (2), acti (1), ad (6), add (1), additiions (1), additions (2), addrove (2), ade (2), adequate (1), adfor (1), adiacent (1), adj (1), adjoining (1), adjustedi (1), administered (1), admintstil (1), ado (2), adobe (2), advances (6), adve (1), advertis (1), advertiser (1), adviso (1), ae (1), aee (3), aeoo (1), aerle (1), aetiolo (1), aett (1), affiliate (1), affords (1), afore (1), agcept (1), agement (1), aggregate (1), aggregated (1), aggregatiing (1), aggregating (1), aging (1), ago (2), agood (1), agree (1), ahead (1), ahount (2), ai (4), aiirlines (1), ain (1), aini (1), ains (1), ake (2), alani'fisch (1), alculat (1), ald (1), aleor (1), ali (1), alll (1), alllslze (1), allow (2), allowance (1), allty (1), ally (1), allzed (1), already (4), alte (1), alterna (1), alue (1), alyse (1), amadcan (1), amadorl (1), aman (1), amd (2), ame (1), amedcan (1), amenem (1), amenement (1), amerlca (1), aminofluorene (1), amo (1), amortizatiion (1), amortization (4), amortize (1), amou (2), ample (1), analysts (1), ance (3), ancer (1), ancial (1), ancil (1), and'building (1), andard (1), andatory (1), andi (3), andiextraordinaryitems (1), andingi (1), andl (8), andlfourth (1), andlpayment (1), andlspring (1), andsubsiidiariies (1), anesthetized (1), ange (1), angust (1), anl (1), ann (1), anning (1), annuall (1), anotser (1), anozher (1), ans (1), anthracene (1), antit (1), anto (1), anyagainst (1), anypne (1), aoce (1), aold (1), aooo (1), aordinary (1), aotu (1), aotur (1), apartments (4), aper (1), app (1), appli (1), applicabje (1), applicablei (1), apply (1), appr (1), approx (1), approxi (2), approxilmately (1), approximatel (1), approximatelyas (1), approxl (1), ar (5), arc (1), arch (1), architectu (1), architectural (1), arclt (1), ard (4), ard's (1), ares (1), argentina (1), argents (1), aries (1), ark (1), arke (1), armajor (1), arn (1), aro (1), arotte (1), arou (1), arr (1), arrange (1), ars (1), arthu (1), arti (1), articl (1), arvard (1), ary (2), ase (2), asily (1), aso (1), asoci (1), ass (1), assenbled (1), asset (2), assetsacquired (1), assoc (1), associat (1), associatedlcompany (1), assooi (1), assur (1), assuring (1), ast (1), ata (1), atatistlco (1), ate (2), ated (2), ater (1), athe (1), athens (3), ather (1), aticost (1), atin (1), atinglrevenues (1), ation (1), ations (1), atistic (1), atlanta (1), atlstics (1), atmospheric (1), attenti (1), atter (1), atthe (1), attr (1), attractive (1), attributable (1), aub (1), audiences (2), auditoria (1), auditorium (2), auditoriums (2), augusti (1), auok (1), aup (1), aurence (1), austin (1), australia (1), automatic (1), autopsystudy (1), avai (1), availasle (1), ave (2), aver (1), away (1), awila (1), azer (1), azes (1), ba (1), bacco (2), back (1), background (2), bad (1), bala (1), balcony (1), ban (3), banks (1), banquet (1), barger (1), barsi (1), basicdiet (1), bay (2), bby (1), bcse (1), bcx (1), bdh (1), be'stressed (1), beach (3), beautiful (1), befo (1), beg (1), begini (1), beires (1), ben (2), bene (1), beo (1), ber (3), bers (1), beterocycles (1), betweenmarch (1), betweon (1), beyonidl (1), bezween (1), bf (1), biefore (1), bilities (1), biologicalassociates (1), bits (1), ble (2), blen (1), bles (1), bo (1), boatel (1), bodified (1), bolivla (1), boll (1), bone (1), booked (1), boratcriez (1), boratories (1), bore (1), bos (1), br (1), brad (1), bradfo (1), brand (1), brit (1), brlt (1), bro (2), broadcast (1), broken (1), bron (1), bronc (1), bronx (1), brookhaven (1), brunswick (1), brunswicki (1), bts (1), building (3), buildings (1), bv (1),

by’inhalation (1), calculations (1), call (1), camino (1), canc (1), cance (1), cancelled (1), candy (1), cane (1), cannula (1), cannulated (1), cant (2), capacities (1), capitalize (1), capitall (1), capllal (1), capsule (1), carcin (1), carcinogen (1), carcinogens (3), carcinoma (1), carlo (4), carlos (1), caroino (1), carol (1), cartoon (1), casas (1), cashi (1), casino (1), casitas (1), cated (1), cathedral (1), cation (1), cations (1), causation (1), causcl (1), causez (1), causing (1), cava (3), cb (1), cbnununities (1), ccfttmon (1), ccm (1), ccmpcurd (1), ccncep (1), ccncer (1), cction (1), cctu (1), ccunted (1), cczonar (1), cd (3), ce (7), ced (1), cede (1), ceived (2), cellu (1), cencer (1), cenfar (1), cenla (1), censtructi (1), centeri (1), centers (1), centrations (1), centrifugal (1), cer (13), certaiin (1), certaini (1), certificate (1), certiflied (1), cess (2), cessive (1), cf (2), cfe (1), cg (1), ch (3), cha (1), chai (1), chains (1), chaln (1), charaoterlstlo (1), chased (1), che (1), chemicals (1), chest (1), chiefi (1), chino (2), chromatogr (1), churchilllin (1), chus (1), ci (7), cial (1), cian (1), ciates (1), cidenoe (1), cienti (1), ciga (3), cigarc (1), cigarettea (1), cigarsisupported (1), cil (4), cilliln (1), cinema (1), cipally (1), ciples (1), cised (1), cited (1), cityassayarerecordedintable (1), ckeun (1), cking (1), cl (1), cla (1), clacs (1), clamp (1), clamped (1), class (1), classifications (1), classlfied (1), cle (2), cleveland (2), clga (1), clgare (1), clgaret (1), clgarettes (2), cliffs (2), cline (1), clint (1), closetoa (1), club (2), cluded (1), cludes (2), clysis (1), cmr (1), cmuses (1), cn (3), cnce (1), cnd (1), cnnoor (1), cns (1), co’oil (1), coco (2), coded (2), cof (1), cohor (1), col (1), colbmn (1), colinted (1), coll (1), collagen (1), collateral (1), colle (2), collection (1), com (10), comb (1), combi (1), coming (1), coml (1), comm (1), commencing (1), commerciali (1), communication (1), compa (3), compan (1), comparabfllty (1), complements (1), completiondate (1), compliance (1), comprise (1), con (11), concentr (1), concepti (1), concernin (1), concession (1), concl (1), conclude (1), cond (1), condii (1), condo (1), confidence (2), conflictin (1), conform (1), confusion (1), conne (1), conpa (1), cons (1), consi (2), consitions (1), consoli (1), consolidatedlfinancial (1), consolidatien (1), const’ructioni (1), constltuents (1), constructed (1), constructing (1), consu (1), cont (1), containi (1), containinq (1), contalnlng (1), continents (1), continu (1), continuiing (1), contpact (1), contr (2), contra (1), contribu (1), contro (1), controls (2), contrtsut (1), contw (1), convertible (1), convertin (1), coo (1), coordinator (1), cootors (1), cor (2), cormuercial (1), corn (2), coron (1), coronado (3), corpo (1), corporatedi (1), corporatiion (2), costa (2), costin (1), cot (1), cou (1), courted (1), cov (1), covena (1), coverglass (1), cr (1), creasing (1), created (2), createdl (1), creative (1), cred (1), credit (2), creditag (1), credits (3), criticisas (1), cse (1), cssocio (1), ct (3), cted (1), ctio (1), ction (2), ctivities (1), ctr (1), cu (3), cunditicns (1), cur (1), currency (1), cus (1), cut (1), cuverslips (1), cx (1), cy (1), cyt (2), cytmtcxicity (1), cyto (1), cytot (1), cytotcxicity (1), d’s (1), da (1), dan (1), darkly (1), darvl (3), dat (1), dayof (1), dc (1), dce (1), dcity (1), de (10), dea (2), deben (1), debtl (1), dec (1), decemter (1), decision (1), decllined (1), decrease (1), ded (2), dedication (1), deduction (1), deferred (9), defint (1), deflcienoy (1), del (1), delilvered (1), dell (1), dell’s (1), delow (1), deluxe (2), demandlng (1), den (1), deo (1), deposits (2), depreciation (2), depreciiation (1), der (2), dered (1), deserres (1), deslgned (1), destination (1), detaile (1), dete (3), detection (1), deter (1), deterl (1), deveeopments (1), develc (1), develo (2), develof (1), developaent (1), develope (1), developmenl (1), developmenti (1), deviatien (1), devised (1), devoted (1), dhso (2), diae (1), die (1), diego (2), differs (1), differsnce (1), diirectory (1), dil (1), dilutions (1), dimethylbenz (1), dimethylsu (1), diminution (1), dinal (1), dipped (1), directions (1), director’s (1), direo (1), discounts (1), disney (1), disousse (1), dispositions (2), distracted (1), ditf (1), dividen (1), dlfferent (1), dlinaryitems (1), dlstillation (1), dlstrlbut (1), dny (1), doctecs (1), docters (1), documenta (1), doe (1), doi (2), doii (1), doli (2), dolphin (1), dome (1), domest (1), dominion (1), donation (1), doo (1), dooember (1), doollne (1), dos (1), doselevelsw (1), doubled (1), doubtfuli (1), downtown (1), drain (1), draw (1), drba (1), dril (1), dritish (1), drld (1), dse (1), dtary (1), du (2), dua (1), ducation (1), ducti (2), duction (1), dueprincipallly (1), duo (1), dur (1), duration (1), durln (1), dustin (1), dvance (1), dvice (1), dvisory (1), dy (2), dyine (1), dyno (1), dzshes (1), e’p (1), ea (5), ealth (3), ean (1), eansstive (1), ear (4), eardlezs (1), earl (1), earlymenopause (1), earn (9), earned (2), earniings (4), earnilngs (1), eas (1), ease (1), eastl (1), eastwood (1), eat (1), eath (2), eation (1), ec (3), eccion (1), ece (1), ecerde (1), eci (1), ecidsili (1), ecoepto (1), ecoh (1), economigs (1), ecoo (1), ecqth (1), ect (1), ecth (1), ecti (1), ection (1), ective (1), ects (1), ed (9), ediately (1), edic (1), edical (1), ediccl (1), edium (1), eds (1), edure (2), edverti (1), ee (1), eem (1), eement (1), een (7), eeoon (1), ees (1), eested (1), eet (1), eetl (1), ef (2), efe (1), eferred (1), effectiveness (1), effectl (1), efforts (2), efinite (1), efo (2), efore (1), egor (1), ei (3), eifsctive (1), eigcrette (1), eigh (1), ein (1), ek (1), el (9), ela (2), elated (1), elationship (1), elimi (1), ell (1), ella (1), ely (1), emba (4), eme (1), ement (1), emhausticn (1), emical (1), employing (1), emso (3), en (13), enacted (1), enc (1), ence (2), endln (1), endol (1), ene (1), enent (1), ener (1), eneral (1), enerelly (1), engineering (1), enjoy (2), enjoyed (1), enlf (1), enoe (3), enoke (1), enont (1), enoour (1), enous (1), enq (3), enqui (1), enquiries (2), enqul (1), enqut (1), ens (1), ent (14), ente (1), entertain (1), enti (1), entin (1), entries (1), ents (5), enty (1), environment (2), enys (1), eo (2), eoews (5), eon (3), eonce (1), eor (1), eori (1), eorillard (2), eot (2), eotive (1), eotives (1), eoturers (1), epidemiolozical (1), epidemological (1), eq (1), equ (2), equency (1), equenl (1), eques (1), equiip (1), equily (1), equityi (1), equityiin (1), equityin (1), equrry (1), er (14), era (1), eral (1), ercial (2), ere (6), erence (2), erfc (1), eri (1), erial (1), erience (1), eriially (1), erik (1), ering (1), ermlne (1), erou (2), eroups (2), erplas (1), erran (1), erred (1), ers (2), erved (1), es (21), ese (3), esearc (1), esearch (3), esent (1), esir (1), eso (1), ess (1), essees (1), essent (1), essess (1), esta (1), estale (1), esthetes’of (1), estiimated (1), estimated (4), estimatedl (1), estlon (1), ests (2),

etained (1), etate (1), etatement (1), ethanollglacial (1), ethe (1), ethyl (1), etiolo (1), ett (2), etween (4), etwo (1), ety (1), eu (1), eup (1), europeantravel (1), eut (1), eval (1), eveloped (1), eventful (1), events (1), everyone (1), evetuatien (1), evi (1), evideno (1), ew (1), examinatiion (1), examinatlon (1), exceed (2), exceeds (1), exceptionally (1), excess (11), excessof (1), excise (3), excitement (1), exciting (4), excluding (1), exclusion (2), exer (1), exercisable (1), exorcist (1), expanding (2), expenses (3), experienced (2), expir (1), expo (1), exported (2), exports (4), exposed (2), express (1), exquisite (1), ext (2), extr (1), extraor (1), extraord (2), extraordi (1), extraordilna (1), extraordina (1), f'rom (1), fac (1), facilily (1), facillities (1), factu (1), familiar (1), families (1), family (2), family'of (1), famous (4), faoto (1), farrow (1), fastest (1), fat (1), faxes (1), fcr (2), fe (1), feature (1), feb (1), feet (1), fence (1), ference (1), ferre (1), fessor (1), fewe (1), ffioer (1), fi (4), fic (1), fied (1), fiel (1), fift (1), fifteen (2), figt (1), fiin (1), fiinancial (2), fiirst (2), films (2), filter (1), finan (1), financi (1), financia (1), financiall (1), financla (1), fine (1), finland (1), fir (1), fits (1), fjcation (1), flavorlngbetween (1), flcauce (1), flirst (1), fllamatlc (1), fllm (1), florida (1), flscal (1), fluctuations (1), fo (11), foil (1), fol (1), foll (1), folllows (1), follo (1), followiing (1), followin (1), foot (1), for'a (1), force (1), foreign (2), fori (1), forme (1), forml (1), forrner (1), fort (1), foxide (1), fr (5), fraction (1), francaise (1), franiklin (1), frankfu (1), frankfurt (3), frankliin (2), fromsprague (1), fron (2), fu (2), fulfilll (1), fullerton (1), fulll (1), func (1), fundingi (1), funs (1), fur (2), furthor (1), fx (1), ga (1), gai (1), gaiins (1), gaines (1), gainsi (1), gaming (1), gan (1), garden (1), garett (1), gars (1), gator (1), gatsby (1), gauge (1), gcims (1), gen (1), genemal's (1), generation (1), generdl (1), gentai (1), gentle (1), georg (1), gerial (1), germany's (1), gg (1), giant (1), gical (1), ginning (1), gives (1), glen (1), glish (1), gmoups (1), go (1), godfather (1), gold (1), golf (1), goncorde (1), gove (3), govern (1), gowr (1), gra (1), grant (1), gratitude (1), grc (1), greens (1), griffitii (1), grou (1), grow (2), gt (1), gta (1), guaranteed (1), guests (1), gure (1), gz (1), ha (3), hadl (1), hamburgi (1), hamburgts (1), han (3), hand (1), haskiins (1), hass (1), hat (3), havaviewed (1), hay (1), haz (1), haza (1), hcr (1), hcs (1), hd (1), he's (1), hea (1), headq (1), healthy (1), hearthe (1), heath (2), heatres (1), hec (1), hediu (1), hediumcontrot (1), hee (1), heelth (1), hehighestconcentration (1), heights (1), heiir (1), hel (2), heldljanuary (1), helld (1), helmet (1), helo (1), hen (2), heno (1), hepa (1), hepat (1), hepatic (1), hepato (2), hercury (1), herefo (1), herefore (1), herero (1), heresultsoftheprelimlnary (1), hese (3), het (3), hether (1), hetwee (1), hfll (1), hgs (1), hi (4), hich (1), hiehly (1), highestl (1), hiii (1), hil (1), hills (1), hillsi (1), hing (2), hisstlrlywas (1), historically (1), hitdr (1), hitis (1), hle (1), hls (1), hn (1), ho (7), hoc (1), hofl (1), holders (1), holdiing (1), holding (1), holdings (1), hole (1), hollaender (1), homes (2), hood (1), hools (1), hopne (1), horeover (1), hoteils (1), hoteli (1), hotell (1), hotellwilt (1), hotelsbrought (1), hour (1), houses (2), hronic (1), hs (3), hua (1), huachuca (1), huge (1), humeotants (1), hut (1), hy (1), hypothalamic (1), hyslc (1), ia (1), iaaf (1), iaries (1), iary (1), ible (1), ibutable (1), iby (2), icable (1), ical (2), ically (1), icance (1), icapital (1), ice (2), icel (1), ich (1), icians (1), icipation (1), iclear (1), icles (1), icompany (1), icreative (1), icsl (1), icte (1), icula (1), iculate (1), icy (1), id (3), idcation (1), idence (1), identify (2), idepend (1), idevelopment (1), idine (1), idiscount (2), idity (1), idren'a (1), idventral (1), idy (1), ie (2), iea (1), iealth (1), iearnings (1), iemt (1), ient (1), ientific (1), ier (1), ies (3), iesof (1), ifa (1), ifeat (1), ific (1), ifio (2), ifity (1), ifrankli (1), ig (2), igqrette (1), igs (1), iiabiilities (1), iies (1), iin (3), iincludes (1), iincludiing (1), iincluding (1), iincome (1), iindependent (1), iingsbefore (1), iinvestments (1), iis (1), il (1), ile (3), ileaf (1), ilerns (1), iliab (1), ilities (1), ilives (1), ill (2), illard (1), illustrate (2), illustrates (1), iln (1), ily (1), image (1), imately (2), imcati (1), imediately (1), immeksion (1), impact (1), imple (1), imply (1), improvement (1), improvements (2), imyerso (1), inactiv (1), inally (1), inanced (1), inary (3), incision (1), incl (1), incllusion (1), includedl (1), includi (1), income (15), incr (1), incre (1), increasedl (1), increcse (1), incro (1), incurred (1), indefinitely (1), indenture (1), indentures (1), indepenc (1), indian (1), induces (2), industrialiand (1), ine (3), ined (1), inel (1), inent (1), inferior (2), informer (1), ing (16), ingl (1), ings (6), ini (3), inicrease (1), inister (2), initially (1), initiative (1), inl (1), inn (1), innet (1), innovation (1), innovations (1), ino (2), inoi (2), inoisenoe (2), inolude (1), inoontrov (1), inore (1), inorecso (1), inotes (1), inserted (1), installed (1), installment (1), instrt (1), int (1), intangi (1), inte (4), intends (1), intepnzt (1), interaction (1), interestand (1), interesti (1), interfaced (1), internationall (1), inthe (2), inti (1), intoner (1), introducedlin (1), intst (1), inventories (4), inventory (1), invested (1), investi (1), investmen'rs (1), investmenl (1), investments (5), inz (1), io (3), ioa (1), iocally (1), iocated (1), iod (1), iol (1), iolo (1), ion (5), iones (1), iong (4), ions (1), ionsh (1), iorillard (1), iothor (1), ioverseas (1), ipment (1), ipool (1), ipreference (1), ir (6), ira (1), irdicate (1), irdicates (1), ire (1), irectly (1), ireit (1), irevenues (1), irregularly (1), irritate (1), irtems (1), isb (1), ise (5), isecu (1), ish (1), isin (1), isl (1), island (1), islandi (1), iso (1), isolatip (1), issua (1), issuance (1), ist (1), istatements (1), ister (2), isting (1), istlcal (1), istrap (1), istry (1), isutions (1), ite (1), iten (1), iterns (1), ith (4), ithe (3), itho (1), ito (1), itod (1), itower (1), itrade (1), itransformed (1), itron (1), itselt (1), itsmanufacturing (1), itte (1), ittee (10), ittoe (1), ity (3), ityi (1), ityiin (1), iu (2), iupon (1), ivbs (1), ive (3), ively (1), iven (1), ivin (1), ivlsion (1), iw (1), iwas (1), ix (4), ixtures (1), iysi (1), iz (1), ized (2), izsue (1), japan (1), jlames (1), jnted (1), jointl (1), jor (1), jou (1), jur (1), ka (1), kat (1), ke (2), kee (1), kefe (1), kent (10), kent's (1), ket (1), ketches (1), kha (1), kidney (1), kiing (1), kingsi (1), kno (1), known (3), krou (1), kylated (1), l'he (1), labelled (3), lag (1), lagoon (1), lagrange (1), lah (1), lahead (1), lanai (1), lancaster (1), land (5), landmark (1), lao (1), lard (2), lareest (1),

larger (2), larly (1), las (1), lasm (1), lastl (1), lat (2), late (1), latedidepreciation (1), lathe (1), lation (1), lau (1), lauuched (1), ld (1), lde (1), ldhiactiv (1), lduring (1), lea (1), leaf (1), learly (1), leasehold (1), leaseholds (2), leasing (1), lected (1), leen (1), lequest (1), lequipment (1), les (4), lett (1), letter (2), lev (1), lewe (1), lex (1), lflc (1), lg (3), li (27), liabi (1), liabil (1), liabili (1), liabilities (5), lic (1), licensees (1), licenses (1), lieht (1), lillsi (1), lilver (1), lin (2), lincrease (1), linvestments (1), liqui (1), liquidation (1), lis (1), lishe (3), litt (1), lity (3), lives (1), llebanon (1), lll (1), lln (1), llne (1), llong (1), llorillard's (1), lmeetings (1), ln (2), lnter (1), lnteres (1), lo (3), loaws (1), lobby (3), lobes (1), lodging (1), lodllard (2), loew (3), loew's (1), loewsleconcorde (1), log (1), loglcal (1), loh (1), lon (2), lone (1), lookine (1), loop (1), loows (1), loril (1), lorill (2), lorilla (1), lorillardls (2), lorillardts (1), lorrain (1), los (2), lose (1), loss (2), lowed (1), lows (1), loyalty (1), lrfield (1), lri (1), ls (1), lsland (1), lste (1), lster (1), ltd (1), lth (1), lthe (3), lue (1), lun (1), luntary (1), luo (1), luotio (1), luotion (1), lusion (1), lusions (1), lutamic (1), luxemburg (1), luxu (1), ly (4), lys (2), m'cqueen (1), m'editerranean (1), machines (1), madison (1), maeh (1), mafkets (1), magnitude (1), magnum (1), mai (1), mainta (1), majori (1), maker (2), malyzed (1), mamin (1), man (3), managetwo (1), manera (1), manu (1), manufact (2), manufactures (1), manufacturiing (1), mar (1), marbella (1), marble (1), marina (1), mark (2), markaret (1), marketingprogram (1), marne (1), mart (2), mary (1), maryiand (1), master (3), mat (1), mately (1), mates (2), matogram (1), matu (1), matudng (1), maturiitiies (1), maturing (1), maturities (2), mber (1), meanwhile (1), measul (1), medi (1), medio (1), mee (1), mef (1), menced (1), mendations (1), meno (1), ment (8), mentioned (1), ments (5), mer (2), mereiv (1), merged (1), merits (1), mesa (1), metabolite (1), mewhat (1), mi (2), mia (1), mice (1), micomoles (1), micr (1), microbio (1), microbioi (1), microbiologic (2), microbiologica (5), microbiologici (1), micronite (1), micropipettes (1), midwesti (1), mieratlon (1), miini (1), mild (1), miles (2), mill (1), minable (1), mination (1), minations (1), minimum (3), minist (1), minium (2), minu (1), mir (1), mirmeapolis (1), mister (1), mith (1), mitm (1), mittee's (1), mlnutes (1), mlttee (1), mn (1), mo (5), mode (1), modernization (1), modest (1), mong (1), moni (1), monte (4), montreal (1), mor (2), morality (4), morandum (1), mortal (1), mortclity (1), mortem (1), mortgages (2), mos (2), mostl (1), mount (1), mountains (1), mounting (1), mounts (1), move (1), movie (1), moving (1), mow (1), mp (1), mpany (1), mpc (1), mpl (1), msddd (1), mted (1), mti (1), mtion (1), multi (1), mum (1), muoh (1), muta (1), mutagenesis (1), mycin (1), mye (1), n'l (1), n's (1), na (1), nal (2), nanciali (1), nary (5), nating (1), natio (1), nation (1), nationall (1), native (1), nave (1), ncd (1), nce (2), nceol (1), ncept (1), ncer (4), ncn (1), ncrease (1), ncreasing (1), ndsu (1), ndvisi (1), ne (11), neasures (1), ned (2), needle (1), neeeed (1), neees (1), neeessary (1), nefere (1), neither'offer (1), nele (1), nend (1), nent (1), neo (1), neoplasr (1), nercre (1), nes (1), nesis (1), netl (2), netlc (1), neue (1), neva (1), newman (1), newport (6), newporti (1), newyork (1), ney (1), nf (1), nferences (1), ng (7), nge (1), ngs (3), nhe (1), nibtry (1), nica (1), nichols (1), nificant (1), nimum (1), ninety (1), nings (2), nioot (1), nip (1), nips (1), nister (1), nistryof (1), nits (1), nittee (1), nittoe (1), nkhouse (1), nknown (1), nltrlles (1), nly (1), nmonly (1), nn (3), nnln (1), nohitis (1), nonclinical (1), nons (1), notebook (1), notes (9), noti (1), notlce (1), novel (1), nses (1), nsor (1), nt (5), ntaining (1), ntbi (2), nted (1), ntelligence (1), ntensi (1), nteresting (1), nterpretation (1), ntinue (1), ntl (1), ntout (1), ntroduc (1), ntrol (1), nts (2), nty (1), ntylf (1), nu (2), nuc (1), nue (1), numbe (1), numder (1), nurber (1), nust (1), nv (2), nvestments (1), nz (2), oa (1), oarcin (1), oases (1), oau (2), ob (1), obacco (1), obl (1), obse (1), observations (1), observcd (1), observe (1), oc (2), occu (1), occupancy (1), occupational (1), occupied (1), occut (1), ocedures (1), oclclly (1), oction (1), ocuses (1), odu (1), oducts (1), ody (1), oe (4), oeeneous (1), oein (1), oek (1), oells (1), oeme (1), oen (1), oer (12), oes (1), oews (2), of'america (1), of'beverly (1), of'the (2), ofc (1), ofconsolidatedlearnings (1), ofessor (1), off'taste (1), offe (1), ofi (11), ofibalboa (1), ofidirectors (1), ofiearnings (1), ofiforrner (1), ofiinvestment (1), ofithe (2), ofivest (1), ofl (4), ofldebt (1), ofllabor (1), oflnet (1), ofpiolc (1), ofthe (1), ofthi (1), ofwarrantsthrough (1), ogreater (1), oh (2), ohie (1), ohil (1), ohio (3), ohr (1), oi (3), oiga (1), oigarettes (1), oil (4), oim (2), oisarottea (1), ok (3), oke (3), oker (1), okera (1), okers (5), okin (4), okine (2), oking (2), okir (1), okiu (1), okln (1), ol (1), olders (2), olgarettes (1), olishos (1), oll (1), ollo (1), ollty (1), ollution (1), ollutlon (1), ology (1), olvent (1), om (3), omed (1), omega (1), omen (1), omitted (1), omls (1), ommercial (1), omnission (1), onal (1), onary (1), once (1), oncerncng (1), oned (1), oneof (1), oner (2), ongshore (1), ons (4), onsolidated (1), onstitution (1), onths (1), ontreal (1), ontribueton (1), onts (1), onu (1), oo (9), oolotion (1), oom (1), oomplalnt (2), oompletely (1), oompo (1), oon (1), oonaisers (1), oondition (1), oonditions (1), oons (1), oont (1), oontens (1), oontr (1), oonversion (1), ooo (4), oor (1), ooronnry (1), ootion (1), oounts (1), ope (1), opens (1), oper (1), operati (1), opgt (1), opiiniioin (1), oporative (1), oppor (1), opti (1), optional (1), opulatlon (1), ords (1), ore (3), oreat (1), organization (1), orider (1), oridisplay'in (1), oriented (1), orillard (2), orilllard (1), oritioism (1), orits (1), oriuard (1), orlg (2), oro (1), ort (3), ory (1), ose (1), oses (1), osg (1), osha (1), osis (3), osoo (1), oss (1), osses (1), ossi (2), ossible (1), osslbly (1), ost (1), ostmenopausal (1), ot (7), otal (3), otell (1), othe (1), other'assets (1), otheri (1), others (2), otherwise (3), othore (1), othpr (1), otion (1), oto (1), otur (1), ou (5), oubt (1), ough (1), ouly (1), oumpatible (1), oumpletion (1), oungstown (1), ount (1), ounts (1), oup (4), our'revenues (1), ouse (1), outd (1), outl (1), outlays (1), outlook (1), outperformed (1), outst (1), ov (1), ove (2), overlooking (1), ovor (1), ow (1), owing (1), owned (3), ows (1), ox (1), oxicity (1), pa (2), pacific (1), pacificheights (1), packaging (1), paeke (1), paffen (1), paffenbar (1), pan (1), pancy (1), pany (4), papillon (1), paradise (3), paraguayan (1), pare (1), pared (1), paris (1), parmele (1), parsippany (1), partiial (1), parts (1),

passing (1), pasti (1), patch (2), pause (1), pay (1), pbs (2), pc (2), pcpulation (1), pe (2), pead (1), peal (1), pease (1), peckers (1), pedera (1), pedestrian (1), pelat (1), peni (1), penicilli (1), peninsula (1), pensiion (1), pensioni (1), per'share (1), per'snare (1), perfezti (1), perfo (1), perfusi (1), peri (3), periiod (1), periodla (1), periods (2), permanently (1), persor (1), pes (1), pestered (1), petitive (1), pez (1), pfiudential (1), pfmical (1), pharmacological (1), philippinesi (1), phosphate (1), phot (1), phys (1), physem (1), physlc (1), physlcian (1), pital (1), pl (2), pla (1), plaint (1), planl (1), plants (1), plasmic (1), plated (1), ple (1), pleas (1), pled (1), pledged (2), plenum (1), pleted (1), pletely (1), plex (2), plicable (1), pllans (1), pmo (1), po (4), poa (1), policies (1), polluti (1), ponents (1), pooled (1), pop (1), popu (1), popula (1), populatzons (1), populazlon (1), portal (1), portance (1), ported (1), ports (1), posi (1), positiio (1), posltioned (1), possess (1), possib (1), postmcster (1), postpone (1), pp (2), pplements (1), pproved (1), pr (5), pract (1), practices (1), prc (1), pre (1), preconditioned (1), preferred (2), prel (2), prelimi (1), prelimimary (1), premiere (3), prepaid (1), prepal (1), pres (2), presen (1), presenky (1), presentl (1), presidenti (1), presidentl (1), preted (1), previously (1), priate (1), prices (1), priincipal (1), priinciparly (1), pril (1), prima (1), prin (1), principality (2), prirmiplesof (1), prl (1), prll (1), prlncipal (1), pro (7), probebly (1), probr (1), procedu (1), procedureemployedwas (1), proceeded (1), processes (1), processing (1), prod (1), produ (1), produc (1), producedi (1), produces (3), productions (1), proflt (1), profusion (1), progra (1), programwas (1), progressingl (1), proidool (1), projected (1), prom (1), promis (1), promising (1), prono (1), prooabil (1), proper (1), property (6), propertyi (1), proposa (1), prostate (1), protoc (1), proval (1), providedl (1), proximately (1), prozrams (1), pted (1), ptio (1), pu (5), pubic (1), pubiisbed (1), pubije (1), publi (2), publishe (2), publishedbyt (1), publlo (1), publlshe (1), publlzhed (1), punctured (1), puncturingthe (1), pur (3), purer (1), puted (2), putem (1), puter (1), py (1), qcc (1), qs (1), qu (1), qual (1), qualltatively (1), qualltatlvely (1), quan (1), quar (1), quarter (1), quarters (1), quently (1), quested (1), quiries (1), quiry (1), rac (1), rade (1), radio (1), radiotherapy (1), ram (1), randomly (1), rant (2), rao (1), rapidly (1), rasente (1), raservatlonsi (1), rating (1), ration (1), rb (1), rc (1), rchase (1), rd (1), rdad (1), rds (1), rea (1), reache (1), reali (1), realizable (2), realize (1), reall (1), reaohe (1), reater (1), rec (1), receipts (1), receivable (2), receivables (2), reciat (1), reclassi (1), reclassified (1), recment (1), reco (3), recognition (1), recognizing (1), recon (1), recoveryof (1), recreational (2), rected (1), red (5), redeemable (1), redf (1), redford (1), redl (1), ree (2), referer (1), referredl (1), regency (1), regu (1), reguests (1), regular (1), regulatory (2), reholders (1), reims (1), rel (1), rela (2), relaced (1), relained (1), relat (2), relatedl (1), relatedlto (1), relatiing (2), relatink (1), release (1), released (2), releasi (1), relev (1), rem (1), remen (1), remittance (1), remitted (1), remllts (1), ren (1), renc (1), rencn (1), renderings (1), rendition (1), renewall (1), rentals (3), rep (1), repai (1), repcred (1), replaced (1), repo (5), repoirit (1), repor (1), reposed (1), republic (1), req (1), require (2), requirements (2), reri (1), res (1), resci (1), researc (1), reservation (1), reserve (4), resets (2), residentiial (1), resort (5), resso (1), rest (1), restau (1), restrictive (2), retaiined (2), reties (2), rette (1), rettes (3), returned (1), reve (2), revenuas (1), revenue (1), revesle (1), revol (1), revolvir (1), rf (1), rger (1), rh (1), rhis (1), ri (5), ributory (1), richmond (1), ridid (1), riding (1), rie (1), rietiriemen'r (1), rig (1), ril (1), rily (1), rimary (1), rin (1), rincipally (1), ring (3), rinti (1), ris (1), rit (2), ritcin (1), rite (1), rities (1), ritish (1), rity (3), rks (1), rles (1), rlier (1), rlo (1), rlor (1), rlt (1), rly (1), rn (2), rninistratiive (1), ro (3), robed (1), robin (1), rocedure (1), rochester (1), rocop (1), romquebec (1), ron (1), ronchiti (1), rono (1), roomgoif (1), roomsi (1), rop (1), rope (1), rose (1), rosf (1), rospeotive (1), rote (2), rou (1), route (2), rove (1), roy'posner (1), rporation (1), rposes (1), rrants (1), rren (1), rrent (2), rrod (1), rs (3), rt (1), rted (1), rtfkind (1), rther (2), rthor (1), rtis (1), rty (1), ruildings (1), rushing (1), rustl (1), ry (7), s't (1), sa (4), safe (1), safety (1), saline (1), salles (1), sam (1), samnle (1), sampie (1), sampl (1), sands (1), sanp (1), santa (1), sapple (1), sarple (1), sary (1), sate (1), satisfactorily (1), satisfy (1), sauad (1), sb (2), scc (1), sclentlflc (1), scoo (1), scored (1), scorers (1), scotti (1), screen (1), screpa (1), scrided (1), se (13), sea (2), search (1), sease (1), seating (1), sec (1), secncd (1), secnrilies (1), secon (1), sectien (1), secu (5), secur (1), secure (2), securi (1), sed (2), seeks (1), sel (1), selllng (1), sen (4), separate (3), separatecaption (1), septem (1), serumlfree (1), servd (1), served (1), serviceoffices (1), ses (2), settings (1), sev (1), seventh (1), severall (2), sh (4), sha (1), shaireh (1), shakers (1), shaking (1), sham (1), shaped (1), shar (1), sharehold (1), sharel (1), sharing (1), sharp (1), shc (1), sheuldn't (1), ship (1), sho (4), shore (1), shores (4), shorn (1), shortl (1), shouldi (1), si (10), sidereal (1), sidiaryofthe (1), sien (1), sieve (1), sig (1), signficant (1), simultaneou (1), sinking (3), sion (2), sions (1), sir (1), siren (1), sisted (1), sium (1), sl (1), slatement (1), slates (1), sleadlng (1), sleek (1), slgni (1), slgnt (1), sli (1), slier (1), slla (1), sln (1), slnce (1), slon (1), slowely (1), sm (3), sma (1), smckers (1), smckin (1), smo (2), smohing (1), smok (6), smoki (1), smokin (2), smokine (1), smokln (3), smoklnz (1), smoxing (1), smust (1), sno (1), snok (2), snyder's (1), soccer (1), social'survey (1), soe (1), sol (4), soltent (1), solubil (1), solubili (1), soluble (1), soluticns (1), son (2), sons (1), sor (2), sot (1), sou (1), sound (1), sp (1), spec (1), specified (3), specimens (2), spg (1), spots (1), spptemher (1), spragueidawley (1), springs (2), sq (1), sr (1), ss (4), ssating (1), sses (1), ssist (1), ssoci (1), ssoolation (1), sta (2), staffs (1), stainin (1), stances (1), standa (1), standardized (1), star (2), stardard (1), starring (3), stat (2), statas (1), staten (1), statis (1), statist (1), statistio (1), statlstic (1), stc (1), stcted (1), stctistical (1), ste (3), stea (1), steep (1), sternum (1), sters (1), stilla (1), sting (1), stitute (1), stituted (1), stjh (1), stl (1), stndy (1), stnrte (1), stocl (1), stoke (1), stoups (1), str (1), stra (1),

straight (1), strains (1), strategies (1), stre (1), stream (1), strengthening (1), streptc (1), striction (1), strone (1), sttrly (1), stu (1), stud (2), stufy (1), su (4), subleases (1), submis (1), subnitte (1), subordinatedl (1), subs (1), subse (1), subsequent (2), subsid (1), subsidi (1), subsidiariies (1), subsidiary's (1), subsldi (1), substanti (1), subu (1), suburb (1), suburban (2), suc (2), sucb (1), suo (1), superstar (1), supp (1), supplies (1), suppose (1), supraoptlc (1), sure (1), surrounded (1), surviva (1), susjects (1), sustained (1), sut (1), suz (1), swer (1), swill (1), t'heatresi (1), ta (1), tab (1), tain (2), tained (1), taining (1), tal (2), talenti (1), tali (1), tallest (1), tants (1), tation (1), tatistio (2), taxesi (1), tb (2), tbwh (1), tc (2), tcke (1), tco (1), tcxicities (1), tde (1), te (13), tec (1), tech (1), techn (1), technicall (1), ted (1), tel (2), tele (1), televlslon (1), tely (1), tember (1), temperature (1), tene (1), tenent (1), tenneco (1), tenti (1), ter (2), terial (1), termdebt (2), termdebti (1), terminating (1), ternatlonsl (1), terrified (1), tes (4), tess (1), testarticleselected (1), testresults (1), tetal (1), tews (1), tha (2), thai (1), thar (1), thc (1), theat (1), theatlre (1), theatrein (1), theedge (1), thehope (1), thei (1), theopini (1), ther (3), thereafter (1), therefo (2), therei (1), thesalaried (1), thespecificprutccol (1), thi (3), thingsi (1), thiocya (1), thisstudy (1), thls (1), tho (7), thoir (1), thor (1), thoracic (1), thorax (1), thorities (1), thorough (1), thou (4), thousand (1), thousands (1), thr (3), thriller (1), throughou (1), ths (4), thym (1), ti (5), tic (1), tid (1), ties (3), til (1), tile (3), tin (1), tins (1), tinued (1), tio (2), tiol (1), tion (16), tions (5), tis (1), tissue (1), tistic (1), titstive (1), tive (2), tively (1), tj (1), tlcular (1), tle (2), tlm (1), tlon (1), tltatlvely (1), tltled (1), tlve (1), tly (2), tm (1), tnoreasing (1), to'cacao (1), toconsolid (1), together (1), tohin (1), toits (1), tola (1), tolal (1), tom (1), tomay (1), tomer (1), ton (1), too (1), tor (3), torto (1), tot (1), totall (4), totallchewing (1), totallnumber (1), tothe (1), tots (1), touref (1), towa (1), tower (3), towers (3), town (4), toxi (1), toxiciity (1), tr (1), traclae (1), tracted (1), tradema (1), tradepress (1), transac (1), transformation (1), traordi (1), travelers (1), tre (1), trea'ii (1), treasu (1), treasury (4), treatheht (2), treathent (1), treatrent (1), trend (1), trent (2), treo (1), trer (1), tributed (1), triers (1), tries (1), trigs (1), trimmed (1), trios (1), triton (4), trols (1), tros (1), troy (1), truei (1), try (2), trypsin (2), trztor (1), ts (5), tt (3), tte (3), ttee (2), tto (1), ttrl (1), ttte (1), tu (2), tuberoulosls (1), tudy (1), tumicities (1), tur (1), ture (1), tures (1), turn (1), tv (1), tvito (1), tw (1), twinnedi (1), twinning (2), ty (7), typlcal (1), typloal (2), ual (1), uarially (1), uarning (1), uarters (1), ubero (1), ubjeot (1), uc (1), ucation (1), uce (1), uced (2), uch (1), uclei (1), uction (1), ucts (1), udged (1), udy (2), udyno (1), uenfant (1), uenlations (1), ues (3), uf (1), ui (1), uicsoent (1), uine (1), uir (1), uirements (1), uisition (1), uk (1), ul (2), ul'iplv (1), ulat (1), ulatios (1), umon (1), umted (1), un (1), unamortized (5), uncertainty (1), unco (1), unconsolidated (1), undergoing (1), unestel (1), ung (1), uniformity (1), unin (1), unit (6), unitedl (1), unities (1), unnec (1), unobstructed (1), unrela (1), unscheck (1), unscheduledenasynthesisassaywas (1), unschei (1), untilithe (1), uo (4), uoh (1), uokin (1), uokor (1), upcoming (1), upto (1), ur (4), ure (1), urera (1), ures (1), urgecns (1), usary (1), usedunusually (1), usefu (1), uses (1), ust (2), ustlfiahly (1), ut (6), utabagas (1), ute (1), uthority (1), uthorltles (1), utilization (1), utj (1), utlml (1), utstanding (2), uver (1), va (1), vacation (1), vailc (1), valence (1), vallue (1), valued (2), van (1), variety (1), vation (1), vde (1), ve (6), ved (1), vehicle (1), vein (1), vena (3), vendor (1), vent (1), venturer's (1), ventures (1), ver (2), version (1), vevor (1), vi (1), vidual (1), vie (1), vieux (1), viewed (1), vii (1), viking (1), village (2), villages (1), villas (1), vioe (1), visions (1), vislon (1), vn (1), vo (3), vol (1), vola (1), volatlle (1), volume (4), voluntarily (1), vortexed (1), vsil (1), wa (2), waiks (1), wal'king (1), walt (1), ware (1), wasa (1), washlng (1), waspresentatthehighestccncentrationoftestarticle (1), wcs (2), wee (1), ween (1), weign (1), welll (1), wer (2), westborough (2), wgs (1), wh (1), whereas (1), whiile (1), whioh (1), whleh (1), whlle (1), whloh (1), wholly (2), wi (5), wili (1), will'be (1), willl (4), willows (1), wilt (1), wit (1), withdrawing (1), withe (1), withi (1), wley (1), wlt (1), wn (1), wns (1), wo (9), wore (1), world's (1), worldi (1), worth (2), ws (1), wyler (1), wzth (1), xev (1), xhibiting (1), xic (1), xith (1), xpected (1), yachtsman's (1), ycin (1), ye (2), yea (3), yearns (1), yearsending (1), yecrs (1), yell (1), yen (1), yerk (1), yeu (1), yhave (1), yle (1), yno (1), yo (1), younge (1), yr (1), ysic (1), ytes (1), ythat (1), yzai (1), z (5), zczty (1), zdl (1), zenes (1), zhe (2), zhouqh (1), zlcnough (1), zne (2), zo (3), zoric (1), zprin (1), zroup (1), zt (1), ztant (1), zzee (1)

Appendix D

Tobacco Documents Project Archiving Data

D.1 Tobacco Documents Project DTD

The following is final Document Type Definition (DTD) followed while XML encoding the documents in the Tobacco Documents Corpus. The editor used for encoding performed on-the-fly validation which insured a strict adherence to this DTD.

```
<!-- This is the DTD being used currently by the Tobacco Documents
Grant at the University of Georgia for individual document entries.

Primary Investigator: Don Rubin.
Data Master: Bill Kretzschmar.
DTD Author: Clayton Darwin.

Definition of Tags:

tobaccodocs = root tag (even for single doc xml files)
document    = top level tag for a single document, attributes:
              decade = uga grouping code
                1900-1959 = 1950
                1960-1969 = 1960
                1970-1979 = 1970
                1980-1989 = 1980
                1990-1999 = 1990
                1900 no date = 19xx
                Bliley set = Bliley
              isource = which industry group created the doc
                RJ Reynolds = rjr
                American Tobacco = atc
                Brown and Williamson = bw
                Center for Tobacco Research = ctr
                Lorillard = ll
                Philip Morris = pm
                Tobacco Institute = ti
              class = uga classification
                named/internal audience = ni
                named/external audience = ne
                unnamed/internal audience = ui
                unnamed/external audience = ue
              rcase = case number for matching to other documents
              rclass = rhetorical class for rhetorical cases (not required)
                A = cross Audience case
                D = cross Draft case
              raudience = rhetorical case sub specification (not required)
                internal = internal audience
                external = external audience
              rnumber = rhetorical case draft number (not required)
```

```
                first = first draft
                last = last draft
                1-10 = draft number
metadata    = any data not found within the original document (prior to Bates Numbering)
docdata     = data recovered from the original document
bates_start = Bates Number of the first page of the doc
bates_end   = Bates Number of the last page in the doc
uga         = metadata added by the uga folks
external    = metadata recovered from non uga sources
date        = date of document origin (not a filing date),
              following the standard date indexing methods
pages       = number of pages in the doc
words       = number of words used (max of @ 2000)
section     = whether the doc was used in its entirety or not
                attribute = amount (all or part)
encoded_by  = person who codes the doc in xml
verified_by = person who checks coding - validate, spell check, check assignments
note        = any type of note about the current section needed for clarification or
              description
predoc      = an intro to maindoc within a doc (optional), must have type (see maindoc),
              examples are a cover sheet, intro letter, etc., generally short and clearly
              not the main document of the set of Bates Numbers that represent the whole
              document
maindoc     = the main document in a document (required), the main item of focus, must have
                attribute = type
                  text = normal
                  form = a form
                  image = an image
                  table = a table
                  xtype = unknown type
postdoc     = an attachment to a maindoc but within a single doc, a concatenated document,
              a doc that could stand-alone, (optional), not an appendix of the maindoc,
              must have type (see maindoc)
xdoc        = an extra position for an unclassified doc attachment to the maindoc,
              (optional), not an appendix to the maindoc, must have type (see maindoc)
appendix    = an appendix of the maindoc, formatted (by pagination and typesetting) to
              indicate that it belongs within the main document, not stand-alone
pretext     = non-analyzable text that precedes the 'text' tag set, headers and marginal
              information that precedes the title (not a memo title/subject, but a paper or
              presentation title) but includes the opening salutation of a letter
text        = analyzable text, from the title or following the opening salutation of a
              letter, to the end of the analyzable text (closing salutation if no
              postscripts, or end of the text body) must have a type:
                text = normal text
                form = text within a form
                image = text within an image
                table = text within a table
                xtype = unknown type
footend     = to designate a footnote or endnote
                type = ref (reference) no data needed
                type = comment, enter all data
                num = foot/endnote number, should reference the appropriate anc tag
anc         = anchor or superscript that denotes a footnote or endnote (empty tag)
                num = foot/endnote number
part        = used to mark sections for documents greater than 2000 words, requires attribute
                section = begin, middle or end
                bates = "start #"
              if used instead of text in the maindoc tag, must appear three times in a row
              (begin middle end).
posttext    = non-analyzable text following the 'text' tag set
p           = empty tag, paragraph beginning, for all paragraphs not following npb and to begin
              a text section (always start with a paragraph break)
npb         = empty tag, noteworthy non-paragraph break: lines, big white spaces, box or
              form boundaries, use to denote text boxes
page        = marks a page change, includes all non-analyzable text associated with the
              page change (numbers, footers, headers, dates, etc.)
h           = header of any type or typesetting, can include most text tags,
              make a new h tag based on relationship rather than font or typesetting
```

```
quote       = quoted material, text identified with another person or document,
              not for "odd" or "foreign" words (use double quote marks "" for these), can
              contain most text tags to include quotes
margin      = marginalia, to be inserted in the most appropriate place, nearest to where it
              applies, not for editing marks, but does include comments or suggestions about
              editing, includes all non-editing additions to the document (filing marks and
              numbers, stamps, checks etc.) must have attribute type:
                label = notes used for filing
                comment = comments
image       = designates an image within text, must be described, may contain text if more than
              50 words of running/continuous text are within the image
form        = designates a form within text, must be described, may contain text if more than
              50 words of running/continuous text are within the form
table       = designates tabular data within text, must be described, may contain text if more
              than 50 words of running/continuous text are within the table
xitem       = designates an unidentified item within text, must be described, may contain text
              if more than 50 words of running/continuous text are within the item
insert      = designates inserted text within text, denotes editing
lineout     = text that has been marked or crossed out, denotes editing
marked      = text that has been marked by a reader or editor after the completion of the
              original doc (circled, underlined, etc.), done by hand
emph        = emphasis denoted by typesetting (bold, underline that is not a header, italics),
              not for headers, This is for rhetorically significant emphasis, not just
              typesetting.
symbol      = designates any symbol within text (check, star, circle)
illegible   = denotes illegible text, put what you estimate it says
formula     = use for scientific or mathematical formula notation that is impossible to
              represent correctly
description = used to describe the contents of tables, forms, and images
caption     = used to denote captions and their content, used for tables, images, etc.

BEGIN DTD -->

<!ELEMENT tobaccodocs                   (document+)>
<!ELEMENT document                      (metadata, docdata)>
    <!ATTLIST document decade           (1950|1960|1970|1980|1990|19xx|Bliley) #REQUIRED>
    <!ATTLIST document isource          (rjr|atc|bw|ctr|ll|pm|ti) #REQUIRED>
    <!ATTLIST document class            (ni|ne|ui|ue) #REQUIRED>
    <!ATTLIST document rcase            CDATA #IMPLIED>
    <!ATTLIST document rclass           (A|D) #IMPLIED>
    <!ATTLIST document raudience        (internal|external) #IMPLIED>
    <!ATTLIST document rnumber          (earliest|lastest|2|3|4|5|6|7|8|9|10|11|12|13|14|15) #IMPLIED>
<!ELEMENT metadata                      (bates_start,bates_end,uga,external)>
    <!ELEMENT bates_start               (#PCDATA)>
    <!ELEMENT bates_end                 (#PCDATA)>
    <!ELEMENT uga                       (note?,date,pages,words,section,encoded_by,verified_by,image_file)>
        <!ELEMENT note                  (#PCDATA)>
        <!ELEMENT date                  (#PCDATA)>
        <!ELEMENT pages                 (#PCDATA)>
        <!ELEMENT words                 (#PCDATA)>
        <!ELEMENT section               (#PCDATA)>
            <!ATTLIST section amount    (all|part) "all">
        <!ELEMENT encoded_by            (#PCDATA)>
        <!ELEMENT verified_by           (#PCDATA)>
        <!ELEMENT image_file            (#PCDATA)>
    <!ELEMENT external                  ANY>
<!ELEMENT docdata                       (note?,predoc*,maindoc,postdoc*,xdoc*)>
    <!ELEMENT predoc                    (note?,pretext?,(text|(part,part,part)),posttext?,appendix*)>
        <!ATTLIST predoc type           (text|image|form|table|xtype) #REQUIRED>
    <!ELEMENT maindoc                   (note?,pretext?,(text|(part,part,part)),posttext?,appendix*)>
        <!ATTLIST maindoc type          (text|image|form|table|xtype) #REQUIRED>
        <!ELEMENT pretext               (#PCDATA|page|anc|h|margin|image|table|form|xitem|insert|lineout|
                                         marked|emph|symbol|illegible|formula|note)*>
        <!ELEMENT text                  (page|p|npb|image|table|form|xitem|note)*>
            <!ATTLIST text type         (text|image|form|table|xtype) #REQUIRED>
            <!ELEMENT page              (#PCDATA|page|anc|h|margin|image|table|form|xitem|insert|lineout|
                                         marked|emph|symbol|illegible|formula|note)*>
            <!ELEMENT npb               (#PCDATA)>
```

```
            <!ELEMENT image             (description,caption?,text?)>
                <!ELEMENT description   (#PCDATA|note)*>
                <!ELEMENT caption       (text|note)*>
            <!ELEMENT form              (description,caption?,text?)>
            <!ELEMENT table             (description,caption?,text?)>
            <!ELEMENT xitem             (description,caption?,text?)>
            <!ELEMENT p                 (#PCDATA|footend|anc|h|quote|margin|insert|lineout|marked|emph|symbol|
                                         illegible|formula|note)*>
            <!ELEMENT footend           (#PCDATA|p|npb|h|quote|margin|table|xitem|insert|lineout|marked|emph|
                                         symbol|illegible|formula|note)*>
                <!ATTLIST footend type  (ref|comment) #REQUIRED>
                <!ATTLIST footend num   CDATA #REQUIRED>
            <!ELEMENT anc               (#PCDATA)>
                <!ATTLIST anc num       CDATA #REQUIRED>
            <!ELEMENT h                 (#PCDATA|quote|margin|insert|lineout|marked|emph|symbol|illegible|
                                         formula|note)*>
            <!ELEMENT quote             (#PCDATA|page|p|quote|margin|image|table|form|xitem|insert|lineout|
                                         marked|emph|symbol|illegible|formula|note)*>
            <!ELEMENT margin            (#PCDATA|quote|image|table|form|xitem|insert|lineout|marked|emph|
                                         symbol|illegible|formula|note)*>
                <!ATTLIST margin type   (label|comment) #REQUIRED>
            <!ELEMENT insert            (#PCDATA|h|quote|margin|image|table|form|xitem|insert|lineout|marked|
                                         emph|symbol|illegible|formula|note)*>
            <!ELEMENT lineout           (#PCDATA|page|p|quote|margin|image|table|form|xitem|insert|marked|
                                         emph|symbol|illegible|formula|note)*>
            <!ELEMENT marked            (#PCDATA|page|p|quote|anc|margin|image|table|form|xitem|insert|
                                         lineout|marked|emph|symbol|illegible|formula|note)*>
            <!ELEMENT emph              (#PCDATA|page|p|quote|margin|image|table|form|xitem|insert|lineout|
                                         marked|emph|symbol|illegible|formula|note)*>
            <!ELEMENT symbol            (#PCDATA|note)*>
            <!ELEMENT illegible         (#PCDATA|page|p|note)*>
            <!ELEMENT formula           (#PCDATA|note|description)*>
        <!ELEMENT part                  (note?|text)>
            <!ATTLIST part section      (begin|middle|end) #REQUIRED>
            <!ATTLIST part bates        CDATA #REQUIRED>
        <!ELEMENT posttext              (#PCDATA|page|h|margin|image|table|form|xitem|insert|lineout|marked|
                                         emph|symbol|illegible|formula|note)*>
    <!ELEMENT postdoc                   (note?,pretext?,(text|(part,part,part)),posttext?,appendix*)>
        <!ATTLIST postdoc type          (text|image|form|table|xtype) #REQUIRED>
    <!ELEMENT xdoc                      (note?,pretext?,(text|(part,part,part)),posttext?,appendix*)>
        <!ATTLIST xdoc type             (text|image|form|table|xtype) #REQUIRED>
    <!ELEMENT appendix                  (note?,predoc*,maindoc,postdoc*,xdoc*)>

<!-- END DTD -->
```

D.2 Stylesheet: Metadata Word Count

The following is the XSL stylesheet used to extract text for word counts to be added into the metadata of the archives. Refer to Chapter 4.5.1 for more information.

```
<?xml version="1.0"?>

<!-- name space declarations -->

<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

<!-- options -->

<xsl:output method="text" omit-xml-declaration="yes" indent="no"/>

<!-- root tag selection - required tags -->
```

```
<xsl:template match="tobaccodocs"><xsl:apply-templates/></xsl:template>

<!-- root tag divisions -->

<xsl:template match="document">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="metadata">&#032;</xsl:template>
<xsl:template match="docdata">&#032;<xsl:apply-templates/></xsl:template>

<!-- metadata tags -->
<!-- Metadata not selected above.  Available, but not used. -->

<xsl:template match="bates_start">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="bates_end">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="uga">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="external">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="date">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="pages">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="words">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="section">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="note">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="encoded_by">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="verified_by">&#032;<xsl:apply-templates/></xsl:template>

<!-- docdata tags -->

<xsl:template match="predoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="maindoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="postdoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="xdoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="appendix">&#032;<xsl:apply-templates/></xsl:template>

<!-- visible data -->

<xsl:template match="text">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="part">&#032;<xsl:apply-templates/></xsl:template>

<xsl:template match="image">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="form">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="table">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="xitem">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="formula">&#032;form_ula&#032;</xsl:template>
<xsl:template match="insert">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="marked">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="quote">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="emph">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="h">&#032;<xsl:apply-templates/></xsl:template>

<xsl:template match="margin">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="footend">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="lineout">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="illegible">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="caption">&#032;<xsl:apply-templates/></xsl:template>

<!-- hidden data -->

<xsl:template match="pretext">&#032;</xsl:template>
<xsl:template match="posttext">&#032;</xsl:template>

<xsl:template match="p">&#032;</xsl:template>
<xsl:template match="npb">&#032;</xsl:template>
<xsl:template match="page">&#032;</xsl:template>
<xsl:template match="note">&#032;</xsl:template>
<xsl:template match="anc">&#032;</xsl:template>
<xsl:template match="symbol">&#032;</xsl:template>
<xsl:template match="description">&#032;</xsl:template>

</xsl:stylesheet>
```

D.3 Stylesheet: OCR Study

The following is the XSL stylesheet used for the transformation of the typed text documents during the OCR study. Refer to Chapter 4.3 for more information. Unused tags (meaning the data was ignored) have been removed.

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0">
<xsl:template match="docdata"><xsl:apply-templates/></xsl:template>
<xsl:template match="predoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="maindoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="postdoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="xdoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="appendix">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="pretext">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="text">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="posttext">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="image">&#032;<xsl:apply-templates select="text"/></xsl:template>
<xsl:template match="form">&#032;<xsl:apply-templates select="text"/></xsl:template>
<xsl:template match="table">&#032;<xsl:apply-templates select="text"/></xsl:template>
<xsl:template match="xitem">&#032;<xsl:apply-templates select="text"/></xsl:template>
<xsl:template match="formula">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="part">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="insert">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="marked">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="quote">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="emph">&#032;<xsl:apply-templates/><xsl:apply-templates/></xsl:template>
<xsl:template match="h">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="page">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="margin">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="footend">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="anc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="lineout">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="illegible">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="caption">&#032;<xsl:apply-templates/></xsl:template>
</xsl:stylesheet>
```

Appendix E

Tobacco Documents Project Analysis Data

E.1 Text Extraction Stylesheet

The following is the XSL stylesheet used to extract text for general analysis. The primary difference between this stylesheet and the one in SectionD.2 above is the inclusion of a header line in the output that contained classification data (this line was removed during processing). Refer to Chapter 5.2 for more information.

```
<?xml version="1.0"?>

<!-- name space declarations -->

<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

<!-- options -->

<xsl:output method="text" omit-xml-declaration="yes" indent="no"/>

<!-- root tag selection - required tags -->

<xsl:template match="tobaccodocs"><xsl:apply-templates/></xsl:template>

<!-- root tag divisions -->

<xsl:template match="document">&lt;note&gt; sample="<xsl:value-of select="@sample"/>"
                                          bates="<xsl:value-of select=".//bates_start"/>"
                                          isource="<xsl:value-of select="@isource"/>"
                                          decade="<xsl:value-of select="@decade"/>"
                                          class="<xsl:value-of select="@class"/>"
                                          date="<xsl:value-of select=".//date"/>" &lt;/note&gt;
<xsl:apply-templates/></xsl:template>
<xsl:template match="metadata"> </xsl:template>
<xsl:template match="docdata">&#032;<xsl:apply-templates/></xsl:template>

<!-- metadata tags -->
<!-- Metadata not selected above.  Available, but not used. -->

<xsl:template match="bates_start">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="bates_end">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="uga">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="external">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="date">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="pages">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="words">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="section">&#032;<xsl:apply-templates/></xsl:template>
```

```
<xsl:template match="note">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="encoded_by">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="verified_by">&#032;<xsl:apply-templates/></xsl:template>

<!-- docdata tags -->

<xsl:template match="predoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="maindoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="postdoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="xdoc">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="appendix">&#032;<xsl:apply-templates/></xsl:template>

<!-- visible data -->

<xsl:template match="text">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="part">&#032;<xsl:apply-templates/></xsl:template>

<xsl:template match="image">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="form">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="table">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="xitem">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="formula">&#032;form_ula&#032;</xsl:template>
<xsl:template match="insert">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="marked">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="quote">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="emph">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="h">&#032;<xsl:apply-templates/></xsl:template>

<xsl:template match="margin">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="footend">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="lineout">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="illegible">&#032;<xsl:apply-templates/></xsl:template>
<xsl:template match="caption">&#032;<xsl:apply-templates/></xsl:template>

<!-- hidden data -->

<xsl:template match="pretext">&#032;</xsl:template>
<xsl:template match="posttext">&#032;</xsl:template>

<xsl:template match="p">&#032;</xsl:template>
<xsl:template match="npb">&#032;</xsl:template>
<xsl:template match="page">&#032;</xsl:template>
<xsl:template match="note">&#032;</xsl:template>
<xsl:template match="anc">&#032;</xsl:template>
<xsl:template match="symbol">&#032;</xsl:template>
<xsl:template match="description">&#032;</xsl:template>

</xsl:stylesheet>
```

E.2 Tokenizing Script

The following is the actual code (unmodified) from the Python module used to parse/tokenize text prior to counting and analysis. It could certainly have been scripted more efficiently, but it represents the state of Python and the author's coding ability at the time.

```
def tokenize(s1):
    "used for prepping strings in compare programs, returns list"
    s1 = ' ' + s1
    # ascii shown in sets of 10
    tlist = [' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', "'", # keep '
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', '0', '1', # keep numbers
             '2', '3', '4', '5', '6', '7', '8', '9', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', 'a', 'b', 'c', 'd', 'e', # lower
             'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o',
             'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y',
             'z', ' ', ' ', ' ', ' ', ' ', ' ', ' ', 'a', 'b', 'c',
             'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm',
             'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w',
             'x', 'y', 'z', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ',
             ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', '0', '0', # replaces fraction with '0'
             '0', ' ', '\xe0', '\xe1', '\xe2', '\xe3', '\xe4', '\xe5', '\xe6', '\xe7',
             '\xe8', '\xe9', '\xea', '\xeb', '\xec', '\xed', '\xee', '\xef', '\xf0', '\xf1',
             '\xf2', '\xf3', '\xf4', '\xf5', '\xf6', ' ', '\xf8', '\xf9', '\xfa', '\xfb',
             '\xfc', '\xfd', '\xfe', '\xff', '\xe0', '\xe1', '\xe2', '\xe3', '\xe4', '\xe5',
             '\xe6', '\xe7', '\xe8', '\xe9', '\xea', '\xeb', '\xec', '\xed', '\xee', '\xef',
             '\xf0', '\xf1', '\xf2', '\xf3', '\xf4', '\xf5', '\xf6', ' ', '\xf8', '\xf9',
             '\xfa', '\xfb', '\xfc', '\xfd', '\xfe', '\xff']
    # remove all . , : preceded and followed by a digit
    # this joins numbers that are comma or colon divided, etc
    # numbers are important in tobacco doc names and descriptions
    # the period may need to be left in some corpora
    s1 = re.sub('(?<=\\d)[.,:](?=\\d)','',s1)
    # the translate lowers ascii and ansi chars,
    # relpaces non-alphanumerics with 32, leaves '
    s1 = s1.translate(''.join(tlist))
    # remove all but word-internal '
    s1 = s1.replace(' \'',' ') # remove leading
    s1 = s1.replace('\' ',' ') # remove trailing
    return s1.split()
```

The following number pairs provide the translation of characters used in tokenizing. This was taken directly from the tokenizing script above and reformatted. Using the ordinal representation of ASCII and ANSI characters, the left number in the pair is the original character and the right is its value after translation.

```
 0=32   32=32   64=32    96=32   128=32  160=32  192=224  224=224
 1=32   33=32   65=97    97=97   129=32  161=32  193=225  225=225
 2=32   34=32   66=98    98=98   130=32  162=32  194=226  226=226
 3=32   35=32   67=99    99=99   131=32  163=32  195=227  227=227
 4=32   36=32   68=100  100=100  132=32  164=32  196=228  228=228
 5=32   37=32   69=101  101=101  133=32  165=32  197=229  229=229
 6=32   38=32   70=102  102=102  134=32  166=32  198=230  230=230
 7=32   39=39   71=103  103=103  135=32  167=32  199=231  231=231
 8=32   40=32   72=104  104=104  136=32  168=32  200=232  232=232
 9=32   41=32   73=105  105=105  137=32  169=32  201=233  233=233
10=32   42=32   74=106  106=106  138=32  170=32  202=234  234=234
11=32   43=32   75=107  107=107  139=32  171=32  203=235  235=235
12=32   44=32   76=108  108=108  140=32  172=32  204=236  236=236
13=32   45=32   77=109  109=109  141=32  173=32  205=237  237=237
14=32   46=32   78=110  110=110  142=32  174=32  206=238  238=238
15=32   47=32   79=111  111=111  143=32  175=32  207=239  239=239
16=32   48=48   80=112  112=112  144=32  176=32  208=240  240=240
17=32   49=49   81=113  113=113  145=32  177=32  209=241  241=241
18=32   50=50   82=114  114=114  146=32  178=32  210=242  242=242
19=32   51=51   83=115  115=115  147=32  179=32  211=243  243=243
20=32   52=52   84=116  116=116  148=32  180=32  212=244  244=244
21=32   53=53   85=117  117=117  149=32  181=32  213=245  245=245
22=32   54=54   86=118  118=118  150=32  182=32  214=246  246=246
23=32   55=55   87=119  119=119  151=32  183=32  215=32   247=32
24=32   56=56   88=120  120=120  152=32  184=32  216=248  248=248
25=32   57=57   89=121  121=121  153=32  185=32  217=249  249=249
26=32   58=32   90=122  122=122  154=32  186=32  218=250  250=250
27=32   59=32   91=32   123=32   155=32  187=32  219=251  251=251
28=32   60=32   92=32   124=32   156=32  188=48  220=252  252=252
29=32   61=32   93=32   125=32   157=32  189=48  221=253  253=253
30=32   62=32   94=32   126=32   158=32  190=48  222=254  254=254
31=32   63=32   95=32   127=32   159=32  191=32  223=255  255=255
```

E.3 Introduction to Quota Sample Data

The Quota Sample data are divided into two parts. The first contains raw count data derived from the sample itself (the *Count* data). The individual data sets in this part are as follows:

The second part contains data derived from comparison to the combined text of the Brown and the Freiburg-Brown Corpra (the *Z-score* data). The individual data sets in this part are as follows:

Top 400 and Bottom 100 Collocations Ranked by File Z-score - Page 359

Please refer to Chapter 5 for explanations of all data and terms.

E.4 Quota Sample Count Data

E.4.1 Top 500 Tokens Ranked by Frequency Count

Rank	Token	Freq	%Total	Files	%Total
1	the	30090	5.5317	804	99.505
2	of	18596	3.4186	786	97.2772
3	to	14095	2.5912	790	97.7723
4	and	14029	2.5791	778	96.2871
5	in	11401	2.0959	764	94.5545
6	a	9393	1.7268	740	91.5842
7	for	6137	1.1282	734	90.8416
8	is	5513	1.0135	681	84.2822
9	be	4686	0.8615	673	83.2921
10	that	4482	0.824	632	78.2178
11	on	4246	0.7806	658	81.4356
12	with	3807	0.6999	662	81.9307
13	this	3410	0.6269	641	79.3317
14	as	3331	0.6124	610	75.495
15	are	3187	0.5859	585	72.401
16	by	3032	0.5574	601	74.3812
17	will	2727	0.5013	517	63.9851
18	was	2579	0.4741	423	52.3515
19	have	2348	0.4317	561	69.4307
20	from	2313	0.4252	550	68.0693
21	or	2216	0.4074	472	58.4158
22	at	2214	0.407	530	65.5941
23	it	2211	0.4065	492	60.8911
24	tobacco	2079	0.3822	363	44.9257
25	not	2057	0.3782	485	60.0248
26	were	1968	0.3618	359	44.4307
27	we	1947	0.3579	413	51.1139
28	i	1917	0.3524	444	54.9505
29	an	1597	0.2936	464	57.4257
30	has	1595	0.2932	433	53.5891
31	1	1561	0.287	402	49.7525
32	you	1559	0.2866	339	41.9554
33	smoking	1473	0.2708	262	32.4257
34	which	1472	0.2706	467	57.797
35	been	1404	0.2581	406	50.2475
36	all	1365	0.2509	415	51.3614
37	cigarette	1296	0.2383	286	35.396
38	2	1237	0.2274	392	48.5149
39	smokers	1172	0.2155	188	23.2673
40	these	1148	0.211	421	52.104
41	smoke	1109	0.2039	228	28.2178
42	no	1095	0.2013	365	45.1733
43	new	1079	0.1984	319	39.4802
44	more	1071	0.1969	334	41.3366
45	would	1028	0.189	323	39.9752
46	than	1010	0.1857	312	38.6139
47	our	1010	0.1857	315	38.9851
48	other	993	0.1826	356	44.0594
49	3	985	0.1811	326	40.3465
50	they	977	0.1796	290	35.8911
51	one	932	0.1713	341	42.203
52	cigarettes	926	0.1702	249	30.8168

53	but	899	0.1653	326	40.3465
54	product	843	0.155	221	27.3515
55	if	842	0.1548	364	45.0495
56	any	838	0.1541	343	42.4505
57	research	836	0.1537	227	28.0941
58	nicotine	827	0.152	147	18.1931
59	test	823	0.1513	193	23.8861
60	may	821	0.1509	321	39.7277
61	their	814	0.1496	309	38.2426
62	two	796	0.1463	308	38.1188
63	4	792	0.1456	300	37.1287
64	study	746	0.1371	208	25.7426
65	also	739	0.1359	338	41.8317
66	can	738	0.1357	307	37.995
67	your	731	0.1344	261	32.302
68	should	722	0.1327	278	34.4059
69	brand	713	0.1311	140	17.3267
70	there	710	0.1305	283	35.0248
71	each	707	0.13	246	30.4455
72	up	696	0.128	258	31.9307
73	5	691	0.127	278	34.4059
74	about	676	0.1243	248	30.6931
75	c	673	0.1237	225	27.8465
76	use	669	0.123	269	33.2921
77	results	662	0.1217	258	31.9307
78	used	661	0.1215	273	33.7871
79	time	653	0.12	302	37.3762
80	year	641	0.1178	198	24.505
81	data	641	0.1178	194	24.0099
82	program	634	0.1166	175	21.6584
83	only	631	0.116	285	35.2723
84	who	629	0.1156	203	25.1238
85	some	619	0.1138	270	33.4158
86	its	618	0.1136	237	29.3317
87	products	597	0.1098	188	23.2673
88	10	579	0.1064	241	29.8267
89	filter	577	0.1061	123	15.2228
90	during	577	0.1061	247	30.5693
91	made	576	0.1059	282	34.901
92	when	563	0.1035	239	29.5792
93	brands	560	0.1029	128	15.8416
94	b	559	0.1028	211	26.1139
95	do	549	0.1009	218	26.9802
96	s	547	0.1006	197	24.3812
97	per	541	0.0995	177	21.9059
98	such	537	0.0987	228	28.2178
99	low	535	0.0984	167	20.6683
100	between	529	0.0972	229	28.3416
101	menthol	528	0.0971	91	11.2624
102	group	528	0.0971	173	21.4109
103	among	527	0.0969	151	18.6881
104	d	525	0.0965	208	25.7426
105	share	522	0.096	85	10.5198
106	market	521	0.0958	153	18.9356
107	6	519	0.0954	218	26.9802
108	first	512	0.0941	253	31.3119
109	both	508	0.0934	262	32.4257
110	had	507	0.0932	215	26.6089
111	what	505	0.0928	171	21.1634
112	health	495	0.091	136	16.8317
113	being	483	0.0888	224	27.7228
114	r	479	0.0881	173	21.4109
115	so	476	0.0875	226	27.9703
116	total	475	0.0873	171	21.1634
117	advertising	474	0.0871	121	14.9752
118	into	473	0.087	247	30.5693
119	after	468	0.086	209	25.8663
120	well	467	0.0859	223	27.599

121	p	466	0.0857	135	16.7079
122	marlboro	464	0.0853	89	11.0149
123	high	463	0.0851	175	21.6584
124	report	462	0.0849	203	25.1238
125	out	461	0.0847	237	29.3317
126	sales	459	0.0844	99	12.2525
127	e	459	0.0844	207	25.6188
128	he	453	0.0833	159	19.6782
129	non	450	0.0827	157	19.4307
130	dr	446	0.082	121	14.9752
131	number	443	0.0814	195	24.1337
132	12	440	0.0809	170	21.0396
133	tar	435	0.08	108	13.3663
134	work	432	0.0794	211	26.1139
135	most	428	0.0787	217	26.8564
136	following	426	0.0783	269	33.2921
137	those	425	0.0781	196	24.2574
138	8	425	0.0781	190	23.5149
139	very	424	0.0779	209	25.8663
140	m	420	0.0772	142	17.5743
141	lights	414	0.0761	82	10.1485
142	information	413	0.0759	211	26.1139
143	cancer	412	0.0757	85	10.5198
144	camel	408	0.075	60	7.4257
145	now	407	0.0748	186	23.0198
146	g	407	0.0748	145	17.9455
147	three	406	0.0746	208	25.7426
148	pack	402	0.0739	117	14.4802
149	over	402	0.0739	193	23.8861
150	level	400	0.0735	161	19.9257
151	studies	398	0.0732	141	17.4505
152	levels	398	0.0732	141	17.4505
153	control	396	0.0728	138	17.0792
154	7	393	0.0722	183	22.6485
155	analysis	388	0.0713	144	17.8218
156	people	381	0.07	123	15.2228
157	through	379	0.0697	215	26.6089
158	exposure	379	0.0697	88	10.8911
159	however	378	0.0695	215	26.6089
160	j	375	0.0689	132	16.3366
161	current	372	0.0684	168	20.7921
162	samples	371	0.0682	121	14.9752
163	air	370	0.068	105	12.995
164	them	366	0.0673	158	19.5545
165	volume	363	0.0667	88	10.8911
166	company	363	0.0667	143	17.698
167	increase	361	0.0664	152	18.8119
168	could	360	0.0662	211	26.1139
169	week	357	0.0656	144	17.8218
170	table	353	0.0649	86	10.6436
171	industry	353	0.0649	125	15.4703
172	using	349	0.0642	156	19.3069
173	u	348	0.064	133	16.4604
174	period	348	0.064	136	16.8317
175	years	346	0.0636	161	19.9257
176	100	345	0.0634	126	15.5941
177	project	342	0.0629	141	17.4505
178	because	341	0.0627	172	21.2871
179	development	340	0.0625	153	18.9356
180	public	339	0.0623	108	13.3663
181	less	339	0.0623	156	19.3069
182	meeting	338	0.0621	146	18.0693
183	found	332	0.061	148	18.3168
184	sample	330	0.0607	114	14.1089
185	under	328	0.0603	190	23.5149
186	morris	328	0.0603	84	10.396
187	mr	327	0.0601	85	10.5198
188	good	326	0.0599	163	20.1733

189	taste	323	0.0594	74	9.1584
190	based	322	0.0592	171	21.1634
191	30	322	0.0592	159	19.6782
192	same	321	0.059	206	25.495
193	many	317	0.0583	144	17.8218
194	philip	316	0.0581	79	9.7772
195	20	316	0.0581	168	20.7921
196	state	315	0.0579	119	14.7277
197	testing	314	0.0577	100	12.3762
198	general	314	0.0577	154	19.0594
199	15	314	0.0577	170	21.0396
200	order	313	0.0575	157	19.4307
201	see	311	0.0572	163	20.1733
202	paper	309	0.0568	110	13.6139
203	system	308	0.0566	118	14.604
204	production	305	0.0561	113	13.9851
205	like	304	0.0559	147	18.1931
206	due	304	0.0559	153	18.9356
207	quality	300	0.0552	115	14.2327
208	please	300	0.0552	182	22.5248
209	my	300	0.0552	135	16.7079
210	flavor	300	0.0552	92	11.3861
211	must	299	0.055	133	16.4604
212	how	299	0.055	130	16.0891
213	make	297	0.0546	178	22.0297
214	date	296	0.0544	154	19.0594
215	then	295	0.0542	153	18.9356
216	his	295	0.0542	123	15.2228
217	us	294	0.054	154	19.0594
218	before	294	0.054	166	20.5446
219	reported	293	0.0539	120	14.8515
220	groups	291	0.0535	99	12.2525
221	support	289	0.0531	151	18.6881
222	available	288	0.0529	158	19.5545
223	american	287	0.0528	111	13.7376
224	next	285	0.0524	156	19.3069
225	rjr	281	0.0517	66	8.1683
226	different	281	0.0517	156	19.3069
227	process	280	0.0515	99	12.2525
228	further	279	0.0513	180	22.2772
229	effect	277	0.0509	124	15.3465
230	rate	276	0.0507	111	13.7376
231	possible	274	0.0504	161	19.9257
232	method	274	0.0504	101	12.5
233	material	274	0.0504	121	14.9752
234	since	273	0.0502	162	20.0495
235	day	273	0.0502	138	17.0792
236	materials	272	0.05	119	14.7277
237	full	272	0.05	114	14.1089
238	95	272	0.05	30	3.7129
239	while	271	0.0498	163	20.1733
240	four	271	0.0498	146	18.0693
241	effects	271	0.0498	111	13.7376
242	11	271	0.0498	136	16.8317
243	weight	270	0.0496	80	9.901
244	where	267	0.0491	155	19.1832
245	much	267	0.0491	156	19.3069
246	present	266	0.0489	143	17.698
247	higher	266	0.0489	121	14.9752
248	lung	265	0.0487	74	9.1584
249	provide	264	0.0485	161	19.9257
250	me	264	0.0485	158	19.5545
251	point	263	0.0483	113	13.9851
252	smoker	262	0.0482	89	11.0149
253	significant	262	0.0482	134	16.5842
254	know	262	0.0482	146	18.0693
255	said	259	0.0476	85	10.5198
256	within	256	0.0471	144	17.8218

257	cost	256	0.0471	106	13.1188
258	changes	256	0.0471	133	16.4604
259	winston	255	0.0469	73	9.0347
260	last	255	0.0469	148	18.3168
261	18	253	0.0465	126	15.5941
262	does	252	0.0463	160	19.802
263	activity	252	0.0463	97	12.005
264	9	252	0.0463	143	17.698
265	t	250	0.046	96	11.8812
266	area	250	0.046	133	16.4604
267	standard	249	0.0458	94	11.6337
268	major	248	0.0456	148	18.3168
269	increased	247	0.0454	111	13.7376
270	free	247	0.0454	114	14.1089
271	did	245	0.045	126	15.5941
272	price	244	0.0449	78	9.6535
273	plan	243	0.0447	98	12.1287
274	animals	242	0.0445	41	5.0743
275	l	241	0.0443	104	12.8713
276	various	240	0.0441	139	17.203
277	ml	240	0.0441	35	4.3317
278	above	240	0.0441	166	20.5446
279	section	239	0.0439	75	9.2822
280	25	239	0.0439	137	16.9554
281	retail	237	0.0436	67	8.2921
282	given	237	0.0436	149	18.4406
283	type	236	0.0434	124	15.3465
284	need	235	0.0432	134	16.5842
285	business	235	0.0432	98	12.1287
286	review	234	0.043	112	13.8614
287	promotion	234	0.043	71	8.7871
288	competitive	234	0.043	88	10.8911
289	blend	234	0.043	78	9.6535
290	h	233	0.0428	100	12.3762
291	scientific	232	0.0427	85	10.5198
292	additional	231	0.0425	156	19.3069
293	national	230	0.0423	120	14.8515
294	shown	229	0.0421	112	13.8614
295	age	229	0.0421	85	10.5198
296	carton	228	0.0419	66	8.1683
297	w	227	0.0417	108	13.3663
298	continue	227	0.0417	129	15.9653
299	committee	226	0.0415	80	9.901
300	mg	225	0.0414	70	8.6634
301	13	225	0.0414	113	13.9851
302	set	222	0.0408	130	16.0891
303	addition	222	0.0408	135	16.7079
304	similar	221	0.0406	141	17.4505
305	lower	221	0.0406	115	14.2327
306	think	220	0.0404	84	10.396
307	case	219	0.0403	121	14.9752
308	prior	218	0.0401	121	14.9752
309	marketing	217	0.0399	85	10.5198
310	change	217	0.0399	117	14.4802
311	kool	216	0.0397	41	5.0743
312	second	215	0.0395	137	16.9554
313	final	215	0.0395	120	14.8515
314	received	212	0.039	131	16.2129
315	shall	210	0.0386	48	5.9406
316	long	210	0.0386	119	14.7277
317	several	209	0.0384	143	17.698
318	consumer	208	0.0382	112	13.8614
319	co	207	0.0381	80	9.901
320	white	206	0.0379	70	8.6634
321	range	206	0.0379	88	10.8911
322	phase	206	0.0379	78	9.6535
323	part	206	0.0379	129	15.9653
324	include	206	0.0379	133	16.4604

325	important	206	0.0379	130	16.0891
326	f	205	0.0377	90	11.1386
327	ets	205	0.0377	27	3.3416
328	completed	205	0.0377	95	11.7574
329	line	203	0.0373	102	12.6238
330	50	203	0.0373	121	14.9752
331	special	202	0.0371	100	12.3762
332	performance	202	0.0371	77	9.5297
333	differences	201	0.037	102	12.6238
334	salem	200	0.0368	65	8.0446
335	quarter	200	0.0368	51	6.3119
336	show	199	0.0366	114	14.1089
337	off	199	0.0366	100	12.3762
338	get	199	0.0366	108	13.3663
339	copy	199	0.0366	121	14.9752
340	reynolds	198	0.0364	65	8.0446
341	media	198	0.0364	83	10.2723
342	states	197	0.0362	90	11.1386
343	including	197	0.0362	130	16.0891
344	attached	197	0.0362	152	18.8119
345	size	196	0.036	92	11.3861
346	form	196	0.036	102	12.6238
347	just	195	0.0358	106	13.1188
348	obtained	194	0.0357	105	12.995
349	virginia	193	0.0355	64	7.9208
350	units	193	0.0355	39	4.8267
351	determine	193	0.0355	134	16.5842
352	women	192	0.0353	43	5.3218
353	questions	192	0.0353	135	16.7079
354	name	192	0.0353	92	11.3861
355	n	192	0.0353	94	11.6337
356	days	192	0.0353	95	11.7574
357	average	192	0.0353	79	9.7772
358	action	191	0.0351	103	12.7475
359	14	191	0.0351	115	14.2327
360	month	190	0.0349	85	10.5198
361	five	190	0.0349	118	14.604
362	display	190	0.0349	48	5.9406
363	below	190	0.0349	104	12.8713
364	against	190	0.0349	122	15.099
365	offer	189	0.0347	72	8.9109
366	presented	188	0.0346	106	13.1188
367	issue	188	0.0346	103	12.7475
368	included	188	0.0346	108	13.3663
369	specific	187	0.0344	117	14.4802
370	even	187	0.0344	115	14.2327
371	ii	186	0.0342	103	12.7475
372	programs	185	0.034	77	9.5297
373	end	185	0.034	132	16.3366
374	distribution	185	0.034	96	11.8812
375	21	185	0.034	100	12.3762
376	cells	184	0.0338	41	5.0743
377	box	184	0.0338	66	8.1683
378	24	184	0.0338	103	12.7475
379	months	181	0.0333	99	12.2525
380	developed	181	0.0333	103	12.7475
381	amount	181	0.0333	115	14.2327
382	tests	180	0.0331	80	9.901
383	small	180	0.0331	124	15.3465
384	value	179	0.0329	88	10.8911
385	q	179	0.0329	18	2.2277
386	points	179	0.0329	64	7.9208
387	place	179	0.0329	116	14.3564
388	it’s	179	0.0329	48	5.9406
389	water	178	0.0327	62	7.6733
390	response	178	0.0327	88	10.8911
391	july	177	0.0325	98	12.1287
392	might	176	0.0324	99	12.2525

393	chemical	176	0.0324	70	8.6634
394	risk	174	0.032	61	7.5495
395	june	174	0.032	83	10.2723
396	down	174	0.032	77	9.5297
397	delivery	174	0.032	63	7.797
398	16	174	0.032	112	13.8614
399	purpose	173	0.0318	85	10.5198
400	considered	173	0.0318	108	13.3663
401	conducted	173	0.0318	116	14.3564
402	better	173	0.0318	116	14.3564
403	areas	173	0.0318	95	11.7574
404	status	172	0.0316	54	6.6832
405	medical	172	0.0316	69	8.5396
406	january	172	0.0316	98	12.1287
407	experimental	172	0.0316	74	9.1584
408	direct	172	0.0316	77	9.5297
409	design	172	0.0316	70	8.6634
410	don’t	171	0.0314	63	7.797
411	subject	170	0.0313	98	12.1287
412	evaluation	170	0.0313	90	11.1386
413	upon	169	0.0311	104	12.8713
414	question	169	0.0311	82	10.1485
415	issues	169	0.0311	81	10.0248
416	required	167	0.0307	103	12.7475
417	basis	167	0.0307	125	15.4703
418	am	167	0.0307	113	13.9851
419	without	166	0.0305	126	15.5941
420	march	166	0.0305	90	11.1386
421	key	166	0.0305	80	9.901
422	kent	166	0.0305	31	3.8366
423	human	166	0.0305	72	8.9109
424	disease	166	0.0305	60	7.4257
425	approximately	166	0.0305	91	11.2624
426	although	166	0.0305	111	13.7376
427	take	165	0.0303	119	14.7277
428	procedure	165	0.0303	72	8.9109
429	potential	165	0.0303	87	10.7673
430	difference	165	0.0303	93	11.5099
431	develop	165	0.0303	83	10.2723
432	campaign	165	0.0303	63	7.797
433	added	165	0.0303	91	11.2624
434	versus	164	0.0301	63	7.797
435	segment	164	0.0301	38	4.703
436	methods	164	0.0301	81	10.0248
437	conditions	164	0.0301	91	11.2624
438	overall	163	0.03	88	10.8911
439	members	163	0.03	79	9.7772
440	interest	163	0.03	97	12.005
441	institute	163	0.03	74	9.1584
442	expected	163	0.03	93	11.5099
443	right	162	0.0298	76	9.4059
444	gas	162	0.0298	57	7.0545
445	tax	161	0.0296	52	6.4356
446	result	161	0.0296	103	12.7475
447	million	161	0.0296	57	7.0545
448	image	161	0.0296	39	4.8267
449	fact	161	0.0296	104	12.8713
450	viceroy	160	0.0294	19	2.3515
451	light	160	0.0294	78	9.6535
452	carbon	160	0.0294	52	6.4356
453	summary	159	0.0292	114	14.1089
454	give	158	0.029	105	12.995
455	best	158	0.029	113	13.9851
456	association	158	0.029	78	9.6535
457	tested	157	0.0289	85	10.5198
458	target	157	0.0289	72	8.9109
459	markets	157	0.0289	52	6.4356
460	little	157	0.0289	96	11.8812

461	compared	157	0.0289	87	10.7673
462	article	157	0.0289	41	5.0743
463	agreement	157	0.0289	67	8.2921
464	letter	156	0.0287	107	13.2426
465	example	156	0.0287	88	10.8911
466	activities	156	0.0287	83	10.2723
467	sampling	155	0.0285	57	7.0545
468	least	155	0.0285	96	11.8812
469	way	154	0.0283	98	12.1287
470	still	154	0.0283	112	13.8614
471	necessary	154	0.0283	106	13.1188
472	treated	153	0.0281	53	6.5594
473	six	153	0.0281	94	11.6337
474	list	153	0.0281	80	9.901
475	going	153	0.0281	73	9.0347
476	approved	153	0.0281	68	8.4158
477	ks	152	0.0279	18	2.2277
478	content	152	0.0279	64	7.9208
479	companies	152	0.0279	74	9.1584
480	slims	151	0.0278	39	4.8267
481	scheduled	151	0.0278	66	8.1683
482	reference	151	0.0278	62	7.6733
483	proposed	151	0.0278	88	10.8911
484	iii	151	0.0278	84	10.396
485	growth	151	0.0278	65	8.0446
486	until	150	0.0276	106	13.1188
487	too	150	0.0276	88	10.8911
488	problem	150	0.0276	91	11.2624
489	pm	150	0.0276	49	6.0644
490	percent	150	0.0276	52	6.4356
491	temperature	149	0.0274	50	6.1881
492	run	149	0.0274	86	10.6436
493	relative	149	0.0274	77	9.5297
494	related	149	0.0274	100	12.3762
495	old	149	0.0274	74	9.1584
496	having	149	0.0274	98	12.1287
497	currently	149	0.0274	98	12.1287
498	concentration	149	0.0274	55	6.8069
499	trial	148	0.0272	48	5.9406
500	primary	148	0.0272	71	8.7871

E.4.2 Top 500 Tokens Ranked by File Count

Rank	Token	Freq	%Total	Files	%Total
1	the	30090	5.5317	804	99.505
2	to	14095	2.5912	790	97.7723
3	of	18596	3.4186	786	97.2772
4	and	14029	2.5791	778	96.2871
5	in	11401	2.0959	764	94.5545
6	a	9393	1.7268	740	91.5842
7	for	6137	1.1282	734	90.8416
8	is	5513	1.0135	681	84.2822
9	be	4686	0.8615	673	83.2921
10	with	3807	0.6999	662	81.9307
11	on	4246	0.7806	658	81.4356
12	this	3410	0.6269	641	79.3317
13	that	4482	0.824	632	78.2178
14	as	3331	0.6124	610	75.495
15	by	3032	0.5574	601	74.3812
16	are	3187	0.5859	585	72.401
17	have	2348	0.4317	561	69.4307
18	from	2313	0.4252	550	68.0693
19	at	2214	0.407	530	65.5941

20	will	2727	0.5013	517	63.9851
21	it	2211	0.4065	492	60.8911
22	not	2057	0.3782	485	60.0248
23	or	2216	0.4074	472	58.4158
24	which	1472	0.2706	467	57.797
25	an	1597	0.2936	464	57.4257
26	i	1917	0.3524	444	54.9505
27	has	1595	0.2932	433	53.5891
28	was	2579	0.4741	423	52.3515
29	these	1148	0.211	421	52.104
30	all	1365	0.2509	415	51.3614
31	we	1947	0.3579	413	51.1139
32	been	1404	0.2581	406	50.2475
33	1	1561	0.287	402	49.7525
34	2	1237	0.2274	392	48.5149
35	no	1095	0.2013	365	45.1733
36	if	842	0.1548	364	45.0495
37	tobacco	2079	0.3822	363	44.9257
38	were	1968	0.3618	359	44.4307
39	other	993	0.1826	356	44.0594
40	any	838	0.1541	343	42.4505
41	one	932	0.1713	341	42.203
42	you	1559	0.2866	339	41.9554
43	also	739	0.1359	338	41.8317
44	more	1071	0.1969	334	41.3366
45	but	899	0.1653	326	40.3465
46	3	985	0.1811	326	40.3465
47	would	1028	0.189	323	39.9752
48	may	821	0.1509	321	39.7277
49	new	1079	0.1984	319	39.4802
50	our	1010	0.1857	315	38.9851
51	than	1010	0.1857	312	38.6139
52	their	814	0.1496	309	38.2426
53	two	796	0.1463	308	38.1188
54	can	738	0.1357	307	37.995
55	time	653	0.12	302	37.3762
56	4	792	0.1456	300	37.1287
57	they	977	0.1796	290	35.8911
58	cigarette	1296	0.2383	286	35.396
59	only	631	0.116	285	35.2723
60	there	710	0.1305	283	35.0248
61	made	576	0.1059	282	34.901
62	should	722	0.1327	278	34.4059
63	5	691	0.127	278	34.4059
64	used	661	0.1215	273	33.7871
65	some	619	0.1138	270	33.4158
66	use	669	0.123	269	33.2921
67	following	426	0.0783	269	33.2921
68	smoking	1473	0.2708	262	32.4257
69	both	508	0.0934	262	32.4257
70	your	731	0.1344	261	32.302
71	up	696	0.128	258	31.9307
72	results	662	0.1217	258	31.9307
73	first	512	0.0941	253	31.3119
74	cigarettes	926	0.1702	249	30.8168
75	about	676	0.1243	248	30.6931
76	into	473	0.087	247	30.5693
77	during	577	0.1061	247	30.5693
78	each	707	0.13	246	30.4455
79	10	579	0.1064	241	29.8267
80	when	563	0.1035	239	29.5792
81	out	461	0.0847	237	29.3317
82	its	618	0.1136	237	29.3317
83	between	529	0.0972	229	28.3416
84	such	537	0.0987	228	28.2178
85	smoke	1109	0.2039	228	28.2178
86	research	836	0.1537	227	28.0941
87	so	476	0.0875	226	27.9703

88	c	673	0.1237	225	27.8465
89	being	483	0.0888	224	27.7228
90	well	467	0.0859	223	27.599
91	product	843	0.155	221	27.3515
92	do	549	0.1009	218	26.9802
93	6	519	0.0954	218	26.9802
94	most	428	0.0787	217	26.8564
95	through	379	0.0697	215	26.6089
96	however	378	0.0695	215	26.6089
97	had	507	0.0932	215	26.6089
98	work	432	0.0794	211	26.1139
99	information	413	0.0759	211	26.1139
100	could	360	0.0662	211	26.1139
101	b	559	0.1028	211	26.1139
102	very	424	0.0779	209	25.8663
103	after	468	0.086	209	25.8663
104	three	406	0.0746	208	25.7426
105	study	746	0.1371	208	25.7426
106	d	525	0.0965	208	25.7426
107	e	459	0.0844	207	25.6188
108	same	321	0.059	206	25.495
109	who	629	0.1156	203	25.1238
110	report	462	0.0849	203	25.1238
111	year	641	0.1178	198	24.505
112	s	547	0.1006	197	24.3812
113	those	425	0.0781	196	24.2574
114	number	443	0.0814	195	24.1337
115	data	641	0.1178	194	24.0099
116	test	823	0.1513	193	23.8861
117	over	402	0.0739	193	23.8861
118	under	328	0.0603	190	23.5149
119	8	425	0.0781	190	23.5149
120	smokers	1172	0.2155	188	23.2673
121	products	597	0.1098	188	23.2673
122	now	407	0.0748	186	23.0198
123	7	393	0.0722	183	22.6485
124	please	300	0.0552	182	22.5248
125	further	279	0.0513	180	22.2772
126	make	297	0.0546	178	22.0297
127	per	541	0.0995	177	21.9059
128	program	634	0.1166	175	21.6584
129	high	463	0.0851	175	21.6584
130	r	479	0.0881	173	21.4109
131	group	528	0.0971	173	21.4109
132	because	341	0.0627	172	21.2871
133	what	505	0.0928	171	21.1634
134	total	475	0.0873	171	21.1634
135	based	322	0.0592	171	21.1634
136	15	314	0.0577	170	21.0396
137	12	440	0.0809	170	21.0396
138	current	372	0.0684	168	20.7921
139	20	316	0.0581	168	20.7921
140	low	535	0.0984	167	20.6683
141	before	294	0.054	166	20.5446
142	above	240	0.0441	166	20.5446
143	while	271	0.0498	163	20.1733
144	see	311	0.0572	163	20.1733
145	good	326	0.0599	163	20.1733
146	since	273	0.0502	162	20.0495
147	years	346	0.0636	161	19.9257
148	provide	264	0.0485	161	19.9257
149	possible	274	0.0504	161	19.9257
150	level	400	0.0735	161	19.9257
151	does	252	0.0463	160	19.802
152	he	453	0.0833	159	19.6782
153	30	322	0.0592	159	19.6782
154	them	366	0.0673	158	19.5545
155	me	264	0.0485	158	19.5545

156	available	288	0.0529	158	19.5545
157	order	313	0.0575	157	19.4307
158	non	450	0.0827	157	19.4307
159	using	349	0.0642	156	19.3069
160	next	285	0.0524	156	19.3069
161	much	267	0.0491	156	19.3069
162	less	339	0.0623	156	19.3069
163	different	281	0.0517	156	19.3069
164	additional	231	0.0425	156	19.3069
165	where	267	0.0491	155	19.1832
166	us	294	0.054	154	19.0594
167	general	314	0.0577	154	19.0594
168	date	296	0.0544	154	19.0594
169	then	295	0.0542	153	18.9356
170	market	521	0.0958	153	18.9356
171	due	304	0.0559	153	18.9356
172	development	340	0.0625	153	18.9356
173	increase	361	0.0664	152	18.8119
174	attached	197	0.0362	152	18.8119
175	support	289	0.0531	151	18.6881
176	among	527	0.0969	151	18.6881
177	given	237	0.0436	149	18.4406
178	major	248	0.0456	148	18.3168
179	last	255	0.0469	148	18.3168
180	found	332	0.061	148	18.3168
181	nicotine	827	0.152	147	18.1931
182	like	304	0.0559	147	18.1931
183	meeting	338	0.0621	146	18.0693
184	know	262	0.0482	146	18.0693
185	four	271	0.0498	146	18.0693
186	g	407	0.0748	145	17.9455
187	within	256	0.0471	144	17.8218
188	week	357	0.0656	144	17.8218
189	many	317	0.0583	144	17.8218
190	analysis	388	0.0713	144	17.8218
191	several	209	0.0384	143	17.698
192	present	266	0.0489	143	17.698
193	company	363	0.0667	143	17.698
194	9	252	0.0463	143	17.698
195	m	420	0.0772	142	17.5743
196	studies	398	0.0732	141	17.4505
197	similar	221	0.0406	141	17.4505
198	project	342	0.0629	141	17.4505
199	levels	398	0.0732	141	17.4505
200	brand	713	0.1311	140	17.3267
201	various	240	0.0441	139	17.203
202	day	273	0.0502	138	17.0792
203	control	396	0.0728	138	17.0792
204	second	215	0.0395	137	16.9554
205	25	239	0.0439	137	16.9554
206	period	348	0.064	136	16.8317
207	health	495	0.091	136	16.8317
208	11	271	0.0498	136	16.8317
209	questions	192	0.0353	135	16.7079
210	p	466	0.0857	135	16.7079
211	my	300	0.0552	135	16.7079
212	addition	222	0.0408	135	16.7079
213	significant	262	0.0482	134	16.5842
214	need	235	0.0432	134	16.5842
215	determine	193	0.0355	134	16.5842
216	u	348	0.064	133	16.4604
217	must	299	0.055	133	16.4604
218	include	206	0.0379	133	16.4604
219	changes	256	0.0471	133	16.4604
220	area	250	0.046	133	16.4604
221	j	375	0.0689	132	16.3366
222	end	185	0.034	132	16.3366
223	received	212	0.039	131	16.2129

224	set	222	0.0408	130	16.0891
225	including	197	0.0362	130	16.0891
226	important	206	0.0379	130	16.0891
227	how	299	0.055	130	16.0891
228	part	206	0.0379	129	15.9653
229	continue	227	0.0417	129	15.9653
230	brands	560	0.1029	128	15.8416
231	without	166	0.0305	126	15.5941
232	did	245	0.045	126	15.5941
233	100	345	0.0634	126	15.5941
234	18	253	0.0465	126	15.5941
235	industry	353	0.0649	125	15.4703
236	basis	167	0.0307	125	15.4703
237	type	236	0.0434	124	15.3465
238	small	180	0.0331	124	15.3465
239	effect	277	0.0509	124	15.3465
240	people	381	0.07	123	15.2228
241	his	295	0.0542	123	15.2228
242	filter	577	0.1061	123	15.2228
243	against	190	0.0349	122	15.099
244	samples	371	0.0682	121	14.9752
245	prior	218	0.0401	121	14.9752
246	material	274	0.0504	121	14.9752
247	higher	266	0.0489	121	14.9752
248	dr	446	0.082	121	14.9752
249	copy	199	0.0366	121	14.9752
250	case	219	0.0403	121	14.9752
251	advertising	474	0.0871	121	14.9752
252	50	203	0.0373	121	14.9752
253	reported	293	0.0539	120	14.8515
254	national	230	0.0423	120	14.8515
255	final	215	0.0395	120	14.8515
256	take	165	0.0303	119	14.7277
257	state	315	0.0579	119	14.7277
258	materials	272	0.05	119	14.7277
259	long	210	0.0386	119	14.7277
260	system	308	0.0566	118	14.604
261	five	190	0.0349	118	14.604
262	specific	187	0.0344	117	14.4802
263	pack	402	0.0739	117	14.4802
264	change	217	0.0399	117	14.4802
265	whether	147	0.027	116	14.3564
266	place	179	0.0329	116	14.3564
267	conducted	173	0.0318	116	14.3564
268	better	173	0.0318	116	14.3564
269	quality	300	0.0552	115	14.2327
270	lower	221	0.0406	115	14.2327
271	even	187	0.0344	115	14.2327
272	amount	181	0.0333	115	14.2327
273	14	191	0.0351	115	14.2327
274	summary	159	0.0292	114	14.1089
275	show	199	0.0366	114	14.1089
276	sample	330	0.0607	114	14.1089
277	full	272	0.05	114	14.1089
278	free	247	0.0454	114	14.1089
279	production	305	0.0561	113	13.9851
280	point	263	0.0483	113	13.9851
281	best	158	0.029	113	13.9851
282	am	167	0.0307	113	13.9851
283	13	225	0.0414	113	13.9851
284	still	154	0.0283	112	13.8614
285	shown	229	0.0421	112	13.8614
286	review	234	0.043	112	13.8614
287	consumer	208	0.0382	112	13.8614
288	16	174	0.032	112	13.8614
289	rate	276	0.0507	111	13.7376
290	increased	247	0.0454	111	13.7376
291	effects	271	0.0498	111	13.7376

292	american	287	0.0528	111	13.7376
293	although	166	0.0305	111	13.7376
294	paper	309	0.0568	110	13.6139
295	w	227	0.0417	108	13.3663
296	tar	435	0.08	108	13.3663
297	public	339	0.0623	108	13.3663
298	included	188	0.0346	108	13.3663
299	get	199	0.0366	108	13.3663
300	considered	173	0.0318	108	13.3663
301	letter	156	0.0287	107	13.2426
302	either	146	0.0268	107	13.2426
303	until	150	0.0276	106	13.1188
304	presented	188	0.0346	106	13.1188
305	necessary	154	0.0283	106	13.1188
306	just	195	0.0358	106	13.1188
307	cost	256	0.0471	106	13.1188
308	obtained	194	0.0357	105	12.995
309	give	158	0.029	105	12.995
310	another	134	0.0246	105	12.995
311	air	370	0.068	105	12.995
312	upon	169	0.0311	104	12.8713
313	l	241	0.0443	104	12.8713
314	fact	161	0.0296	104	12.8713
315	below	190	0.0349	104	12.8713
316	again	143	0.0263	104	12.8713
317	result	161	0.0296	103	12.7475
318	required	167	0.0307	103	12.7475
319	issue	188	0.0346	103	12.7475
320	ii	186	0.0342	103	12.7475
321	done	148	0.0272	103	12.7475
322	developed	181	0.0333	103	12.7475
323	action	191	0.0351	103	12.7475
324	24	184	0.0338	103	12.7475
325	line	203	0.0373	102	12.6238
326	form	196	0.036	102	12.6238
327	differences	201	0.037	102	12.6238
328	therefore	136	0.025	101	12.5
329	method	274	0.0504	101	12.5
330	testing	314	0.0577	100	12.3762
331	special	202	0.0371	100	12.3762
332	related	149	0.0274	100	12.3762
333	off	199	0.0366	100	12.3762
334	h	233	0.0428	100	12.3762
335	21	185	0.034	100	12.3762
336	sales	459	0.0844	99	12.2525
337	process	280	0.0515	99	12.2525
338	months	181	0.0333	99	12.2525
339	might	176	0.0324	99	12.2525
340	indicated	140	0.0257	99	12.2525
341	groups	291	0.0535	99	12.2525
342	way	154	0.0283	98	12.1287
343	subject	170	0.0313	98	12.1287
344	recent	144	0.0265	98	12.1287
345	plan	243	0.0447	98	12.1287
346	july	177	0.0325	98	12.1287
347	january	172	0.0316	98	12.1287
348	held	140	0.0257	98	12.1287
349	having	149	0.0274	98	12.1287
350	currently	149	0.0274	98	12.1287
351	call	130	0.0239	98	12.1287
352	business	235	0.0432	98	12.1287
353	interest	163	0.03	97	12.005
354	future	138	0.0254	97	12.005
355	activity	252	0.0463	97	12.005
356	t	250	0.046	96	11.8812
357	little	157	0.0289	96	11.8812
358	least	155	0.0285	96	11.8812
359	distribution	185	0.034	96	11.8812

360	days	192	0.0353	95	11.7574
361	completed	205	0.0377	95	11.7574
362	areas	173	0.0318	95	11.7574
363	standard	249	0.0458	94	11.6337
364	six	153	0.0281	94	11.6337
365	n	192	0.0353	94	11.6337
366	follows	118	0.0217	93	11.5099
367	expected	163	0.03	93	11.5099
368	difference	165	0.0303	93	11.5099
369	17	142	0.0261	93	11.5099
370	taken	126	0.0232	92	11.3861
371	size	196	0.036	92	11.3861
372	provided	136	0.025	92	11.3861
373	name	192	0.0353	92	11.3861
374	large	133	0.0245	92	11.3861
375	flavor	300	0.0552	92	11.3861
376	produced	146	0.0268	91	11.2624
377	problem	150	0.0276	91	11.2624
378	menthol	528	0.0971	91	11.2624
379	find	132	0.0243	91	11.2624
380	conditions	164	0.0301	91	11.2624
381	approximately	166	0.0305	91	11.2624
382	appropriate	135	0.0248	91	11.2624
383	added	165	0.0303	91	11.2624
384	states	197	0.0362	90	11.1386
385	march	166	0.0305	90	11.1386
386	few	142	0.0261	90	11.1386
387	f	205	0.0377	90	11.1386
388	evaluation	170	0.0313	90	11.1386
389	40	148	0.0272	90	11.1386
390	smoker	262	0.0482	89	11.0149
391	marlboro	464	0.0853	89	11.0149
392	individual	131	0.0241	89	11.0149
393	etc	119	0.0219	89	11.0149
394	complete	141	0.0259	89	11.0149
395	volume	363	0.0667	88	10.8911
396	value	179	0.0329	88	10.8911
397	too	150	0.0276	88	10.8911
398	response	178	0.0327	88	10.8911
399	regarding	109	0.02	88	10.8911
400	range	206	0.0379	88	10.8911
401	proposed	151	0.0278	88	10.8911
402	problems	141	0.0259	88	10.8911
403	overall	163	0.03	88	10.8911
404	exposure	379	0.0697	88	10.8911
405	example	156	0.0287	88	10.8911
406	competitive	234	0.043	88	10.8911
407	22	128	0.0235	88	10.8911
408	potential	165	0.0303	87	10.7673
409	early	124	0.0228	87	10.7673
410	compared	157	0.0289	87	10.7673
411	april	141	0.0259	87	10.7673
412	terms	123	0.0226	86	10.6436
413	table	353	0.0649	86	10.6436
414	run	149	0.0274	86	10.6436
415	position	134	0.0246	86	10.6436
416	own	137	0.0252	86	10.6436
417	initial	147	0.027	86	10.6436
418	indicate	120	0.0221	86	10.6436
419	department	134	0.0246	86	10.6436
420	believe	142	0.0261	86	10.6436
421	19	141	0.0259	86	10.6436
422	tested	157	0.0289	85	10.5198
423	share	522	0.096	85	10.5198
424	scientific	232	0.0427	85	10.5198
425	said	259	0.0476	85	10.5198
426	purpose	173	0.0318	85	10.5198
427	mr	327	0.0601	85	10.5198

428	month	190	0.0349	85	10.5198
429	marketing	217	0.0399	85	10.5198
430	determined	146	0.0268	85	10.5198
431	cancer	412	0.0757	85	10.5198
432	age	229	0.0421	85	10.5198
433	think	220	0.0404	84	10.396
434	morris	328	0.0603	84	10.396
435	likely	128	0.0235	84	10.396
436	iii	151	0.0278	84	10.396
437	established	104	0.0191	84	10.396
438	discussion	111	0.0204	84	10.396
439	v	135	0.0248	83	10.2723
440	rather	128	0.0235	83	10.2723
441	media	198	0.0364	83	10.2723
442	june	174	0.032	83	10.2723
443	here	140	0.0257	83	10.2723
444	develop	165	0.0303	83	10.2723
445	activities	156	0.0287	83	10.2723
446	york	141	0.0259	82	10.1485
447	question	169	0.0311	82	10.1485
448	lights	414	0.0761	82	10.1485
449	far	104	0.0191	82	10.1485
450	efforts	112	0.0206	82	10.1485
451	discussed	113	0.0208	82	10.1485
452	certain	121	0.0222	82	10.1485
453	back	134	0.0246	82	10.1485
454	times	147	0.027	81	10.0248
455	selected	126	0.0232	81	10.0248
456	past	124	0.0228	81	10.0248
457	methods	164	0.0301	81	10.0248
458	issues	169	0.0311	81	10.0248
459	introduction	136	0.025	81	10.0248
460	weight	270	0.0496	80	9.901
461	tests	180	0.0331	80	9.901
462	seen	139	0.0256	80	9.901
463	reduced	143	0.0263	80	9.901
464	previous	107	0.0197	80	9.901
465	list	153	0.0281	80	9.901
466	limited	110	0.0202	80	9.901
467	key	166	0.0305	80	9.901
468	committee	226	0.0415	80	9.901
469	co	207	0.0381	80	9.901
470	working	124	0.0228	79	9.7772
471	philip	316	0.0581	79	9.7772
472	members	163	0.03	79	9.7772
473	greater	125	0.023	79	9.7772
474	except	98	0.018	79	9.7772
475	come	101	0.0186	79	9.7772
476	average	192	0.0353	79	9.7772
477	showed	139	0.0256	78	9.6535
478	request	123	0.0226	78	9.6535
479	price	244	0.0449	78	9.6535
480	phase	206	0.0379	78	9.6535
481	light	160	0.0294	78	9.6535
482	great	125	0.023	78	9.6535
483	blend	234	0.043	78	9.6535
484	association	158	0.029	78	9.6535
485	want	143	0.0263	77	9.5297
486	submitted	110	0.0202	77	9.5297
487	relative	149	0.0274	77	9.5297
488	programs	185	0.034	77	9.5297
489	produce	108	0.0199	77	9.5297
490	performance	202	0.0371	77	9.5297
491	down	174	0.032	77	9.5297
492	direct	172	0.0316	77	9.5297
493	designed	123	0.0226	77	9.5297
494	described	130	0.0239	77	9.5297
495	comments	118	0.0217	77	9.5297

496	cause	119	0.0219	77	9.5297
497	appear	111	0.0204	77	9.5297
498	right	162	0.0298	76	9.4059
499	reports	124	0.0228	76	9.4059
500	means	124	0.0228	76	9.4059

E.4.3 Top 500 Collocations Ranked by Frequency Count

Rank	Collocation	Freq	%Total	Files	%Total
1	philip /3 morris	315	0.0194	76	9.4059
2	u /3 s	295	0.0181	113	13.9851
3	more /3 than	251	0.0154	133	16.4604
4	1 /3 2	149	0.0092	91	11.2624
5	non /3 smokers	145	0.0089	35	4.3317
6	virginia /3 slims	144	0.0089	35	4.3317
7	new /3 york	142	0.0087	79	9.7772
8	tobacco /3 smoke	139	0.0085	59	7.302
9	cigarette /3 smoking	132	0.0081	58	7.1782
10	r /3 d	131	0.0081	57	7.0545
11	lung /3 cancer	127	0.0078	49	6.0644
12	there /3 no	124	0.0076	90	11.1386
13	r /3 j	122	0.0075	56	6.9307
14	less /3 than	121	0.0074	78	9.6535
15	low /3 tar	120	0.0074	45	5.5693
16	tobacco /3 products	118	0.0073	52	6.4356
17	cigarette /3 smoke	117	0.0072	58	7.1782
18	r /3 reynolds	107	0.0066	45	5.5693
19	j /3 reynolds	106	0.0065	45	5.5693
20	tobacco /3 company	103	0.0063	48	5.9406
21	tobacco /3 industry	100	0.0061	46	5.6931
22	12 /3 95	99	0.0061	5	0.6188
23	billion /3 units	97	0.006	11	1.3614
24	united /3 states	96	0.0059	61	7.5495
25	those /3 who	96	0.0059	45	5.5693
26	tar /3 nicotine	96	0.0059	45	5.5693
27	smoking /3 health	96	0.0059	42	5.198
28	2 /3 3	96	0.0059	60	7.4257
29	if /3 any	93	0.0057	84	10.396
30	among /3 smokers	89	0.0055	46	5.6931
31	e /3 g	88	0.0054	58	7.1782
32	b /3 w	88	0.0054	38	4.703
33	flue /3 cured	86	0.0053	23	2.8465
34	would /3 like	85	0.0052	60	7.4257
35	full /3 flavor	85	0.0052	26	3.2178
36	et /3 al	85	0.0052	36	4.4554
37	higher /3 than	84	0.0052	51	6.3119
38	rather /3 than	82	0.005	61	7.5495
39	non /3 menthol	80	0.0049	23	2.8465
40	had /3 been	79	0.0049	55	6.8069
41	carbon /3 monoxide	76	0.0047	31	3.8366
42	long /3 term	75	0.0046	40	4.9505
43	birth /3 weight	73	0.0045	2	0.2475
44	tobacco /3 taste	71	0.0044	19	2.3515
45	p /3 p	70	0.0043	10	1.2376
46	tobacco /3 research	69	0.0042	40	4.9505
47	marlboro /3 lights	69	0.0042	19	2.3515
48	brown /3 williamson	69	0.0042	29	3.5891
49	low /3 weight	68	0.0042	5	0.6188
50	3 /3 4	68	0.0042	51	6.3119
51	ultra /3 lights	67	0.0041	12	1.4851
52	passive /3 smoking	67	0.0041	16	1.9802
53	competitive /3 smokers	67	0.0041	24	2.9703
54	tobacco /3 institute	66	0.0041	34	4.2079

55	lucky /3 strike	66	0.0041	21	2.599
56	flare /3 up	66	0.0041	3	0.3713
57	american /3 tobacco	66	0.0041	30	3.7129
58	any /3 questions	65	0.004	62	7.6733
59	any /3 other	64	0.0039	46	5.6931
60	p /3 m	63	0.0039	28	3.4653
61	year /3 period	61	0.0037	16	1.9802
62	table /3 1	61	0.0037	33	4.0842
63	smokers /3 smokers	59	0.0036	26	3.2178
64	new /3 product	59	0.0036	26	3.2178
65	last /3 year	59	0.0036	33	4.0842
66	no /3 significant	57	0.0035	32	3.9604
67	first /3 quarter	57	0.0035	19	2.3515
68	differences /3 between	57	0.0035	37	4.5792
69	test /3 market	56	0.0034	30	3.7129
70	please /3 me	56	0.0034	54	6.6832
71	pall /3 mall	55	0.0034	25	3.0941
72	1 /3 1	55	0.0034	38	4.703
73	test /3 article	54	0.0033	4	0.495
74	nicotine /3 salicylate	54	0.0033	1	0.1238
75	1 /3 3	53	0.0033	38	4.703
76	share /3 points	52	0.0032	7	0.8663
77	set /3 up	52	0.0032	38	4.703
78	other /3 than	52	0.0032	45	5.5693
79	dual /3 filter	52	0.0032	9	1.1139
80	under /3 conditions	51	0.0031	37	4.5792
81	reynolds /3 tobacco	51	0.0031	29	3.5891
82	smokers /3 who	50	0.0031	27	3.3416
83	people /3 who	50	0.0031	26	3.2178
84	g /3 13	50	0.0031	12	1.4851
85	during /3 period	50	0.0031	36	4.4554
86	benson /3 hedges	50	0.0031	19	2.3515
87	share /3 point	49	0.003	4	0.495
88	new /3 products	49	0.003	27	3.3416
89	indoor /3 air	49	0.003	17	2.104
90	ng /3 ml	48	0.003	2	0.2475
91	anti /3 smoking	48	0.003	24	2.9703
92	year /3 old	47	0.0029	19	2.3515
93	smoke /3 exposure	47	0.0029	17	2.104
94	r /3 tobacco	47	0.0029	28	3.4653
95	form /3 ula	47	0.0029	20	2.4752
96	d /3 c	47	0.0029	30	3.7129
97	carbon /3 dioxide	47	0.0029	15	1.8564
98	board /3 directors	47	0.0029	17	2.104
99	07 /3 95	47	0.0029	1	0.1238
100	which /3 been	46	0.0028	39	4.8267
101	vice /3 president	46	0.0028	20	2.4752
102	than /3 other	46	0.0028	33	4.0842
103	been /3 made	46	0.0028	38	4.703
104	12 /3 07	46	0.0028	1	0.1238
105	which /3 would	45	0.0028	39	4.8267
106	their /3 own	45	0.0028	33	4.0842
107	smoking /3 cessation	45	0.0028	6	0.7426
108	smoke /3 condensate	45	0.0028	19	2.3515
109	ring /3 around	45	0.0028	1	0.1238
110	prior /3 year	45	0.0028	6	0.7426
111	other /3 brands	45	0.0028	24	2.9703
112	let /3 know	45	0.0028	42	5.198
113	j /3 tobacco	45	0.0028	26	3.2178
114	if /3 they	45	0.0028	34	4.2079
115	filling /3 capacity	45	0.0028	10	1.2376
116	what /3 do	44	0.0027	24	2.9703
117	mainstream /3 smoke	44	0.0027	26	3.2178
118	only /3 one	43	0.0026	29	3.5891
119	no /3 difference	43	0.0026	28	3.4653
120	let /3 me	43	0.0026	40	4.9505
121	carried /3 out	43	0.0026	28	3.4653
122	warning /3 statement	42	0.0026	4	0.495

123	units /3 share	42	0.0026	1	0.1238
124	there /3 been	42	0.0026	37	4.5792
125	north /3 carolina	42	0.0026	19	2.3515
126	m /3 d	42	0.0026	18	2.2277
127	billion /3 share	42	0.0026	2	0.2475
128	2 /3 1	42	0.0026	30	3.7129
129	1 /3 4	42	0.0026	27	3.3416
130	public /3 health	41	0.0025	25	3.0941
131	menthol /3 smokers	41	0.0025	13	1.6089
132	low /3 delivery	41	0.0025	8	0.9901
133	heart /3 disease	41	0.0025	18	2.2277
134	female /3 smokers	41	0.0025	16	1.9802
135	effects /3 smoking	41	0.0025	27	3.3416
136	cigarette /3 advertising	41	0.0025	19	2.3515
137	camel /3 cash	41	0.0025	7	0.8663
138	c /3 c	41	0.0025	25	3.0941
139	your /3 letter	40	0.0025	34	4.2079
140	same /3 time	40	0.0025	34	4.2079
141	filter /3 up	40	0.0025	4	0.495
142	these /3 two	39	0.0024	35	4.3317
143	see /3 table	39	0.0024	14	1.7327
144	public /3 smoking	39	0.0024	19	2.3515
145	marlboro /3 smokers	39	0.0024	15	1.8564
146	lower /3 than	39	0.0024	30	3.7129
147	greater /3 than	39	0.0024	26	3.2178
148	fourth /3 quarter	39	0.0024	11	1.3614
149	filter /3 flare	39	0.0024	3	0.3713
150	difference /3 between	39	0.0024	32	3.9604
151	who /3 smoke	38	0.0023	23	2.8465
152	tobacco /3 use	38	0.0023	16	1.9802
153	they /3 would	38	0.0023	29	3.5891
154	share /3 down	38	0.0023	5	0.6188
155	low /3 birth	38	0.0023	2	0.2475
156	gas /3 phase	38	0.0023	17	2.104
157	claim /3 wherein	38	0.0023	2	0.2475
158	cigarette /3 paper	38	0.0023	22	2.7228
159	all /3 these	38	0.0023	37	4.5792
160	18 /3 24	38	0.0023	15	1.8564
161	11 /3 95	38	0.0023	4	0.495
162	which /3 may	37	0.0023	35	4.3317
163	ultra /3 low	37	0.0023	16	1.9802
164	tobacco /3 companies	37	0.0023	15	1.8564
165	table /3 2	37	0.0023	22	2.7228
166	smoke /3 cigarettes	37	0.0023	23	2.8465
167	sidestream /3 smoke	37	0.0023	19	2.3515
168	low /3 low	37	0.0023	8	0.9901
169	king /3 size	37	0.0023	15	1.8564
170	following /3 1	37	0.0023	32	3.9604
171	air /3 quality	37	0.0023	15	1.8564
172	winston /3 salem	36	0.0022	26	3.2178
173	what /3 they	36	0.0022	25	3.0941
174	they /3 do	36	0.0022	27	3.3416
175	smokers /3 non	36	0.0022	17	2.104
176	share /3 market	36	0.0022	17	2.104
177	prior /3 period	36	0.0022	1	0.1238
178	l /3 m	36	0.0022	25	3.0941
179	b /3 h	36	0.0022	14	1.7327
180	american /3 company	36	0.0022	17	2.104
181	adult /3 smokers	36	0.0022	21	2.599
182	4 /3 5	36	0.0022	30	3.7129
183	point /3 sale	35	0.0022	17	2.104
184	ph /3 d	35	0.0022	14	1.7327
185	no /3 than	35	0.0022	23	2.8465
186	no /3 differences	35	0.0022	24	2.9703
187	new /3 brand	35	0.0022	15	1.8564
188	he /3 said	35	0.0022	19	2.3515
189	council /3 tobacco	35	0.0022	19	2.3515
190	council /3 research	35	0.0022	19	2.3515

191	among /3 blacks	35	0.0022	2	0.2475
192	two /3 pack	34	0.0021	12	1.4851
193	reynolds /3 company	34	0.0021	19	2.3515
194	please /3 let	34	0.0021	33	4.0842
195	photoacoustic /3 spectroscopy	34	0.0021	1	0.1238
196	per /3 pack	34	0.0021	20	2.4752
197	per /3 ml	34	0.0021	5	0.6188
198	per /3 day	34	0.0021	23	2.8465
199	nicotine /3 cotinine	34	0.0021	11	1.3614
200	much /3 more	34	0.0021	27	3.3416
201	filter /3 tareyton	34	0.0021	7	0.8663
202	camel /3 wides	34	0.0021	5	0.6188
203	between /3 smoking	34	0.0021	19	2.3515
204	been /3 completed	34	0.0021	23	2.8465
205	any /3 please	34	0.0021	34	4.2079
206	1 /3 10	34	0.0021	15	1.8564
207	who /3 had	33	0.002	20	2.4752
208	these /3 results	33	0.002	26	3.2178
209	these /3 products	33	0.002	20	2.4752
210	share /3 up	33	0.002	5	0.6188
211	relationship /3 between	33	0.002	26	3.2178
212	reference /3 cigarette	33	0.002	7	0.8663
213	outer /3 wrap	33	0.002	2	0.2475
214	no /3 no	33	0.002	14	1.7327
215	menthol /3 taste	33	0.002	8	0.9901
216	me /3 know	33	0.002	32	3.9604
217	ets /3 exposure	33	0.002	6	0.7426
218	do /3 think	33	0.002	16	1.9802
219	cured /3 tobacco	33	0.002	15	1.8564
220	body /3 odor	33	0.002	1	0.1238
221	all /3 other	33	0.002	26	3.2178
222	air /3 pollution	33	0.002	15	1.8564
223	6 /3 7	33	0.002	19	2.3515
224	very /3 much	32	0.002	26	3.2178
225	salem /3 lights	32	0.002	7	0.8663
226	particle /3 size	32	0.002	12	1.4851
227	next /3 week	32	0.002	22	2.7228
228	new /3 cigarette	32	0.002	17	2.104
229	menthol /3 tobacco	32	0.002	6	0.7426
230	m /3 1	32	0.002	6	0.7426
231	last /3 week	32	0.002	21	2.599
232	j /3 company	32	0.002	18	2.2277
233	focus /3 group	32	0.002	7	0.8663
234	final /3 budget	32	0.002	1	0.1238
235	exposure /3 smoke	32	0.002	14	1.7327
236	dual /3 tareyton	32	0.002	6	0.7426
237	charcoal /3 filter	32	0.002	11	1.3614
238	but /3 they	32	0.002	27	3.3416
239	better /3 than	32	0.002	26	3.2178
240	all /3 smokers	32	0.002	19	2.3515
241	2 /3 4	32	0.002	21	2.599
242	test /3 cigarette	31	0.0019	10	1.2376
243	surgeon /3 general's	31	0.0019	18	2.2277
244	sales /3 force	31	0.0019	19	2.3515
245	questions /3 please	31	0.0019	31	3.8366
246	particulate /3 matter	31	0.0019	17	2.104
247	n /3 c	31	0.0019	16	1.9802
248	mg /3 tar	31	0.0019	17	2.104
249	follow /3 up	31	0.0019	26	3.2178
250	final /3 report	31	0.0019	20	2.4752
251	each /3 group	31	0.0019	17	2.104
252	cigarettes /3 made	31	0.0019	17	2.104
253	but /3 also	31	0.0019	27	3.3416
254	11 /3 27	31	0.0019	3	0.3713
255	years /3 ago	30	0.0018	25	3.0941
256	year /3 date	30	0.0018	6	0.7426
257	who /3 smoked	30	0.0018	14	1.7327
258	which /3 can	30	0.0018	28	3.4653

259	ug /3 ml	30	0.0018	3	0.3713
260	there /3 any	30	0.0018	25	3.0941
261	their /3 brand	30	0.0018	17	2.104
262	smoke /3 exposed	30	0.0018	7	0.8663
263	six /3 months	30	0.0018	22	2.7228
264	please /3 know	30	0.0018	29	3.5891
265	per /3 cigarette	30	0.0018	22	2.7228
266	p /3 s	30	0.0018	13	1.6089
267	light /3 green	30	0.0018	2	0.2475
268	coronary /3 disease	30	0.0018	11	1.3614
269	burley /3 tobacco	30	0.0018	12	1.4851
270	body /3 weight	30	0.0018	14	1.7327
271	all /3 brands	30	0.0018	19	2.3515
272	5 /3 10	30	0.0018	21	2.599
273	27 /3 95	30	0.0018	2	0.2475
274	2 /3 2	30	0.0018	26	3.2178
275	18 /3 year	30	0.0018	11	1.3614
276	young /3 adult	29	0.0018	10	1.2376
277	urinary /3 nicotine	29	0.0018	2	0.2475
278	two /3 years	29	0.0018	22	2.7228
279	tobacco /3 which	29	0.0018	25	3.0941
280	than /3 one	29	0.0018	22	2.7228
281	task /3 force	29	0.0018	15	1.8564
282	tar /3 cigarettes	29	0.0018	18	2.2277
283	smoking /3 cancer	29	0.0018	19	2.3515
284	slow /3 72	29	0.0018	1	0.1238
285	short /3 term	29	0.0018	23	2.8465
286	see /3 attached	29	0.0018	26	3.2178
287	much /3 than	29	0.0018	26	3.2178
288	million /3 units	29	0.0018	4	0.495
289	male /3 female	29	0.0018	16	1.9802
290	low /3 high	29	0.0018	14	1.7327
291	full /3 price	29	0.0018	11	1.3614
292	environmental /3 tobacco	29	0.0018	19	2.3515
293	environmental /3 smoke	29	0.0018	19	2.3515
294	direct /3 mail	29	0.0018	17	2.104
295	cigarette /3 manufacturers	29	0.0018	13	1.6089
296	camel /3 lights	29	0.0018	6	0.7426
297	c /3 d	29	0.0018	20	2.4752
298	american /3 association	29	0.0018	22	2.7228
299	age /3 group	29	0.0018	11	1.3614
300	5 /3 6	29	0.0018	20	2.4752
301	who /3 smoking	28	0.0017	14	1.7327
302	viceroy /3 lights	28	0.0017	5	0.6188
303	table /3 3	28	0.0017	18	2.2277
304	surgeon /3 general	28	0.0017	18	2.2277
305	quality /3 control	28	0.0017	12	1.4851
306	per /3 year	28	0.0017	16	1.9802
307	p /3 o	28	0.0017	15	1.8564
308	nicotine /3 levels	28	0.0017	15	1.8564
309	more /3 likely	28	0.0017	20	2.4752
310	lights /3 smokers	28	0.0017	11	1.3614
311	lights /3 85	28	0.0017	7	0.8663
312	laser /3 photoacoustic	28	0.0017	2	0.2475
313	figure /3 1	28	0.0017	19	2.3515
314	each /3 year	28	0.0017	15	1.8564
315	don’t /3 know	28	0.0017	20	2.4752
316	division /3 manager	28	0.0017	10	1.2376
317	determine /3 if	28	0.0017	24	2.9703
318	d /3 d	28	0.0017	16	1.9802
319	can /3 used	28	0.0017	20	2.4752
320	both /3 sexes	28	0.0017	6	0.7426
321	18 /3 34	28	0.0017	13	1.6089
322	younger /3 adult	27	0.0017	10	1.2376
323	white /3 filter	27	0.0017	7	0.8663
324	those /3 which	27	0.0017	21	2.599
325	said /3 they	27	0.0017	10	1.2376
326	ring /3 filter	27	0.0017	1	0.1238

327	purpose /3 develop	27	0.0017	3	0.3713
328	public /3 service	27	0.0017	13	1.6089
329	public /3 places	27	0.0017	15	1.8564
330	new /3 brands	27	0.0017	16	1.9802
331	menthol /3 menthol	27	0.0017	18	2.2277
332	know /3 if	27	0.0017	25	3.0941
333	during /3 week	27	0.0017	23	2.8465
334	annual /3 meeting	27	0.0017	15	1.8564
335	about /3 about	27	0.0017	9	1.1139
336	3 /3 hydroxycotinine	27	0.0017	2	0.2475
337	wrap /3 paper	26	0.0016	1	0.1238
338	women /3 who	26	0.0016	12	1.4851
339	white /3 froeb	26	0.0016	3	0.3713
340	ultra /3 tar	26	0.0016	13	1.6089
341	these /3 cigarettes	26	0.0016	21	2.599
342	smoking /3 tobacco	26	0.0016	14	1.7327
343	shown /3 table	26	0.0016	16	1.9802
344	people /3 smoke	26	0.0016	17	2.104
345	past /3 years	26	0.0016	19	2.3515
346	nicotine /3 nicotine	26	0.0016	13	1.6089
347	ng /3 per	26	0.0016	1	0.1238
348	mr /3 mr	26	0.0016	7	0.8663
349	more /3 tobacco	26	0.0016	7	0.8663
350	mg /3 per	26	0.0016	15	1.8564
351	low /3 births	26	0.0016	1	0.1238
352	last /3 years	26	0.0016	21	2.599
353	information /3 about	26	0.0016	22	2.7228
354	if /3 there	26	0.0016	26	3.2178
355	hong /3 kong	26	0.0016	8	0.9901
356	do /3 so	26	0.0016	19	2.3515
357	competitive /3 brands	26	0.0016	16	1.9802
358	been /3 used	26	0.0016	24	2.9703
359	because /3 they	26	0.0016	19	2.3515
360	around /3 filter	26	0.0016	2	0.2475
361	activated /3 charcoal	26	0.0016	4	0.495
362	about /3 smoking	26	0.0016	19	2.3515
363	12 /3 months	26	0.0016	6	0.7426
364	12 /3 31	26	0.0016	9	1.1139
365	1 /3 ml	26	0.0016	6	0.7426
366	which /3 they	25	0.0015	23	2.8465
367	which /3 could	25	0.0015	23	2.8465
368	u /3 k	25	0.0015	12	1.4851
369	tobacco /3 tobacco	25	0.0015	14	1.7327
370	they /3 had	25	0.0015	20	2.4752
371	these /3 data	25	0.0015	21	2.599
372	test /3 results	25	0.0015	17	2.104
373	statistically /3 significant	25	0.0015	11	1.3614
374	smoking /3 machine	25	0.0015	15	1.8564
375	smoking /3 habits	25	0.0015	15	1.8564
376	smokers /3 nonsmokers	25	0.0015	15	1.8564
377	smoke /3 inhalation	25	0.0015	9	1.1139
378	shown /3 figure	25	0.0015	9	1.1139
379	set /3 forth	25	0.0015	18	2.2277
380	please /3 call	25	0.0015	25	3.0941
381	outer /3 paper	25	0.0015	1	0.1238
382	next /3 meeting	25	0.0015	16	1.9802
383	most /3 important	25	0.0015	21	2.599
384	more /3 one	25	0.0015	19	2.3515
385	middle /3 ear	25	0.0015	2	0.2475
386	market /3 share	25	0.0015	16	1.9802
387	know /3 what	25	0.0015	22	2.7228
388	inhalation /3 studies	25	0.0015	11	1.3614
389	indoor /3 quality	25	0.0015	11	1.3614
390	if /3 can	25	0.0015	21	2.599
391	how /3 much	25	0.0015	17	2.104
392	flavor /3 smokers	25	0.0015	10	1.2376
393	field /3 sales	25	0.0015	13	1.6089
394	e /3 c	25	0.0015	8	0.9901

395	down /3 points	25	0.0015	5	0.6188
396	don’t /3 think	25	0.0015	13	1.6089
397	correlation /3 between	25	0.0015	15	1.8564
398	cigarette /3 slow	25	0.0015	1	0.1238
399	cigarette /3 brands	25	0.0015	12	1.4851
400	being /3 made	25	0.0015	20	2.4752
401	among /3 year	25	0.0015	5	0.6188
402	after /3 exposure	25	0.0015	10	1.2376
403	about /3 10	25	0.0015	8	0.9901
404	7 /3 1	25	0.0015	17	2.104
405	3 /3 5	25	0.0015	22	2.7228
406	1 /3 mg	25	0.0015	14	1.7327
407	york /3 city	24	0.0015	18	2.2277
408	year /3 share	24	0.0015	5	0.6188
409	virginia /3 lights	24	0.0015	7	0.8663
410	two /3 weeks	24	0.0015	22	2.7228
411	they /3 don’t	24	0.0015	15	1.8564
412	smoking /3 pregnancy	24	0.0015	6	0.7426
413	slims /3 lights	24	0.0015	7	0.8663
414	significant /3 differences	24	0.0015	18	2.2277
415	s /3 p	24	0.0015	12	1.4851
416	results /3 obtained	24	0.0015	18	2.2277
417	reported /3 volume	24	0.0015	1	0.1238
418	quality /3 assurance	24	0.0015	15	1.8564
419	pack /3 carton	24	0.0015	15	1.8564
420	p /3 vol	24	0.0015	1	0.1238
421	nicotine /3 content	24	0.0015	16	1.9802
422	new /3 city	24	0.0015	18	2.2277
423	mg /3 nicotine	24	0.0015	14	1.7327
424	may /3 also	24	0.0015	21	2.599
425	marlboro /3 marlboro	24	0.0015	10	1.2376
426	make /3 sure	24	0.0015	17	2.104
427	main /3 claim	24	0.0015	2	0.2475
428	kool /3 kool	24	0.0015	10	1.2376
429	great /3 deal	24	0.0015	16	1.9802
430	golden /3 lights	24	0.0015	4	0.495
431	gas /3 chromatography	24	0.0015	14	1.7327
432	full /3 taste	24	0.0015	7	0.8663
433	flavor /3 generating	24	0.0015	1	0.1238
434	during /3 quarter	24	0.0015	13	1.6089
435	current /3 product	24	0.0015	12	1.4851
436	cigarette /3 smokers	24	0.0015	17	2.104
437	cancer /3 research	24	0.0015	14	1.7327
438	between /3 two	24	0.0015	19	2.3515
439	8 /3 9	24	0.0015	16	1.9802
440	4 /3 6	24	0.0015	19	2.3515
441	10 /3 95	24	0.0015	4	0.495
442	which /3 had	23	0.0014	18	2.2277
443	when /3 they	23	0.0014	21	2.599
444	tobacco /3 s	23	0.0014	13	1.6089
445	thank /3 your	23	0.0014	23	2.8465
446	soon /3 possible	23	0.0014	20	2.4752
447	range /3 finding	23	0.0014	1	0.1238
448	product /3 which	23	0.0014	17	2.104
449	product /3 development	23	0.0014	16	1.9802
450	pointed /3 out	23	0.0014	18	2.2277
451	per /3 week	23	0.0014	17	2.104
452	over /3 period	23	0.0014	19	2.3515
453	one /3 more	23	0.0014	19	2.3515
454	no /3 longer	23	0.0014	20	2.4752
455	no /3 1	23	0.0014	18	2.2277
456	more /3 more	23	0.0014	15	1.8564
457	medical /3 research	23	0.0014	16	1.9802
458	m /3 3	23	0.0014	4	0.495
459	lights /3 100mm	23	0.0014	4	0.495
460	kool /3 milds	23	0.0014	8	0.9901
461	journal /3 article	23	0.0014	3	0.3713
462	january /3 1	23	0.0014	17	2.104

463	heater /3 claim	23	0.0014	1	0.1238
464	filter /3 filter	23	0.0014	12	1.4851
465	filter /3 cigarettes	23	0.0014	19	2.3515
466	during /3 pregnancy	23	0.0014	4	0.495
467	control /3 cigarettes	23	0.0014	7	0.8663
468	confidential /3 minnesota	23	0.0014	6	0.7426
469	being /3 used	23	0.0014	19	2.3515
470	before /3 after	23	0.0014	13	1.6089
471	association /3 between	23	0.0014	14	1.7327
472	among /3 females	23	0.0014	8	0.9901
473	among /3 competitive	23	0.0014	12	1.4851
474	american /3 society	23	0.0014	16	1.9802
475	all /3 three	23	0.0014	21	2.599
476	all /3 groups	23	0.0014	14	1.7327
477	95 /3 amac	23	0.0014	1	0.1238
478	1 /3 per	23	0.0014	10	1.2376
479	years /3 age	22	0.0014	13	1.6089
480	two /3 groups	22	0.0014	13	1.6089
481	tobacco /3 u	22	0.0014	13	1.6089
482	three /3 years	22	0.0014	19	2.3515
483	they /3 like	22	0.0014	16	1.9802
484	there /3 evidence	22	0.0014	18	2.2277
485	than /3 tobacco	22	0.0014	10	1.2376
486	than /3 smokers	22	0.0014	15	1.8564
487	some /3 other	22	0.0014	19	2.3515
488	smokers /3 18	22	0.0014	9	1.1139
489	smoke /3 than	22	0.0014	16	1.9802
490	significant /3 difference	22	0.0014	13	1.6089
491	research /3 institute	22	0.0014	11	1.3614
492	rdr /3 no	22	0.0014	1	0.1238
493	present /3 invention	22	0.0014	2	0.2475
494	pesticide /3 residues	22	0.0014	5	0.6188
495	over /3 time	22	0.0014	16	1.9802
496	one /3 year	22	0.0014	16	1.9802
497	no /3 effect	22	0.0014	16	1.9802
498	mr /3 lewis	22	0.0014	1	0.1238
499	more /3 people	22	0.0014	10	1.2376
500	month /3 period	22	0.0014	7	0.8663

E.4.4 Top 500 Collocations Ranked by File Count

Rank	Collocation	Freq	%Total	Files	%Total
1	more /3 than	251	0.0154	133	16.4604
2	u /3 s	295	0.0181	113	13.9851
3	1 /3 2	149	0.0092	91	11.2624
4	there /3 no	124	0.0076	90	11.1386
5	if /3 any	93	0.0057	84	10.396
6	new /3 york	142	0.0087	79	9.7772
7	less /3 than	121	0.0074	78	9.6535
8	philip /3 morris	315	0.0194	76	9.4059
9	any /3 questions	65	0.004	62	7.6733
10	united /3 states	96	0.0059	61	7.5495
11	rather /3 than	82	0.005	61	7.5495
12	would /3 like	85	0.0052	60	7.4257
13	2 /3 3	96	0.0059	60	7.4257
14	tobacco /3 smoke	139	0.0085	59	7.302
15	e /3 g	88	0.0054	58	7.1782
16	cigarette /3 smoking	132	0.0081	58	7.1782
17	cigarette /3 smoke	117	0.0072	58	7.1782
18	r /3 d	131	0.0081	57	7.0545
19	r /3 j	122	0.0075	56	6.9307
20	had /3 been	79	0.0049	55	6.8069
21	please /3 me	56	0.0034	54	6.6832

22	tobacco /3 products	118	0.0073	52	6.4356
23	higher /3 than	84	0.0052	51	6.3119
24	3 /3 4	68	0.0042	51	6.3119
25	lung /3 cancer	127	0.0078	49	6.0644
26	tobacco /3 company	103	0.0063	48	5.9406
27	tobacco /3 industry	100	0.0061	46	5.6931
28	any /3 other	64	0.0039	46	5.6931
29	among /3 smokers	89	0.0055	46	5.6931
30	those /3 who	96	0.0059	45	5.5693
31	tar /3 nicotine	96	0.0059	45	5.5693
32	r /3 reynolds	107	0.0066	45	5.5693
33	other /3 than	52	0.0032	45	5.5693
34	low /3 tar	120	0.0074	45	5.5693
35	j /3 reynolds	106	0.0065	45	5.5693
36	smoking /3 health	96	0.0059	42	5.198
37	let /3 know	45	0.0028	42	5.198
38	tobacco /3 research	69	0.0042	40	4.9505
39	long /3 term	75	0.0046	40	4.9505
40	let /3 me	43	0.0026	40	4.9505
41	which /3 would	45	0.0028	39	4.8267
42	which /3 been	46	0.0028	39	4.8267
43	set /3 up	52	0.0032	38	4.703
44	been /3 made	46	0.0028	38	4.703
45	b /3 w	88	0.0054	38	4.703
46	1 /3 3	53	0.0033	38	4.703
47	1 /3 1	55	0.0034	38	4.703
48	under /3 conditions	51	0.0031	37	4.5792
49	there /3 been	42	0.0026	37	4.5792
50	differences /3 between	57	0.0035	37	4.5792
51	all /3 these	38	0.0023	37	4.5792
52	et /3 al	85	0.0052	36	4.4554
53	during /3 period	50	0.0031	36	4.4554
54	which /3 may	37	0.0023	35	4.3317
55	virginia /3 slims	144	0.0089	35	4.3317
56	these /3 two	39	0.0024	35	4.3317
57	non /3 smokers	145	0.0089	35	4.3317
58	your /3 letter	40	0.0025	34	4.2079
59	tobacco /3 institute	66	0.0041	34	4.2079
60	same /3 time	40	0.0025	34	4.2079
61	if /3 they	45	0.0028	34	4.2079
62	any /3 please	34	0.0021	34	4.2079
63	their /3 own	45	0.0028	33	4.0842
64	than /3 other	46	0.0028	33	4.0842
65	table /3 1	61	0.0037	33	4.0842
66	please /3 let	34	0.0021	33	4.0842
67	last /3 year	59	0.0036	33	4.0842
68	no /3 significant	57	0.0035	32	3.9604
69	me /3 know	33	0.002	32	3.9604
70	following /3 1	37	0.0023	32	3.9604
71	difference /3 between	39	0.0024	32	3.9604
72	questions /3 please	31	0.0019	31	3.8366
73	carbon /3 monoxide	76	0.0047	31	3.8366
74	test /3 market	56	0.0034	30	3.7129
75	lower /3 than	39	0.0024	30	3.7129
76	d /3 c	47	0.0029	30	3.7129
77	american /3 tobacco	66	0.0041	30	3.7129
78	4 /3 5	36	0.0022	30	3.7129
79	2 /3 1	42	0.0026	30	3.7129
80	they /3 would	38	0.0023	29	3.5891
81	reynolds /3 tobacco	51	0.0031	29	3.5891
82	please /3 know	30	0.0018	29	3.5891
83	only /3 one	43	0.0026	29	3.5891
84	brown /3 williamson	69	0.0042	29	3.5891
85	which /3 can	30	0.0018	28	3.4653
86	r /3 tobacco	47	0.0029	28	3.4653
87	p /3 m	63	0.0039	28	3.4653
88	no /3 difference	43	0.0026	28	3.4653
89	carried /3 out	43	0.0026	28	3.4653

90	they /3 do	36	0.0022	27	3.3416
91	smokers /3 who	50	0.0031	27	3.3416
92	new /3 products	49	0.003	27	3.3416
93	much /3 more	34	0.0021	27	3.3416
94	effects /3 smoking	41	0.0025	27	3.3416
95	but /3 they	32	0.002	27	3.3416
96	but /3 also	31	0.0019	27	3.3416
97	1 /3 4	42	0.0026	27	3.3416
98	winston /3 salem	36	0.0022	26	3.2178
99	very /3 much	32	0.002	26	3.2178
100	these /3 results	33	0.002	26	3.2178
101	smokers /3 smokers	59	0.0036	26	3.2178
102	see /3 attached	29	0.0018	26	3.2178
103	relationship /3 between	33	0.002	26	3.2178
104	people /3 who	50	0.0031	26	3.2178
105	new /3 product	59	0.0036	26	3.2178
106	much /3 than	29	0.0018	26	3.2178
107	mainstream /3 smoke	44	0.0027	26	3.2178
108	j /3 tobacco	45	0.0028	26	3.2178
109	if /3 there	26	0.0016	26	3.2178
110	greater /3 than	39	0.0024	26	3.2178
111	full /3 flavor	85	0.0052	26	3.2178
112	follow /3 up	31	0.0019	26	3.2178
113	better /3 than	32	0.002	26	3.2178
114	all /3 other	33	0.002	26	3.2178
115	2 /3 2	30	0.0018	26	3.2178
116	years /3 ago	30	0.0018	25	3.0941
117	what /3 they	36	0.0022	25	3.0941
118	tobacco /3 which	29	0.0018	25	3.0941
119	there /3 any	30	0.0018	25	3.0941
120	public /3 health	41	0.0025	25	3.0941
121	please /3 call	25	0.0015	25	3.0941
122	pall /3 mall	55	0.0034	25	3.0941
123	l /3 m	36	0.0022	25	3.0941
124	know /3 if	27	0.0017	25	3.0941
125	c /3 c	41	0.0025	25	3.0941
126	what /3 do	44	0.0027	24	2.9703
127	other /3 brands	45	0.0028	24	2.9703
128	no /3 differences	35	0.0022	24	2.9703
129	determine /3 if	28	0.0017	24	2.9703
130	competitive /3 smokers	67	0.0041	24	2.9703
131	been /3 used	26	0.0016	24	2.9703
132	anti /3 smoking	48	0.003	24	2.9703
133	who /3 smoke	38	0.0023	23	2.8465
134	which /3 they	25	0.0015	23	2.8465
135	which /3 could	25	0.0015	23	2.8465
136	thank /3 your	23	0.0014	23	2.8465
137	smoke /3 cigarettes	37	0.0023	23	2.8465
138	short /3 term	29	0.0018	23	2.8465
139	per /3 day	34	0.0021	23	2.8465
140	non /3 menthol	80	0.0049	23	2.8465
141	no /3 than	35	0.0022	23	2.8465
142	flue /3 cured	86	0.0053	23	2.8465
143	during /3 week	27	0.0017	23	2.8465
144	been /3 completed	34	0.0021	23	2.8465
145	two /3 years	29	0.0018	22	2.7228
146	two /3 weeks	24	0.0015	22	2.7228
147	than /3 one	29	0.0018	22	2.7228
148	table /3 2	37	0.0023	22	2.7228
149	six /3 months	30	0.0018	22	2.7228
150	per /3 cigarette	30	0.0018	22	2.7228
151	next /3 week	32	0.002	22	2.7228
152	know /3 what	25	0.0015	22	2.7228
153	information /3 about	26	0.0016	22	2.7228
154	first /3 time	22	0.0014	22	2.7228
155	cigarette /3 paper	38	0.0023	22	2.7228
156	american /3 association	29	0.0018	22	2.7228
157	3 /3 5	25	0.0015	22	2.7228

158	when /3 they	23	0.0014	21	2.599
159	those /3 which	27	0.0017	21	2.599
160	these /3 data	25	0.0015	21	2.599
161	these /3 cigarettes	26	0.0016	21	2.599
162	most /3 important	25	0.0015	21	2.599
163	may /3 also	24	0.0015	21	2.599
164	lucky /3 strike	66	0.0041	21	2.599
165	last /3 years	26	0.0016	21	2.599
166	last /3 week	32	0.002	21	2.599
167	if /3 can	25	0.0015	21	2.599
168	all /3 three	23	0.0014	21	2.599
169	adult /3 smokers	36	0.0022	21	2.599
170	5 /3 10	30	0.0018	21	2.599
171	2 /3 4	32	0.002	21	2.599
172	who /3 had	33	0.002	20	2.4752
173	vice /3 president	46	0.0028	20	2.4752
174	they /3 had	25	0.0015	20	2.4752
175	these /3 products	33	0.002	20	2.4752
176	soon /3 possible	23	0.0014	20	2.4752
177	per /3 pack	34	0.0021	20	2.4752
178	no /3 longer	23	0.0014	20	2.4752
179	more /3 likely	28	0.0017	20	2.4752
180	form /3 ula	47	0.0029	20	2.4752
181	final /3 report	31	0.0019	20	2.4752
182	don’t /3 know	28	0.0017	20	2.4752
183	can /3 used	28	0.0017	20	2.4752
184	c /3 d	29	0.0018	20	2.4752
185	being /3 made	25	0.0015	20	2.4752
186	5 /3 6	29	0.0018	20	2.4752
187	year /3 old	47	0.0029	19	2.3515
188	tobacco /3 taste	71	0.0044	19	2.3515
189	three /3 years	22	0.0014	19	2.3515
190	they /3 can	20	0.0012	19	2.3515
191	some /3 other	22	0.0014	19	2.3515
192	smoking /3 cancer	29	0.0018	19	2.3515
193	smoke /3 condensate	45	0.0028	19	2.3515
194	sidestream /3 smoke	37	0.0023	19	2.3515
195	sales /3 force	31	0.0019	19	2.3515
196	reynolds /3 company	34	0.0021	19	2.3515
197	public /3 smoking	39	0.0024	19	2.3515
198	past /3 years	26	0.0016	19	2.3515
199	over /3 period	23	0.0014	19	2.3515
200	one /3 more	23	0.0014	19	2.3515
201	north /3 carolina	42	0.0026	19	2.3515
202	more /3 one	25	0.0015	19	2.3515
203	me /3 if	19	0.0012	19	2.3515
204	marlboro /3 lights	69	0.0042	19	2.3515
205	how /3 can	21	0.0013	19	2.3515
206	he /3 said	35	0.0022	19	2.3515
207	first /3 quarter	57	0.0035	19	2.3515
208	filter /3 cigarettes	23	0.0014	19	2.3515
209	figure /3 1	28	0.0017	19	2.3515
210	environmental /3 tobacco	29	0.0018	19	2.3515
211	environmental /3 smoke	29	0.0018	19	2.3515
212	do /3 so	26	0.0016	19	2.3515
213	council /3 tobacco	35	0.0022	19	2.3515
214	council /3 research	35	0.0022	19	2.3515
215	cigarette /3 advertising	41	0.0025	19	2.3515
216	but /3 no	19	0.0012	19	2.3515
217	between /3 two	24	0.0015	19	2.3515
218	between /3 smoking	34	0.0021	19	2.3515
219	benson /3 hedges	50	0.0031	19	2.3515
220	being /3 used	23	0.0014	19	2.3515
221	because /3 they	26	0.0016	19	2.3515
222	all /3 smokers	32	0.002	19	2.3515
223	all /3 brands	30	0.0018	19	2.3515
224	about /3 smoking	26	0.0016	19	2.3515
225	6 /3 7	33	0.002	19	2.3515

226	4 /3 6	24	0.0015	19	2.3515
227	2 /3 5	20	0.0012	19	2.3515
228	york /3 city	24	0.0015	18	2.2277
229	which /3 had	23	0.0014	18	2.2277
230	there /3 some	19	0.0012	18	2.2277
231	there /3 evidence	22	0.0014	18	2.2277
232	tar /3 cigarettes	29	0.0018	18	2.2277
233	take /3 place	21	0.0013	18	2.2277
234	table /3 3	28	0.0017	18	2.2277
235	surgeon /3 general's	31	0.0019	18	2.2277
236	surgeon /3 general	28	0.0017	18	2.2277
237	significant /3 differences	24	0.0015	18	2.2277
238	should /3 noted	21	0.0013	18	2.2277
239	set /3 forth	25	0.0015	18	2.2277
240	results /3 obtained	24	0.0015	18	2.2277
241	research /3 which	20	0.0012	18	2.2277
242	pointed /3 out	23	0.0014	18	2.2277
243	no /3 1	23	0.0014	18	2.2277
244	new /3 city	24	0.0015	18	2.2277
245	menthol /3 menthol	27	0.0017	18	2.2277
246	m /3 d	42	0.0026	18	2.2277
247	j /3 company	32	0.002	18	2.2277
248	heart /3 disease	41	0.0025	18	2.2277
249	divided /3 into	22	0.0014	18	2.2277
250	which /3 he	17	0.001	17	2.104
251	what /3 would	19	0.0012	17	2.104
252	they /3 should	19	0.0012	17	2.104
253	there /3 significant	18	0.0011	17	2.104
254	there /3 little	18	0.0011	17	2.104
255	their /3 brand	30	0.0018	17	2.104
256	than /3 any	18	0.0011	17	2.104
257	test /3 results	25	0.0015	17	2.104
258	smokers /3 non	36	0.0022	17	2.104
259	smoke /3 exposure	47	0.0029	17	2.104
260	share /3 market	36	0.0022	17	2.104
261	research /3 program	20	0.0012	17	2.104
262	product /3 which	23	0.0014	17	2.104
263	point /3 sale	35	0.0022	17	2.104
264	per /3 week	23	0.0014	17	2.104
265	people /3 smoke	26	0.0016	17	2.104
266	particulate /3 matter	31	0.0019	17	2.104
267	other /3 factors	18	0.0011	17	2.104
268	next /3 steps	20	0.0012	17	2.104
269	new /3 cigarette	32	0.002	17	2.104
270	mg /3 tar	31	0.0019	17	2.104
271	meeting /3 held	19	0.0012	17	2.104
272	make /3 sure	24	0.0015	17	2.104
273	look /3 forward	17	0.001	17	2.104
274	january /3 1	23	0.0014	17	2.104
275	indoor /3 air	49	0.003	17	2.104
276	if /3 could	18	0.0011	17	2.104
277	how /3 much	25	0.0015	17	2.104
278	high /3 level	21	0.0013	17	2.104
279	gas /3 phase	38	0.0023	17	2.104
280	follows /3 1	19	0.0012	17	2.104
281	each /3 group	31	0.0019	17	2.104
282	direct /3 mail	29	0.0018	17	2.104
283	cigarettes /3 made	31	0.0019	17	2.104
284	cigarette /3 smokers	24	0.0015	17	2.104
285	but /3 do	17	0.001	17	2.104
286	board /3 directors	47	0.0029	17	2.104
287	between /3 smokers	18	0.0011	17	2.104
288	been /3 reported	21	0.0013	17	2.104
289	been /3 all	17	0.001	17	2.104
290	based /3 upon	21	0.0013	17	2.104
291	american /3 company	36	0.0022	17	2.104
292	already /3 been	17	0.001	17	2.104
293	all /3 cigarettes	18	0.0011	17	2.104

294	7 /3 1	25	0.0015	17	2.104
295	3 /3 1	21	0.0013	17	2.104
296	1 /3 5	21	0.0013	17	2.104
297	year /3 period	61	0.0037	16	1.9802
298	would /3 if	17	0.001	16	1.9802
299	work /3 done	20	0.0012	16	1.9802
300	washington /3 d	20	0.0012	16	1.9802
301	very /3 low	19	0.0012	16	1.9802
302	ultra /3 low	37	0.0023	16	1.9802
303	tobacco /3 use	38	0.0023	16	1.9802
304	tobacco /3 blend	21	0.0013	16	1.9802
305	they /3 their	19	0.0012	16	1.9802
306	they /3 like	22	0.0014	16	1.9802
307	these /3 studies	18	0.0011	16	1.9802
308	so /3 can	18	0.0011	16	1.9802
309	smoke /3 than	22	0.0014	16	1.9802
310	shown /3 table	26	0.0016	16	1.9802
311	questions /3 about	17	0.001	16	1.9802
312	product /3 development	23	0.0014	16	1.9802
313	please /3 contact	19	0.0012	16	1.9802
314	period /3 time	19	0.0012	16	1.9802
315	per /3 year	28	0.0017	16	1.9802
316	passive /3 smoking	67	0.0041	16	1.9802
317	over /3 years	17	0.001	16	1.9802
318	over /3 time	22	0.0014	16	1.9802
319	our /3 own	18	0.0011	16	1.9802
320	one /3 year	22	0.0014	16	1.9802
321	one /3 two	19	0.0012	16	1.9802
322	one /3 one	21	0.0013	16	1.9802
323	no /3 other	16	0.001	16	1.9802
324	no /3 effect	22	0.0014	16	1.9802
325	no /3 between	20	0.0012	16	1.9802
326	nicotine /3 content	24	0.0015	16	1.9802
327	next /3 meeting	25	0.0015	16	1.9802
328	new /3 brands	27	0.0017	16	1.9802
329	n /3 c	31	0.0019	16	1.9802
330	medical /3 research	23	0.0014	16	1.9802
331	market /3 share	25	0.0015	16	1.9802
332	male /3 female	29	0.0018	16	1.9802
333	let /3 us	16	0.001	16	1.9802
334	let /3 if	16	0.001	16	1.9802
335	least /3 one	19	0.0012	16	1.9802
336	if /3 would	17	0.001	16	1.9802
337	if /3 had	20	0.0012	16	1.9802
338	how /3 they	21	0.0013	16	1.9802
339	great /3 deal	24	0.0015	16	1.9802
340	five /3 year	20	0.0012	16	1.9802
341	female /3 smokers	41	0.0025	16	1.9802
342	during /3 first	17	0.001	16	1.9802
343	do /3 think	33	0.002	16	1.9802
344	do /3 smoke	16	0.001	16	1.9802
345	d /3 d	28	0.0017	16	1.9802
346	competitive /3 brands	26	0.0016	16	1.9802
347	cigarettes /3 tobacco	20	0.0012	16	1.9802
348	call /3 me	16	0.001	16	1.9802
349	appreciate /3 your	17	0.001	16	1.9802
350	any /3 information	16	0.001	16	1.9802
351	american /3 society	23	0.0014	16	1.9802
352	all /3 tobacco	18	0.0011	16	1.9802
353	about /3 1	22	0.0014	16	1.9802
354	8 /3 9	24	0.0015	16	1.9802
355	would /3 more	18	0.0011	15	1.8564
356	would /3 appear	21	0.0013	15	1.8564
357	when /3 he	15	0.0009	15	1.8564
358	washington /3 c	19	0.0012	15	1.8564
359	tobacco /3 companies	37	0.0023	15	1.8564
360	they /3 smoke	18	0.0011	15	1.8564
361	they /3 don’t	24	0.0015	15	1.8564

362	these /3 would	16	0.001	15	1.8564
363	than /3 smokers	22	0.0014	15	1.8564
364	task /3 force	29	0.0018	15	1.8564
365	so /3 far	19	0.0012	15	1.8564
366	smoking /3 smoking	20	0.0012	15	1.8564
367	smoking /3 machine	25	0.0015	15	1.8564
368	smoking /3 habits	25	0.0015	15	1.8564
369	smokers /3 nonsmokers	25	0.0015	15	1.8564
370	results /3 indicate	18	0.0011	15	1.8564
371	research /3 development	19	0.0012	15	1.8564
372	research /3 center	21	0.0013	15	1.8564
373	quality /3 assurance	24	0.0015	15	1.8564
374	public /3 places	27	0.0017	15	1.8564
375	please /3 find	15	0.0009	15	1.8564
376	per /3 carton	20	0.0012	15	1.8564
377	pack /3 carton	24	0.0015	15	1.8564
378	p /3 o	28	0.0017	15	1.8564
379	other /3 such	15	0.0009	15	1.8564
380	other /3 hand	18	0.0011	15	1.8564
381	other /3 groups	21	0.0013	15	1.8564
382	only /3 1	21	0.0013	15	1.8564
383	non /3 smoking	16	0.001	15	1.8564
384	no /3 evidence	19	0.0012	15	1.8564
385	nicotine /3 levels	28	0.0017	15	1.8564
386	new /3 brand	35	0.0022	15	1.8564
387	near /3 future	15	0.0009	15	1.8564
388	national /3 institute	20	0.0012	15	1.8564
389	most /3 likely	18	0.0011	15	1.8564
390	more /3 more	23	0.0014	15	1.8564
391	mg /3 per	26	0.0016	15	1.8564
392	marlboro /3 smokers	39	0.0024	15	1.8564
393	many /3 other	15	0.0009	15	1.8564
394	low /3 nicotine	22	0.0014	15	1.8564
395	last /3 two	16	0.001	15	1.8564
396	king /3 size	37	0.0023	15	1.8564
397	health /3 research	19	0.0012	15	1.8564
398	five /3 years	18	0.0011	15	1.8564
399	even /3 more	19	0.0012	15	1.8564
400	even /3 if	19	0.0012	15	1.8564
401	each /3 year	28	0.0017	15	1.8564
402	data /3 obtained	18	0.0011	15	1.8564
403	cured /3 tobacco	33	0.002	15	1.8564
404	correlation /3 between	25	0.0015	15	1.8564
405	cigarettes /3 smoked	18	0.0011	15	1.8564
406	carbon /3 dioxide	47	0.0029	15	1.8564
407	but /3 only	16	0.001	15	1.8564
408	been /3 found	16	0.001	15	1.8564
409	been /3 established	15	0.0009	15	1.8564
410	annual /3 meeting	27	0.0017	15	1.8564
411	all /3 materials	19	0.0012	15	1.8564
412	all /3 information	18	0.0011	15	1.8564
413	all /3 data	21	0.0013	15	1.8564
414	all /3 cigarette	19	0.0012	15	1.8564
415	air /3 quality	37	0.0023	15	1.8564
416	air /3 pollution	33	0.002	15	1.8564
417	4 /3 1	17	0.001	15	1.8564
418	18 /3 24	38	0.0023	15	1.8564
419	1 /3 no	20	0.0012	15	1.8564
420	1 /3 all	15	0.0009	15	1.8564
421	1 /3 10	34	0.0021	15	1.8564
422	would /3 appreciate	14	0.0009	14	1.7327
423	who /3 smoking	28	0.0017	14	1.7327
424	who /3 smoked	30	0.0018	14	1.7327
425	who /3 do	16	0.001	14	1.7327
426	which /3 only	14	0.0009	14	1.7327
427	very /3 little	15	0.0009	14	1.7327
428	very /3 important	14	0.0009	14	1.7327
429	use /3 tobacco	19	0.0012	14	1.7327

430	up /3 date	14	0.0009	14	1.7327
431	tobacco /3 tobacco	25	0.0015	14	1.7327
432	tobacco /3 cigarettes	20	0.0012	14	1.7327
433	tobacco /3 been	16	0.001	14	1.7327
434	they /3 want	18	0.0011	14	1.7327
435	they /3 did	19	0.0012	14	1.7327
436	they /3 also	16	0.001	14	1.7327
437	these /3 should	16	0.001	14	1.7327
438	these /3 may	14	0.0009	14	1.7327
439	than /3 those	18	0.0011	14	1.7327
440	than /3 10	15	0.0009	14	1.7327
441	talk /3 about	16	0.001	14	1.7327
442	so /3 they	15	0.0009	14	1.7327
443	so /3 much	14	0.0009	14	1.7327
444	smoking /3 tobacco	26	0.0016	14	1.7327
445	smoking /3 lung	18	0.0011	14	1.7327
446	smoking /3 but	17	0.001	14	1.7327
447	smoke /3 cigarette	16	0.001	14	1.7327
448	significantly /3 than	21	0.0013	14	1.7327
449	see /3 table	39	0.0024	14	1.7327
450	said /3 he	19	0.0012	14	1.7327
451	pick /3 up	17	0.001	14	1.7327
452	ph /3 d	35	0.0022	14	1.7327
453	our /3 brands	20	0.0012	14	1.7327
454	one /3 these	14	0.0009	14	1.7327
455	one /3 per	14	0.0009	14	1.7327
456	no /3 no	33	0.002	14	1.7327
457	no /3 more	16	0.001	14	1.7327
458	next /3 year	20	0.0012	14	1.7327
459	new /3 england	18	0.0011	14	1.7327
460	mg /3 nicotine	24	0.0015	14	1.7327
461	method /3 used	15	0.0009	14	1.7327
462	made /3 available	14	0.0009	14	1.7327
463	low /3 high	29	0.0018	14	1.7327
464	los /3 angeles	17	0.001	14	1.7327
465	its /3 own	14	0.0009	14	1.7327
466	into /3 two	14	0.0009	14	1.7327
467	into /3 account	16	0.001	14	1.7327
468	incorporated /3 into	17	0.001	14	1.7327
469	if /3 one	17	0.001	14	1.7327
470	if /3 all	14	0.0009	14	1.7327
471	health /3 smoking	17	0.001	14	1.7327
472	he /3 would	16	0.001	14	1.7327
473	gas /3 chromatography	24	0.0015	14	1.7327
474	filter /3 cigarette	18	0.0011	14	1.7327
475	few /3 years	17	0.001	14	1.7327
476	exposure /3 smoke	32	0.002	14	1.7327
477	even /3 though	16	0.001	14	1.7327
478	each /3 these	15	0.0009	14	1.7327
479	do /3 they	16	0.001	14	1.7327
480	development /3 new	15	0.0009	14	1.7327
481	cigarettes /3 per	19	0.0012	14	1.7327
482	cigarettes /3 been	16	0.001	14	1.7327
483	cancer /3 research	24	0.0015	14	1.7327
484	cancer /3 institute	17	0.001	14	1.7327
485	body /3 weight	30	0.0018	14	1.7327
486	been /3 received	16	0.001	14	1.7327
487	been /3 done	16	0.001	14	1.7327
488	because /3 its	15	0.0009	14	1.7327
489	b /3 h	36	0.0022	14	1.7327
490	association /3 between	23	0.0014	14	1.7327
491	answer /3 questions	15	0.0009	14	1.7327
492	among /3 all	15	0.0009	14	1.7327
493	all /3 groups	23	0.0014	14	1.7327
494	additional /3 information	14	0.0009	14	1.7327
495	3 /3 2	18	0.0011	14	1.7327
496	10 /3 12	19	0.0012	14	1.7327
497	10 /3 1	17	0.001	14	1.7327

498	1 /3 mg	25	0.0015	14	1.7327
499	years /3 age	22	0.0014	13	1.6089
500	would /3 make	14	0.0009	13	1.6089

E.5 Quota Sample Comparison Data

E.5.1 Top 400 and Bottom 100 Tokens Ranked by Frequency Z-score

Rank	Token	Freq-Z	Freq-V	File-Z	File-V
1	tobacco	87.323	1	22.208	1
2	smoking	73.206	1	16.985	1
3	cigarette	67.372	1	17.038	1
4	smokers	66.307	1	16.008	1
5	smoke	59.818	1	12.31	1
6	cigarettes	57.965	1	17.204	1
7	nicotine	55.68	1	13.952	1
8	brand	49.293	1	10.179	1
9	product	46.628	1	9.333	1
10	filter	45.818	1	11.761	1
11	brands	45.547	1	12.423	1
12	test	44.644	1	6.07	1
13	menthol	44.529	1	10.89	1
14	1	44.484	1	6.548	1
15	marlboro	41.468	1	10.307	1
16	3	40.098	1	6.145	1
17	2	39.625	1	5.632	1
18	tar	39.373	1	11.104	1
19	camel	38.633	1	8.038	1
20	will	38.439	1	-7.717	1
21	research	37.778	1	5.442	1
22	advertising	37.664	1	8.146	1
23	products	37.619	1	8.864	1
24	4	36.876	1	6.712	1
25	pack	35.683	1	8.065	1
26	results	35.636	1	6.439	1
27	5	34.8	1	6.332	1
28	lights	33.221	1	2.069	1
29	samples	33.075	1	9.572	1
30	share	32.928	1	-3.329	1
31	rjr	32.483	1	9.207	1
32	cancer	32.46	1	5.961	1
33	morris	32.269	1	7.253	1
34	exposure	32.235	1	5.476	1
35	sales	31.289	1	3.897	1
36	smoker	31.274	1	10.609	1
37	philip	31.077	1	6.487	1
38	data	31.027	1	6.702	1
39	6	30.695	1	4.915	1
40	flavor	30.408	1	7.636	1
41	95	30.078	1	2.059	1
42	r	29.76	1	5.334	1
43	lung	29.397	1	7.904	1
44	carton	29.26	1	9.207	1
45	10	28.952	1	3.745	1
46	kool	28.378	1	6.981	1
47	market	28.024	1	2.527	1
48	ml	27.861	1	4.708	1
49	testing	27.842	1	6.425	1
50	ets	27.744	1	5.824	1
51	8	27.708	1	5.086	1
52	d	27.593	1	2.904	1
53	blend	27.459	1	7.536	1

54	c	27.449	1	2.921	1
55	retail	27.407	1	6.621	1
56	winston	27.339	1	8.415	1
57	sample	27.226	1	6.252	1
58	non	27.154	1	2.199	1
59	dr	26.602	1	2.315	1
60	12	26.43	1	2.47	1
61	promotion	25.989	1	5.445	1
62	levels	25.917	1	3.213	1
63	studies	25.713	1	3.608	1
64	b	25.612	1	3.566	1
65	reynolds	25.568	1	7.78	1
66	salem	25.341	1	7.468	1
67	mg	24.966	1	8.337	1
68	low	24.939	1	-2.654	1
69	report	24.781	1	3.963	1
70	7	24.709	1	4.29	1
71	be	24.489	1	-13.064	1
72	please	24.427	1	6.621	1
73	viceroy	24.394	1	4.551	1
74	analysis	23.957	1	2.654	1
75	slims	23.811	1	7.024	1
76	ks	23.651	1	4.411	1
77	kent	23.437	1	4.419	1
78	pm	23.373	1	7.291	1
79	100	23.35	1	2.021	1
80	per	23.018	1	2.241	1
81	marketing	22.971	1	7.332	1
82	100’s	22.803	1	7.518	1
83	competitive	22.547	1	4.131	1
84	attached	22.259	1	9.182	1
85	consumer	22.149	1	7.913	1
86	date	22.08	1	4.227	1
87	delivery	21.985	1	4.711	1
88	g	21.935	1	2.045	1
89	burley	21.878	1	6.981	1
90	segment	21.52	1	2.675	1
91	ultra	21.28	1	3.542	1
92	18	21.245	1	3.157	1
93	rats	21.215	1	4.028	1
94	project	21.196	1	3.019	1
95	cotinine	20.96	1	4.744	1
96	versus	20.892	1	4.832	1
97	11	20.781	1	2.541	1
98	sampling	20.756	1	6.154	1
99	copy	20.614	1	6.582	1
100	following	20.602	1	2.386	1
101	discount	20.526	1	2.396	1
102	30	20.459	1	2.39	1
103	smoked	20.398	1	6.986	1
104	carbon	20.376	1	5.523	1
105	prior	20.369	1	4.384	1
106	inhalation	20.367	1	6.601	1
107	packs	20.249	1	7.026	1
108	promotions	20.238	1	7.112	1
109	9	20.09	1	4.701	1
110	tobaccos	19.95	1	7.728	1
111	summary	19.776	1	7.854	1
112	iii	19.589	1	6.433	1
113	nonsmokers	19.57	1	5.933	1
114	lorillard	19.474	1	6.837	1
115	materials	19.466	1	2.922	1
116	group	19.435	1	-5.642	1
117	respondents	19.418	1	3.116	1
118	cured	19.403	1	5.049	1
119	tareyton	19.377	1	4.744	1
120	ventilation	19.372	1	4.062	1
121	packaging	19.346	1	6.123	1

122	cartons	19.28	1	6.762	1
123	evaluation	19.152	1	5.945	1
124	review	19.087	1	2.537	1
125	media	18.99	1	3.258	1
126	13	18.829	1	1.987	1
127	15	18.816	1	2.073	1
128	phase	18.794	1	3.18	1
129	experimental	18.744	1	3.772	1
130	franchise	18.713	1	4.589	1
131	tested	18.707	1	5.433	1
132	ads	18.696	1	3.291	1
133	puff	18.687	1	5.852	1
134	w	18.668	1	2.16	1
135	overall	18.473	1	4.131	1
136	carlton	18.294	1	5.504	1
137	flue	18.177	1	5.601	1
138	doral	18.177	1	6.04	1
139	paper	18.062	1	-2.034	1
140	condensate	17.97	1	5.487	1
141	concentrations	17.919	1	3.975	1
142	100mm	17.759	1	4.47	1
143	coupon	17.705	1	6.601	1
144	analytical	17.557	1	6.776	1
145	filters	17.52	1	6.464	1
146	moisture	17.482	1	3.202	1
147	leaf	17.434	1	4.385	1
148	prototype	17.33	1	3.075	1
149	institute	17.271	1	3.382	1
150	compounds	17.264	1	6.292	1
151	tests	17.149	1	3.265	1
152	ad	17.101	1	3.718	1
153	mm	16.988	1	5.273	1
154	scheduled	16.961	1	2.072	1
155	ammonia	16.947	1	4.686	1
156	1994	16.845	1	4.91	1
157	pos	16.781	1	5.003	1
158	assay	16.734	1	5.044	1
159	monoxide	16.724	1	5.991	1
160	fda	16.688	1	2.346	1
161	additional	16.678	1	3.217	1
162	85	16.597	1	2.346	1
163	table	16.545	1	-6.546	1
164	this	16.458	1	-13.78	1
165	coupons	16.444	1	5.743	1
166	groups	16.407	1	-2.033	1
167	24	16.39	1	2.491	1
168	7	16.382	1	3.681	1
169	package	16.304	1	2.758	1
170	among	16.297	1	-11.808	1
171	inc	16.245	1	5.3	1
172	conducted	16.198	1	3.939	1
173	level	16.117	1	-2.414	1
174	1995	16.117	1	3.066	1
175	ii	16.067	1	3.564	1
176	promotional	15.98	1	5.727	1
177	ftc	15.979	1	5.601	1
178	nitrogen	15.969	1	3.61	1
179	air	15.961	1	-6.712	1
180	consumers	15.855	1	4.203	1
181	purchase	15.839	1	3.23	1
182	units	15.81	1	-2.555	1
183	use	15.779	1	-7.435	1
184	requested	15.756	1	5.621	1
185	january	15.709	1	2.697	1
186	analyses	15.602	1	4.224	1
187	chemical	15.497	1	3.111	1
188	july	15.473	1	2.316	1
189	blends	15.446	1	4.586	1

190	obtained	15.444	1	2.723	1
191	category	15.409	1	2.012	1
192	evaluated	15.406	1	4.876	1
193	laboratory	15.289	1	4.827	1
194	vs	15.271	1	3.263	1
195	toxicity	15.195	1	5.196	1
196	objectives	15.193	1	3.153	1
197	submitted	15.165	1	5.288	1
198	currently	15.153	1	2.467	1
199	are	15.145	1	-11.236	1
200	metabolism	15.138	1	4.791	1
201	control	15.08	1	-4.031	1
202	approximately	15.075	1	2.201	1
203	comments	15.061	1	4.04	1
204	tipping	14.996	1	4.79	1
205	period	14.962	1	-4.309	1
206	filtration	14.961	1	5.073	1
207	camels	14.961	1	2.56	1
208	16	14.898	1	2.326	1
209	respiratory	14.827	1	5.322	1
210	week	14.791	1	-4.197	1
211	outlets	14.766	1	4.372	1
212	findings	14.69	1	3.188	1
213	panel	14.631	1	2.241	1
214	evaluate	14.615	1	4.224	1
215	used	14.594	1	-6.762	1
216	gc	14.581	1	4.686	1
217	prototypes	14.564	1	3.639	1
218	distribution	14.562	1	3.491	1
219	17	14.497	1	2.76	1
220	increased	14.483	1	-2.014	1
221	incidence	14.468	1	4.516	1
222	92	14.467	1	2.106	1
223	19	14.465	1	2.496	1
224	acid	14.449	1	2.724	1
225	williamson	14.44	1	5.156	1
226	ph	14.332	1	3.362	1
227	31	14.296	1	2.731	1
228	for	14.187	1	-9.772	1
229	filler	14.173	1	4.686	1
230	higher	14.138	1	-2.519	1
231	sidestream	14.107	1	5.824	1
232	urine	14.1	1	3.259	1
233	particulate	14.059	1	4.522	1
234	35	14.013	1	2.068	1
235	determine	14.01	1	2.209	1
236	charcoal	13.978	1	2.441	1
237	request	13.972	1	2.998	1
238	newport	13.927	1	3.116	1
239	kg	13.838	1	5.128	1
240	hedges	13.79	1	4.103	1
241	residues	13.753	1	2.361	1
242	offer	13.731	1	-3.922	1
243	sheet	13.691	1	2.442	1
244	cells	13.636	1	2.383	1
245	pall	13.606	1	4.533	1
246	co2	13.579	1	4.245	1
247	barclay	13.572	1	4.686	1
248	mall	13.55	1	4.155	1
249	methodology	13.525	1	5.049	1
250	1996	13.525	1	3.484	1
251	attributes	13.442	1	2.81	1
252	90	13.408	1	2.056	1
253	28	13.408	1	3.291	1
254	reduction	13.383	1	2.453	1
255	specifications	13.35	1	4.339	1
256	ppm	13.301	1	4.103	1
257	acceptable	13.271	1	2.668	1

258	nitrate	13.249	1	3.628	1
259	effect	13.2	1	-5.038	1
260	significantly	13.162	1	2.827	1
261	during	13.13	1	-9.126	1
262	retailers	13.124	1	4.711	1
263	launch	13.115	1	2.051	1
264	extraction	13.109	1	3.171	1
265	surgeon	13.098	1	3.963	1
266	1	13.096	1	2.757	1
267	91	13.089	1	2.27	1
268	merchandising	13.048	1	4.111	1
269	warning	13.04	1	-2.385	1
270	s	13.035	1	-3.16	1
271	shipped	13.024	1	4.542	1
272	cell	12.991	1	2.817	1
273	flavors	12.967	1	3.42	1
274	reps	12.952	1	2.56	1
275	indoor	12.946	1	2.403	1
276	confidential	12.942	1	3.234	1
277	presentation	12.883	1	2.866	1
278	tpm	12.853	1	4.179	1
279	rjrt	12.853	1	4.026	1
280	treated	12.84	1	-2.407	1
281	1981	12.8	1	3.118	1
282	ctr	12.706	1	4.026	1
283	1st	12.612	1	4.525	1
284	ula	12.558	1	5.003	1
285	29	12.552	1	2.26	1
286	etc	12.541	1	3.164	1
287	further	12.536	1	-3.194	1
288	min	12.529	1	4.017	1
289	nm	12.502	1	4.525	1
290	addition	12.433	1	-2.206	1
291	ca	12.408	1	4.093	1
292	brand’s	12.407	1	4.744	1
293	iv	12.393	1	5.037	1
294	draft	12.353	1	2.571	1
295	5	12.311	1	3.566	1
296	process	12.287	1	-5.111	1
297	liggett	12.255	1	4.609	1
298	34	12.244	1	2.241	1
299	our	12.216	1	-7.956	1
300	parity	12.21	1	3.483	1
301	calibration	12.179	1	3.803	1
302	cc	12.162	1	4.791	1
303	each	12.112	1	-13.436	1
304	regarding	12.083	1	3.421	1
305	article	12.019	1	-2.624	1
306	support	12.004	1	-2.874	1
307	vending	11.931	1	4.098	1
308	sugars	11.901	1	3.46	1
309	enclosed	11.837	1	4.888	1
310	83	11.821	1	2.762	1
311	2nd	11.811	1	4.124	1
312	deluxe	11.743	1	2.906	1
313	ms	11.729	1	3.566	1
314	uk	11.708	1	3.466	1
315	feasibility	11.695	1	3.681	1
316	guidelines	11.681	1	2.314	1
317	dioxide	11.678	1	2.593	1
318	casing	11.654	1	3.981	1
319	month	11.629	1	-3.936	1
320	analyzed	11.58	1	2.376	1
321	manufacture	11.518	1	4.01	1
322	modifications	11.514	1	3.116	1
323	irritation	11.514	1	2.403	1
324	wynder	11.464	1	3.867	1
325	memorandum	11.461	1	5.388	1

326	medium	11.441	1	-2.097	1
327	reviewed	11.432	1	2.685	1
328	store	11.409	1	-1.963	1
329	performance	11.389	1	-2.548	1
330	deliveries	11.374	1	4.654	1
331	constituents	11.364	1	3.121	1
332	should	11.311	1	-12.302	1
333	number	11.264	1	-7.18	1
334	usa	11.263	1	3.304	1
335	status	11.261	1	-4.593	1
336	retailer	11.258	1	4.386	1
337	page	11.231	1	-2.647	1
338	qualitative	11.221	1	4.093	1
339	profile	11.198	1	2.902	1
340	mainstream	11.194	1	2.579	1
341	section	11.18	1	-3.49	1
342	distributors	11.137	1	3.171	1
343	final	11.135	1	-2.748	1
344	v	11.111	1	2.402	1
345	positioning	11.11	1	4.293	1
346	excise	11.11	1	3.194	1
347	coronary	11.11	1	3.46	1
348	below	11.104	1	-2.544	1
349	alveolar	11.073	1	2.701	1
350	harshness	11.049	1	3.805	1
351	approx	11.049	1	3.639	1
352	statistical	11.03	1	2.02	1
353	compound	10.997	1	2.81	1
354	plan	10.979	1	-3.494	1
355	shipment	10.971	1	4.33	1
356	pyrolysis	10.961	1	4.327	1
357	packings	10.961	1	4.026	1
358	preliminary	10.941	1	3.015	1
359	methods	10.911	1	-2.265	1
360	these	10.909	1	-13.053	1
361	ti	10.893	1	3.362	1
362	size	10.855	1	-3.317	1
363	various	10.831	1	-3.868	1
364	revised	10.823	1	2.65	1
365	booklet	10.808	1	3.791	1
366	birth	10.789	1	-6.512	1
367	cost	10.743	1	-2.447	1
368	harmful	10.731	1	2.828	1
369	respondent	10.679	1	3.284	1
370	key	10.674	1	-3.463	1
371	diseases	10.649	1	3.482	1
372	similar	10.641	1	-3.444	1
373	average	10.614	1	-4.05	1
374	raleigh	10.613	1	4.327	1
375	we	10.587	1	-11.021	1
376	additive	10.576	1	4.411	1
377	humidity	10.569	1	2.535	1
378	carcinogenic	10.491	1	2.92	1
379	modified	10.471	1	2.02	1
380	solvent	10.438	1	3.483	1
381	aftertaste	10.435	1	4.179	1
382	respectively	10.432	1	2.029	1
383	labels	10.411	1	2.762	1
384	generic	10.411	1	3.404	1
385	workplace	10.402	1	2.868	1
386	memo	10.402	1	4.623	1
387	additionally	10.4	1	3.914	1
388	pricing	10.399	1	2.485	1
389	flavoring	10.399	1	3.284	1
390	preference	10.364	1	2.169	1
391	cancers	10.353	1	2.119	1
392	rjr's	10.348	1	4.411	1
393	chronic	10.337	1	3.146	1

394	type	10.292	1	-3.433	1
395	cardiovascular	10.277	1	4.245	1
396	introductory	10.264	1	2.593	1
397	invention	10.263	1	-2.095	1
398	exposures	10.237	1	2.704	1
399	benson	10.237	1	3.33	1
400	decreased	10.165	1	2.699	1
100	over	-9.875	1	-25.12	1
99	bush	-9.881	1	-7.066	1
98	boy	-9.881	1	-12.872	1
97	across	-9.915	1	-16.555	1
96	seemed	-10.138	1	-16.701	1
95	nothing	-10.153	1	-18.307	1
94	says	-10.189	1	-13.018	1
93	long	-10.193	1	-24.116	1
92	get	-10.222	1	-18.255	1
91	moment	-10.237	1	-16.061	1
90	political	-10.271	1	-13.7	1
89	once	-10.48	1	-19.952	1
88	woman	-10.566	1	-14.587	1
87	told	-10.582	1	-16.104	1
86	hands	-10.611	1	-15.757	1
85	formula	-10.64	1	-3.912	1
84	something	-10.678	1	-19.403	1
83	mother	-10.78	1	-12.99	1
82	love	-10.817	1	-13.978	1
81	big	-10.825	1	-16.769	1
80	toward	-10.969	1	-18.396	1
79	school	-10.987	1	-13.329	1
78	head	-11.044	1	-17.003	1
77	little	-11.101	1	-21.133	1
76	about	-11.126	1	-24.958	1
75	did	-11.132	1	-20.63	1
74	door	-11.153	1	-14.288	1
73	another	-11.176	1	-23.03	1
72	go	-11.206	1	-20.484	1
71	power	-11.311	1	-15.842	1
70	house	-11.378	1	-15.982	1
69	turned	-11.388	1	-17.928	1
68	old	-11.394	1	-20.458	1
67	got	-11.396	1	-17.959	1
66	men	-11.543	1	-18.161	1
65	still	-11.554	1	-22.792	1
64	left	-11.717	1	-21.045	1
63	father	-11.731	1	-14.024	1
62	church	-11.817	1	-11.067	1
61	night	-11.87	1	-17.243	1
60	saw	-11.91	1	-17.386	1
59	though	-12.008	1	-20.67	1
58	here	-12.059	1	-21.422	1
57	always	-12.065	1	-20.646	1
56	face	-12.075	1	-18.135	1
55	come	-12.086	1	-21.813	1
54	mrs	-12.15	1	-9.931	1
53	god	-12.201	1	-13.182	1
52	eyes	-12.388	1	-16.556	1
51	away	-12.481	1	-20.42	1
50	thought	-12.591	1	-19.932	1
49	home	-12.655	1	-20.071	1
48	own	-12.798	1	-23.656	1
47	took	-12.905	1	-20.751	1
46	down	-12.959	1	-22.566	1
45	knew	-12.986	1	-17.668	1
44	looked	-13.091	1	-16.614	1
43	too	-13.114	1	-23.693	1
42	went	-13.504	1	-19.48	1
41	into	-13.746	1	-27.201	1

40	just	-13.925	1	-23.797	1
39	never	-14.109	1	-23.5	1
38	came	-14.303	1	-21.496	1
37	there	-14.589	1	-25.692	1
36	didn’t	-14.663	1	-17.155	1
35	way	-14.667	1	-25.748	1
34	life	-14.764	1	-23.077	1
33	then	-15.004	1	-26.608	1
32	world	-15.058	1	-21.107	1
31	could	-15.159	1	-23.311	1
30	himself	-15.32	1	-21.278	1
29	their	-15.389	1	-24.926	1
28	like	-15.958	1	-24.866	1
27	who	-15.98	1	-25.667	1
26	back	-16.051	1	-22.567	1
25	war	-16.089	1	-17.208	1
24	me	-16.248	1	-13.179	1
23	what	-16.467	1	-27.876	1
22	them	-17.018	1	-28.18	1
21	that	-17.166	1	-15.534	1
20	even	-17.548	1	-28.539	1
19	so	-17.569	1	-28.15	1
18	when	-18.0	1	-28.101	1
17	my	-18.564	1	-16.066	1
16	one	-18.607	1	-27.046	1
15	out	-18.859	1	-26.912	1
14	man	-19.086	1	-22.586	1
13	they	-19.442	1	-26.978	1
12	i	-21.24	1	-10.271	1
11	a	-25.344	1	-9.351	1
10	said	-26.668	1	-22.313	1
9	but	-28.21	1	-27.151	1
8	was	-29.965	1	-20.183	1
7	it	-30.092	1	-21.523	1
6	him	-31.66	1	-25.986	1
5	had	-38.986	1	-25.089	1
4	she	-40.719	1	-21.359	1
3	her	-42.078	1	-23.404	1
2	his	-52.752	1	-29.678	1
1	he	-59.486	1	-27.21	1

E.5.2 Top 400 and Bottom 100 Tokens Ranked by File Z-score

Rank	Token	Freq-Z	Freq-V	File-Z	File-V
1	tobacco	87.323	1	22.208	1
2	cigarettes	57.965	1	17.204	1
3	cigarette	67.372	1	17.038	1
4	smoking	73.206	1	16.985	1
5	smokers	66.307	1	16.008	1
6	nicotine	55.68	1	13.952	1
7	brands	45.547	1	12.423	1
8	smoke	59.818	1	12.31	1
9	filter	45.818	1	11.761	1
10	tar	39.373	1	11.104	1
11	menthol	44.529	1	10.89	1
12	smoker	31.274	1	10.609	1
13	marlboro	41.468	1	10.307	1
14	brand	49.293	1	10.179	1
15	samples	33.075	1	9.572	1
16	product	46.628	1	9.333	1
17	rjr	32.483	1	9.207	1
18	carton	29.26	1	9.207	1
19	attached	22.259	1	9.182	1

20	products	37.619	1	8.864	1
21	winston	27.339	1	8.415	1
22	mg	24.966	1	8.337	1
23	advertising	37.664	1	8.146	1
24	pack	35.683	1	8.065	1
25	camel	38.633	1	8.038	1
26	consumer	22.149	1	7.913	1
27	lung	29.397	1	7.904	1
28	summary	19.776	1	7.854	1
29	reynolds	25.568	1	7.78	1
30	tobaccos	19.95	1	7.728	1
31	flavor	30.408	1	7.636	1
32	blend	27.459	1	7.536	1
33	100’s	22.803	1	7.518	1
34	salem	25.341	1	7.468	1
35	marketing	22.971	1	7.332	1
36	pm	23.373	1	7.291	1
37	morris	32.269	1	7.253	1
38	promotions	20.238	1	7.112	1
39	packs	20.249	1	7.026	1
40	slims	23.811	1	7.024	1
41	smoked	20.398	1	6.986	1
42	kool	28.378	1	6.981	1
43	burley	21.878	1	6.981	1
44	lorillard	19.474	1	6.837	1
45	analytical	17.557	1	6.776	1
46	cartons	19.28	1	6.762	1
47	4	36.876	1	6.712	1
48	data	31.027	1	6.702	1
49	retail	27.407	1	6.621	1
50	please	24.427	1	6.621	1
51	inhalation	20.367	1	6.601	1
52	coupon	17.705	1	6.601	1
53	copy	20.614	1	6.582	1
54	1	44.484	1	6.548	1
55	philip	31.077	1	6.487	1
56	filters	17.52	1	6.464	1
57	results	35.636	1	6.439	1
58	iii	19.589	1	6.433	1
59	testing	27.842	1	6.425	1
60	5	34.8	1	6.332	1
61	compounds	17.264	1	6.292	1
62	sample	27.226	1	6.252	1
63	sampling	20.756	1	6.154	1
64	3	40.098	1	6.145	1
65	packaging	19.346	1	6.123	1
66	test	44.644	1	6.07	1
67	doral	18.177	1	6.04	1
68	monoxide	16.724	1	5.991	1
69	cancer	32.46	1	5.961	1
70	evaluation	19.152	1	5.945	1
71	nonsmokers	19.57	1	5.933	1
72	puff	18.687	1	5.852	1
73	sidestream	14.107	1	5.824	1
74	ets	27.744	1	5.824	1
75	coupons	16.444	1	5.743	1
76	promotional	15.98	1	5.727	1
77	2	39.625	1	5.632	1
78	requested	15.756	1	5.621	1
79	ftc	15.979	1	5.601	1
80	flue	18.177	1	5.601	1
81	carbon	20.376	1	5.523	1
82	carlton	18.294	1	5.504	1
83	condensate	17.97	1	5.487	1
84	exposure	32.235	1	5.476	1
85	promotion	25.989	1	5.445	1
86	research	37.778	1	5.442	1
87	tested	18.707	1	5.433	1

88	memorandum	11.461	1	5.388	1
89	r	29.76	1	5.334	1
90	respiratory	14.827	1	5.322	1
91	inc	16.245	1	5.3	1
92	submitted	15.165	1	5.288	1
93	mm	16.988	1	5.273	1
94	toxicity	15.195	1	5.196	1
95	williamson	14.44	1	5.156	1
96	kg	13.838	1	5.128	1
97	8	27.708	1	5.086	1
98	filtration	14.961	1	5.073	1
99	methodology	13.525	1	5.049	1
100	cured	19.403	1	5.049	1
101	assay	16.734	1	5.044	1
102	iv	12.393	1	5.037	1
103	ula	12.558	1	5.003	1
104	pos	16.781	1	5.003	1
105	pulmonary	7.945	1	4.919	1
106	6	30.695	1	4.915	1
107	1994	16.845	1	4.91	1
108	enclosed	11.837	1	4.888	1
109	evaluated	15.406	1	4.876	1
110	versus	20.892	1	4.832	1
111	laboratory	15.289	1	4.827	1
112	metabolism	15.138	1	4.791	1
113	cc	12.162	1	4.791	1
114	tipping	14.996	1	4.79	1
115	tareyton	19.377	1	4.744	1
116	cotinine	20.96	1	4.744	1
117	brand’s	12.407	1	4.744	1
118	retailers	13.124	1	4.711	1
119	delivery	21.985	1	4.711	1
120	ml	27.861	1	4.708	1
121	9	20.09	1	4.701	1
122	gc	14.581	1	4.686	1
123	filler	14.173	1	4.686	1
124	barclay	13.572	1	4.686	1
125	ammonia	16.947	1	4.686	1
126	deliveries	11.374	1	4.654	1
127	memo	10.402	1	4.623	1
128	liggett	12.255	1	4.609	1
129	franchise	18.713	1	4.589	1
130	blends	15.446	1	4.586	1
131	viceroy	24.394	1	4.551	1
132	shipped	13.024	1	4.542	1
133	pall	13.606	1	4.533	1
134	nm	12.502	1	4.525	1
135	1st	12.612	1	4.525	1
136	particulate	14.059	1	4.522	1
137	incidence	14.468	1	4.516	1
138	100mm	17.759	1	4.47	1
139	dilution	8.273	1	4.426	1
140	kent	23.437	1	4.419	1
141	rjr’s	10.348	1	4.411	1
142	ks	23.651	1	4.411	1
143	additive	10.576	1	4.411	1
144	retailer	11.258	1	4.386	1
145	leaf	17.434	1	4.385	1
146	prior	20.369	1	4.384	1
147	outlets	14.766	1	4.372	1
148	specifications	13.35	1	4.339	1
149	shipment	10.971	1	4.33	1
150	raleigh	10.613	1	4.327	1
151	pyrolysis	10.961	1	4.327	1
152	positioning	11.11	1	4.293	1
153	7	24.709	1	4.29	1
154	smokes	9.466	1	4.247	1
155	additives	9.954	1	4.247	1

156	co2	13.579	1	4.245	1
157	cardiovascular	10.277	1	4.245	1
158	acceptability	8.56	1	4.245	1
159	date	22.08	1	4.227	1
160	evaluate	14.615	1	4.224	1
161	analyses	15.602	1	4.224	1
162	consumers	15.855	1	4.203	1
163	tpm	12.853	1	4.179	1
164	aftertaste	10.435	1	4.179	1
165	mall	13.55	1	4.155	1
166	overall	18.473	1	4.131	1
167	competitive	22.547	1	4.131	1
168	2nd	11.811	1	4.124	1
169	merchandising	13.048	1	4.111	1
170	ppm	13.301	1	4.103	1
171	hedges	13.79	1	4.103	1
172	vending	11.931	1	4.098	1
173	qualitative	11.221	1	4.093	1
174	ca	12.408	1	4.093	1
175	ventilation	19.372	1	4.062	1
176	comments	15.061	1	4.04	1
177	rats	21.215	1	4.028	1
178	rjrt	12.853	1	4.026	1
179	packings	10.961	1	4.026	1
180	mildness	9.493	1	4.026	1
181	ctr	12.706	1	4.026	1
182	assays	9.88	1	4.026	1
183	min	12.529	1	4.017	1
184	manufacture	11.518	1	4.01	1
185	casing	11.654	1	3.981	1
186	concentrations	17.919	1	3.975	1
187	surgeon	13.098	1	3.963	1
188	report	24.781	1	3.963	1
189	determinations	10.094	1	3.947	1
190	conducted	16.198	1	3.939	1
191	additionally	10.4	1	3.914	1
192	sales	31.289	1	3.897	1
193	vitro	9.339	1	3.874	1
194	cellulose	9.63	1	3.874	1
195	wynder	11.464	1	3.867	1
196	acetate	8.172	1	3.833	1
197	harshness	11.049	1	3.805	1
198	vii	8.612	1	3.803	1
199	calibration	12.179	1	3.803	1
200	booklet	10.808	1	3.791	1
201	experimental	18.744	1	3.772	1
202	10	28.952	1	3.745	1
203	1997	9.907	1	3.727	1
204	ad	17.101	1	3.718	1
205	feasibility	11.695	1	3.681	1
206	ambient	9.194	1	3.681	1
207	7	16.382	1	3.681	1
208	emphysema	9.954	1	3.646	1
209	reconstituted	9.79	1	3.639	1
210	prototypes	14.564	1	3.639	1
211	approx	11.049	1	3.639	1
212	vi	9.388	1	3.633	1
213	nitrate	13.249	1	3.628	1
214	nitrogen	15.969	1	3.61	1
215	studies	25.713	1	3.608	1
216	summarized	8.59	1	3.577	1
217	enclosing	7.263	1	3.576	1
218	ms	11.729	1	3.566	1
219	b	25.612	1	3.566	1
220	5	12.311	1	3.566	1
221	ii	16.067	1	3.564	1
222	ultra	21.28	1	3.542	1
223	chromatography	9.537	1	3.524	1

224	distribution	14.562	1	3.491	1
225	usage	9.994	1	3.488	1
226	1996	13.525	1	3.484	1
227	99	8.496	1	3.484	1
228	solvent	10.438	1	3.483	1
229	parity	12.21	1	3.483	1
230	diseases	10.649	1	3.482	1
231	uk	11.708	1	3.466	1
232	toxicology	9.326	1	3.46	1
233	sugars	11.901	1	3.46	1
234	coronary	11.11	1	3.46	1
235	regarding	12.083	1	3.421	1
236	mailing	6.901	1	3.42	1
237	flavors	12.967	1	3.42	1
238	questionnaire	5.979	1	3.415	1
239	generic	10.411	1	3.404	1
240	forwarded	9.158	1	3.404	1
241	institute	17.271	1	3.382	1
242	ti	10.893	1	3.362	1
243	ph	14.332	1	3.362	1
244	96	9.254	1	3.356	1
245	overview	8.774	1	3.33	1
246	benson	10.237	1	3.33	1
247	nc	8.156	1	3.314	1
248	epidemiological	9.026	1	3.314	1
249	usa	11.263	1	3.304	1
250	miscellaneous	8.656	1	3.304	1
251	ads	18.696	1	3.291	1
252	28	13.408	1	3.291	1
253	tars	8.992	1	3.284	1
254	respondent	10.679	1	3.284	1
255	flavoring	10.399	1	3.284	1
256	bronchial	4.991	1	3.284	1
257	tests	17.149	1	3.265	1
258	vs	15.271	1	3.263	1
259	urine	14.1	1	3.259	1
260	media	18.99	1	3.258	1
261	evaluating	8.077	1	3.256	1
262	confidential	12.942	1	3.234	1
263	purchase	15.839	1	3.23	1
264	additional	16.678	1	3.217	1
265	monitoring	9.392	1	3.213	1
266	levels	25.917	1	3.213	1
267	implemented	9.19	1	3.213	1
268	moisture	17.482	1	3.202	1
269	excise	11.11	1	3.194	1
270	4	8.774	1	3.194	1
271	findings	14.69	1	3.188	1
272	phase	18.794	1	3.18	1
273	puffs	9.556	1	3.171	1
274	extraction	13.109	1	3.171	1
275	distributors	11.137	1	3.171	1
276	etc	12.541	1	3.164	1
277	screening	9.981	1	3.162	1
278	18	21.245	1	3.157	1
279	objectives	15.193	1	3.153	1
280	initiated	8.711	1	3.146	1
281	chronic	10.337	1	3.146	1
282	cigars	8.437	1	3.138	1
283	constituents	11.364	1	3.121	1
284	1981	12.8	1	3.118	1
285	respondents	19.418	1	3.116	1
286	receipt	8.496	1	3.116	1
287	newport	13.927	1	3.116	1
288	modifications	11.514	1	3.116	1
289	chemical	15.497	1	3.111	1
290	silica	8.156	1	3.099	1
291	qa	8.398	1	3.099	1

292	confidentiality	8.707	1	3.099	1
293	carcinoma	8.918	1	3.099	1
294	recommendation	7.882	1	3.097	1
295	extracts	9.39	1	3.093	1
296	prototype	17.33	1	3.075	1
297	dated	7.825	1	3.074	1
298	1995	16.117	1	3.066	1
299	project	21.196	1	3.019	1
300	epa	8.958	1	3.019	1
301	distributor	7.057	1	3.019	1
302	toxic	7.1	1	3.018	1
303	preliminary	10.941	1	3.015	1
304	request	13.972	1	2.998	1
305	copies	9.463	1	2.974	1
306	quantitative	8.674	1	2.971	1
307	supplier	9.706	1	2.933	1
308	clearance	9.151	1	2.933	1
309	behavioral	7.425	1	2.933	1
310	materials	19.466	1	2.922	1
311	c	27.449	1	2.921	1
312	carcinogenic	10.491	1	2.92	1
313	81	7.987	1	2.916	1
314	methanol	9.254	1	2.906	1
315	lbs	8.774	1	2.906	1
316	deluxe	11.743	1	2.906	1
317	d	27.593	1	2.904	1
318	profile	11.198	1	2.902	1
319	substances	9.918	1	2.899	1
320	workplace	10.402	1	2.868	1
321	presentation	12.883	1	2.866	1
322	vivo	6.744	1	2.838	1
323	3rd	10.108	1	2.838	1
324	investigated	6.893	1	2.837	1
325	modification	8.306	1	2.828	1
326	harmful	10.731	1	2.828	1
327	significantly	13.162	1	2.827	1
328	cell	12.991	1	2.817	1
329	93	9.339	1	2.812	1
330	compound	10.997	1	2.81	1
331	attributes	13.442	1	2.81	1
332	specification	8.359	1	2.762	1
333	labels	10.411	1	2.762	1
334	89	7.636	1	2.762	1
335	83	11.821	1	2.762	1
336	17	14.497	1	2.76	1
337	package	16.304	1	2.758	1
338	1	13.096	1	2.757	1
339	monitor	7.667	1	2.742	1
340	labeling	8.141	1	2.742	1
341	4th	8.559	1	2.742	1
342	97	6.744	1	2.742	1
343	79	7.235	1	2.742	1
344	8	8.353	1	2.742	1
345	31	14.296	1	2.731	1
346	acid	14.449	1	2.724	1
347	obtained	15.444	1	2.723	1
348	exposures	10.237	1	2.704	1
349	advertisements	8.964	1	2.704	1
350	ky	3.443	1	2.701	1
351	glycol	9.466	1	2.701	1
352	ct	9.523	1	2.701	1
353	benzene	7.489	1	2.701	1
354	alveolar	11.073	1	2.701	1
355	decreased	10.165	1	2.699	1
356	january	15.709	1	2.697	1
357	reviewed	11.432	1	2.685	1
358	segment	21.52	1	2.675	1
359	richmond	9.165	1	2.674	1

360	acceptable	13.271	1	2.668	1
361	analysis	23.957	1	2.654	1
362	revised	10.823	1	2.65	1
363	tow	9.423	1	2.648	1
364	positioned	8.18	1	2.648	1
365	inhaled	9.423	1	2.648	1
366	84	9.23	1	2.644	1
367	0	9.851	1	2.618	1
368	updated	5.762	1	2.593	1
369	introductory	10.264	1	2.593	1
370	dioxide	11.678	1	2.593	1
371	3	8.958	1	2.593	1
372	2	9.341	1	2.593	1
373	1970	10.137	1	2.584	1
374	mainstream	11.194	1	2.579	1
375	draft	12.353	1	2.571	1
376	reps	12.952	1	2.56	1
377	camels	14.961	1	2.56	1
378	acetone	6.988	1	2.56	1
379	myers	7.245	1	2.555	1
380	11	20.781	1	2.541	1
381	review	19.087	1	2.537	1
382	protocol	9.414	1	2.535	1
383	humidity	10.569	1	2.535	1
384	acids	9.91	1	2.535	1
385	market	28.024	1	2.527	1
386	completion	6.307	1	2.509	1
387	laboratories	8.447	1	2.497	1
388	19	14.465	1	2.496	1
389	24	16.39	1	2.491	1
390	pricing	10.399	1	2.485	1
391	requests	9.655	1	2.48	1
392	hereby	5.685	1	2.479	1
393	82	5.272	1	2.479	1
394	12	26.43	1	2.47	1
395	currently	15.153	1	2.467	1
396	processed	9.025	1	2.465	1
397	lab	8.415	1	2.465	1
398	reduction	13.383	1	2.453	1
399	residual	7.955	1	2.449	1
400	mail	9.581	1	2.449	1
100	enough	-8.698	1	-19.507	1
99	see	-5.5	1	-19.535	1
98	take	-8.034	1	-19.555	1
97	great	-9.519	1	-19.617	1
96	people	-5.095	1	-19.905	1
95	thought	-12.591	1	-19.932	1
94	once	-10.48	1	-19.952	1
93	at	-8.87	1	-19.988	1
92	home	-12.655	1	-20.071	1
91	say	-8.737	1	-20.138	1
90	was	-29.965	1	-20.183	1
89	every	-8.348	1	-20.185	1
88	right	-9.246	1	-20.259	1
87	away	-12.481	1	-20.42	1
86	old	-11.394	1	-20.458	1
85	go	-11.206	1	-20.484	1
84	while	-4.92	1	-20.513	1
83	did	-11.132	1	-20.63	1
82	always	-12.065	1	-20.646	1
81	those	-2.392	1	-20.657	1
80	few	-8.371	1	-20.662	1
79	though	-12.008	1	-20.67	1
78	took	-12.905	1	-20.751	1
77	make	-5.288	1	-20.792	1
76	through	-5.629	1	-20.947	1
75	were	6.385	1	-20.95	1

74	not	-6.066	1	-21.019	1
73	left	-11.717	1	-21.045	1
72	world	-15.058	1	-21.107	1
71	after	-4.619	1	-21.119	1
70	little	-11.101	1	-21.133	1
69	if	-8.68	1	-21.178	1
68	been	4.618	1	-21.206	1
67	himself	-15.32	1	-21.278	1
66	she	-40.719	1	-21.359	1
65	such	-5.113	1	-21.37	1
64	here	-12.059	1	-21.422	1
63	might	-8.704	1	-21.45	1
62	came	-14.303	1	-21.496	1
61	it	-30.092	1	-21.523	1
60	years	-7.526	1	-21.65	1
59	well	-2.642	1	-21.756	1
58	come	-12.086	1	-21.813	1
57	because	-6.89	1	-21.949	1
56	before	-8.338	1	-22.011	1
55	said	-26.668	1	-22.313	1
54	an	-7.807	1	-22.38	1
53	other	2.941	1	-22.416	1
52	down	-12.959	1	-22.566	1
51	back	-16.051	1	-22.567	1
50	man	-19.086	1	-22.586	1
49	where	-9.689	1	-22.759	1
48	much	-8.863	1	-22.79	1
47	still	-11.554	1	-22.792	1
46	time	-5.825	1	-22.866	1
45	would	-8.411	1	-22.882	1
44	most	-8.221	1	-22.922	1
43	all	-4.237	1	-22.928	1
42	now	-9.212	1	-22.992	1
41	another	-11.176	1	-23.03	1
40	how	-8.631	1	-23.032	1
39	life	-14.764	1	-23.077	1
38	do	-7.251	1	-23.197	1
37	could	-15.159	1	-23.311	1
36	many	-9.011	1	-23.384	1
35	her	-42.078	1	-23.404	1
34	up	-9.726	1	-23.429	1
33	never	-14.109	1	-23.5	1
32	own	-12.798	1	-23.656	1
31	too	-13.114	1	-23.693	1
30	its	-9.824	1	-23.791	1
29	just	-13.925	1	-23.797	1
28	first	-6.909	1	-23.916	1
27	long	-10.193	1	-24.116	1
26	than	2.271	1	-24.222	1
25	some	-7.475	1	-24.714	1
24	more	-3.328	1	-24.824	1
23	like	-15.958	1	-24.866	1
22	their	-15.389	1	-24.926	1
21	about	-11.126	1	-24.958	1
20	had	-38.986	1	-25.089	1
19	over	-9.875	1	-25.12	1
18	only	-7.95	1	-25.27	1
17	who	-15.98	1	-25.667	1
16	there	-14.589	1	-25.692	1
15	way	-14.667	1	-25.748	1
14	him	-31.66	1	-25.986	1
13	then	-15.004	1	-26.608	1
12	out	-18.859	1	-26.912	1
11	they	-19.442	1	-26.978	1
10	one	-18.607	1	-27.046	1
9	but	-28.21	1	-27.151	1
8	into	-13.746	1	-27.201	1
7	he	-59.486	1	-27.21	1

6	what	-16.467	1	-27.876	1
5	when	-18.0	1	-28.101	1
4	so	-17.569	1	-28.15	1
3	them	-17.018	1	-28.18	1
2	even	-17.548	1	-28.539	1
1	his	-52.752	1	-29.678	1

E.5.3 Top 400 and Bottom 100 Collocations Ranked by Frequency Z-score

Rank	Collocation	Freq-Z	Freq-V	File-Z	File-V
1	philip /3 morris	34.392	1	9.909	1
2	non /3 smokers	23.333	1	6.646	1
3	virginia /3 slims	23.253	1	6.646	1
4	tobacco /3 smoke	22.72	1	8.5	1
5	cigarette /3 smoking	22.263	1	8.612	1
6	lung /3 cancer	21.576	1	7.488	1
7	low /3 tar	21.227	1	7.557	1
8	tobacco /3 products	21.049	1	8.14	1
9	r /3 d	20.943	1	7.44	1
10	r /3 j	20.876	1	7.708	1
11	cigarette /3 smoke	20.688	1	8.236	1
12	r /3 reynolds	19.901	1	7.343	1
13	j /3 reynolds	19.807	1	7.343	1
14	tobacco /3 company	19.666	1	7.812	1
15	tobacco /3 industry	19.377	1	7.643	1
16	tar /3 nicotine	18.986	1	7.557	1
17	smoking /3 health	18.986	1	7.295	1
18	among /3 smokers	18.28	1	7.643	1
19	b /3 w	18.177	1	6.931	1
20	flue /3 cured	17.97	1	5.37	1
21	full /3 flavor	17.705	1	5.434	1
22	non /3 menthol	17.331	1	5.37	1
23	carbon /3 monoxide	16.724	1	5.991	1
24	tobacco /3 taste	16.327	1	4.875	1
25	tobacco /3 research	16.096	1	7.115	1
26	marlboro /3 lights	16.096	1	4.875	1
27	brown /3 williamson	16.096	1	6.04	1
28	ultra /3 lights	15.861	1	3.867	1
29	passive /3 smoking	15.861	1	4.47	1
30	competitive /3 smokers	15.861	1	5.487	1
31	tobacco /3 institute	15.742	1	6.549	1
32	lucky /3 strike	15.742	1	5.128	1
33	american /3 tobacco	15.742	1	6.145	1
34	any /3 questions	15.261	1	8.55	1
35	smokers /3 smokers	14.884	1	5.714	1
36	test /3 market	14.5	1	6.145	1
37	pall /3 mall	14.173	1	5.316	1
38	smokers /3 who	13.702	1	5.824	1
39	g /3 13	13.702	1	3.867	1
40	benson /3 hedges	13.702	1	4.875	1
41	reynolds /3 tobacco	13.633	1	5.774	1
42	no /3 significant	13.525	1	4.952	1
43	anti /3 smoking	13.425	1	5.487	1
44	indoor /3 air	13.355	1	4.267	1
45	new /3 product	13.301	1	4.91	1
46	smoke /3 exposure	13.284	1	4.609	1
47	form /3 ula	13.284	1	5.003	1
48	first /3 quarter	13.187	1	3.42	1
49	r /3 tobacco	13.071	1	5.663	1
50	smoke /3 condensate	12.998	1	4.875	1
51	other /3 brands	12.998	1	5.487	1
52	mainstream /3 smoke	12.853	1	5.714	1
53	j /3 tobacco	12.78	1	5.434	1

54	menthol /3 smokers	12.407	1	4.026	1
55	female /3 smokers	12.407	1	4.47	1
56	effects /3 smoking	12.407	1	5.824	1
57	cigarette /3 advertising	12.407	1	4.875	1
58	public /3 smoking	12.101	1	4.875	1
59	marlboro /3 smokers	12.101	1	4.327	1
60	carbon /3 dioxide	12.084	1	2.742	1
61	please /3 me	11.959	1	5.719	1
62	who /3 smoke	11.945	1	5.37	1
63	sidestream /3 smoke	11.787	1	4.875	1
64	king /3 size	11.787	1	4.327	1
65	tobacco /3 use	11.708	1	4.118	1
66	cigarette /3 paper	11.708	1	4.947	1
67	winston /3 salem	11.626	1	5.714	1
68	smokers /3 non	11.626	1	4.609	1
69	adult /3 smokers	11.626	1	5.128	1
70	ultra /3 low	11.547	1	4.118	1
71	tobacco /3 companies	11.547	1	3.965	1
72	new /3 products	11.525	1	3.695	1
73	point /3 sale	11.464	1	4.609	1
74	new /3 brand	11.464	1	4.327	1
75	council /3 tobacco	11.464	1	4.875	1
76	fourth /3 quarter	11.422	1	2.906	1
77	your /3 letter	11.374	1	5.612	1
78	two /3 pack	11.299	1	3.867	1
79	reynolds /3 company	11.299	1	4.875	1
80	per /3 pack	11.299	1	5.003	1
81	between /3 smoking	11.299	1	4.875	1
82	cured /3 tobacco	11.131	1	4.327	1
83	smoke /3 cigarettes	11.091	1	4.791	1
84	any /3 please	11.049	1	6.303	1
85	if /3 any	10.99	1	2.655	1
86	new /3 cigarette	10.961	1	4.609	1
87	j /3 company	10.961	1	4.744	1
88	exposure /3 smoke	10.961	1	4.179	1
89	all /3 smokers	10.961	1	4.875	1
90	share /3 market	10.922	1	3.947	1
91	b /3 h	10.922	1	3.138	1
92	air /3 quality	10.874	1	3.314	1
93	surgeon /3 general's	10.789	1	4.744	1
94	questions /3 please	10.789	1	6.248	1
95	particulate /3 matter	10.789	1	4.609	1
96	mg /3 tar	10.789	1	4.609	1
97	cigarettes /3 made	10.789	1	4.609	1
98	council /3 research	10.751	1	3.956	1
99	their /3 brand	10.613	1	4.609	1
100	please /3 know	10.613	1	6.04	1
101	per /3 cigarette	10.613	1	5.25	1
102	burley /3 tobacco	10.613	1	3.867	1
103	all /3 brands	10.613	1	4.875	1
104	board /3 directors	10.583	1	2.138	1
105	no /3 differences	10.529	1	4.4	1
106	tobacco /3 which	10.435	1	5.601	1
107	tar /3 cigarettes	10.435	1	4.744	1
108	smoking /3 cancer	10.435	1	4.875	1
109	environmental /3 tobacco	10.435	1	4.875	1
110	environmental /3 smoke	10.435	1	4.875	1
111	cigarette /3 manufacturers	10.435	1	4.026	1
112	heart /3 disease	10.402	1	3.259	1
113	who /3 smoked	10.348	1	3.805	1
114	18 /3 year	10.348	1	3.284	1
115	american /3 company	10.286	1	3.093	1
116	who /3 smoking	10.253	1	4.179	1
117	surgeon /3 general	10.253	1	4.744	1
118	nicotine /3 levels	10.253	1	4.327	1
119	gas /3 phase	10.248	1	3.947	1
120	these /3 products	10.173	1	4.103	1
121	air /3 pollution	10.173	1	3.628	1

122	please /3 let	10.136	1	5.501	1
123	new /3 brands	10.069	1	4.47	1
124	menthol /3 menthol	10.069	1	4.744	1
125	differences /3 between	10.068	1	2.15	1
126	p /3 o	9.979	1	3.965	1
127	see /3 attached	9.907	1	5.167	1
128	ultra /3 tar	9.88	1	4.026	1
129	these /3 cigarettes	9.88	1	5.128	1
130	smoking /3 tobacco	9.88	1	4.179	1
131	nicotine /3 nicotine	9.88	1	4.026	1
132	competitive /3 brands	9.88	1	4.47	1
133	coronary /3 disease	9.85	1	3.284	1
134	been /3 completed	9.725	1	3.552	1
135	tobacco /3 tobacco	9.688	1	4.179	1
136	smoking /3 machine	9.688	1	4.327	1
137	smoking /3 habits	9.688	1	4.327	1
138	smokers /3 nonsmokers	9.688	1	4.327	1
139	field /3 sales	9.688	1	4.026	1
140	cigarette /3 brands	9.688	1	3.867	1
141	1 /3 mg	9.688	1	4.179	1
142	about /3 smoking	9.596	1	4.551	1
143	sales /3 force	9.581	1	4.245	1
144	final /3 report	9.581	1	3.833	1
145	public /3 places	9.523	1	3.628	1
146	pack /3 carton	9.493	1	4.327	1
147	nicotine /3 content	9.493	1	4.47	1
148	mg /3 nicotine	9.493	1	4.179	1
149	gas /3 chromatography	9.493	1	4.179	1
150	current /3 product	9.493	1	3.867	1
151	cigarette /3 smokers	9.493	1	4.609	1
152	table /3 3	9.466	1	3.803	1
153	next /3 meeting	9.399	1	4.118	1
154	market /3 share	9.399	1	4.118	1
155	particle /3 size	9.35	1	2.762	1
156	people /3 smoke	9.326	1	4.267	1
157	no /3 difference	9.197	1	2.396	1
158	u /3 k	9.124	1	3.099	1
159	please /3 call	9.124	1	5.044	1
160	following /3 1	9.028	1	3.437	1
161	m /3 d	9.025	1	2.102	1
162	filter /3 filter	8.992	1	3.466	1
163	public /3 health	8.99	1	3.018	1
164	low /3 high	8.975	1	2.555	1
165	direct /3 mail	8.975	1	3.093	1
166	age /3 group	8.975	1	2.241	1
167	quality /3 assurance	8.918	1	3.965	1
168	during /3 quarter	8.918	1	3.284	1
169	mg /3 per	8.821	1	3.628	1
170	quality /3 control	8.774	1	3.099	1
171	each /3 group	8.763	1	2.361	1
172	thank /3 your	8.707	1	4.791	1
173	product /3 which	8.707	1	3.947	1
174	l /3 m	8.528	1	4.784	1
175	january /3 1	8.437	1	3.646	1
176	cancer /3 research	8.398	1	2.838	1
177	d /3 d	8.354	1	2.661	1
178	lower /3 than	8.222	1	2.208	1
179	body /3 weight	8.192	1	3.46	1
180	association /3 between	8.18	1	2.838	1
181	table /3 2	8.129	1	2.563	1
182	these /3 results	8.112	1	2.791	1
183	cancer /3 society	7.99	1	2.762	1
184	determine /3 if	7.964	1	3.48	1
185	c /3 c	7.944	1	2.828	1
186	all /3 groups	7.935	1	2.838	1
187	shown /3 table	7.932	1	2.661	1
188	test /3 results	7.925	1	2.837	1
189	significant /3 differences	7.925	1	3.259	1

190	let /3 know	7.836	1	2.253	1
191	significantly /3 than	7.725	1	2.838	1
192	data /3 collected	7.725	1	2.56	1
193	product /3 development	7.701	1	3.483	1
194	correlation /3 between	7.511	1	2.479	1
195	national /3 institute	7.489	1	3.314	1
196	health /3 service	7.489	1	2.241	1
197	results /3 obtained	7.287	1	2.314	1
198	follow /3 up	7.254	1	2.791	1
199	research /3 which	7.235	1	3.259	1
200	our /3 research	7.235	1	2.449	1
201	j /3 r	7.007	1	2.241	1
202	research /3 center	6.79	1	2.479	1
203	1 /3 no	6.763	1	2.479	1
204	all /3 data	6.381	1	1.993	1
205	pilot /3 plant	6.334	1	2.346	1
206	medical /3 research	6.31	1	2.415	1
207	follows /3 1	5.878	1	2.138	1
208	per /3 week	5.814	1	2.138	1
209	u /3 s	5.785	1	-3.464	1
210	higher /3 level	5.762	1	2.361	1
211	any /3 information	5.486	1	2.415	1
212	research /3 program	5.242	1	2.138	1
213	1 /3 all	4.781	1	1.993	1
214	than /3 other	4.263	1	-2.971	1
215	vice /3 president	4.015	1	-2.946	1
216	advisory /3 board	3.983	1	2.701	1
217	mr /3 mr	3.752	1	-2.112	1
218	set /3 up	3.612	1	-3.472	1
219	greater /3 than	3.516	1	-3.012	1
220	can /3 used	3.464	1	-2.103	1
221	any /3 other	3.358	1	-5.153	1
222	do /3 think	3.139	1	-3.578	1
223	other /3 than	3.106	1	-3.844	1
224	more /3 people	3.064	1	-2.645	1
225	ever /3 before	3.064	1	-2.878	1
226	relationship /3 between	3.003	1	-2.507	1
227	more /3 likely	2.989	1	-2.103	1
228	next /3 year	2.984	1	-1.97	1
229	among /3 other	2.944	1	-2.855	1
230	let /3 me	2.919	1	-2.208	1
231	each /3 year	2.914	1	-3.026	1
232	n /3 j	2.882	1	-2.095	1
233	less /3 than	2.721	1	-8.374	1
234	women /3 who	2.666	1	-3.647	1
235	last /3 year	2.595	1	-4.33	1
236	june /3 30	2.513	1	2.346	1
237	these /3 two	2.509	1	-3.104	1
238	difference /3 between	2.454	1	-2.941	1
239	three /3 weeks	2.445	1	-1.973	1
240	new /3 new	2.403	1	-2.046	1
241	control /3 over	2.391	1	-2.366	1
242	no /3 no	2.308	1	-4.397	1
243	all /3 three	2.272	1	-2.143	1
244	person /3 who	2.231	1	-3.277	1
245	those /3 which	2.221	1	-2.882	1
246	days /3 after	2.169	1	-2.353	1
247	which /3 been	2.158	1	-4.366	1
248	past /3 years	2.094	1	-2.647	1
249	make /3 sure	2.039	1	-2.877	1
250	which /3 may	2.026	1	-2.725	1
251	put /3 together	1.993	1	-2.031	1
252	last /3 month	1.993	1	-2.373	1
253	would /3 than	1.98	1	-2.083	1
100	but /3 never	-4.67	1	-8.419	1
99	she /3 out	-4.694	1	-7.771	1
98	but /3 didn’t	-4.694	1	-7.657	1

97	my /3 life	-4.723	1	-7.066	1
96	time /3 had	-4.73	1	-7.896	1
95	her /3 when	-4.73	1	-7.142	1
94	would /3 her	-4.758	1	-7.671	1
93	man /3 had	-4.758	1	-7.498	1
92	but /3 her	-4.758	1	-7.671	1
91	had /3 left	-4.786	1	-7.84	1
90	even /3 though	-4.794	1	-9.952	1
89	had /3 never	-4.809	1	-8.687	1
88	years /3 ago	-4.82	1	-10.87	1
87	but /3 if	-4.822	1	-9.497	1
86	had /3 made	-4.836	1	-9.186	1
85	young /3 man	-4.841	1	-7.44	1
84	man /3 who	-4.847	1	-9.514	1
83	had /3 all	-4.848	1	-8.096	1
82	could /3 do	-4.848	1	-8.575	1
81	had /3 seen	-4.849	1	-9.191	1
80	what /3 happened	-4.866	1	-8.261	1
79	had /3 out	-4.888	1	-8.531	1
78	had /3 said	-4.896	1	-7.498	1
77	united /3 nations	-4.916	1	-5.947	1
76	per /3 cent	-4.923	1	-5.233	1
75	she /3 herself	-4.95	1	-7.727	1
74	then /3 she	-5.004	1	-8.116	1
73	had /3 into	-5.004	1	-8.224	1
72	had /3 taken	-5.024	1	-8.681	1
71	him /3 she	-5.03	1	-7.498	1
70	her /3 head	-5.03	1	-7.44	1
69	had /3 she	-5.03	1	-7.671	1
68	but /3 they	-5.032	1	-12.35	1
67	one /3 had	-5.075	1	-8.426	1
66	had /3 gone	-5.075	1	-8.635	1
65	she /3 says	-5.083	1	-5.894	1
64	nineteenth /3 century	-5.083	1	-6.833	1
63	they /3 all	-5.095	1	-8.942	1
62	about /3 her	-5.127	1	-8.049	1
61	she /3 didn’t	-5.135	1	-7.498	1
60	she /3 up	-5.161	1	-7.84	1
59	her /3 face	-5.161	1	-7.896	1
58	more /3 than	-5.169	1	-21.503	1
57	said /3 she	-5.188	1	-7.763	1
56	said /3 but	-5.214	1	-8.986	1
55	didn’t /3 know	-5.238	1	-8.437	1
54	when /3 came	-5.255	1	-8.84	1
53	told /3 him	-5.264	1	-8.204	1
52	had /3 up	-5.302	1	-9.085	1
51	when /3 had	-5.306	1	-8.885	1
50	so /3 she	-5.313	1	-8.278	1
49	him /3 but	-5.313	1	-8.594	1
48	so /3 much	-5.362	1	-10.851	1
47	their /3 they	-5.464	1	-9.723	1
46	after /3 all	-5.489	1	-10.352	1
45	if /3 had	-5.494	1	-11.409	1
44	what /3 she	-5.502	1	-8.84	1
43	my /3 father	-5.527	1	-6.428	1
42	when /3 they	-5.558	1	-11.485	1
41	each /3 other	-5.584	1	-10.928	1
40	she /3 been	-5.63	1	-8.331	1
39	her /3 own	-5.645	1	-9.086	1
38	old /3 man	-5.654	1	-7.896	1
37	my /3 mother	-5.654	1	-5.525	1
36	her /3 but	-5.693	1	-9.141	1
35	her /3 eyes	-5.77	1	-7.951	1
34	which /3 had	-5.782	1	-11.538	1
33	if /3 she	-5.831	1	-8.738	1
32	what /3 had	-5.959	1	-9.918	1
31	white /3 house	-5.998	1	-6.554	1
30	one /3 another	-6.053	1	-10.55	1

29	had /3 had	-6.1	1	-9.716	1
28	her /3 husband	-6.143	1	-8.418	1
27	but /3 had	-6.33	1	-11.264	1
26	could /3 see	-6.355	1	-9.861	1
25	world /3 war	-6.363	1	-8.951	1
24	her /3 mother	-6.405	1	-7.262	1
23	but /3 there	-6.453	1	-12.068	1
22	had /3 come	-6.508	1	-10.416	1
21	united /3 states	-6.525	1	-10.25	1
20	had /3 no	-6.79	1	-11.919	1
19	her /3 had	-6.847	1	-9.634	1
18	f /3 f	-6.86	1	-3.484	1
17	who /3 had	-6.882	1	-13.708	1
16	she /3 him	-7.076	1	-9.346	1
15	she /3 could	-7.207	1	-9.775	1
14	she /3 would	-7.251	1	-9.764	1
13	no /3 one	-7.523	1	-13.507	1
12	but /3 she	-7.574	1	-10.851	1
11	had /3 him	-7.62	1	-11.116	1
10	she /3 she	-7.734	1	-10.589	1
9	had /3 her	-7.786	1	-10.275	1
8	when /3 she	-8.343	1	-11.366	1
7	they /3 had	-8.697	1	-13.785	1
6	her /3 her	-9.586	1	-12.233	1
5	she /3 said	-9.633	1	-11.213	1
4	she /3 her	-10.653	1	-13.442	1
3	her /3 she	-11.927	1	-14.226	1
2	she /3 had	-13.494	1	-13.67	1
1	had /3 been	-16.055	1	-20.443	1

E.5.4 Top 400 and Bottom 100 Collocations Ranked by File Z-score

Rank	Collocation	Freq-Z	Freq-V	File-Z	File-V
1	philip /3 morris	34.392	1	9.909	1
2	cigarette /3 smoking	22.263	1	8.612	1
3	any /3 questions	15.261	1	8.55	1
4	tobacco /3 smoke	22.72	1	8.5	1
5	cigarette /3 smoke	20.688	1	8.236	1
6	tobacco /3 products	21.049	1	8.14	1
7	tobacco /3 company	19.666	1	7.812	1
8	r /3 j	20.876	1	7.708	1
9	tobacco /3 industry	19.377	1	7.643	1
10	among /3 smokers	18.28	1	7.643	1
11	tar /3 nicotine	18.986	1	7.557	1
12	low /3 tar	21.227	1	7.557	1
13	lung /3 cancer	21.576	1	7.488	1
14	r /3 d	20.943	1	7.44	1
15	r /3 reynolds	19.901	1	7.343	1
16	j /3 reynolds	19.807	1	7.343	1
17	smoking /3 health	18.986	1	7.295	1
18	tobacco /3 research	16.096	1	7.115	1
19	b /3 w	18.177	1	6.931	1
20	virginia /3 slims	23.253	1	6.646	1
21	non /3 smokers	23.333	1	6.646	1
22	tobacco /3 institute	15.742	1	6.549	1
23	any /3 please	11.049	1	6.303	1
24	questions /3 please	10.789	1	6.248	1
25	test /3 market	14.5	1	6.145	1
26	american /3 tobacco	15.742	1	6.145	1
27	please /3 know	10.613	1	6.04	1
28	brown /3 williamson	16.096	1	6.04	1
29	carbon /3 monoxide	16.724	1	5.991	1
30	smokers /3 who	13.702	1	5.824	1

31	effects /3 smoking	12.407	1	5.824	1
32	reynolds /3 tobacco	13.633	1	5.774	1
33	please /3 me	11.959	1	5.719	1
34	winston /3 salem	11.626	1	5.714	1
35	smokers /3 smokers	14.884	1	5.714	1
36	mainstream /3 smoke	12.853	1	5.714	1
37	r /3 tobacco	13.071	1	5.663	1
38	your /3 letter	11.374	1	5.612	1
39	tobacco /3 which	10.435	1	5.601	1
40	please /3 let	10.136	1	5.501	1
41	other /3 brands	12.998	1	5.487	1
42	competitive /3 smokers	15.861	1	5.487	1
43	anti /3 smoking	13.425	1	5.487	1
44	j /3 tobacco	12.78	1	5.434	1
45	full /3 flavor	17.705	1	5.434	1
46	who /3 smoke	11.945	1	5.37	1
47	non /3 menthol	17.331	1	5.37	1
48	flue /3 cured	17.97	1	5.37	1
49	pall /3 mall	14.173	1	5.316	1
50	per /3 cigarette	10.613	1	5.25	1
51	see /3 attached	9.907	1	5.167	1
52	these /3 cigarettes	9.88	1	5.128	1
53	lucky /3 strike	15.742	1	5.128	1
54	adult /3 smokers	11.626	1	5.128	1
55	please /3 call	9.124	1	5.044	1
56	per /3 pack	11.299	1	5.003	1
57	form /3 ula	13.284	1	5.003	1
58	no /3 significant	13.525	1	4.952	1
59	cigarette /3 paper	11.708	1	4.947	1
60	new /3 product	13.301	1	4.91	1
61	tobacco /3 taste	16.327	1	4.875	1
62	smoking /3 cancer	10.435	1	4.875	1
63	smoke /3 condensate	12.998	1	4.875	1
64	sidestream /3 smoke	11.787	1	4.875	1
65	reynolds /3 company	11.299	1	4.875	1
66	public /3 smoking	12.101	1	4.875	1
67	marlboro /3 lights	16.096	1	4.875	1
68	environmental /3 tobacco	10.435	1	4.875	1
69	environmental /3 smoke	10.435	1	4.875	1
70	council /3 tobacco	11.464	1	4.875	1
71	cigarette /3 advertising	12.407	1	4.875	1
72	between /3 smoking	11.299	1	4.875	1
73	benson /3 hedges	13.702	1	4.875	1
74	all /3 smokers	10.961	1	4.875	1
75	all /3 brands	10.613	1	4.875	1
76	thank /3 your	8.707	1	4.791	1
77	smoke /3 cigarettes	11.091	1	4.791	1
78	l /3 m	8.528	1	4.784	1
79	tar /3 cigarettes	10.435	1	4.744	1
80	surgeon /3 general’s	10.789	1	4.744	1
81	surgeon /3 general	10.253	1	4.744	1
82	menthol /3 menthol	10.069	1	4.744	1
83	j /3 company	10.961	1	4.744	1
84	their /3 brand	10.613	1	4.609	1
85	smokers /3 non	11.626	1	4.609	1
86	smoke /3 exposure	13.284	1	4.609	1
87	point /3 sale	11.464	1	4.609	1
88	particulate /3 matter	10.789	1	4.609	1
89	new /3 cigarette	10.961	1	4.609	1
90	mg /3 tar	10.789	1	4.609	1
91	cigarettes /3 made	10.789	1	4.609	1
92	cigarette /3 smokers	9.493	1	4.609	1
93	about /3 smoking	9.596	1	4.551	1
94	passive /3 smoking	15.861	1	4.47	1
95	nicotine /3 content	9.493	1	4.47	1
96	new /3 brands	10.069	1	4.47	1
97	female /3 smokers	12.407	1	4.47	1
98	competitive /3 brands	9.88	1	4.47	1

99	no /3 differences	10.529	1	4.4	1
100	smoking /3 machine	9.688	1	4.327	1
101	smoking /3 habits	9.688	1	4.327	1
102	smokers /3 nonsmokers	9.688	1	4.327	1
103	pack /3 carton	9.493	1	4.327	1
104	nicotine /3 levels	10.253	1	4.327	1
105	new /3 brand	11.464	1	4.327	1
106	marlboro /3 smokers	12.101	1	4.327	1
107	king /3 size	11.787	1	4.327	1
108	cured /3 tobacco	11.131	1	4.327	1
109	people /3 smoke	9.326	1	4.267	1
110	indoor /3 air	13.355	1	4.267	1
111	sales /3 force	9.581	1	4.245	1
112	who /3 smoking	10.253	1	4.179	1
113	tobacco /3 tobacco	9.688	1	4.179	1
114	smoking /3 tobacco	9.88	1	4.179	1
115	mg /3 nicotine	9.493	1	4.179	1
116	gas /3 chromatography	9.493	1	4.179	1
117	exposure /3 smoke	10.961	1	4.179	1
118	1 /3 mg	9.688	1	4.179	1
119	ultra /3 low	11.547	1	4.118	1
120	tobacco /3 use	11.708	1	4.118	1
121	next /3 meeting	9.399	1	4.118	1
122	market /3 share	9.399	1	4.118	1
123	these /3 products	10.173	1	4.103	1
124	ultra /3 tar	9.88	1	4.026	1
125	nicotine /3 nicotine	9.88	1	4.026	1
126	menthol /3 smokers	12.407	1	4.026	1
127	field /3 sales	9.688	1	4.026	1
128	cigarette /3 manufacturers	10.435	1	4.026	1
129	tobacco /3 companies	11.547	1	3.965	1
130	quality /3 assurance	8.918	1	3.965	1
131	p /3 o	9.979	1	3.965	1
132	council /3 research	10.751	1	3.956	1
133	share /3 market	10.922	1	3.947	1
134	product /3 which	8.707	1	3.947	1
135	gas /3 phase	10.248	1	3.947	1
136	ultra /3 lights	15.861	1	3.867	1
137	two /3 pack	11.299	1	3.867	1
138	g /3 13	13.702	1	3.867	1
139	current /3 product	9.493	1	3.867	1
140	cigarette /3 brands	9.688	1	3.867	1
141	burley /3 tobacco	10.613	1	3.867	1
142	final /3 report	9.581	1	3.833	1
143	who /3 smoked	10.348	1	3.805	1
144	table /3 3	9.466	1	3.803	1
145	new /3 products	11.525	1	3.695	1
146	january /3 1	8.437	1	3.646	1
147	public /3 places	9.523	1	3.628	1
148	mg /3 per	8.821	1	3.628	1
149	air /3 pollution	10.173	1	3.628	1
150	been /3 completed	9.725	1	3.552	1
151	product /3 development	7.701	1	3.483	1
152	determine /3 if	7.964	1	3.48	1
153	filter /3 filter	8.992	1	3.466	1
154	body /3 weight	8.192	1	3.46	1
155	following /3 1	9.028	1	3.437	1
156	first /3 quarter	13.187	1	3.42	1
157	national /3 institute	7.489	1	3.314	1
158	air /3 quality	10.874	1	3.314	1
159	during /3 quarter	8.918	1	3.284	1
160	coronary /3 disease	9.85	1	3.284	1
161	18 /3 year	10.348	1	3.284	1
162	significant /3 differences	7.925	1	3.259	1
163	research /3 which	7.235	1	3.259	1
164	heart /3 disease	10.402	1	3.259	1
165	b /3 h	10.922	1	3.138	1
166	u /3 k	9.124	1	3.099	1

167	quality /3 control	8.774	1	3.099	1
168	direct /3 mail	8.975	1	3.093	1
169	american /3 company	10.286	1	3.093	1
170	public /3 health	8.99	1	3.018	1
171	fourth /3 quarter	11.422	1	2.906	1
172	significantly /3 than	7.725	1	2.838	1
173	cancer /3 research	8.398	1	2.838	1
174	association /3 between	8.18	1	2.838	1
175	all /3 groups	7.935	1	2.838	1
176	test /3 results	7.925	1	2.837	1
177	c /3 c	7.944	1	2.828	1
178	these /3 results	8.112	1	2.791	1
179	follow /3 up	7.254	1	2.791	1
180	particle /3 size	9.35	1	2.762	1
181	cancer /3 society	7.99	1	2.762	1
182	carbon /3 dioxide	12.084	1	2.742	1
183	advisory /3 board	3.983	1	2.701	1
184	shown /3 table	7.932	1	2.661	1
185	d /3 d	8.354	1	2.661	1
186	if /3 any	10.99	1	2.655	1
187	table /3 2	8.129	1	2.563	1
188	data /3 collected	7.725	1	2.56	1
189	low /3 high	8.975	1	2.555	1
190	research /3 center	6.79	1	2.479	1
191	correlation /3 between	7.511	1	2.479	1
192	1 /3 no	6.763	1	2.479	1
193	our /3 research	7.235	1	2.449	1
194	medical /3 research	6.31	1	2.415	1
195	any /3 information	5.486	1	2.415	1
196	no /3 difference	9.197	1	2.396	1
197	higher /3 level	5.762	1	2.361	1
198	each /3 group	8.763	1	2.361	1
199	pilot /3 plant	6.334	1	2.346	1
200	june /3 30	2.513	1	2.346	1
201	results /3 obtained	7.287	1	2.314	1
202	let /3 know	7.836	1	2.253	1
203	j /3 r	7.007	1	2.241	1
204	health /3 service	7.489	1	2.241	1
205	age /3 group	8.975	1	2.241	1
206	lower /3 than	8.222	1	2.208	1
207	differences /3 between	10.068	1	2.15	1
208	research /3 program	5.242	1	2.138	1
209	per /3 week	5.814	1	2.138	1
210	follows /3 1	5.878	1	2.138	1
211	board /3 directors	10.583	1	2.138	1
212	m /3 d	9.025	1	2.102	1
213	all /3 data	6.381	1	1.993	1
214	1 /3 all	4.781	1	1.993	1
100	one /3 most	-3.887	1	-8.056	1
99	don’t /3 know	-2.679	1	-8.066	1
98	had /3 all	-4.848	1	-8.096	1
97	them /3 they	-3.147	1	-8.099	1
96	then /3 she	-5.004	1	-8.116	1
95	so /3 they	-3.316	1	-8.128	1
94	told /3 him	-5.264	1	-8.204	1
93	had /3 into	-5.004	1	-8.224	1
92	so /3 many	-3.915	1	-8.238	1
91	but /3 what	-3.917	1	-8.252	1
90	what /3 happened	-4.866	1	-8.261	1
89	they /3 would	-2.178	1	-8.278	1
88	so /3 she	-5.313	1	-8.278	1
87	what /3 did	-3.839	1	-8.312	1
86	she /3 been	-5.63	1	-8.331	1
85	less /3 than	2.721	1	-8.374	1
84	her /3 husband	-6.143	1	-8.418	1
83	but /3 never	-4.67	1	-8.419	1
82	one /3 had	-5.075	1	-8.426	1

81	but /3 when	-4.066	1	-8.428	1
80	didn't /3 know	-5.238	1	-8.437	1
79	all /3 over	-4.125	1	-8.481	1
78	know /3 what	-2.724	1	-8.499	1
77	had /3 out	-4.888	1	-8.531	1
76	could /3 do	-4.848	1	-8.575	1
75	him /3 but	-5.313	1	-8.594	1
74	even /3 if	-3.184	1	-8.623	1
73	had /3 gone	-5.075	1	-8.635	1
72	had /3 taken	-5.024	1	-8.681	1
71	had /3 never	-4.809	1	-8.687	1
70	if /3 she	-5.831	1	-8.738	1
69	when /3 came	-5.255	1	-8.84	1
68	what /3 she	-5.502	1	-8.84	1
67	its /3 own	-4.494	1	-8.848	1
66	when /3 had	-5.306	1	-8.885	1
65	one /3 who	-4.25	1	-8.886	1
64	even /3 more	-2.874	1	-8.924	1
63	they /3 all	-5.095	1	-8.942	1
62	world /3 war	-6.363	1	-8.951	1
61	said /3 but	-5.214	1	-8.986	1
60	no /3 more	-3.97	1	-9.046	1
59	all /3 but	-4.174	1	-9.071	1
58	had /3 up	-5.302	1	-9.085	1
57	her /3 own	-5.645	1	-9.086	1
56	her /3 but	-5.693	1	-9.141	1
55	who /3 been	-4.173	1	-9.174	1
54	had /3 made	-4.836	1	-9.186	1
53	had /3 seen	-4.849	1	-9.191	1
52	too /3 much	-4.416	1	-9.303	1
51	she /3 him	-7.076	1	-9.346	1
50	but /3 if	-4.822	1	-9.497	1
49	man /3 who	-4.847	1	-9.514	1
48	only /3 one	-1.976	1	-9.528	1
47	but /3 no	-4.159	1	-9.614	1
46	her /3 had	-6.847	1	-9.634	1
45	they /3 could	-4.472	1	-9.697	1
44	had /3 had	-6.1	1	-9.716	1
43	their /3 they	-5.464	1	-9.723	1
42	she /3 would	-7.251	1	-9.764	1
41	if /3 they	-2.199	1	-9.765	1
40	she /3 could	-7.207	1	-9.775	1
39	could /3 see	-6.355	1	-9.861	1
38	but /3 one	-4.497	1	-9.892	1
37	what /3 had	-5.959	1	-9.918	1
36	even /3 though	-4.794	1	-9.952	1
35	what /3 do	-2.831	1	-10.042	1
34	they /3 their	-4.363	1	-10.09	1
33	united /3 states	-6.525	1	-10.25	1
32	had /3 her	-7.786	1	-10.275	1
31	after /3 all	-5.489	1	-10.352	1
30	no /3 longer	-4.52	1	-10.374	1
29	had /3 come	-6.508	1	-10.416	1
28	one /3 another	-6.053	1	-10.55	1
27	she /3 she	-7.734	1	-10.589	1
26	so /3 much	-5.362	1	-10.851	1
25	but /3 she	-7.574	1	-10.851	1
24	years /3 ago	-4.82	1	-10.87	1
23	their /3 own	-3.797	1	-10.874	1
22	each /3 other	-5.584	1	-10.928	1
21	had /3 him	-7.62	1	-11.116	1
20	she /3 said	-9.633	1	-11.213	1
19	but /3 also	-4.592	1	-11.232	1
18	but /3 had	-6.33	1	-11.264	1
17	when /3 she	-8.343	1	-11.366	1
16	if /3 had	-5.494	1	-11.409	1
15	when /3 they	-5.558	1	-11.485	1
14	which /3 had	-5.782	1	-11.538	1

13	had /3 no	-6.79	1	-11.919	1
12	but /3 there	-6.453	1	-12.068	1
11	her /3 her	-9.586	1	-12.233	1
10	but /3 they	-5.032	1	-12.35	1
9	she /3 her	-10.653	1	-13.442	1
8	no /3 one	-7.523	1	-13.507	1
7	there /3 no	-2.474	1	-13.571	1
6	she /3 had	-13.494	1	-13.67	1
5	who /3 had	-6.882	1	-13.708	1
4	they /3 had	-8.697	1	-13.785	1
3	her /3 she	-11.927	1	-14.226	1
2	had /3 been	-16.055	1	-20.443	1
1	more /3 than	-5.169	1	-21.503	1