Looking for the Smoking Gun : Principled Sampling in Creating the Tobacco Industry Documents Corpus

William A. Kretzschmar, Jr., Clayton Darwin, Cati Brown, Donald L. Rubin and Douglas Biber

Journal of English Linguistics 2004 32: 31
DOI: 10.1177/0075424204263024

Looking for the Smoking Gun

Principled Sampling in Creating the Tobacco Industry Documents Corpus

WILLIAM A. KRETZSCHMAR, JR.
CLAYTON DARWIN
CATI BROWN
DONALD L. RUBIN
University of Georgia

DOUGLAS BIBER
Northern Arizona University

As a result of litigation over the past decade, major tobacco companies were compelled to make public a broad range of previously confidential documents. We have created a series of corpora from the tobacco industry documents (TIDs) for three purposes: (1) to establish baseline descriptions of various linguistic features of this unique set of texts; (2) to identify TIDs in which rhetorical manipulation ("deception") may have occurred and to estimate the extent and prevalence of manipulation; (3) to analyze manipulation in order to classify it and develop means to identify similar manipulation in other industry document sets. Our three-part corpus creation strategy employed rigorous sampling methods. First, we drew a limited sample from the largest collection of TIDs, to determine a representative classification of text types and to estimate their proportions within the overall body of texts. Then, we created a reference corpus (500,000+ words) constituting a stratified random sample of all TIDs, whether or not they exhibit manipulation. Finally, we compiled a corpus of texts presumed to exhibit rhetorical manipulation. We assumed that multiple drafts of a text or versions of a text prepared for different audiences constituted rhetorical manipulation. This article presents our experience with the sampling methods utilized in this corpus-building process and our findings regarding text types comprising the reference corpus.

Keywords: *corpus linguistics; rhetorical manipulation; text sampling methods; tobacco control*

The tobacco industry documents (TIDs), which number more than four million, were produced by tobacco industry defendants as a result of state and federal litigation and legislative hearings (Hurt and Robertson 1998; Malone and Balbach 2000) and cover the complete range of corporate operations in the tobacco companies,

Journal of English Linguistics, Vol. 32 / No. 1, March 2004 31-47
DOI: 10.1177/0075424204263024

from letters and memoranda to research papers to procurement invoices. The remarkable nature of this document set lies not only in its astonishing size but also in the rather candid and comprehensive window it opens on the workings of an entire industry. While some document export and destruction no doubt took place when executives realized that their files were likely to be made public, the vociferous opposition to disclosure mounted by industry attorneys bears testament to the uncensored nature of the set as a whole. The tobacco industry documents are stored physically in depositories in Minneapolis (the site of the original trial) and Guildford, England. Large collections of them are now also available in electronic form on the Web.

While most current research on TIDs pertains to revelations concerning marketing strategies, product design, or deception in reporting scientific evidence (Bero 2003), the National Cancer Institute has granted support for our multidisciplinary team to conduct research on the TIDs qua language artifacts. Our group's aim is to detect deceptive and manipulative language strategies possibly employed by the tobacco industry in its efforts to obstruct public understanding of tobacco company activities. In short, we were funded to look for the linguistic smoking gun in the long-running conflict between tobacco interests and public health.

The first and still most well-known book about TIDs, Glantz et al.'s *The Cigarette Papers* (1996), is based on a small set of Brown and Williamson company documents (1,383 documents, under 10,000 pages) that was leaked to the press and to Glantz in 1994. The public furor caused by this act of whistle-blowing eventually led to the Master Settlement Agreement, which was signed in 1998 by the attorneys general of forty-six states and the nation's seven major tobacco industry organizations: Philip Morris, R. J. Reynolds, Brown and Williamson, the American Tobacco Company, the Council for Tobacco Research, and the Tobacco Institute (Fisher 2000). Early analyses of TIDs by Glantz et al. (1996) were quickly followed by a wealth of subsequent studies (see, e.g., papers collected in a special issue of *Tobacco Control* 2002; Cummings and Pollay 2002). This body of research has pursued questions such as whether tobacco company officials knew about the adverse health effects of smoking at the same time they were denying that smoking caused disease, whether the TIDs revealed deliberate efforts to promote addiction, and whether tobacco companies had attempted to exert undue influence on public policy in an illegal manner. For example, the Glantz et al. exposé begins with a foreword by former Surgeon General C. Everett Koop, in which he states that "the contrast of public and private statements from the tobacco industry reveals their

AUTHORS' NOTE: This research was supported by a grant titled 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, however, are solely those of the authors.

deceit," and further comments on "the sleazy behavior of the tobacco industry" and how "the medical and public health professions were misled by the tobacco industry" (p. xiv). In his own preface, Glantz leads with the fact that "cigarettes and other tobacco products kill 420,000 American smokers and 53,000 non-smokers every year," and closes with the assertion that "these documents provide an opportunity to see firsthand how the brown plague of tobacco has been allowed to flourish and spread. . . . Perhaps this understanding will finally lead the public and public policy makers to deal with the tobacco industry in a manner appropriate to the amount of death and suffering it knowingly creates" (pp. xvii, xix).

This kind of analysis, while certainly of profound import, addresses the content of selected TIDs rather than the language which represents that entire universe of discourse. That is, it can establish neither a pattern of behavior within tobacco companies nor the patterns of language supporting such behavior. However damning these particular documents may appear to be, they constitute only isolated instances, which we cannot assume to be indicative of the whole. As Glantz et al. (1996) admit of the Brown and Williamson document set, "One of its limitations has to do with the possibility of selection bias; that is, the documents may have been picked by a whistle-blower with an eye toward smoking guns" (p. 11). Linguistic and rhetorical research conducted on such grounds, as opposed to legal or policy analysis designed to persuade adjudicators and regulators, would leave itself open to attack owing to highly selective use of data. For this reason, the premise of our work is to treat the TIDs as a corpus and to apply accepted methods of corpus and forensic linguistics and rhetorical analysis in order to

(1) establish baseline descriptions of various linguistic features of this unique set of texts;
(2) identify TIDs in which rhetorical manipulation ("deception") may have occurred and to estimate the extent and prevalence of manipulation; and
(3) analyze manipulation in order to classify it and develop means to identify similar manipulation in other industry document sets.

Of course, these objectives required sampling from the TID set to create a reasonable corpus for study, since we do not have the resources to include the entire set of several million documents. We report here our experience with this process, our findings regarding the distribution of document types sampled, and some notions of how we proceeded with our analysis given those findings. We believe that this experience has value for corpus linguistics and other linguistic researchers, not just for tobacco control investigators, because it offers a model for large-scale forensic investigation of corporate documents with rigorous sampling, and its results begin to show previously undocumented aspects of corporate discourse.

We began our work with a three-part strategy for corpus creation that employed well-defined sampling methods. We first drew a limited sample from the entire body of TIDs (an exploratory "core sample"), so that we could determine the best classification of text types and estimate their proportions within the overall body of texts. We then created a reference corpus from those text types we considered relevant to (i.e., subject to) rhetorical manipulation. This reference corpus is a stratified random sample of all eligible TIDs, regardless of whether they were expected to contain any manipulation. Finally, we compiled a corpus of parallel texts of particular rhetorical interest because they are subject cross-draft and cross-audience analysis.

Exploratory Core Sample of TIDs

The extant set of TIDs comprises millions of documents, ranging in length from just a few words to hundreds of pages, and it is clearly not possible to inspect every document. Yet we did need to know what types of documents exist in the set of TIDs and, more specifically, what types of documents exist that would be relevant to the project. Furthermore, we needed to know the extent of those documents, both the quantity of relevant documents and how long they tend to be. We could not create a representative sample of relevant documents without this information. To accomplish these goals, we drilled deeply into the larger document set, sampling the body of TIDs according to a fixed random-sampling frame, a procedure which gave every document in the collection an equal chance of selection. (For general issues of sampling in a corpus situation, see Kretzschmar, Meyer, and Ingegneri 1997.)

For pragmatic reasons, we decided to draw for this exploratory core sample a randomly drawn set of about 300 documents from the body of TIDs, each of which would be closely inspected. To define our sampling universe, we based our sample on the seven industry organizations and documents found in the National Association of Attorneys General (NAAG), Master Settlement Agreement, 1999 "Digital Snapshot" (see Office of the Attorney General, State of California 2001: http://caag.state.ca.us/tobacco/resources/msasumm.htm). The NAAG set consists of the documents of The Tobacco Institute, The Council for Tobacco Research, Lorillard, R. J. Reynolds, Philip Morris, Brown and Williamson, and The American Tobacco Company prior to January 1999. Because of their special value for the tobacco-control community, we also included the roughly 33,000 documents of the so-called Bliley Collection, most of which were excluded from the NAAG set because they were earlier withheld as trial evidence on the basis of attorney-client privilege. Eventually, these very sensitive documents were subpoenaed by Chairman Bliley of the House Committee on Commerce (1997-1998, see The House Committee on Commerce Web site: http://www.house.gov/commerce/TobaccoDocs/documents.html) and made publicly available. The NAAG set as well as the Bliley Collection is

TABLE 1
Sampling Targets for the Exploratory Core Sample

Group Decade	Total Documents	Year/Month	Documents	Need	Taken	%/Sample
1900-1959	103,574	1/August	1,193	10	10	3.055
1960	223,544	0/April	1,136	22	22	6.593
1970	660,223	12/September	5,895	66	66	19.473
1980	1,318,813	1/January	7,185	132	132	33.898
1990	988,793	6/June	1,679	99	99	29.164
Undated	62,494	n/a	n/a	6	10	1.843
Bliley	33,003	n/a	n/a	3	10	0.973
Total	3,390,444		338	349		

indexed and searchable by an attorneys' document index (but not by text) at two consolidated online archives (Tobacco Control 2002).

The advantages of the electronic processing and fixed boundaries of these online collections rendered the logistics far more manageable than trying to enumerate the approximately 27 million pages of hard-copy documents residing in cardboard file boxes at the Minnesota Tobacco Document Depository. It is known that the NAAG set consists of approximately 3.4 million documents. Therefore, we drew a .001 percent sample to constitute a 340-document core sample.

To begin this exploratory core sampling, we used the online document search engines to determine how many documents were indexed within each decade. This information allowed us to determine the proportion of documents that should be drawn from each decade, as represented in Table 1. Because there are relatively few documents from the early decades (1900-1950), we regrouped the documents prior to 1960 into a single frame. Next, we randomly selected one year within each decade (after 1900, since only very few TIDs are available for the nineteenth century), and for each year selected, we randomly drew a single month. The Bliley Collection was treated as if it comprised its own decade frame, as was the set of all undated documents (which could not be placed in decades based on indexing information). We then proceeded to select every *n*th document from each selected month up to the proportional targets shown in Table 1, where *n* equals the last two digits of the year selected for the decade. We deviated from our proportional targets only to take at least ten documents from each group, which meant selecting a few additional documents from the undated and the Bliley sets. This raised the number of documents in the exploratory core sample to 349.

Once the documents for the exploratory core sample were collected using the above procedure, they were classified by a number of criteria that would properly exclude any ineligible documents from being of objects of study in our corpus:

(1) Texts that were authored by sources outside the tobacco industry were excluded. For example, newspaper clippings, government reports, and even antismoking handbills often found their way into the files of tobacco industry employees. We were interested only in archiving language originating within this industry; however, our criterion for industry-internal encompassed any individual or organization known to have received financial support from any tobacco source. Extensive documentation of tobacco funding and organizational affiliation of a "cast of characters" is available at Tobacco Documents Online (http://www.tobaccodocuments.org), and no ambiguous cases arose.
(2) Texts that had no public health significance were excluded. Our definition of public health significance was quite broad, encompassing, for example, memoranda characterizing various politicians as friendly or unfriendly to the industry. On the other hand, information pertaining solely to internal corporate affairs such as executive parking permits or utility bills for manufacturing facilities was not of interest.
(3) Texts that were written in languages other than English were excluded, as they would be misanalyzed by many of our computer-assisted tools.
(4) Short documents, operationalized as those containing fewer than fifty words of continuous discourse, were excluded. Thus, most stand-alone cigarette advertisements were not included in the corpus, since they would likewise defy analysis by computer-assisted tools. On the other hand, if an advertisement was included as part of larger text or if it contained a caption at least fifty words long, it was included in the corpus. (It should be noted that to avoid biasing results due to a few very long documents—documents of fifty pages or more were not rare—only three thousand words were taken from long documents. One thousand of these words were taken from the beginning, one thousand from the middle, and one thousand from the end.)

Two additional coding categories were of interest because of their rhetorical significance. Along with decade, these would generate quotas to drive the stratification of the larger reference sample:

(1) Target audiences for each document were coded as to whether they were (a) industry-internal or (b) industry-external. The same decision rule was used for this categorization as was used for determining whether a document was from an internal or an external source.
(2) Document audiences were further classified as to whether they were (a) named or (b) unnamed. The rationale for this coding was that a named/unnamed audience might roughly distinguish between genres of letters and memoranda on one hand (typically named recipients) and reports (unnamed) on the other. A named audience was operationalized as designating an individual or a distribution list of named individuals. Lists of corporate

TABLE 2
Distribution by Classification Categories

	No Date	Bliley	1950s	1960s	1970s	1980s	1990s	Total
Total documents	10	10	10	22	66	132	99	349
Internal source	8	9	8	20	55	108	93	301
Internal audience	8	9	6	20	53	109	88	293
Named audience	0	7	3	13	27	62	33	145
Public health	9	10	10	22	61	126	96	334
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	8	5	12	34	73	39	176
Short	3	2	2	4	20	37	33	101

titles or roles or organizational units (e.g., "Latin American Sales") were treated as unnamed audiences.

Finally, with respect to categorizing the exploratory core sample, we wished to characterize certain surface features of these documents. Thus, coders also tabulated whether documents (a) consisted primarily of a form (like an invoice), (b) were predominantly graphic or photographic images, (c) contained interlinear editing, and (d) had handwritten marginalia commentary. The latter two items were considered valuable as potential indicators of manipulative intent.

The 349 documents in the core exploratory sample were each examined twice, once for an initial classification that allowed the definitions to be tested for functionality and then a second time to verify that the classification of each document agreed with the final set of definitions. In addition, a set of 50 documents was randomly selected for a check of interrater reliability. The rate of agreement between two independent coders was 145/150 codings. The five discrepancies were examined and resolved for the entire set, resulting in the distribution of classifications shown in Table 2.

As expected, we found that the documents were not evenly distributed among our primary classification categories. Most of the documents in our limited sample (82.8 percent) consisted of industry internal documents with industry internal audiences, fewer than half (41.8 percent) with named recipients. Almost all of the documents examined (98.3 percent) were English. Fourteen percent of the core sample documents were primarily forms, and images comprised just over 1 percent. A substantial number of the documents (28.9 percent) were short and therefore not analyzable by our criteria. About 5 percent of documents showed evidence of editing, but about half (50.4 percent) contained marginalia of some kind. The tendencies

TABLE 3
Stratification Quotas for Reference Corpus (percentages)

Decade	Named + Internal	Named + External	Unnamed + Internal	Unnamed + External
Bliley	3	0	0.5	0
No date	0	0	2.5	0
1950s	1	0	1	1
1960s	4	0.5	3	0.5
1970s	6.5	1	10.5	0
1980s	19	0.5	16	0
1990s	10.5	0	20	0

were generally consistent across all seven of our decade frames (including the Bliley set and the set of undated documents).

Reference Corpus

Descriptive statistics derived from the exploratory core sample were used to generate sampling quotas for a larger reference corpus. The purpose of the reference corpus was to create a comparison set of TIDs from among those in which manipulation potentially occurred (but did not necessarily occur), from which we could estimate the general frequency of occurrence of linguistic characteristics of interest in the analysis of rhetorical manipulation. Because many of these characteristics might occur at low frequencies, the corpus had to be large enough to ensure that they were represented. On the other hand—especially considering that all documents would need to be keyed by hand (documents are available as image files, and image quality is generally too degraded to allow adequate optical character recognition)—the corpus could not be so large as to overrun available resources. Accordingly, a corpus size of about 500,000 words was designed.

After excluding all external source, short, and/or non-English texts, only 57.9 percent (202) of the exploratory core documents remained. Nearly all of these (96.5 percent) were addressed to industry-internal audiences; documents addressed to external audiences proved much rarer than anticipated. The eligible documents were fairly evenly split with respect to named or unnamed audiences (45.5 percent named audiences). The resulting sampling quotas used to stratify the reference corpus by decade, audience affiliation, and audience namedness are shown in Table 3. Applying all the proportions in this sampling plan resulted in a final reference corpus of 808 documents and 529,000 words.

To check the relationship between the reference corpus and the exploratory core sample on which it was based, we calculated a comparative-sampling rejection ratio. In the core sample, 42.12 percent of the documents proved ineligible for inclusion, due to short length, non-English language, or industry-external authorship.

TABLE 4
Supplemental External Audience Corpus (percentges)

Decade	% Documents Overall	Named + External	Unnamed + External	Total
Bliley	1	0	1	1
No date	2	1	0	1
1950s	3	2	2	4
1960s	7	3	3	6
1970s	19	10	10	20
1980s	39	19	19	38
1990s	29	15	15	30

This rejection ratio was almost exactly replicated (42.04 percent) when we sampled documents to fill quotas for the reference corpus, thus lending additional credibility to the sampling process we utilized.

Constructing a Supplemental Stratified Random Sample for External Audience Documents

Our supposition in this project was that comparing documents directed to industry-external audiences with those directed toward industry-internal audiences would prove of heuristic value in identifying potential linguistic markers of manipulation and deception. However, about only 3.5 percent, or 28 documents out of the 808-document reference corpus, were addressed to external audiences. To address the low count of rhetorically significant external audience documents, we constructed an additional corpus of external audience documents only, the supplemental corpus. The supplemental corpus was constructed using the same decade and named/unnamed audience quotas as were employed for the reference corpus. The same procedures for randomly sampling documents within each of the strata were also utilized. The distributions of the strata are shown in Table 4.

Again owing to the demands of hand-keying these texts, it was determined to limit the supplemental sample to 100 documents. The resulting corpus is composed of 49,146 words. Interestingly, the average length of documents directed to external audiences is shorter than that of internal-directed documents, 422 words versus 655 words, owing largely to the absence of long reports in the set of supplemental corpus of external documents.

Corpus of Parallel Rhetorical Texts

The exploratory core sample, the reference corpus, and the supplementary external-document corpus all utilize random-sampling procedures. In contrast, the

corpus of parallel rhetorical texts is sampled deliberately and, to some extent, serendipitously. Two kinds of rhetorical "cases" were sought for this corpus:

(1) Cross-draft cases collect multiple drafts of documents as they evolve over time within the organizations. Thus, one cross-draft case might begin with an early draft of a letter, report, or speech text. This may be followed by several copies of that draft with interlinear editing marks by lawyers, public relations specialists, and executives. Several additional typed drafts may follow in sequence. At a minimum, cross-draft cases must include two parallel documents. Our longest cross-draft rhetorical case to date contains twelve sequential or simultaneous drafts. The value of cross-draft cases is that the nature of the edits frequently reveals industry concerns or efforts to manipulate readers' perceptions.
(2) Cross-audience cases pair an internal-audience document with an external-audience document, controlling for identical content. To authenticate the external member of this pair for these cases, we require that it appear on letterhead, be published in a newspaper or magazine, or bear some other evidence that it was actually issued. The value of the cross-audience cases is that they can provide very clear evidence of how particular issues were spun differently for the public, as compared to audiences presumed to have allegiance to tobacco interests.

Constructing the corpus of parallel rhetorical texts is a work in progress. As of this writing, it is composed of 38 cross-draft cases, in turn consisting of a total of 157 documents; and 24 cross-audience cases, consisting of a total of 49 documents. Altogether, the corpus of parallel rhetorical documents consists of 139,229 words of text.

Examples of Use of the Reference/Sample Corpus

While rhetorical analysis of "deception" awaits further study, we can offer some initial findings on the characteristics of the language of the reference sample. We have, for instance, created word lists from the TIDs and compared them with the benchmarks of the Brown Corpus (1961; as found in the International Computer-Archive of Modern and Medieval English [ICAME] set; see http://www.hit.uib.no/icame/cd/credit.htm). The keyness metric in the Keywords tool of WordSmith Tools (Scott 1998) was used to develop a list of the top fifty "content" words in the tobacco industry documents reference corpus (TIDC). *Keyness* is a measure of whether a word has a higher or lower frequency than expected in a subcorpus, given "normative" frequencies derived from some larger corpus of language (the Brown Corpus in our case). A significant positive jump in the keyness of a word simply means that the word is more frequently used in a particular setting than was ex-

pected. Conversely, a keyness change that is negative indicates a much lower frequency of use for a particular word. Our fifty key content words were all significantly more frequently used ($p < .05$) in the TIDC than in the Brown Corpus. By *content words*, we mean those that are not single letters, not titles like *Mr.* and not mechanical writing terms like *page* or *cc*. We also excluded function words like articles or prepositions and common verbs like *will* or *be*. Because of these exclusions, the keyness rankings reported in Table 5 range from rank 1 (*tobacco*) to rank 98 (*promotion*); that is, we had to consider the first ninety-eight words of all types to build our set of fifty content words.

The fifty content words that we identified in the TIDC fall naturally into four clusters based on their reference: (1) industry terms and company names, (2) marketing terms and brand names, (3) disease terms, and (4) marketing research terms. These clusters are displayed in Table 5.

The first cluster contains the words of the trade: the product and its components, words for the act of using it (*smoke*, *smoking*, *smoker*), and company names. The second group shows the vocabulary of selling the product, including brand names and marketing strategies (*blend*, *flavor*, *lights*, *taste*) as well as business terms like *advertising*, *market*, or *retail*. The third group comes from the industry's attempt to confront the health effects of smoking. The last group consists of words with applications to research, both market research and product research, and thus represents a combination of the marketing and health concerns in the industry. Only one word from the top fifty, *current*, cannot readily be classified into one of the four groups.

Trends across time are an especially telling topic of inquiry in tobacco industry research (e.g., Hilts 1996). Half a century ago, the tobacco industry was a well-respected, corporate, global citizen marketing a product that was often associated with relaxation and hardiness. During the course of those fifty years, however, revelations both internal and external to the industry provided increasing proof that the product was a major menace to human health. Toward the end of the fifty years, the industry experienced pressure from regulators and public health advocates. Tracing diachronic language patterns is a valuable exercise in documenting how the industry responded to changing conditions. Thus, in addition to the words themselves, Table 5 also indicates whether each word is significantly correlated ($p < .05$), either positively (i.e., the word occurred significantly more often in the decade than in the corpus overall) or negatively (i.e., the word occurred significantly less often in the decade than in the corpus overall), as indicated by a comparison of rate of occurrence of the word in the TIDC as a whole with the rate of occurrence in each decade. Of the 22,000 word types in the TIDC, 1,538 were significantly associated with at least one decade. Table 5, in effect, shows a history within industry documents of the words most distinctive of that industry. For example, the company name *RJR* is negatively associated with the first four decades of the period and posi-

TABLE 5
Content Words Occurring More Frequently in the Tobacco Industry Document Reference Corpus than in the Brown and Frown Corpora

Keyness Rank	Word	1950s	1960s	1970s	1980s	1990s
Cluster 1: Industry terms and company names						
69	Carton		negative	positive		positive
4	Cigarette(s)	positive	positive			negative
18	Filter	negative	positive			negative
22	Menthol	negative	negative		positive	
25	Morris	positive		negative		positive
51	Pack		negative	positive	negative	positive
31	Philip	positive		negative		positive
15	Product(s)	negative	negative	negative	positive	positive
58	Reynolds		negative	negative	positive	
40	RJR		negative	negative	negative	positive
7	Smoke	positive	positive		negative	
5	Smoker(s)		negative			
3	Smoking				negative	negative
1	Tobacco		positive		negative	
Cluster 2: Marketing terms and brand names						
44	Advertising		negative			
88	Blend		positive		negative	
12	Brand(s)		negative	negative		
32	Camel	positive	negative	negative	negative	positive
62	Flavor					
72	Kool					
54	Lights	negative	negative	negative	positive	negative
60	Market		negative	negative	positive	positive
21	Marlboro	negative	negative	negative	positive	positive
94	Media					
98	Promotion	negative	negative	negative	negative	positive
92	Retail	negative	negative		negative	positive
91	Salem			negative	positive	
63	Sales	negative				positive
53	Share	negative	negative	negative		positive
78	Taste		positive	negative	positive	
80	Winston		negative	negative		positive
Cluster 3: Disease terms						
49	Cancer	negative	positive	positive	negative	negative
56	Exposure	negative	negative	positive	negative	positive
57	Health	negative	positive		negative	
76	Lung	negative	positive		negative	
10	Nicotine	negative	negative	positive	negative	
37	Tar	negative			positive	negative

TABLE 5 (continued)

Keyness Rank	Word	1950s	1960s	1970s	1980s	1990s
Cluster 4: Marketing reserch terms						
90	Analysis	negative	negative			positive
42	Data	negative	negative			
64	Levels	negative				
84	Low	negative	negative			positive
68	Project		negative	positive	positive	negative
50	Report		positive			negative
13	Research	negative	positive	negative		negative
34	Results					
61	Sample(s)	negative			positive	
36	Study(ies)	negative		negative		
14	Test	negative	negative	negative	positive	negative
65	Testing	negative			positive	
Unclassified Words						
93	Current		negative		positive	

tively associated with the 1990s. This follows from the change in corporate name (originally *RJ Reynolds*) after its merger to form *RJR Nabisco*. The other words in the industry set change their rate of use over time, not always predictably. It is interesting to note that *cigarette(s)* changes from a highly used term in the 1950s to a term used at a significantly lower rate in the 1990s, while the words *carton*, *pack*, and *product* are used at significantly higher rates in the 1990s. This pattern may indicate less focus in the industry on what the product is and more emphasis on its packaging for sale.

The cluster of disease terms shows perhaps the most telling trend. All of the words were negatively associated with the 1950s, before public health became big news. In the 1960s and 1970s, the health terms were used at significantly higher rates in industry documents, first, as they related to cancer, and somewhat later, as associated with nicotine. In the 1980s, the industry used words in this set at significantly lower rates, except for *tar*, which was associated with the marketing of low-tar cigarettes during that period. The same is true for the 1990s, except for *exposure*, which at that time was a word commonly used in discussion of what has come to be called *secondhand smoke*.

The marketing set shows what the industry was talking about in the moments when health was not the topic in the 1980s and 1990s. In the first three decades of the document set, only the occasional brand or market strategy is positively associated with any decade (*camel*, *taste*, *blend*). In the 1980s and 1990s, a large number of marketing terms come to be used at significantly higher rates. The research ter-

minology cluster tends to bear out these trends. These terms may refer to both marketing and health effects research. They are used at significantly lower rates in the 1950s, before the expansion of either health research or marketing efforts, but different words from this set rise to notable rates of use in the following decades.

There is much more to be learned from analysis of the vocabulary of the reference corpus, yet even these early and rough results show that the language of the tobacco industry reveals its concerns and practices. We did not find that the words of industry documents simply replicated normal rates of usage, as found in the Brown Corpus, nor that the lexicon of TIDs merely reflected the names and product terms of the industry. The use of words in industry documents is consistent with facts that have been established and examined via other means (e.g., legal inquiry, instances of whistle-blowers). What this corpus linguistic perspective supplies, something that is largely missing in those other realms of inquiry regarding the tobacco industry, are the virtues of (a) systematic rather than selective sampling and (b) concrete rather than impressionistic indices of document themes and content.

Conclusion

At this stage, we cannot show readers the linguistic smoking gun from the tobacco documents. We have, however, been able to explicate some of the resources becoming available for finding it. Our corpus design strategy was (1) to draw an exploratory core sample, (2) followed by a reference corpus contingent on characteristics of that core sample, (3) to be compared with corpus of parallel rhetorical texts. This strategy has proved both feasible and illuminating. The exploratory core sample allowed us to determine the nature of the document set, and some of the findings surprised us, notably, the extremely high proportion of documents addressed to industry-internal audiences and the prevalence of rather short documents. In addition, we found that the proportions of documents in our different classifications stayed relatively consistent over time, even though the weight of documents in particular decades is highly variable.

Our findings already tell us something about corporate documents and corporate communication in general that, to our knowledge, has not previously been documented. While some researchers are applying corpus linguistic methods to the analysis of business documents (Upton and Connor 2001; Gunnarsson et al. 1997), the TID corpora we describe here are, we believe, unique in that they consist of documents that were released in a nonvoluntary, and therefore relatively noncensored, manner. No previous studies of business corpora have been able to draw these kinds of conclusions about the overall shape of the universe of discourse from which they draw their artifacts. For example, we expected that a large proportion of corporate documents were likely to be composed by industry authors, but we were surprised to see that fully 86.2 percent of the documents had internal sources, as we were also

not anticipating the extreme paucity of external audience documents. Nor did we have any basis for predicting that fewer than half of the documents, only 41.5 percent, would be letters or memoranda addressed to named recipients.

To the extent that the document profile of the tobacco industry is not different from the profile in other industries, these findings may be helpful to those planning linguistic studies of corporate discourse. Corporate communications appear to be internal to such a great degree that it will be difficult for researchers to find examples of documents with external addressees, especially if they are interested in documents long enough to fairly assess their linguistic or rhetorical features. On the other hand, researchers interested in more self-reflexive linguistic or rhetorical study of corporate communications will have rich, perhaps overwhelming quantities of documents to explore. The proportion of short documents seems likely to increase in the future, as e-mail and instant messaging become increasingly important parts of the corporate document profile. The fact that over half of our documents (50.4 percent) contained marginalia suggests the extent to which corporate communication takes place, or is interpreted, literally outside the prepared text of messages, which will present (now-predictable) problems in the construction of new corpora of corporate documents. Finally, even our initial, elementary analyses of word frequency in the Reference Corpus reflect the same kind of issues developed by content analysis of the tobacco documents, only without risk of any charge of selective use of the evidence.

Our findings are of course not unrelated to the purposes of our work. We are in process of developing XSLT-driven interfaces to facilitate public access to these corpora online and on distributed media. If the Master Settlement Agreement has made it possible for linguists to ask and answer questions of interest to them, that will be an additional and welcome, if unexpected, form of compensation from the tobacco industry.

References

Bero, Lisa. 2003. Implications of the Tobacco Industry Documents for Public Health and Policy. *Annual Review of Public Health* 24:267-88.

Cummings, K. Michael, and R. W. Pollay. 2002. Exposing Mr Butts' Tricks of the Trade. *Tobacco Control* 11 (suppl. 1): i1-i4.

Fisher, Laurie. 2000. Update: Master Settlement Agreement between the States and the Tobacco Industry (United States). *Cancer Causes & Control* 11 (3): 285-87.

Francis, N., and H. Kučera, directors. (early 1960s). *The Brown Corpus*. Brown University, Providence, RI.

Glantz, Stanley, John Slade, Lisa Bero, Peter Hannauer, and Deborah Barnes. 1996. *The Cigarette Papers*. Berkeley: University of California Press.

Gunnarsson, Britt-Louise, Per Linell, and Bengt Nordberg, eds. 1997. *The Construction of Professional Discourse*. New York: Longman.

Hilts, Philip J. 1996. *Smokescreen: The Truth Behind the Tobacco Industry Cover-up*. Reading, MA: Addison-Wesley.

The House Committee on Commerce. 1998. Chairman Tom Bliley Releases Subpoenaed Tobacco Documents to the American People. http://www.house.gov/commerce/TobaccoDocs/documents.html.

Hurt, Richard D., and Channing R. Robertson. 1998. Prying Open the Door to the Tobacco Industry's Secrets about Nicotine: the Minnesota Tobacco Trial. *Journal of the American Medical Association* 280:1173-81.

Kretzschmar, William A., Jr., Charles Meyer, and Dominique Ingegneri. 1997. Uses of Inferential Statistics in Corpus Studies. In *Corpus-based Studies in English*, edited by Magnus Ljung, 167-77. Amsterdam: Rodopi.

Malone, Ruth, and Edith Balbach. 2000. Tobacco Industry Documents: Treasure Trove or Quagmire? *Tobacco Control* 9:334-38.

Office of the Attorney General, State of California. 2001. Tobacco Master Settlement Agreement Summary. http://caag.state.ca.us/tobacco/resources/msasumm.htm.

Scott, Michael. 1998. *WordSmith Tools*. Oxford, UK: Oxford University Press.

Tobacco Control. 2002. How to Access Tobacco Industry Documents. *Tobacco Control* 11 (suppl. 1): i39.

Tobacco Documents Online. 1999-2003. http://www.tobaccodocuments.org.

Upton, Thomas, and Connor Ula. 2001. Using Computerized Corpus Analysis to Investigate the Textlinguistic Discourse Moves of a Genre. *English for Specific Purposes: An International Journal* 20:313-29.

William A. Kretzschmar, Jr. is Professor of English and Linguistics at the University of Georgia. His major publications include the Oxford Dictionary of Pronunciation for Current English *(with Clive Upton and Rafal Konopka; Oxford U Press, 2001);* Introduction to Quantitative Analysis of Linguistic Survey Data *(with Edgar Schneider, Sage Publications, 1996);* Handbook of the Linguistic Atlas of the Middle and South Atlantic States *(with Virginia McDavid, Theodore Lerud, and Ellen Johnson; U Chicago Press, 1994). He served as editor of the Journal of English Linguistics for 15 years and now serves as board member for JEngL, the TEI Consortium, and various professional journals, atlases, and dictionaries, including preparation of American pronunciations for the new online Oxford English Dictionary.*

Don Rubin is Professor in the Departments of Speech Communication and Language Education and in the Program in Linguistics at the University of Georgia. He is Principal Investigator for the National Cancer Institute funded grant, "Linguistic Analyses of Tobacco Industry Documents" and also edits Communication Education. *Among his research interests are studies of linguistic adaptation to writers' audiences.*

Clayton Darwin is a Doctoral Candidate in Linguistics at the University of Georgia, where he works as a research assistant on the National Cancer Institute funded grant "Linguistic Analyses of Tobacco Industry Documents."

Cati Brown is a Doctoral student of Linguistics at the University of Georgia. She is research assistant for the National Cancer Institute funded grant "Linguistic Analyses of Tobacco Industry Documents." Her doctoral dissertation will address aspects of language use in the tobacco industry.