RATING TOBACCO INDUSTRY DOCUMENTS FOR CORPORATE DECEPTION AND PUBLIC FRAUD: A CORPUS LINGUISTIC ASSESSMENT OF INTENT

by

CATHERINE GENE BROWN

(Under Direction the of Dr. Don Rubin)

ABSTRACT

Publicly available tobacco industry documents represent a window into an industry that perpetrated corporate deception and fraud that resulted in degraded public health and cost millions of lives. The current study addresses the topic of corporate deception and fraud from a linguistic standpoint, employing corpus methods, text analysis, (critical) discourse analysis and automated computational linguistic methods to assess a selection of six automated linguistic indicators of deceptive corporate strategy.

These six linguistic indicators of deceptive corporate strategy were mined from an extensive body of deception and language research. These indicators represent common themes and observations in the literature and include the following: adversarial language, allness and superlative language, deprofiled agency due to overuse of passive constructions, group mentality, cognitive verbs, and strategically ambiguous language. Computer programs were written and used to assess single documents for the instance of each linguistic indicator of deceptive corporate strategy.

Using the Tobacco Documents Corpus, a specialized full-text corpus representative of the entire body of tobacco industry documents, each indicator was assessed separately by source (company of origin), audience affiliation (internal or external to the tobacco industry), decade and audience type (individual or mass recipients).

Additionally, internal audience documents were automatically ranked for deceptive corporate strategy using a vector model method. Tobacco control literature has demonstrated that external audience documents are deceptive and fraudulent as a whole. Accordingly, the linguistic benchmark for deception was estimated by taking an average external audience document. Internal audience documents were ranked against this benchmark using the vector analysis classification method. To evaluate the efficacy of the indicators and the multivariate method for ranking documents, documents from the highest, middle and lowest rankings were assessed by-hand using (critical) discourse analytic methods.

Analysis validated the automatic ranking algorithm in part. However, statistical tests did not support hypotheses projecting higher instance of the six indicators in external audience documents and certain sources. Rather, deceptive corporate strategy can be better captured by examining potential indicators in concert. The automatic ranking algorithm results demonstrate an avenue for quickly organizing document in a large collection for subsequent discourse analysis.

INDEX WORDS: Corpus linguistics, deception research, business ethics, document ranking, discourse analysis, tobacco industry documents, corporate fraud, document categorization

RATING TOBACCO INDUSTRY DOCUMENTS FOR CORPORATE DECEPTION AND PUBLIC FRAUD: A CORPUS LINGUISTIC ASSESSMENT OF INTENT

by

CATHERINE GENE BROWN

B.A., Agnes Scott College, 2000

Ph. D, University of Georgia, 2006

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

2006

RATING TOBACCO INDUSTRY DOCUMENTS FOR CORPORATE DECEPTION AND PUBLIC FRAUD: A CORPUS LINGUISTIC ASSESSMENT OF INTENT

by

CATHERINE GENE BROWN

Major Professor:	Donald Rubin
Committee:	Michael Covington Bill Kreztschmar Betsy Rymes

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2006

DEDICATION

This is dedicated to my friends and family. To my brother, Nick Brown, who kept me sane during years of hard work. To my mother, Colleen Morich, and her mother (Grandmother) and father (Grandaddy) who were always committed to public education and honesty. To my dad, Paul Vaughn Brown, who would do just about anything to help out anyone (me included).

I would like to dedicate this dissertation to many important Kathryn/ines in my life: Kat Ballard, Kate Daley-Bailey, Kate Zimmerman and Dr. Kate Andersen. I hope a working coffee break with one of them is just around the corner.

Lastly, I would like to dedicate this work (again) to my grandmother, Eugenia Caldwell Johnson. She had never left the southern U.S. when she boarded a train for Chicago to attend graduate school at Garrett (now a part of Northwestern University). She made her own way as a woman, as a southerner and as a scholar. I can only hope to be as blessed.

ACKNOWLEDGEMENTS

I would like to thank my committee and professors. Dr. Don Rubin has been an encouraging advisor, a thorough editor and someone who could keep me on track. My committee, Dr. Betsy Rymes, Dr. Michael Covington and Dr. Bill Kreztschmar always demanded my best work and effort. In addition, devoted instructors and teachers have been a source of sustenance and inspiration throughout my education. Of those many important people in my life, some should be acknowledged by name: Mrs. Abbott, Mrs. Schlumpf, Dr. Mi-Ran Kim, Dr. Marlyse Baptista. I would also like to thank Dr. Rubin for introducing me to Dr. Robbin Derry. She has been an inspiration and in some ways, a guide.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Motivations for the Study

Quantifiable approaches to language have blown open the doors of linguistic research in the past twenty years. An era of linguistic study that was exclusively interested in what the possibilities of language use might be has given way to an era of linguistic study interested in what language use is and how language actually behaves in large samples of language, or corpora. Researchers prior to the advent of corpus and computational methods may have been interested in language use patterns in corpora comprised of spoken or written text, but they did not have the computational resources to investigate language on this scale. Corpus linguistics and its counterpart, computational linguistics, have enabled aggregate studies of language use that were previously tedious at best, or, at worst, impossible.

Early corpus linguistic studies (Biber, 1988; Sinclair, 1991) established that language patterns observed in large bodies of text are fundamental to understanding language use on an individual and aggregate level. Corpus linguistics has moved from strictly linguistic applications in language research areas such as genre studies (Biber, 1988) to widespread use in studies tracking social and psychological phenomena as they are related to and expressed by language. For example, research has tracked interpersonal deception and language use (Newman, Pennebaker, Berry, & Richards, 2003), language use and quality of life (Chung & Pennebaker, 2005), and even language-use correlations with emotional health in general (Nyklâiécek, Temoshok, & Vingerhoets, 2004). As social discourse is more thoroughly explored from a

corpus linguistic point of view, the possibilities of linking behavior with language use have expanded exponentially. At present, researchers are working to productively associate language patterns with human activities and psychological states – such as deception (White & Burgoon, 2001) – that are otherwise not yet fully predictable or wholly accounted for.

From the standpoint of practical applications of linguistics, the sociolinguistic revelation that language patterns can be correlated with groups, activities and even potentially states-of-mind is a cornerstone of powerful new corpus linguistic research. For example, initially, links between language and individual psychology were established using discourse analytic accounts of language use (Bruner, 1990). More recently, incorporation of corpus and computational techniques into studies linking psychology and language has unearthed correlations that have impacts in health and everyday life. Corpus methods have demonstrated a link between language usage patterns and recovery rates in schizophrenia (Breitborde, Lopes, Ochs & Nuechterlein, 2005) as well as links between language patterns and quality of life (Chung & Pennebaker, 2005), just to name a few of the wide range of productive applications.

Language patterns and language usage choices can be correlated not only with group-membership (Labov, 1964; Nevalainen & Raumolin-Brunberg, 1996) and activities (Lave & Wenger, 1991), but possibly also with intentions or goals. This dissertation uses an application of corpus linguistics to move another step towards deciphering the interaction between language use and intentions or goals, such as deflecting responsibility for socially unacceptable actions (Rymes, 1995). To do this, this study investigates correlations between language patterns and meta-discursive traits of corporate deception and fraud.

This dissertation attempts to build on discourse analytic and corpus linguistic research to isolate and test linguistic features of corporate deception. What kinds of language patterns are presently touted as being associated with deception? Focusing on a selection of these, this research uses the documents from the Tobacco Documents Corpus to assess the validity and transferability to corporate deception and fraud. Can these language patterns help us narrow enormous bodies of documents from inside a company or an industry down to a manageable size that could be assessed by a human expert?

Determining whether any particular document is deceptive is not the goal of the algorithms and computational assessments reported in this dissertation. This dissertation will not develop a linguistic lie detector. On the contrary, this dissertation is interested in determining whether links between language patterns and deception that have already been observed and tested stand up to the challenge of a varied corpus of corporate texts. In addition to each indicator being assessed singly, this research investigates how the indicators interact when used together. I anticipate that this multivariate approach will more closely approximate the linguistic complexities associated with deception. However, I want to insist that the final determination of whether or not a document is supporting a corporate agenda of public deception and fraud is a matter better suited to experts and human assessments, not computers.

Language, Deception and Corporate Malfeasance

There is inherent variability in language use that is coupled with necessary creativity and linguistic innovation needed to successfully deceive. This variability and creativity can only be expected to be matched by the variability and creativity of human assessors. Computers are not creative or variable by design and so can only get us so far in the direction of automated

assessments or document rating. The promise of automatic assessments of language is one of time and energy savings, but it should not be expected that computationally-based assessments could soon replace human assessments of this type. Despite this caveat, investigating relations between language and corporate deception and fraud is a pressing and urgent need. Corporate malfeasance and fraud have made headlines again and again during the past ten years. Enron executives have finally been put on trial for the alleged fraud perpetrated against their shareholders and investors in the late 1990s (Barrionuevo, 2006). The gross disregard for investor monies demonstrated by Enron was paralleled in the WorldCom and HealthSouth corporate malfeasance cases. These fraud cases shocked the public with their evident neglect of their duty to shareholders, investors, employees and the public alike. The artifacts available for scrutiny after the fact for these cases were previously unreleased internal corporate documents and recordings. Language artifacts and ruined lives were in the end the only evidence available to investigators and researchers trying to piece together what went wrong in these companies.

The type of monetary fraud involved in recent cases such as those mentioned above, however, is trumped by the financial and personal impact of tobacco industry negligence and deception. Figures from the Center for Disease Control (CDC) indicate that between 1995 and 1999, the United States lost over $150 billion per year in smoking-related economic losses (Center for Disease Control, 2002). In addition, smoking caused 264,087 male and 178,311 female deaths each year between 1995 and 1999 (Center for Disease Control, 2002).

While these figures alone do not necessarily prove deception or unethical behavior on the part of the tobacco industry, ensuing legal action has verified this industry's pattern of public negligence and deception. For example, the Final Proposed Finding of Facts in the United States

v. Philip Morris, et al. RICO racketeering trial characterizes the tobacco industry's behavior as "a massive 50-year scheme to defraud the public" (Kessler, 2004). The tobacco industry has engaged in systemic unethical business practices, in campaigns to mislead the public, and in outright deceit.

The public health tragedy perpetrated by the tobacco industry was one of a handful of events in the 20th Century that undermined trust in business practices for future generations (Wong, 2002). This industry's public image crashed and burned due to unethical business practices. Its actions defined gross negligence and greed-induced public endangerment. The tobacco industry used deception and unethical business and marketing practices to perpetrate a variety of activities that were illegal and endangered public health, including marketing to youth (Cummings, Morley, Horan, Steger, & Leavell, 2002), marketing "light" cigarettes as healthful (Pollay & Dewhirst, 2002), and controlling certain international tobacco markets through black-market trade (Lee, Gilmore, & Collin, 2004). These and other tobacco industry tactics epitomize the use of sustained and strategic deception for industry gain. In particular, these deceitful actions completely disregarded public health and well-being (Koop, 1998).

These deceitful actions were not only well-engineered and complete strategies meant to deceive the public. They were also formed in the context of a tension between socio-political factors (e.g. changing expectations of appropriate public business disclosure) and historical or financial factors. The tension between these two factors is a moral ridgepole that companies have tenuously walked over the last fifty years. As ethical business practice has evolved into Corporate Social Responsibility and further to Triple-Bottom Line reporting, financial and socio-political issues have continued to be at odds (Carroll & Buchholtz, 2006). Of interest in this

research is where tobacco companies and conglomerates cross the line into unethical activities. Deception, as a subset of unethical behavior, is at play in this instance.

Even more specifically, deception is a behavior enacted by individuals within an organization. Though these individuals may operate as small groups (as in the cases of departments, for instance), it is essential an individual choice to deceive. Or is it? Corporate deception has been shown in fact to be a result not only of "rouge employees," but of the whole atmosphere of a company, or in the case of tobacco, an industry. Although in theory this atmosphere could stem from evil lower level employees, this has not been shown to be the case. Higher-up individuals such as Kenneth Lay have in a best case scenario started the ball rolling towards deception in companies. In the worst case, they have not only illustrated how to deceive by example, but have also demanded the same behavior from lower-level employees. For the tobacco industry, the moment that leaders from all large cigarette producing U.S. companies met with their P.R. advisory firm to strategically wage war against public access to information, an infrastructure of deception was established that would wind its way through the cigarette-production industry as a whole, creating the atmosphere of deception that has persisted in some cases to this very day.

The Electronic Repository of Tobacco Industry Documents

During these years of high-profile fraud and negligence cases, and in tandem with a damning report on the health impacts of smoking released by the Surgeon General (Koop, 1998), the tobacco companies in the United States settled a groundbreaking lawsuit (McDaniel & Malone, 2005; Sloan, Mathews, & Trogdon, 2004). Part of their penalty was to make virtually all internal documents available for public scrutiny. Documents in file cabinets and desks at seven

major tobacco companies and tobacco trade organizations were made available to the public and are currently being stored in physical and electronic repositories.

The University of California at San Francisco alone has an electronic repository of over seven million documents that have been made available to the public via the internet (http://legacy.library.ucsf.edu/). These documents represent everyday tobacco industry activity, and include the most sensitive of materials: memos, confidential reports, and secret meeting notes. This dissertation project uses a stratified random sample of these documents which has become the Tobacco Documents Corpus (TDC). The TDC provides a representative view of deception in the tobacco industry without the need to examine all seven million documents. The documents in the TDC provide a linguistic perspective on an industry rolling towards ethical implosion and externally imposed restrictions.

Document samples from inside the tobacco industry provide the perfect test ground for studies of unethical business practices and deception. Only the Enron email corpus (Klimt & Yang, 2004) rivals the scope of the tobacco industry repositories. Alongside the access to Enron-internal emails in the Enron corpus, the access to tobacco industry documents offers an unprecedented view into the inner dialogue of an unethical industry. The tobacco industry documents present a detailed and mostly unexpurgated look at internal accounts of unethical and deceptive business practices (Malone & Balbach, 2000). There is a large and growing number of single-issue studies focusing on the tobacco industry's ethical failings and based on the publicly available tobacco documents (see http://www.library.ucsf.edu/tobacco/docsbiblio.html for a full list). Of these studies, however, there are few that attempt a broader assessment of tobacco

industry discourse. This dissertation intends to remedy this research omission by using corpus linguistic to test and evaluate linguistic indicators correlated with deception.

Deception in Business: The Changing Face of Business Ethics

Deception is a strange beast. From the point of view of veracity, humans could be seen to be constantly deceiving themselves and others. Memory is our only retrospective link with reality, but even it is constantly not veridical. People do not report consistent memories (Koriat, Goldsmith, & Pansky, 2000). Even eye-witnesses' accounts are notoriously open to change by suggestion (K. L. Chambers & Zaragoza, 2001; S. M. Chambers, 1995; Koriat et al., 2000; D. S. Lindsay, Allen, Chan, & Dahl, 2004; R. C. L. Lindsay, Martin, & Webber, 1994; Pryke, Lindsay, Dysart, & Dupuis, 2004; Steblay, Dysart, Fulero, & Lindsay, 2003).

Aside from the shortcomings of veracity in memory, some aspects of daily communication have codified untruths into the fabric of social parameters. Pragmatic and socially accepted exchanges between strangers can depart wildly from truth. In different contexts, the standard, "How are you today?" "I'm fine" can be an untruthful, but more socially acceptable answer. The malleability of language "veracity" to expectation and social acceptability could be distressing for someone who equates deception exclusively with the opposite of truth. Considering this, a more interesting and discernable approach to deception has to incorporate not only attention to truth and falsification, but also to planning and strategizing behind the scenes of a particular context of deception.

The stakes involved in the context of corporate deception make language and communications an important area of study. For instance, the tobacco industry's deception was uncovered most thoroughly through documents. Internal documents revealed that the tobacco

industry has used their communication with the public to knowingly and willingly undermined health officials' communication with the public (Kessler, 2004). In the process, for the benefit of their industry, they have completely disregarded public health. The tobacco industry's corporate fraud and continuing public deception are so horrific because the expense to society is not merely monetary. The tobacco industry has put the health and very lives of individuals at risk.

Research Goals

It is imperative, considering the pervasive quality of corporate deception, that we explore linguistic definitions of corporate deception. The time-bomb quality of the tobacco industry documents should serve as a keen reminder of what policy and attitude advancements can result from documents found even as a result of manual searching efforts. Can observed and reported indicators of deception be shown to correlate with real-world corporate deception? Can these indicators together (in a multivariate approach) be helpful in leading toward pockets of deceptive documents? This study should serve as a step towards validating known indicators of deception. At the same time, it must propose additional insights toward a more nuanced understanding of how deception works in corporate discourse.

This research has been a journey of validation. In order to explore the efficacy and use of certain indicators of deceptive corporate strategy, I first assess these indicators separately. Finding these inadequate, I use all six markers together in a multivariate approach. Using a discourse analytic framework to assess the multivariate results, it has become clear that deception in corporate language is engaging regardless of whether it is instantiated as a lie, a mistruth or even an omission. It seems to me that research looking at language and deception has actually been examining language and *lying*, which is only one type of deception, as will be

outlined in Chapter 2. Other types of deception can be just as insidious, but so far have gotten short shrift in research linking it with language. In the end, I propose this process of iterative corpus and discourse analytic tests as a method for getting as much information about language use or a language structure as possible. Though using one method or the other builds linguistic exploration, a best assessment of a theory, idea or indicator could productively incorporate both approaches for a more thorough account of language in all of its contexts.

Additionally, in attempting to address in some way the issues of corporate fraud in our day-to-day lives, this dissertation has as a goal a more refined understanding of language correlates with corporate fraud and deception. This research will validate and challenge links between language and deception in hopes of a fuller picture of deception in discourse. I hope that these observations could be used to help stem the tidal wash of corporate irresponsibility, fraud and deceit.

Plan of the Dissertation

This research investigates deception in a corporate context using methodological tools derived from corpus linguistics, discourse analysis and computational linguistics. In particular, documents are ranked for deceptive corporate strategy based on an "average deceptive document." There has been quite a bit of research on deception and language and deception. This study takes this language and deception research vein one step further into the background of deception. What does an in-house tobacco industry document supporting a deceptive publicly released document look like? Can principles of deception be applied to these internal documents to sort them by interest for corporate strategy to deceive? In order to answer those questions, I

first lay out the background of corpus linguistics, deception and language research and the tobacco industry.

In Chapter 1, I have presented the overall view of this study and provided an introduction to corpus linguistics. In addition, I introduced some basics about corporate deception and the tobacco industry, and introduced the tobacco industry documents and my dataset, the Tobacco Document Corpus.

In Chapter 2 (Review of Literature), I review literature in three main areas of interest: corpus linguistics, tobacco control and language studies of deception. First, I outline the work of corpus linguistics and sketch the impact and use of corpus linguistic techniques in this research. The second section of Chapter 2 gives a short history of the tobacco industry and makes a case for which areas of the TDC might be candidates for exemplary deceptive documents. In addition, two fundamental hypotheses to be tested using the six linguistic indicators of deception are proposed in this chapter. The third body of literature reviewed is language studies of deception and linguistic modes for identifying instances of deception. I define deception and provide a typology of it. Lastly, I link observations about deception in corporate settings to the set of six linguistic indicators of deception to be further explored and used in this research.

Chapter three (methods) details the computational assessments used in this study. Methods for measuring the instance of the proposed linguistic indicators of deception are reviewed. Also, the methods used for ranking the internal audience TDC documents for corporate strategy to deceive are outlined.

Chapter four reports on results of the two hypotheses and the algorithm used for ranking documents for deceptive corporate strategy. This chapter details the statistical results

corresponding with the formalized hypotheses. It also includes discourse analytic assessments of some ranked tobacco industry documents. These documents are those identified as most and least deceptive as well as a group of middle documents.

Chapter five discusses results of the research and the potential impact of results on the fields of corpus linguistics, discourse analysis and deception research. Chapter five also addresses gaps in the work, potential practical non-academic applications of this research, and further research possibilities in this vein.

CHAPTER 2: REVIEW OF LITERATURE

Corpus Linguistics

Corpus linguistic techniques have progressed rapidly over the past thirty years. Initial KWIC (key word in context) lexical studies are now accompanied by broad descriptions of "dimensional" differences between speech and writing (Biber, 1988) and genre-specific corpus studies of variation in niche discourses like direct-mail philanthropic letters (Connor & Upton, 2004). The field has seen productive syntheses of many traditional sub-fields of linguistics. For instance, discourse analysis has benefited from corpus analysis through a focus on keywords that mirrors the tradition of KWIC studies (Stubbs, 2001). Also, computational techniques used in corpus studies have been applied to traditionally forensic problems such as author identification (Stamatatos, Kokkinakis, & Fakotakis, 2000). These advances in the field of corpus linguistics have been used to inform and enrich other fields of study too, ranging from work in instruction (Partington, 1998; Sinclair, 2004) to expert witnessing (Cicres i Bosch, 2003).

As computers gave us access to previously intractable amounts of language data, linguistic researchers began creating large databases of language such as The Brown Corpus (Francis, 1964; Francis & Kučera, 1979, 1982) and the Lancaster-Oslo/Bergen Corpus, or the "LOB," Corpus (Johansson, Leech, & Goodluck, 1978). Initially, many corpora were used for dictionary creation (e.g. the Collins Birmingham University International Language Database project – COBUILD (Sinclair, 1987)), and also the British National Corpus (Aston & Burnard, 1998). Although dictionary creation and revision initially spurred much of the trend toward larger linguistic datasets, corpus linguistic methodologies quickly spread into other areas of linguistic research, and even beyond academic linguistics altogether.

At this point, corpus linguistic methodologies have been incorporated into every aspect of language research. In comprehensive books about the origins and evolving states of corpus linguistics, McEnery and Wilson (1996) and McEnery, Xiao and Tono (2006) devote entire sections to the use of corpora in language studies and discuss the use of corpora and the following areas of linguistic research: speech research, lexical studies, grammar, semantics, pragmatics, sociolinguistics, stylistics, language teaching, historical linguistics, dialectology, psycholinguistic, cultural studies, and social psychology.

Computerized corpora

Corpus linguistics is not a new approach to language research. Patterns in written language have been tracked and studied in religious texts for many centuries. Concordances based on the bible were available as early as 1769 (McEnery, Xiao & Tono, 2006). For the purposes of this study, however, corpus linguistics becomes valuable with the advent of computers and computational assessments of language. Corpus linguistics has two main areas that feed each other: theories pertaining to the creation of corpora and theories pertaining to the methodology used to assess corpora. This part of the chapter outlines the evolution of the creation of corpora and highlights significant corpora used in corpus studies. A subsequent section traces the development of methodologies and uses of corpora in linguistics and as applied to business language.

Corpora can differ widely in scope, and can focus on a range of linguistic constructions from the minutiae of lexical use up to broad generalizations about language use practices in general. A corpus can be very small and specified (e.g. Connor and Upton's [2003] corpus of

316 direct mail philanthropic letters) or large and multi-use, as in the case of the Brown Corpus (Francis, 1964; Francis & Kučera, 1979, 1982).

The earliest machine-readable corpus was started in1949 and only contained 10,000 sentences from St. Thomas Aquinas and related authors (Busa, 1992). The first machine-readable corpus created for linguistic research was the Brown University Standard Corpus of Present-Day English (the Brown Corpus – Francis & Kučera, 1964). This corpus represented a body of English text samples written in 1961 and originating from within the United States. Each of the 500 text samples was about 2,000 words long. A sister British English corpus, the LOB – Lancaster-Oslo/Bergen (Johansson et al., 1978) – was created about a decade later than the Brown Corpus, but replicated its parameters for date (1961), sampling techniques and length of text sample (2,000 words).

In the 1990s, both of these corpora were replicated for a diachronic addition to the datasets. The Freiberg corpora are more recent instantiations of the Brown and LOB corpora. These corpora were compiled of texts from 1991, exactly 30 years after the Brown and LOB. These corpora, the Freiberg-Brown (Frown) (Hundt, Sand & Skandera, 1999) and the Freiberg-LOB (FLOB) (Hundt, Sand & Simund, 1998) can be used to trace dialect changes between American English and British English over thirty years.

Specialized corpora

Specialized corpora, according to McEnery, Xiao and Tono (2006), are only really specialized "relative to a general corpus" (p. 60). The Tobacco Document Corpus would definitely be categorized as a specialized corpus because it does not represent general language, but rather is limited specifically to the tobacco industry. Other specialized corpora have been

industry-based, as well, though it should be noted that specialized corpora can also be used to represent sub-languages or other specific aspects of language (McEnery, Xiao & Tono, 2006).

The Guangzhou Petroleum English Corpus (Qi-bo, 1989), for example, could be seen as very similar to the TDC. With 411,612 words, it is in the same ballpark as the TDC. This specialized corpus contains 500 to 600 word samples from texts written in Petroleum English between 1975 and 1986. All in all, the corpus contains 411,612 words and was randomly sampled for this complete number. This corpus was developed with a training and vocabulary goal in mind (Qi-bo, 1989).

In theory, there is no limit to what subject matter or domain might be used to create specialized corpora. Many specialized corpora are corporate in nature – the Guangzhou Petroleum corpus, the Corpus of Professional Spoken American English and the JDEST (Jiaotong Daxue English of Science and Technology) (Yang, 1985). In fact, the Professional English Consortium (PERC) is in the process of creating a 100-million-word Corpus of Professional English, according to some sources (McEnery, et al., 2006). These types of corpora have been important in domain-specific research and genre research (McEnery, et al., 2006).

Developmental and Learner Corpora

The CHILDES (Child Language Data Exchange System) database (MacWhinney & Snow, 1990) is made up of child language acquisition data. More than 500 children are represented in this 20+million word corpus. In the 1980s this corpus was begun; it now contains data from many different languages. It has been used not only in first language acquisition studies, but also in research looking at impaired language, and cross-linguistic studies.

Another type of corpora that has been extremely important in linguistic research, and particularly second language acquisition research, has been learner corpora. Learner corpora, though documenting language acquisition is fundamentally different from developmental corpora, such as that of the CHILDES database (MacWhinney & Snow, 1990). Data from children as they learn their first language has been defined as differing from second language acquisition in such fundamental ways that these two types of corpora cannot be lumped together (McEnery et al., 2006).

Studies Using Corpus Linguistics

Corpus linguistic methods are used in many different sub-fields of linguistic research, and even in research outside of linguistics proper. On the level of minutiae, for example, Atkins and Levin (1995) used corpus data to assess whether "quiver" is actually a transitive or intransitive verb. Although it is listed in the Oxford dictionary as being transitive, their corpus approach demonstrated that it can actually be both (Atkins & Levin, 1995). On the broader level of whole-text stylistics, work on genre using corpora has demonstrated that contrary to popular belief, written and spoken language styles are not all that different (Biber, 1988). In fact, certain types of written language, such as narrative fiction, more closely resemble spoken genres than other types of written language (Biber, 1988).

These studies share a feature common to many corpus research studies. Corpus studies have repeatedly shown that what we think of as the canonical language usage paradigm may not be so uniform. Prior to corpus linguistic approaches, it was much more difficult to access the corpus data that have led us to these tradition- and assumption-breaking insights. At all levels of linguistic research, corpus methodologies provide access to linguistic data at a scale that has not

been accessed before. Results from corpus studies are unquestionably necessary for further scholarship in all areas of linguistics.

This section of the chapter will outline some of the major applications and methods of corpus linguistics in order to situate the current study in this growing field. Some major themes in corpus research to be reviewed here include: the idiom principle, ESL applications, discourse analysis applications, multidimensional factor analysis applications to corpus linguistics and corpus applications to business language research.

The Idiom Principle and Corpus Linguistics

For each idea or unit of information that can be conveyed, there are a number of ways to express that idea through language. For example, for a greeting, the basic idea and general meaning can be conveyed (in English) as "Hello," "hi," "how's it goin'," "what's up," "howdy," "hello there," etc. Though we see that language can be highly creative in expressing ideas, work in corpus linguistics shows that regularities in language on the level of word choice are largely habitual and patterned. Sinclair (1992) introduced the idiom principle as a way to explain the tension between creativity and habit in language's effective use.

ESL Applications of Corpus Linguistics

Productive work has been done using corpus linguistic methodologies to enrich English as a Second Language (ESL) scholarship. These studies have tended to elucidate language use thoroughly by only focusing on very specific usage or linguistic construction questions. For example, a typical ESL or ESP (English for Special Purposes) corpus study will pick an element of language use, such as "evaluative that," to study (e.g. We believe *that* this will be an

interesting thing to study) (Hyland & Tse, 2004). In the instance of evaluative that, Hyland and Tse found that evaluative that was a preferred method of attitudinal communication in non-native English speakers' abstracts. In this example, for instance, English language teachers and ESL teachers who have access to these results can focus teaching effort on the best and most appropriate way to introduce evaluative that since it is a preferred mode of communication for their students. This and many other detailed studies of specific language features have been bright spotlights of research. These studies have focused on narrow topics within linguistics including hedging (Hyland, 2000), reporting verbs (G. Thomas & Ye, 1991; S. Thomas & Hawes, 1995) or verbs with inanimate subjects (Master, 1991), just to name a few.

Hyland and Tse's, 2005 work focused on specific language features which could be easily integrated into teaching curriculums. Their commitment to direct application of corpus studies is a recurring theme of corpus linguistic work in general and is a particularly evident feature of ESL and ESP corpus studies. Additionally, an inspiration for the current research is Hyland and Tse's commitment to research making an impact on real-world issues.

Collocation in Corpus Studies

Collocation is the co-location of two words, or a series of words (Sinclair, 1991). How words co-locate helps define our comprehension of language and creates context for words that alone are merely referential. Collocation has been examined with corpora as a way to investigate the semantics of lexemes, and the language contexts of lexical items.

Some collocation corpus studies (Stubbs, 1996, 2001) have focused mainly on specific lexical terms to trace societal reactions and conceptualizations of certain events or ideas. Corpus lexeme studies trace certain words throughout a large corpus in order to extrapolate their

importance, use or social relevance. Michael Stubbs (1996, 2001) in particular argued that word choice patterns reflect personal and societal understandings and assumptions of the world at large. Stubbs has been interested in the mini-world created by a certain term or choice of terms. By looking at these mini-world lexical choices in aggregate using corpus linguistic methodologies, discourse analysts can systematically examine the interplay between society and language. In the introduction to Stubbs's 2001 book, Words and Phrases, he asserted, "our knowledge of a language is not only a knowledge of individual words, but of their predictable combinations, and of the cultural knowledge which these combinations often encapsulate."

Similarly, by reviewing the collocates of a series of key words in a corpus, another analysis reveals some of the fundamental stances of British culture (Stubbs, 1996). The technique used to locate these cultural key words focuses on specific lexical items and their co-occurrences. For example, examining collocates for key national-region words (BRITISH, IRISH, SCOTTISH, WELSH, ENGLISH) unearths cultural prejudices: BRITISH is found in close association with language, literature, heritage, national, and, traditional, while IRISH, for example, collocates with accent, folk, peasantry, and brogue. The difference in collocates, or words that occur frequently within a range of two or three words, reveals glaring inequalities between attitudes toward British versus Irish things. This study demonstrates the internal prejudices and associations of British culture through examinations of singular keywords. Though these prejudices and associations might have been revealed through discourse analysis studies of national-region discourse, the corpus aspect of this work adds to its validity and generalizability by demonstrating these lexical choices and associations as trends instead of simply single occurrences.

The success of corpus linguistic studies suggests that pushing the boundaries of corpus linguistic and computational methods into communication and interactional realms of discourse analysis is very possible. A recent study has used corpus data gathered from interviews to reveal striking information about families with members who have experienced schizophrenic episodes (Breitborde et al., 2005). In a setting where a family member has schizophrenia, types of talk about that family member (not directly from that family member) are associated with successful treatment outcomes. This study demonstrates that we are creating our world and our families'/ neighbors'/ co-workers'/ friends' worlds as we talk or communicate. On a philosophical note, there are researchers and pragmaticists that argue that language and behavior are the same thing essentially, and that people are all the time doing things with language (Austin, 1962; Habermas, 1985).

Corpus Multidimensional Factor Analysis Studies

In order to "do" things with language, communication must have a function or purpose. Corpus linguistic assessments of genre have demonstrated that communication function is a attribute of language, texts and documents that can be accessed through language patterns.

Focus on written academic texts (Hyland & Milton, 1997), English learners' classroom texts (Axelsson & Berglund, 1999), business language (Connor, Precht, & Upton, 2002) or historical English (project ARCHER – (Biber, Finegan, & Atkinson, 1994)) are some examples of exploration of genre from a corpus approach. Genre studies are prime examples of a thriving branch of linguistics that evolved to its current state only as a result of corpus methodological approaches. Works by Doug Biber, Dwight Atkinson and Ken Hyland have contributed greatly to an interest in corpus linguistics as a way to do genre analysis (see Biber, Conrad & Reppen,

1998 or Meyer, 2002 for a full review). Biber (1988) pioneered the use of multidimensional factor analysis in corpus linguistic research within the context of genre exploration. He used that statistical method to map linguistic features that differ between genres.

Much of Biber's early work mapped the use and frequency of a large number (70+) of linguistic features in texts. This feature-based investigation of language assumed "that strong co-occurrence patterns of linguistic features mark underlying functional dimensions" (Biber, 1988). Multidimensional factor analysis has been used in other fields to predict what parts of a data set work together or co-occur to create the effect under study. Multidimensional factor analysis helped Biber group his genres according to seven "dimensions" based on different linguistic features. In particular, Biber's work demonstrated a finding that ran counter to the accepted understandings of genre: there is no clear and absolute difference between spoken and written discourse (1988). Spoken and written language genres, in fact, seem to fall across a continuum based on functional similarities (Biber, 1988). For instance, fiction contains similar dimensional distributions to speech, while prepared speeches more closely resemble written texts.

Biber's (1988) use of multidimensional factor analysis in a linguistic setting was highly innovative and opened the field of corpus linguistics to new inquiries. The method of multidimensional factor analysis has now been employed in many diverse fields outside of linguistics. These fields include marketing (Verbeke & Bagozzi, 2002), public health (Seeman, Seeman, & Sayles, 1985) and even studies of attitudes towards death (Schiappa, Gregg, & Hewes, 2004)

Biber used multidimensional factor analysis to analyze his corpus data. Using multidimensional factor analysis makes it much more difficult to trace back how a specific

feature is placed in one group of features versus another. For example, if a feature falls in the middle between two opposing feature groups, how is that feature's association with a dimension determined? Is it grouped with feature 1 or feature 2 in a dimension? Using multidimensional factor analysis, this cannot always be determined, due to the complexity of the mathematical model used in factor analysis. By contrast, the mathematical model used in this study, the vector method model, has been pared down so that it is as straightforward and transparent as possible.

Also in contrast to Biber's (1988) work, this study uses a few select features instead of a large swath of them. This choice was precipitated by the fact that this study is not only assessing documents with the assistance of indicators, but it is also, and most importantly, assessing the indicators themselves. This focus on indicator validation demands a transparent mathematical model (which will be more fully explained in Chapter 3. The exploratory nature of Biber's initial work was largely a-theoretical and leaves further researchers without a null hypothesis baseline for future genre work. By contrast, the goal of this research is to test hypotheses about these indicators themselves in addition to testing new methods of document categorization and learning more about the documents.

Applications of Corpus Linguistics and Computational Methods to Business Language

Corpus linguistic studies have yielded valuable insights regarding language in the domain of commerce. For example, one application of the multidimensional factor analysis approach developed by Biber (1988) has been in the realm of the type of business language used in direct mail letters (Connor & Upton, 2004). This type of business language is particularly interesting because of its public audience. In this way, it resembles some of the tobacco industry's

documented deceptive discourse. The documents in this class are directed at public, or industry external, audiences.

Connor and Upton (2004) concluded that direct mail letters constitute a genre separate and distinct from previously investigated genres. Connor and Upton's work also solidified Biber's (1988) observations about language patterns interacting with functionality of text. Linguistic features co-occur and reflect the general goals or function of a genre or type of communication. Based on this observation, conversely, communication with similar goals could be expected to have similar co-occurrence of features.

Additional applications of corpus linguistics to investigations of business language follow a precedent within other communication research that stresses the importance of language and communication in business settings (see Cheney, Christensen, Zorn & Ganesh, 2004 for a review of the language and communication in business settings in the organizational communication tradition). Business language operates in the day-to-day interactions of a corporation: in the water-cooler chats, the emails, the memos, phone calls, reports, presentations, gossip, "organizational charts, paperwork, and records" (Eisenberg & Riley, 1988).

Some studies have applied computational methods derived from corpus linguistics to the genre of annual reports. Differences between annual reports in years when a company reports profits compared to years when a company loses money have been observed in quasi-corpus studies that use annual reports as data (Subramanian, Insley, & Blackwell, 1993). One observation from this approach was that annual reports of better-performing companies are more readable (based on computational measures of readability) than annual reports of companies whose performance was not as good as expected (Subramanian et al., 1993). This finding

suggests perhaps an eagerness for companies to reveal good news to their investors, and conversely to inhibit investor comprehension of bad news.

In another study, passive voice verb structures, typically associated with deflection of responsibility ("the passive voice distances the messenger from the message" – Thomas, 1997, p.3) doubled in those years when the company lost money, compared to years when the company reported profits. Thomas (1997) noted that "a close look at the language structures … offers a view of not only what the company want[s] its audience to know but also what the company may not [wish] to reveal" (p. 50). Furthermore, these language structures create "a strong but subtle impression of a factual situation (part of the 'real' world and not to be questioned) caused by circumstances … not attributable to any person or persons who might otherwise be thought responsible" (Thomas, 1997, p. 56).

Beyond simply observing differences in business language when company profits are up or down, some researchers have even applied the use of computational language approaches such as word and sentence counts and content analysis to predict the future performance of companies (Kohut & Segars, 1992). It is possible to assess company reputation (which is part of the accounting value of goodwill) relying in part on measurements of language structures and lexical choices in company documents (Argenti & Druckenmiller, 2004).

These studies demonstrate differences in language strategies associated with success and/or transparency between and within companies. They illustrate just the tip of the iceberg with respect to language variation in business settings. For instance, specialized lexicon is frequently used in different business settings. Not only do individuals have specific linguistic ways to communicate (sometimes called a "linguistic footprint" – Shuy, 1998), but groups who

have something in common also tend to have linguistic similarities. For instance, Malcolm Coulthard has used corpus methods to investigate register features for police and found that as a whole, police workers tend to use “then” more frequently and to place this temporal marker after the subject in a sentence (1999). Most other people, as judged by Coulthard’s large corpus investigation using COBUILD, place “then” at the very beginning of a sentence, before the subject. This set of differences characterizes the entire set of police workers.

Another diachronic corpus-oriented study of gender-marking in general society over thirty years (1961 to 1991) suggests that although the female gender continues to be marked linguistically in business, a positive trend has emerged in business language and discourse about business: occupations that had been exclusively male show an entry of women through their female-gender word use (Holmes & Sigley, 2002). Moving away from corpus studies, language variation has also been demonstrated on a micro level. Studies of differences in individuals’ language use, or idiolect, have shown that language features can be used to determine authorship (Coulthard, 2004). Fundamentally both qualitative and quantitative corpus language studies within the realm of business in specific (but also in language use in general) show that any difference (individual, in-company or cross-company) is reflected in language use.

The studies mentioned above illustrate some interesting conclusions about business language. First, business genres are distinguishable from other high-profile genres such as spoken language (Connor & Upton, 2004). Also, within the world of texts that originate from the business sector, differences can be detected not just between functionally different genres, but also within the same genre when there is a difference of intent. These differences in intent correspond to differences in non-linguistic factors such as success (Subramanian et al., 1993; J.

Thomas, 1997). Fundamentally both qualitative and quantitative language assessments within the realm of business in particular (but also in language use in general) show that any difference in language function or intent is reflected in language use.

Thus language co-occurrence patterns are associated with communicative function and intent in a variety of domains, including the world of commerce. Deception can be regarded as one such communicative function; however no prior research has applied practices of corpus linguistics to discerning the functions of misinforming and deceiving within a corpus of speeches, memos and research reports.

The Tobacco Industry – A Historical Perspective

In order to assess the efficacy of the linguistic indicators of deceptive corporate strategy within the tobacco industry documents, we have to know the history and actual trajectory of the tobacco industry. The next sections review the history of the tobacco industry, the master settlement agreement, the building of the Tobacco Documents Corpus and current work in tobacco control endeavors.

A History of Strategically Deceptive Communication: The Tobacco Industry

Since 1992 when a rogue paralegal leaked the first set of highly incriminating secret documents to tobacco control activists, the tobacco industry's internal strategies to increase addiction and decrease accountability have become, quite literally, an open book (Glantz, 1996). The availability of such a voluminous and comprehensive set of involuntarily disclosed texts gives researchers an unprecedented opportunity to investigate the textualization of corporate deception.

The textual history of the tobacco industry's public deception began in earnest in 1954 with the publishing of one document: A Frank Statement to Cigarette Smokers, or "the Frank Statement." Tobacco industry documents show that the Frank Statement was originally commissioned in 1954 as an industry-wide response to early scientific evidence of serious health risks posed by smoking (Forsberg, 2003). A 1953 literature review of 78 scientific papers on smoking and cancer written by a tobacco industry chemist concluded: "[S]tudies of clinical data tend to confirm the relationship between heavy and prolonged tobacco smoking and incidence of cancer of the lung" (Teague, 1953). Despite Teague's well-informed assertion, the Frank Statement, first published a year later, makes directly contrary claims that have been reiterated throughout the past fifty years: "There is no proof that cigarette smoking is one of the causes [of lung cancer] . . . We believe the products we make are not injurious to health" (Tobacco Industry Research Committee, 1953).

In December 1953, executives from all but one major American tobacco company met with representatives from the public relations firm Hill and Knowlton to address serious concerns about the link between public health issues and stock prices for the tobacco industry. What emerged out of that meeting of tobacco executives and public relations specialists was the Tobacco Industry Research Committee (TIRC), and with it, the Frank Statement. TIRC was billed as a research vehicle for the continuing public health concerns surrounding tobacco smoking. In reality, however, its sole purpose was to disseminate information and rhetoric intended to undermine the validity of links between smoking and increased instance of cancer. TIRC (and its later incarnation, the Center for Tobacco Research – CTR) served as lightning rods for criticism, intentionally drawing attention away from tobacco companies themselves. In the

end, it was determined that the material disseminated by TIRC and CTR was so deceptive and misleading that both associations were disbanded as part of the master settlement agreement (Cummings & Pollay, 2002).

With the creation of TIRC came the publication of the Frank Statement. It appeared in the New York Times and over 400 other newspapers on January 4, 1954. It was reprinted in various versions over a 40-year period. Although the statement was ostensibly an attempt at direct communication with cigarette smokers and the American public, archival evidence has revealed that its primary aim from inception was to deflect criticism and reshape public attitude towards the tobacco industry (Cummings & Pollay, 2002).

The principles articulated in the Frank Statement set the tobacco industry's discursive agenda for the next half century, apropos public health research, by questioning the validity of experimentation, by rejecting the meaningfulness of statistics, and by conveying an image of the tobacco industry as a rational party with an interest in minimizing risk to consumers. The tobacco industry continued to hold this discursive stance, and the position that smoking was not proven to be injurious to public health, until 1999 (Cummings & Pollay, 2002). Tobacco Control researchers further contend that effects of the Frank Statement's strong rhetoric are still felt to this day in continued public misinformation about the connection between smoking and health (Cummings & Pollay, 2002). The discourse strategies of the Frank Statement diffused into the wide-spread campaigns of disinformation and manipulation in the intervening years.

As one of the most widely distributed and deceptive tobacco industry documents, the Frank Statement took a star role in anti-tobacco industry trials of the late 1990s. Plaintiffs' and state lawyers reconstructed the Frank Statement as a banner of shirked duty and unfulfilled

responsibility. Previously unavailable documents from within the tobacco industry situate the Frank Statement in a tobacco industry already aware of the effects of cigarette smoking on health, and within an industry ready to disavow public health in favor of market share and corporate growth. It is a public document, but we would not have known about the circumstances and strategy associated with it without access to previously internal tobacco industry documents.

Texts can be regarded as arrayed along some continuum ranging from primarily communicative to primarily strategic (Habermas, 1984). Considering this spectrum, the Frank Statement is a paragon of a mainly strategic written text. In contrast to its fundamental strategic aim however, it is presented as if it were solely a communicative informational exchange. In fact, a great preponderance of tobacco industry communication with the public over the past 50 years has presented itself as communicative while being primarily strategic. Actions performed by the tobacco industry have been directly contrary to those promised. The deceptive and manipulative nature of research that was financially supported by the tobacco industry and their misinformation campaigns has been amply documented in court case after court case as well as in research literature (Barnoya & Glantz, 2002; Carter & Chapman, 2003; Cummings, 2003; Cummings, Morley, Horan et al., 2002; Cummings, Morley, & Hyland, 2002; Cummings & Pollay, 2002; Dearlove, Bialous, & Glantz, 2002; Eriksen, 2000; Glantz, 1996; Hirschhorn, 2004; Katz, 2005; Koop, 1998; Kozlowski & Edwards, 2005; Landman, 2000; McDaniel, Smith, & Malone, 2006; Novotny & Carlin, 2005; Pilkington & Gilmore, 2004; Pollay & Dewhirst, 2003; Tofler & Chapman, 2003). The gulf between words and actions reveals that many statements by the tobacco industry were strategically engineered; these tobacco industry public

communications were not intended to be referential exchanges (Searle, 1969) as normally constituted.

The Current State of the Tobacco Industry and Tobacco Control

Since the mid 1990s when lawsuits aimed at the tobacco industry began exposing tobacco's dirty underbelly in its own words, the tobacco industry has been on the defensive. Access to these documents has opened a way for tobacco control to pry back the doors around some of the tobacco industry's key defensive strongholds. Documents show that it was common knowledge with the tobacco industry that smoking cigarettes was addictive and that children were being targeted with cartoon-like campaigns such as that of Joe Camel (Cummings, Morley, Horan et al., 2002; Pollay, 2000). The documents that demonstrate these inconsistencies between tobacco industry assertions and their own research and knowledge have been accompanied by documents that also undermined some of the tobacco industry's core propositions.

Consequences of these tobacco-industry-document-based revelations included smoking bans in many places on a state, city and even community level. Between 1999 and 2004, smoking restrictions were strengthened on the state level for worksites, restaurants and bars across the U.S. (Center for Disease Control, 2005). And by 2006, 33 states and the District of Columbia had enacted smokefree ordinances (American Nonsmoker's Rights Foundation, 2006).

On the local front, progress toward better public health is being made as well. As of July 28, 2006, restrictions on smoking in shared spaces were in place for 474 municipalities (381 smokefree workplace ordinances, 305 smokefree restaurant ordinances and, 222 smokefree bars ordinances) (American Nonsmoker's Rights Foundation, 2006). These smoking restrictions were made possible by a platoon of tobacco control researchers and heavy local involvement.

These achievements are evidence that the tobacco industry is in retreat in this country. However, these public health measures have taken (or are continuing to take) more than 10 years to enact. Most of the research supporting public health initiatives with respect to tobacco control has been slow and laborious. Survey studies are time-consuming and costly. Medical research takes a long time as well. For instance, longitudinal studies of the impacts of behaviors or health risks can be decade-long endeavors. Even finding documents of the sort that have been used to inform public action and tobacco control initiatives can be difficult. A quicker and more accurate way to locate key documents would enable smaller teams to more effectively leverage their resources. A verifiable way to locate key documents would put the fire-power of damning documents into the hands of researchers and policy makers and could even speed along public health initiatives and results.

Initially, tobacco control researchers physically reviewed hard copy documents from boxes of subpoenaed material. But even more recently, the only methods for finding documents within the tobacco industry repositories are metadata searches and keyword searches, which are dependent on the accuracy of industry paralegals charged with providing such indexing information. Using metadata and keyword searches, researchers can find documents that are similar with respect to some elements: topic, author, date or other non-textual attributes are searchable. However, researchers have no way of finding documents that are associated with each other for reasons such as style or genre similarity. The research reported in this dissertation makes possible a different kind of document research and document triangulation. Instead of using topics to find similar documents, this project uses similarities in document lexical *and* syntactic information from the texts. Hopefully, the results will demonstrate that the function of

these documents can be related to this lexical and syntactic information as well. Specifically, the function of deceiving and misinforming is targeted. But how exactly are the principles of corpus linguistics and the observations about language attributes of deception leveraged in a tobacco industry backdrop to pinpoint documents of interest for deceptive content? The answer to that question is laid out in the next chapter as part of the technical and methodological section of this dissertation.

The Tobacco Documents Corpus: the Tobacco Industry at a Glance

The Tobacco Document Corpus (Kretzschmar, Darwin, Brown, Rubin, & Biber, 2004) is especially well suited for overview studies of tobacco discourse for a number of reasons. The TDC was rigorously designed to represent the entire body of tobacco documents. The computational processing and manual review time needed to assess the entirety of the seven million documents (3.5 million at the time the corpus was constructed) would have prohibited any but the most extensive research and the most extensive budget. Instead, the creation of the stratified and representative sample in the TDC opens an avenue for researcher to accurately assess the general nature of the tobacco industry documents without looking at each of the seven million documents.

Stringent statistical methods adapted from sociological research protocols were employed to create the TDC. The sampling methods used to make the TDC (outlined in Kretzschmar et al., 2004 and revisited in the next chapter) enable researchers to use observations from the TDC to make assertions about the entirety of tobacco industry documents.

In addition to allowing researchers to make generalizations about the whole body of tobacco industry documents, the TDC also uses an encoding format that is particularly well-

suited for detailed linguistic analysis. The general tobacco industry document files are available electronically thanks to scanning techniques that translate hard paper copies into electronic text via Optical Character Recognition (OCR). Unfortunately, due to quality and font differences in the hard copies, these documents often do not reproduce the text word-for-word. In some extreme cases, these OCR-ed documents cannot accurately reproduce the text of a document at all. Whereas the rest of tobacco industry documents are only available as picture files and OCR (Optical Character Recognition) text files, the TDC was keyed in as full-text. In addition, as part of this re-typing process, XML coding was used to represent potentially important divisions within a single text. The documents in the TDC contain fully represented text as well as various differentiated meta-text including headers, pretext, marginalia, main text, post text, lineouts, etc. This kind of textual detail has great advantages over imperfect OCR for any thorough linguistic analysis of the documents.

Features of the TDC: Some Document Class Differences

Document classes within the TDC were established at the beginning of its development to enable future assessments of language differences between different sections of the corpus. The initial research was interested in differences between documents addressed to internal audiences versus those addressed to external audiences. These two document classes are not classes that originate from the text, but instead are dictated by some external context. This external context can sometimes be provided in the form of meta-data (as in the case of the class source), but it can also be some other external criteria. Source, audience, decade and addressee are the classes of interest available as part of the TDC and used in this study. They are outlined in part below.

Document Class: Source

One document class difference of interest is source. Five major corporate defendants were party to the Master Settlement Agreement that compelled massive disclosure of previously secret documents. These manufacturers were American Tobacco, Brown and Williamson, Lorillard, Philip Morris, and R. J. Reynolds. In addition, the tobacco manufacturers created trade organizations to screen them from certain liabilities and to serve as their promotional organs. Material originating from the two tobacco industry umbrella organizations, the Tobacco Industry Research Committee (TIRC) renamed the Council for Tobacco Research (CTR) in 1964, was arguably more misinformative and deceitful than documents from the other tobacco industry company sources (Yach & Bettcher, 2000). These "front groups set up by industry to block public health policies" were actually disbanded as part of the Minnesota Settlement, due to the pervasive misinformation and deceptive content they consistently espoused (Yach & Bettcher, 2000, p.210).

In order to allow for within-industry comparison, the TDC has used extensible markup language (XML) coding and tagging to retain metadata linked to each document. Within this corpus, Source is a metadata field that indicates from which tobacco company or organization documents originate. As mentioned above, there are six divisions in the TDC for source: American Tobacco, Brown and Williamson, Lorillard, Philip Morris, R. J. Reynolds, Council for Tobacco Research (CTR), Tobacco Institute (TIRC). These different sources have very different backgrounds. While the tobacco companies in the U.S. were going about their daily business of tobacco production, after the creation of TIRC and CTR, the general tobacco companies were not discussing key issues about tobacco with the public. Instead, they were using TIRC and CTR as

smoke-screens to divert attention away from themselves. We can see this type of behavior from a linguistic standpoint by examining the relative distribution of some "hot" terms across sources using the TDC as a corpus. The following figure derived from TDC shows that *cancer, cancers, cancerous, carcinogen, carcinogens* are about 10 times more likely to be found in CTR documents than in any of the tobacco industry companies.

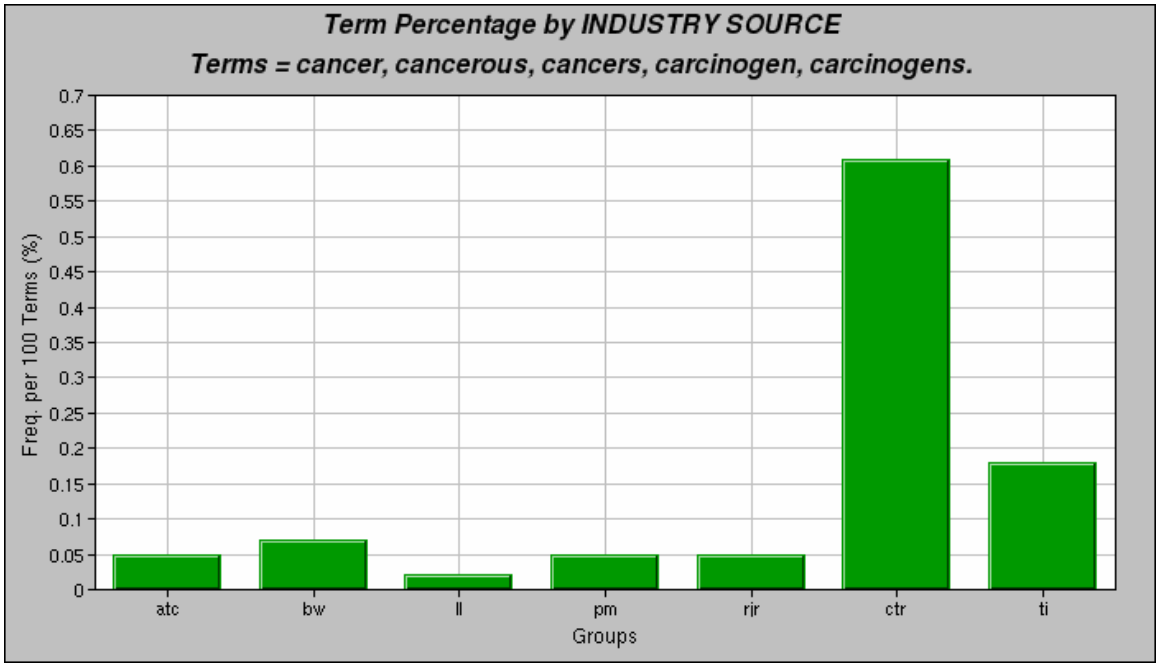


Figure 1. *Term Percentage by Industry Source for* cancer, cancerous, cancers, carcinogen, carcinogens *(UGA Tobacco Document Corpus and Toolkit, 2005)*

These data suggest that the research organizations within the tobacco industry (CTR and TIRC) were talking about sensitive topics like cancer, deflecting that topic from the ken of the tobacco manufacturing companies. Despite the fact that the tobacco manufacturers would have been most

affected by links between cancer and tobacco, the subject seems to have been so sensitive that it was relegated to entities that would be perceived as distinct from the tobacco companies.

Document Class: Audience

Also of interest in this project is a difference in documents that are addressed to industry internal audiences versus documents that are addressed to audiences outside of the tobacco industry. Tobacco control research affirms that tobacco industry documents addressed to external (outside) audiences use manipulative language strategies to deceive (Rubin, Hirschhorn, & Turner, 2002; Shuy, 2003a). These external audience documents can be seen as the more deceptive documents within the industry. Although we cannot truly link deception definitively with the use of one word (or one suite of words), again, the distribution of *cancer* and *carcinogen* across audience is telling. Using a Z-score measurement which represents the proportional frequency of the word based on overall frequency of a word in that language, we see that the instance of *cancer* or *carcinogen* terms is much more likely to occur in internal documents in the TDC (Rubin et al., 2002). One possible explanation for this difference is that the tobacco industry was strategically omitting discussions of non-complementary tobacco-related findings in their externally-facing documents. Both across audience and source, it seems to be the case that discussions of risky topics were either relegated to “safe-zones” like TIRC or CRT, or were not discussed at all.

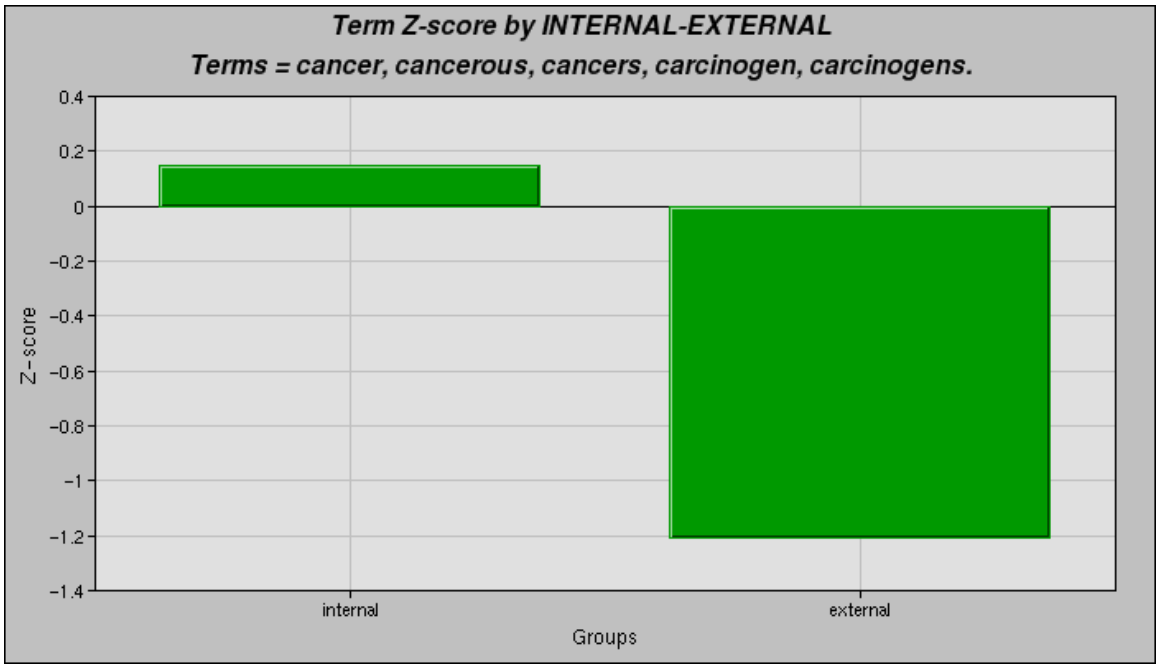


Figure 2. *Z-score by Audience for* cancer, cancerous, cancers, carcinogen, carcinogens *(UGA Tobacco Document Corpus and Toolkit, 2005)*

These corpus attributes of source and audience will be used to test hypotheses about audience- and source-centered deception. Hypotheses concerning source and audience and the distribution of the proposed indicators of deception are laid out in the following section of this chapter. These hypotheses are tested based on methods outlined in chapter 3, and their results are outlined in chapter 4.

H1: TDC Internal Versus External Audience Documents

Compared with documents addressed to audiences internal to tobacco companies, documents addressed to external audiences will exhibit higher relative frequencies of each linguistic indicator of deception. This hypothesis is based on a body of evidence showing that the industry engaged in certain practices deliberately designed to deceive the public, and that internal

information was on occasions at variance with information released to the public. These facts reflect a pattern of deception that can be presumed to pervade industry discourse.

H2: Tobacco Institute and Center for Tobacco Research (CTR) Documents

Compared with other tobacco company documents, Tobacco Institute and CTR documents will demonstrate a significantly higher relative frequency of the above linguistic indicators of deception. In 1954, the tobacco industry banded together to create the Tobacco Institute and CTR. These two entities were created to continue tobacco research *outside* of tobacco companies. The idea behind their creation was to provide a mouth piece for the discussion of the health effects of tobacco and smoking that would not focus on any of the tobacco companies. The Tobacco Institute and CTR strategically misled and deceived the public from 1954 until their court-ordered disbanding in 1998.

LANGUAGE-BASED CORRELATES OF DECEPTIVE COMMUNICATION

Assumption of Truth

Humans assume that in interpersonal communication, our communications are truthful and trustworthy, or at least not consciously intended to mislead. This assumption corresponds to conversational maxim of quality (Grice, 1989). This assumption of truth in language helps individuals defer the social, psychological and cognitive costs of assuming that no one can be trusted (B. M. DePaulo & Jordan, 1982). Because human perception generally operates to confirm expectations (Bruner, Postman, & Rodrigues, 1951), this predisposition to assume truthfulness can also discourage perception of cues that would distinguish a non-trustworthy source from a trustworthy source of information. Human assumption and expectation of

truthfulness is one of the root difficulties with detecting deception. Researchers have addressed this issue of assumption of truth not only from an experimental or observational perspective, but also from a theoretical and typological viewpoint.

In order to begin a study outlining an approach to linguistic correlates of deceptive communication, we have to define deception in the first place. Interpersonal deception involves messages or interactions that are knowingly enacted in order to give a false impression of some true state (Buller & Burgoon, 1996). In a presentation about automated linguistic correlates of deceptive communication, Burgoon (2005) listed the following range of deceitful and intentionally misinformative events as examples of deception: lies, fabrications, concealments, omissions, misdirection, bluffs, fakery, mimicry, tall tales, white lies, deflections, evasions, equivocation, exaggeration, camouflage, strategic ambiguity, hoaxes, charades, and imposters. Even outside of interpersonal aspects of deception, deliberate manipulation of impressions is a cornerstone of deception.

Galasiânski agreed that intentionality is one of three main aspects of deception along with manipulation of others, and imparting false information (2000). In addition, he defined deception first by attributes and then using a structured hierarchy to explicate the different types of deception. A deceptive interaction can be categorized according to Galasiânski's typological breakdown. Deception by evasion, deception by omission, deception by commission (explicit information) and deception by commission (implicit information) comprise Galasiânski's basic typology. While these categories cover most of the topology of deception, deception by obfuscation (e.g., excessive use of jargon or other complex language structures) is also of interest and will be added to the basic typology for this research.

These two language and deception researchers, Galasiânski and Burgoon, explain deception from very different angles despite the fact that their definitions of deception are almost exactly the same. Galasiânski explains deception in a typology based on data-driven observations, while Burgoon and her associates map the contours of deceptive encounters based on an interpersonal communication studies perspective. Even when approaching deception from such different viewpoints, their end results converge. With the basic attributes of deception (intentionality, manipulation and false information) in mind, Galasiânski breaks down the range of deception into distinct different types that correspond to different types of language in use. By contrast, Burgoon has developed her approach to deception within the theoretical frameworks of Interpersonal Deception Theory (Buller & Burgoon, 1996) and Interaction Adaptation Theory (Burgoon, Stern, & Dillman, 1995), focusing less on different types of language use and more on communication style changes within a single interaction. Both approaches have informed the field of study of deception and language. In particular, however, Burgoon has continued to explore her theories within the realm of linguistic correlates of deceptive communication and its relevance to national intelligence (counter-terrorism). The following sections outline the main characteristics and attributes of deception and further discuss Galasiânski and Burgoon's research agendas and results.

Deception Attributes and Typology

Deception by definition involves manipulation. Deliberate manipulation is the key factor in deception (Galasiânski, 2000). Manipulation is "an attempt to affect the target in such a way that her or his behavior/action is an instrument of attaining the goals of the manipulator, who acts

without using force but in such a way that the target does not know the goal of the manipulator's action" (p. 21).

Deception imparts false or misleading information. Deception is "a communicative act that is intended to induce in the addressee a particular belief, by manipulating the truth and falsity of information" (Galasiânski, 2000, p.20). Linguistically, deception is a type of manipulation that involves some form of non-truth, although not necessarily an untrue message. After all, it is conceivable that one could induce a listener to conclude a truthful belief by lying, or induce a false belief by telling the truth. Galasiânski makes this distinction to differentiate deception from persuasion, which he views as manipulation of value or cultural evaluation rather than truth and falsity. The distinction is important because the difference is subtle, particularly in colloquial use of the terms. For instance, convincing someone that it is a good practice to wear seat belts is persuasion because the proposition is true, but convincing the same person that smoking is not detrimental to one's health is deception because the proposition that smoking is not detrimental is patently false.

The speaker/writer's intentionality is another key element in defining deception (Galasiânski, 2000). Conveying false information out of ignorance is rarely considered to be deception; it is just being wrong. Of course, knowing a speaker's intent is never a sure thing, but we often know enough about the context of a statement to make an informed guess (Galasiânski, 2000, p. 51). This aspect of contextualization makes the tobacco industry documents a logical choice for a large-scale investigation of deception and language. Historical analyses—confirmed by court findings—make it clear that deceptive marketing and deceptive public affairs were a

staple of tobacco industry practice at least through 1998 (Cummings, Morley, & Hyland, 2002; Cummings & Pollay, 2002).

There are researchers who contend that all tobacco industry documents are deceptive and unethical because they all contribute to the machinations of the industry in one way or another (Hirschhorn, 2004). Although this position seems defensible on a macro or systems level, looking at one document at a time, it is obvious that there is variation in deceptiveness between tobacco industry documents. Determining and detecting the difference between day-to-day corporate events and strategic attempts to lie is the main goal of this dissertation research. There is a continuum of documents that participate more or less directly in tobacco industry corporate deception. This study hopes to enable a first assessment of which internal audience documents are candidate deceptive documents and not simply documents reflecting the day-to-day operations of the tobacco business and industry.

Typology of Deception

Deception by Evasion

Evasion is generally considered to be a form of deception (Ekman, 2001). Evasion is defined as using utterances "that are intended to be semantically irrelevant" (Galasiânski, 2000, p.59). These irrelevant utterances serve as misdirection or diversions (Burgoon & Buller, 1994). Enacting deception by evasion, a deceiver may ignore a question rather than give a false answer, or may refer a listener to previous statements.

Deception by Omission

Deception by omission is characterized by withholding pertinent information from the target (Galasiânski, 2000) or concealing relevant information (Burgoon & Buller, 1994). Omission is the essence of manipulation: the intent of deception by omission is information control. Manipulation in such situations occurs passively by leading the target to make poor judgments based on inadequate data. This practice is parasitic on the Gricean Maxim of Quantity (Grice, 1989). The maxim of quantity as a pragmatic principle dictates that successful communication will be no more and no less informative than necessary. For example, a professor who writes a letter of recommendation that only says, "Julie is a decent student," would be flaunting the maxim of quantity in order to show displeasure with Julie's performance without directly announcing that displeasure.

Deception by Commission: Explicit Information

Deception by commission represents active participation by the manipulator, who "contributes to the target's acquiring or continuing a belief that suits the purpose of the deceiver" (Galasiânski, 2000, p.22). A deceiver can achieve this end by explicitly or implicitly conveying false information. The prototypical example of misleading by giving explicit information is lying, but Galasiânski also includes any other speech act that violates Grice's other maxims. Buller and Burgoon (1994) define lying and falsification, as they refer to this type of deception, as excluding half-truths (which they designate as a different type of deception). However, within the purview of corporate deception, including half-truths, "white lies," exaggerations, minimizations, and equivocations in this category may be most beneficial. This of course excludes discourse practices that may appear structurally similar, but are not intended to be taken

literally—e.g. jokes, irony and teasing. Many tobacco document research articles highlight the explicitly commissive aspect of the tobacco industry and define tobacco communication as outright lies and denial. For example, the following research article titles all refer to deception by commission perpetrated by the tobacco industry (italics mine): "GASP: picking off the *pack of lies*" (Farren, 2004); "The low tar *lie*" (Leavell, 1999); "Smoking, disease, and *obdurate denial*: the Australian tobacco industry in the 1980s" (Carter & Chapman, 2003); "First, *tell the truth*: a dialogue on human rights, deception, and the use of smokeless tobacco as a substitute for cigarettes" (Kozlowski, 2003). In the context of the tobacco industry, documentary evidence has shown than many propositions denied by the tobacco industry (e.g. the addictive impact of nicotine, the link between smoking and cancer) were actually true. Thus, frequently in the context of the tobacco industry, obdurate denial is an example of explicit deception by commission.

Deception by Commission: Implied Information.

In contrast to deception by commission via explicit information (i.e. lying), deception by commission using implied information is much less direct. For Buller and Burgoon (1994), deception by commission using implied information most closely corresponds with their category of deception by exaggerations. For example, within the tobacco industry, documents intended for public consumption consistently create clear distinctions between "scientists" versus "eminent scientists." A corpus linguistic study showed that the connotations associated with "scientist" were negative or neutral and referred only to non-tobacco-funded scientists, while "eminent … scientists" was associated with both the tobacco industry and excellent scientific work (Brown & Rubin, 2005). This subtle language choice highlights a point: The tobacco

industry has consistently used authoritative qualifiers like “eminent” to legitimize their scientists and their side of any discussion. Other scientists, perhaps scientists that disagree with the tobacco industry’s position on smoking and health, are not “eminent.” Selectively crediting sources in this way is an example of deception by commission via implied information.

Deception by Obfuscation

Deception by obfuscation is the last kind of deception in this typology and includes a mixture of omission and commission. Obfuscation is a regular occurrence in business and law settings that does not fit well into the typology described so far. Deception by obfuscation includes making information incomprehensible by obscuring the meaning through jargon, complex syntax or difficult lexicon. Obfuscation is a particularly linguistic manifestation. It is not as passive as simple deception by omission. Accurate information may be presented, but it is intentionally made difficult or impossible to decipher.

For an individual uninitiated into a community of practice that uses such obfuscating language, this type of deception is evasive and functions in much the same way as omission, by violating one of Grice’s (1989) maxims: the maxim of quality. For instance, high use of nominalized words often correspond with in-group technical jargon and other specific usages of language that hide agency in favor of referring to a process as a static noun (Fairclough, 1989). Jargon and specific usages of language are not inherently deceptive. Indeed, they are often used as highly efficient signifiers within a group of specialists. However, when the audience of a document is no longer within the technical in-group of a specialty, retaining the in-group jargon interferes with clarity. This confusion can be used to deceive by obfuscation. It distracts the hearer/recipient from the actual message and masks the message with technicalities.

Another instance of deception by obfuscation involves the tobacco industry using "smokescreens" or "knowledge shields." Using a "smokescreen" or "knowledge shield" (Messick, 1998) is a way to obfuscate by distributing information that will create doubt or pretending to have amnesia about a certain subject (e.g. the ubiquitous "I don't recall"). Throughout their battle with states and municipalities over banning smoking in public places and businesses, the tobacco industry refused to refer to second-hand smoke by that name, in part to create doubt and shield any discussion of the subject from negative connotations. Instead, they employed an acronym that was not emotionally charged – "ETS" – Environmental Tobacco Smoke. A tobacco industry internal document (Young & Rubicam, 1993) specifically directed tobacco employees to use "ETS" over "second-hand smoke" in order to downplay the volatility of the subject and its impact on public health. Using this acronym could be seen to distract the hearer/recipient with language that is more technical and could even exclude them from the discussion altogether.

To review, deception is manipulative (trying to direct the way someone else is thinking or understanding), intentional (purposefully doing so), and imparts false information. Types of deception include deception by evasion; deception by omission; deception by commission: explicit information; deception by commission: implicit information; and deception by obfuscation. To illustrate the difference between each of these, consider the following scenario: you are Carol's best friend. You know that Carol and Carl are in a supposedly monogamous relationship. You also know that Carol has just left a coffee date with you to go to the grocery store to pick up food for her dinner with Joe, her other man. You run into Carl on the street and he asks: Where did Carol go? You deceive by evasion: *I have to run*, you reply. Or maybe you

deceive by omission: *Oh, she went to the store*. You deceive by commission explicitly and tell a total lie: *I haven't seen Carol today*. Or you imply that she is taking care of her sick friend and deceive by commission implicitly: *Oh, didn't Carol tell you about her sick friend?* Lastly, you could deceive by obfuscation: *Whereas we find Carol not here, we might send out an APB so as to discover her*. Or: *Her present whereabouts is not immediately evident to me* (Covington, p.c., 2006).

This typology helps outline the different possible linguistic manifestations of deception. Of course no typology by itself can predict and explicate which type of deception might be used in any interaction. Research on deception in the context of communication, on the other hand, has worked to outline and explain why certain types of deception are used, in what circumstances they are employed, and what linguistic features they employ.

Identifying Deception in Language

Two major strands of deception studies involve linguistics: discourse analytic case studies and corpus and computational studies. Discourse analytic studies have observed a rich, but intrinsically diverse, set of potential linguistic indicators of deception (Burgoon, Buller, Floyd, & Grandpre, 1996; Burgoon, Buller, Guerrero, Afifi, & Feldman, 1996; Burgoon & Qin, 2006; Dyas, 2003; Galasiânski, 2000; Vrij & Mann, 2001; Lina Zhou, Burgoon, Zhang, & Nunamaker, 2004). At the same time, corpus and computational attempts to characterize and categorize deceptive language using standardized sets of features have been increasingly successful (Fuller et al., 2006; Newman et al., 2003; L. Zhou, Burgoon, Nunamaker, & Twitchell, 2004; Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004). However, these corpus and computational attempts share with case studies the problem of replication: case studies are

inherently impossible to replicate, and the results from corpus and computational studies examining deceptive language have not been replicated with much success.

Approaches to Researching Deception and Language

Vrij and Mann (1998, 2001)

One paradigmatic study of discourse analytic linguistic correlates of deceptive communication was a case study of a high-stakes murder case (Vrij & Mann, 2001). This study analyzed the differences between six fragments of speech from interviews with the yet-to-be-convicted murderer (three provable lies and three provable truths). In this within-subject framework, the following mainly prosodic features indicated deception: the number of pauses, length of pauses, speech rate, ah disturbances, non-ah disturbances, and sentence change.

Number of pauses refers to frequency of noticeable pauses in speech. There were more pauses for the deceptive individual. Speech rate was faster. There were more ah and non-ah disturbances. Sentence change, which refers to sentence incompletion and slips of the tongue, was much higher for the deceptive individual.

Unfortunately, this study only deals with the patterns of one individual. A further study in this vein would entail an application of these elements to a wide range of high- and low-stakes situations, as well as a wider range of individuals (not just murderers).

Shuy

Shuy's 1998 book, *The Language of Confession, Interrogation and Deception* has a clear commitment to justice that he strongly supports. His examples and explanations serve to bolster his position that there are no hard and fast parameters that can be used to linguistically detect

deception. He does, however, assert that "liars are not good at prevarication, especially during complex and pressure-packed interrogation by law enforcement officers. They tend to slip up some-where and become inconsistent. When they do, they can get caught in their inconsistent language (78)." He stresses examining patterns of consistency or inconsistency in language use as the only semi-sure way to determine if an individual is lying.

Shuy gives a pessimistic account of linguistic analysis being used by law enforcement officers. Due to factors of individual difference, it is extraordinarily hard to pinpoint any specific signals that indicate truth telling. I think Shuy's main point is that keeping language contextualized is highly beneficial. As far as the tobacco documents go, this contextualization will make all the difference. It is within the context of public health and public policy that the tobacco industry fails to be truthful. Within the decontextualized market economy, the tobacco industries are only doing what is necessary to grow a business.

Shuy asserts that no studies have shown that any rubric for linguistic measurement of deception works. Despite Shuy's assertion, Pennebaker et al. have produced a lexical-based computational tool that correctly assesses truth in a phrase at about 70% correctness.

Although language reflects cognitive states generally (I.Q.), the time lapse between what we think and what we say (or even more distant, what we write) seems to allow us to be very creative in our deceptive responses.

Shuy lists potential verbal cues for deception:

1. Providing overly detailed statements (in order to measure this, we need a base-line of detail for a truthful speaker)
2. Repeating oneself spontaneously

3. Complicating the story unexpectedly
4. Giving unusual details
5. Providing marginally relevant details
6. Giving related external associations
7. Displaying subjectivity
8. Correcting spontaneously
9. Admitting memory loss
10. Hedging
11. Self-referencing excessively
12. Manifesting verbosity
13. Pausing excessively
14. Using unnecessary connectors
15. Using pronoun deviations such as you for I
16. Producing disproportionate amounts of language in the prologue, central action, or epilogue proportions of the narrative
17. Producing low lexical diversity by means of type-token ratio.

Porter and Yuille's review

Porter and Yuille (1996) did a thorough review of the literature and assert that none of the following verbal indicators from the following programs work across the board: Statement validity analysis (Undeutsch, 1982), reliability monitoring (Leippe & Manion, 1992), Sapir's SCAN training program, lexical diversity. However, they did find that truthful subjects

consistently produced more details, were twice as coherent, and admitted memory loss more often.

IDA (investigative discourse analysis): Rabon (1994) and Sapir (1987)

According to IDA, liars refer to past events not in past tense. They relate events vaguely, avoid self-referencing, and produce a quantitative imbalance between prologue, main event, and epilogue. IDA's four parameters suggest that liars tend to lean more heavily on general schemas for their constructed events in order to compensate for their lack of actual memory for such events.

Lying words: Newman (2003)

Another recent study related to quantitative assessments of the language differences between truth and deception in a large sample group is reported by Newman, Pennebaker, Berry, & Richards (2003). A mixture of written and spoken samples of true versus deliberately false stories made up the data set. After analyzing the text samples using Linguistic Inquiry Word Count (Pennebaker, Francis, & Booth, 2001), the research team used these analyses to derive profiles of deceptive and truthful communications. These profiles were then tested on independent samples and against predictions made by human judges. The profiles accurately predicted the truth of written texts about 67% of the time (Newman et al., 2003).

LIWC is a word-based text analysis program. Each word is compared to a 2,000 word lexicon which is divided into 72 linguistic dimensions. The output is represented by the total number of words in a category per total words in the text sample. LIWC uses its word categories to derive psychometric attributes of a text such as *affective*, *cognitive*, *process*, etc. LIWC links

physical and mental health with the language use of individuals. It was initially created to characterize the psychological state behind a certain language sample. The predictive use of this program necessitated that LIWC be verified and tested for reliability. Accordingly, the groups of words for each psychometric category have been tested for similarity and reliability by experts.

Using LWIC, Newman et al. (2003) concluded the following: liars tend to use more negative emotion words, fewer sensation words, fewer exclusive words (but, except, without), and more action verbs that do truth tellers. These usage tendencies, along with fewer first-person singular pronouns and fewer third person pronouns, predicted deception cross-modally (written and spoken) in diverse contexts at an overall accuracy rate of 61%. Although 61% accuracy is not very statistically impressive, it is strikingly better than the human judges in this study who predicted deception with an accuracy rate of just chance (50%).

Although the Newman et al. (2003) mention two limitations of their study--English-language specificity and inability to control the intensity of speakers' motivation to lie--they fail to mention two other serious issues. First, all text samples were gathered from undergraduate college students. This obvious choice unfortunately does limit the study's applicability to a broader range of individuals. Also, factors of sociolinguistic variation between dialects and jargons again restrict the generalizability of Newman et al.'s (2003) profiles. Research designs building on this research should utilize language samples composed of a diverse group of documents and discourses.

CBSA

CBSA (Criteria-based Statement Analysis) attempts to systematically assess the truthful or deceptive content of statements. Steller and Kohnken (1989) compiled a list of 19 criteria

based on the Undeutsch (1989) hypothesis that a statement originating from meaning in memory will differ in content and quality from a statement derived from invention. This is known as the Undeutsch Hypothesis. Although not widely used in the U.S., CBCA is used in Germany to test statements from potentially sexually abused children. Vrij's 2001 review reveals that that these 11 criteria (of the original 19) are more often present in truthful statements than in deceptive accounts:

1. Logical structure
2. Unstructured production
3. Quantity of details
4. Contextual embedding
5. Interactions
6. Reproduction of speech
7. Unusual details
8. Perpetrator's mental state
9. Spontaneous corrections
10. Admitting lack of memory
11. Raising doubts

Granhag and Stromwall (2002)

In complete contrast to any study that has found a detectable difference between liars and truth-tellers, Granhag and Stromwall's 2002 study found no linguistic difference at all. Richness of detail was not upheld (as opposed to other studies) as an indicative marker of verbal truth. Surprisingly, even though three interviews were conducted over a period of 11 days for each

participant, there was no difference in consistency of truthful and deceptive statements. They propose that the rehearsed nature of deceptive statements counterweights a general tendency of memory erosion in the truthful statements.

Knapp, Hart & Dennis (1974)

Larger studies have observed some of the same types of attributes. The first large sample study listing linguistic indicators of deception incorporated some linguistic indicators akin to those described in case studies (Knapp, Hart, & Dennis, 1974). Knapp, et al. elicited deceptive speeches from male students and veterans by having them role-play a non-favored position with respect to veterans' benefits. All of the participants were in favor of better benefits but had to argue in one instance for better benefits and in one instance against them. Deceptive statements from this study showed more (1) uncertainty, (2) vagueness, (3) nervousness, and (4) unpleasantness than truthful statements (Knapp et al., 1974). Uncertainty included the following elements… Vagueness coincides with a strategic ambiguity or an intentional ambiguity or an ambiguity of intent. Nervousness… unpleasantness… links with adversarial language. Although tangentially related, Vrij and Mann's (2001) deception indicators (pauses, etc.) could be construed as parts of vagueness or nervousness, or at least be understood as being in the same hesitation family as these attributes.

As a follow-up to the Knapp, et al. study, a study using female subjects (and a discussion topic of abortion rights) was designed with identical methodology. This study (Todd, 1977) only confirmed the results of the original (Knapp et al., 1974) with respect to the variable of disparaging statements which was part of the group of features used to determine unpleasantness. Although Todd (1977) suggested that this result may represent a difference in linguistic strategy

between men and women for telling lies, she also notes that a better study would need to incorporate both genders, and divulge itself of the role-playing methodology. A subsequent study did just that (incorporated both genders), but only reproduced three out of fourteen statistically significant results observed in Knapp, Hart & Dennis's work (1974): total words, speech duration and number of clarifications (Todd-Mancillas & Kibler, 1979).

Interpersonal Deception Theory and Burgoon's research

Communication studies have approached deception from a specific angle that focuses on interpersonal communication. Interaction Adaptation Theory (Burgoon, Stern et al., 1995) and Interpersonal Deception Theory (Buller & Burgoon, 1996) kicked off a series of studies that have investigated communication and deception. In the initial study, Interaction Adaptation Theory was proposed to highlight the adaptiveness of communicative interactions. Interaction adaptation theory employs forms of adaptation (approach, avoidance, reciprocity and compensation) to explain communicative interactions and deceptive interpersonal communications. One of the basic tenets of the initial application of Interaction Adaptation Theory is that adapting to another communicator is the most prevalent way to communicate: "synchrony, matching, and reciprocity are the default condition in human interaction" (Burgoon, Stern et al., 1995). The follow-up study (Buller & Burgoon, 1996) further developed Interaction Adaptation Theory into a more specifically deception-relevant Interpersonal Deception Theory (IDT). IDT borrows from IAT in that it highlights the adaptiveness of individuals to each other in any communication. Buller and Burgoon found that not only the senders, but also the receivers are active participants and are both engaged in encoding and decoding deceptive messages

(1996). Both senders and receivers oriented to each other to achieve communication (deceptive or not) with as little friction and abnormality as possible.

Burgoon and her colleagues have published research on detecting deception in experimental settings (Burgoon, Blair, Qin, & Nunamaker, 2003) as well as in computer-mediated settings (Twitchell, Forsgren, Wiers, Burgoon, & Nunamaker, 2005; Twitchell, Nunamaker, & Burgoon, 2004) and high-stakes, real-world settings of security police investigations (Fuller et al., 2006).

Initially, Burgoon et al. (2003) proposed sixteen "linguistic cues" as inputs to detecting deception in an experimental setting. These linguistic cues were number of syllables, number of words, number of sentences, number of short sentences, number of long sentences, number of simple sentences, number of big words, average syllables per word, average words per sentence, Flesh-Kincaid grade level, number of conjunctions, rate of adjectives and adverbs, emotiveness and affect (Burgoon et al., 2003). Although none of these indicators was effective in determining deception by itself, a cluster analysis showed that in aggregate, these cues predicted deception in 60.72% of cases. Sentence level complexity and vocabulary or affect acted as good classifiers, while other cues (emotiveness) were not helpful in classifying deception at all. However, this is not particularly surprising, considering that many of these indicators are possibly overlapping (e.g. number of syllables and number of big words). Despite this confounding of indicators, the study in question does demonstrate that using cues in aggregate (as opposed to using them separately) helps deception detection. The effectiveness of this aggregate approach despite single classifiers not assisting deception classification at all is the reason behind the use of the vector model method in this study. The vector model method (outlined in Chapter 3) uses all six

linguistic indicators of deceptive corporate strategy together to rank TDC documents for deceptive corporate strategy.

Despite promising results, Burgoon et al.'s, 2003 study was limited significantly by the nature and scope of their research. Data were gathered in experimental instead of naturally occurring settings, and even those data were limited (there were 41 participants). However, Burgoon's group followed this research with a suite of text-based automated linguistic correlates of deceptive communication research publications that remedied some of these shortcomings. Using task-based email communications, this group found that language diversity of deceivers was significantly lower than that of truth-tellers (Zhou, Burgoon, Nunamaker, et al., 2004; Zhou, Burgoon, & Twitchell, 2003; Zhou, Burgoon, Twitchell, et al., 2004). However, in a following study examining naturalistically sampled data from chat rooms and instant messaging, very short turns in their streams of interaction showed that the observation about language diversity of deceivers was entirely dependant on communication medium. Data from chat rooms and instant messaging used in that follow-up study (Twitchell et al., 2005) did not pass 20 words, while data from the previous email communication studies had an average length of 133 words. This difference underscores a continuing issue in linguistic correlates of deceptive communication studies, that linguistic correlates of deceptive communication studies "seem to be very sensitive to medium, context, culture and individual differences" (Twitchell, 2005, p. 471).

In aggregate, studies from Burgoon's group have uncovered insights into deceptive communication. Most importantly, perhaps, these studies have forced a conception of deception as a complex phenomenon best assessed with a simultaneous barrage of tests, cues and theoretical constructs including the cues mentioned above

Among the problems Burgoon's research group has encountered, differences in medium and mode of communication, as well as limited data set issues have been largely unresolved. The largest data set of the above studies consists of 13,243 words (Twitchell et al., 2005), which is a relatively large collection of spontaneous text in a research setting, but still very small in comparison to corpus linguistic studies. Examining deception in the TDC in this study will be one of the first times that a corpus of standard size has been used for research in the area of deception and language. Also, the focal measures for deception employed by this group are very well suited to their data, which is mostly spoken or real-time text and computer media.

Listener/reader perceptions of deception: Dyas (2003)

Most of the deception studies highlighted to this point have addressed aspects of observed deception production. Markers of deceptive language identified in these studies have ranged from low frequency of exclusive words (e.g. "but," "except," "without") (Newman et al., 2003) to high frequency of unfilled pauses (B. M. DePaulo, Lindsay et al., 2003; Ekman, 1992; O'Sullivan, Ekman, Friesen, & Scherer, 1985). A very different approach to studying the language of deception utilizes listener/readers' perceptions of deception as a criterion variable. One deception perception study, for example, focused on perjured courtroom testimony (Dyas, 2003). Dyas (2003) claimed that her study was the single study of deception and language, at the time of publication, to use a scientific approach to subject selection. She sampled discourse by tracking down Texas perjury cases where the jury had found the defendants guilty. That is, those jurors perceived the defendants to be lying on the witness stand. Dyas (2003) used transcripts from a sample of these cases to create a small corpus of non-solicited deceptive language. Apart from her sample selection, Dyas' (2003) research also represents an innovative step in studies

integrating corpus studies and perceptions of deception (in this case, jurors' perceptions). In a first part of her study, Dyas created a small corpus of eleven perjured testimony excerpts and analyzed them for deception indicators. She then manipulated key linguistic features within the perjured testimony excerpts in order to produce texts with high, mid and low levels of the following features of interest: filled pauses, imperfective aspect, and first-person deixis in possessive adjectives. Perceptually, the testimony with the highest number of filled pauses (e.g. *um, eh, wellllll*) was more often seen as deceptive. Dyas (2003) demonstrated a definite link between language production and perception of deceit by showing that changes in linguistic features of know perjured statements also changes participants' perceptions of falsehood and truth.

Additional studies have proposed a vast number of supportable and theoretical linguistic indicators of deceptive communication in comparison with truthful communication. A review of deception clues include the following traits of deception compared to non-deceptive communication among others: fewer words, fewer unique words, changes in Type-Token Ratio, fewer past tense verbs, greater indicative mood sentences, fewer phrases with subjunctive mood, shorter duration of speech, more frequent reference to others, fewer references to a group, more disparaging remarks, more hesitation, more repetition, more unfilled pauses, more filled pauses, more negative words, more passive verbs, more future verbs, more intransitive verbs, more lexical terms of non-existence, more conditional verbs, more gaps before responding, and more lexical diversity (B. M. DePaulo, Lindsay et al., 2003). Although this laundry list has provided (and continues to provide) a good jumping-off point for further studies, in many cases these potential features of deception have been ill-defined, poorly explained or only tangentially linked

with any theory or explanation of deception. In fact, in general, there is an underlying split between features of deception in language that remains largely unaddressed in the deceptive communication research literature. Features present in deceptive language by virtue of the cognitive constraints of deception (e.g. filled and unfilled pauses) have been lumped together with features that are present because they are discursive strategies people use to deceive. For instance, we might see that in the instance of written deception, features resulting from cognitive constraints would not be as productive as features linked with known discursive strategies, due to the highly planned and non-immediate aspects of written language. This study addresses theoretical gaps in deception studies present in the corpus linguistic literature on this subject.

Are Features of Deceptive Language Results of Strategic Communication or Cognitive Constraints?

Some deception researchers have already parsed out deceptive cues into productive categories. Ekman (1985, 1992) categorized cues into thinking and feeling cues. Thinking cues include the results of what happens with liars over-think their lies. The linguistic results of over-thinking are Ekman's "thinking" cues: over-prepared renditions of the truth that seem too rehearsed, slower speech as a result of thinking harder about what to say, betraying deceit with factual inconsistencies. Ekman's "feeling" cues focus on emotional impacts of deception and are demonstrated by fear indicators such as higher pitch, faster speech, pauses and speech errors. This distinction between thinking and feeling cues is productive within the realm of analysis of spoken deception because it postulates that deception occurs on a continuum of detection apprehension. This focus on detection apprehension allows for predictions that higher-stakes situations will produce greater feeling cues in a liar.

Both of these types of cues result from cognitive factors. Slower speech (a thinking cue) is a result of a taxed cognitive system, along with speech errors (a feeling cue) and pauses (also a feeling cue). Perhaps, outside of the realm of spontaneously spoken lies, a different split of features of deception can be explored. Cognitive processing constraints may be less at play during conscious deception. For example, Ekman's concept of "leakage," that indications of deception are unwittingly revealed in deceptive communication, coincides more closely with cognitive constraints of telling lies. By contrast, some deceptive messages are not unconsciously "leaked," but instead are purposefully used to deceive other people. This distinction could be key to a thorough account of deception in business language and corporate fraud.

Cognitive stresses revealed as a consequence of the time and interpersonal expectation constraints imposed by spoken language become less marked in corporate discourse—which is most often composed over time, with oversight from a variety of individuals and groups, and in writing. In the literature on deception communication, this distinction between cognitive constraints of spontaneous deception and deliberate deceptive strategies is not always made. However, it is a useful distinction to keep in mind because cognitive constraints may be less present in more planned language of written text.

The following section outlines research observations about deceptive communication attributes and continues to elaborate this cognitive constraint versus deceptive strategy distinction.

Observations about Deceptive Communication

Deceptive Communication is More Cognitively Taxing

Deceptive communication is more cognitively difficult. This is the fundamental underpinning of the cognitive constraint side of my proposed deception paradigm. Popular theories (Berger, Karol, & Jordan, 1989; Vrij, 2000) suggest that since deceiving extrapolates beyond reality, it takes more cognitive processing to deceive than to just represent the patent truth: more memory to remember how one has departed from reality in weaving one's lie, and more cognitive bandwidth to produce a deceptive alter-world. Brain imaging studies have demonstrated that being truthful involves a cognitive "baseline," while being deceptive is harder cognitive work (Spence et al., 2004).

From a production standpoint, this cognitive difficulty manifests in deceptive communication that is less syntactically complex, more hesitant, shorter, and full of errors than truthful communication. Buller, Burgoon, Buslig and Roiger (1994) found general "performance detriments" in deceptive communication; Newman et al. (2003) found deceptive communication to have "lower cognitive complexity."

A meta-analysis of "paraverbal" indicators of deception showed that higher speech errors were a reliable indicator of deception (Sporer & Schwandt, 2006). In fact, higher speech errors were one of only four reliable indicators that included pitch, response latency and shorter message duration (Sporer & Schwandt, 2006). Even in non-speech written communication such as electronic texting and instant messaging, typographical errors are more commonly present in deceptive communication (Zhou, Burgoon, Nunamaker et al., 2004).

Deceptive Communication Engenders Hesitation

Although a body of research has shown manifestations of cognitive difficulty in deceptive communication on an aggregate level, other studies have focused on more specific aspects of degradations in performance, including hesitation, that result from the effort of maintaining deception. Deception is revealed by hesitation (Harrison, Hwalek, Raney, & Fritz, 1978; Mehrabian, 1971a, 1971b). Interviewers associated hesitancy with deception (Harrison et al., 1978). Hesitation is also an indicator that listeners use to clue them in to deceit (Baskett & Freedle, 1974).

Deceptive Communication is Relatively Negative

Deceptive communication includes more disparaging remarks and more negative affect as part of a general negativity associated with deceiving.

Buller and Burgoon (1994) propose an "image- and relationship-protecting behavior" aspect of deception. This behavior results in negative affect of a deceptive speaker when deceptive communication is present, ostensibly for the purpose of covering up any leakage of guilty or fearful feelings (Zhou, Burgoon, Nunamaker et al., 2004). Thinking more broadly about the nature of instinctual *fight-or-flight* mechanisms, in a situation where someone is communicating deceptively, the deceiver effectively chooses a verbal strategy akin to *fighting*. Instead of running away, a deceiver uses negative language and expectations to fight an opponent. Deception as a fighting mechanism logically should have attributes associated with physical conflict such as aggression, negative affect and aversion. This is the case. In discourse studies of deception, statements indicating targeted aversion (Vrij, 2000), higher instance of negative affect (Buller, Burgoon, Buslig, & Roiger, 1996) and high levels of general negative

emotion words (Newman et al., 2003) have been observed. In some studies, linguistic measurements of negativity effectively discriminated between deceptive and truthful communication (Newman et al., 2003; Zhou, Burgoon, Nunamaker et al., 2004).

The increased negativity of deception straddles the strategic/cognitive division in traits of deceptive communication. On one hand, negativity is a direct result of the cognitive impacts of feelings or fear and guilt. However, this aspect of deceptive communication is not a matter of "leakage" but instead is a rather obvious trait. In some cases, in fact, the highly negative and adversarial aspect of deceptive communication can be used deliberately and strategically as a smokescreen to put message recipients off the trail of deception (Lebaron, 1996). I think we could expect that negativity in deceptive corporate documents would be more closely related to a strategic desire to distract recipients. Negativity would be less closely related to cognitive cues of deception (except perhaps in emails, which can be more "off-the-cuff" in composition).

Deceptive Communication is Relatively Impersonal

Besides negativity, other aspects of "image- and relationship-protecting behavior" (Buller & Burgoon, 1994, p. 204) that can be associated with deception are expressions of doubt and expressions of distance (Zhou, Burgoon, Nunamaker et al., 2004). Deceptive talk has been observed to be more "distant, impersonal, evasive, and unclear" (ibid. p.25). Deceivers are less concrete and less direct than truth-tellers, using more passive voice than in normal speech (Buller & Burgoon, 1996; George et al., 2004). Deceptive communicators use more distancing linguistic mechanisms than truthful, presumably to distance the deceptive communicator from his or her deceptive message (Zhou, Burgoon, Nunamaker et al., 2004). A possible instantiation of this distanced mindset is manifested in relatively fewer self-references (Newman et al., 2003). As

part of deflecting responsibility and protecting relationships, deceptive communicators also use more passive sentence structures than truthful communicators (B. M. DePaulo, Wetzel, Sternglanz, & Wilson, 2003).

Linguistic Indicators of Deception and Unethical Business Discourse

Corporate communication by and large takes place in the medium of writing. Writing eliminates some of the memory burdens on cognition inherent in speaking (Olson, 1977), and writing inherently distances from the message source (Ong, 1982). Accordingly it might be expected that some of the language features that mark deceptive corporate strategies in corporate discourse may be different than the language of interpersonal deception. Despite this fact, linguistic indicators of deception and unethical business discourse look in some ways very similar to general observations about interpersonal language and deception. The following potential indicators of deceptive corporate strategies in corporate fraud settings have parallels in the previously discussed attributes of deception.

These indicators were chosen in particular to cut through variations in genre and audience of the documents in the TDC. Extensive work has been done in corpus linguistics linking language features and genre. One risk of this study was that these linguistic indicators might only reproduce the divisions of genre that have been previously observed, and that at the same time, the linguistic indicators of this study would not be linked with deceptive corporate strategy at all. In order to come to conclusions about the efficacy of these indicators, the indicators needed to be few in number (I chose six indicators) and very specific. A large number of indicators would have risked replicating previous genre work instead of shedding any light on the efficacy of these

specific indicators. These indicators needed to be as deception-specific as possible, lest the study simply rank the TDC documents inadvertently along some pre-existing genre.

1) High Use of Adversarial Language

One supposition of research on deception is that truth-tellers expect others to believe them, whereas liars are more adversarial and challenging to their audiences because they expect challenges to their position (Shuy, 1998). As discussed previously, a study of deception in the Clarence Thomas Supreme Court nomination hearings notes that "aggression is always a part of deception to some degree" (Lebaron, 1996, p.119). Vrij (2000) calls this adversarial position a "negative stance" and gives examples of statements indicating aversion toward person and opinion "targets."

Within the tobacco industry, tobacco document research has shown that the tobacco industry actually threatened companies and individuals that supported anti-tobacco initiatives. Landman (2000) reported, "This aggression took many forms: threatening letters, economic attacks or threats thereof, orchestration of congressional investigations, and the formation of 'front groups'" (p. 340). The stonewalling approach of the tobacco industry utilized planned adversarial communication to distract (or intimidate) recipients. This strategic approach assisted deception and corporate fraud by forcing recipients to reckon first with the negative and aggressive aspects of any tobacco industry communication before they might be able to parse out deceptive elements of that same communication.

How the strategic trait of adversarial language is actually instantiated in a discourse community could be in part dependant on the culture of that corporate setting. A fuzzy word search for *shit*, *crap* and *damn* in the Tobacco Documents Corpus only yielded references to

s*crap*, skys*crap*ers, and Yo*shit*omi. We might hypothesize from this quick investigation that in a well-established and stable corporate environment, blatant vulgarity is rare because it is not seen as professional. On the other hand, metaphoric adversarial language, including "war metaphors," may be more indicative of negative and adversarial communication in most corporate settings (Koller, 2004).

In TDC documents, statements made by Philip Morris president Geoffrey Bible about the Environmental Protection Agency (EPA) and "Highly Vocal Extremists," as he calls anti-tobacco activists, are laden with rhetorical and metaphorical references to battle and war. These statements include the following: "They say that there is nothing like a battle to bring a team together… Let's fight hard. Let's fight smart. And let's fight using the ethics and the reasonableness that we know are our greatest weapons..." (Bible, 1996). These metaphors underscore the adversarial position taken by the tobacco industry as a whole.

2) Marked by Allness and Superlative Terms

Along with adversarial language, deception research asserts that allness terms and superlatives often mark efforts to deceive via exaggeration, a form of deception by explicit commission linked to bald-faced lying. *Always*, *never*, *nobody*, and *everybody* are mentioned as trigger words by Vrij (2000). Also, Lebaron mentions the use of *worst* (1996). Allness and superlative terms mark communication that has an absolutist position. Tobacco industry decision-makers often appear to have regarded themselves and their industry as 'singled out' for legislative and judicial attention.

3) Manifesting Group Mentality

Adversarial battle language underscores the inherent violence in corporate deception. One more way that deceivers' language reflects this violence is by constructing the deceiver as part of a group. According to guidelines used by FBI investigators, deceivers often try to enlarge the group they are in (Hess, 1997). They use more group inclusion in addition to fewer self-references to distribute responsibility and distance or cushion themselves from blame (Hess, 1997).

Linguistically, an enlargement of the group that you are in is most readily expressed by the use of first person plural pronouns (*we, us, our*). Of course many referentially appropriate uses of first person plural pronouns are present in everyday communication and speech. However, a non-specific *we* can also diffuse responsibility for negative outcomes and can be associated with a kind of identity of victimization. Seemingly in contrast, other research projects have singled out a lower usage of all pronouns (Lina Zhou et al., 2004), and specifically all first person pronouns, in deceptive compared with truthful communication.

4) Strategically Ambiguous Intent: Few Commissives

Being strategically ambiguous about intentions may be one of the most effective and easiest ways to avoid responsibility and to deceive. It allows the speaker or the writer to avoid being specific about his or her involvement in any situation. Lack of obligation deflects responsibility since responsibility is directly related to attribution of causality and involvement.

Roger Shuy (2003) has pointed to the use of ambiguous intention in language as a strategy employed by the tobacco industry in their public announcements and websites. Theoretically from a linguistics pragmatics standpoint, all communication should be as specific

as necessary to impart the relevant information (Grice, 1989). With this in mind, Shuy (2003) notes that strategic and pervasive ambiguity runs contrary to a kind of “best practice” of informative communication.

Shuy associated ambiguity, and by extension lack of obligation, with rephrasing of future tense verb phrases as hypothetical phrases to obscure intention (Shuy, 2003a). For instance, ‘would’ or ‘want to’ does not have the same specificity as ‘will.’ This choice of words defines the difference between “desire without obligation” versus “commitment” (Shuy, 2003a). Shuy reports: “the text of the Philip Morris’ Options message includes no promises and no verbs of futurity” (2003). His examples give a clear picture of the tobacco industry (or at least Philip Morris) as avoiding references to concrete or future events:

"We are working with business owners..,"

"Options provides information and resources..,"

"Options is involved in a number of practical activities..,”

"We sponsor forums ..,"

"We also invest...in educational programs..,"

"Options offers business owners access to..," [sic]

All six expressions are vague. While this kind of non-specific language is not inherently unethical, when it is used inappropriately or excessively, it can lead an unsavvy reader to the conclusion that actions have been and will be taken, when there may be no such actions in the works. At the heart of the matter, Shuy is giving specific tobacco industry evidence of strategic manipulation of what Searle discusses as *commissives* (1979). Commissives are a form of performative speech-acts that promise future action. These commissives commit the speaker to

some future action, creating an obligation which is sometimes so strong as to be enforceable by law.

Using commissives such as *will*, *shall*, *promise* or *should* creates an environment of obligation (Searle, 1969). Corporations involved in deception and fraud are keen to avoid obligation and obscure their future intentions so that they cannot be held responsible for corporate actions. Considering Searle, strategic ambiguity about intent might actually be best indicated by a lack of commissive statements. Commissive linguistic constructions that directly or indirectly promise definite future-states are at odds with strategically ambiguous intent. *Will, shall,* and *going to* are definite future modal verbs (Kahn, 2004). *Will* and *shall* are both colored by volition and obligation (Jespersen, 1909-49; Wekker, 1976), in contrast to intentionally ambiguous and obligation-avoiding language. Leech (1971) agrees that *be going to + verb* is a "future fulfillment of present" intentions. Future modal verbs have a heavy mark of definite anticipatory action makes them antithetical to statements that are strategically ambiguous about future intentions. Statements that use *will, shall* or *going* from a pragmatic perspective, fall into the commissive realm of indirect promise. The frequency of future modal verbs as linguistic features may be the best measure of strategically ambiguous intent. Lower instances of these commissive construction verbs may be part of what differentiates deceptive communication from ordinary communication.

5) Deprofiling Agency with Passive Constructions

Passive constructions have been observed to be a major element in rationalizing arguments (Drozhashchikh, 1991; Lachowicz, 1981). Passive voice sentences deflect responsibility and deprofile human agency (Ilie, 1998; M. Manning, Rymes, Weninger, &

Brown, 2004; Rymes, 1995). In a study on the discourse patterns of high-school delinquents, participant narratives of events that reported unscrupulous behavior used passive voice to de-agentivize their actions (Rymes, 1995). The individuals recast themselves, not as the aggressors they have been accused of, but instead as passive responders who are not responsible for their actions and have no agency in the situation.

Using passive voice in contrast to active voice rearranges emphasis in a phrase. For instance, the passive voice sentence *Julie was hit by Frank* shifts the emphasis to Julie while understating the action of Frank. Another version of this passive sentence is agentless: *Julie was hit.* It in fact doesn't even have to mention Frank at all. Using the passive voice contributes to depersonalism (disassociation) manipulations that distance the speaker from their deceptive message. Using passive structures when active structures would be possible "reduces a sender's ownership of a statement" (Zhou, Burgoon, Nunamaker & Twitchell, 2004, p.87) and "distances the messenger from the message." In effect, passive constructions can contribute to deceptive communication by deflecting responsibility (or simply not attributing it at all) and by deprofiling the agency of actors involved in whatever action is being reported or communicated.

In addressing the issue of deception by omission, a type of particularly insidious deception, Thomas (1997, p.50) notes that "a close look at the language structures … offers a view of not only what the company wanted its audience to know but also what the company may not have wished to reveal." Specifically, the Thomas (1997) study showed that in a series of annual reports from the same business, the frequency of passive voice verb structures, which is associated with deflection of responsibility, doubled in the years when the company lost money compared to years when the company made money (Thomas, 1997, p.53).

Frequency of passive constructions was calculated as the total passive phrases divided by the total number of phrases. These calculations are based on integrating a function of MontyTagger (Liu, 2000) that tags phrases for passives. See Appendix A for the programming module for passive constructions.

6) Managing Image with Cognitive Verbs (as Opposed to Content Verbs)

Another linguistic resource specifically observed in the area of deceptive image-management for the tobacco industry is the use of cognitive (in contrast to content or behavior) verbs (Shuy, 2003). The continued use of cognitive verbs privileges an emotion-oriented instead of fact-oriented account of communicated information. Does privileging emotion over fact and promise in the context of business communication constitute deception? Not necessarily. In isolated instances, use of cognitive verbs would probably not indicate deception and unethical business practices. However, this type of communication limits the informative impact of a message.

The use of *believe* and other cognitive verbs such as *consider* (Lebaron, 1996) also continues to be used to create an emotional stance for Philip Morris USA (Brown & Derry, 2004). For example, in an address about helping to reduce tobacco sales to minors, Mr. Szmanczyk, CEO of Philip Morris USA says, "We *support* the states in their enforcement of existing laws and agreements and we *stand ready* to share our knowledge of the industry if needed" (cognitive verbs italicized) (Szymanczyk, 1996). These cognitive verbs orient Philip Morris USA towards a position on the topic that highlights emotional or cognitive. In addition, this focus on cognitive and emotive aspects of existence steers Philip Morris USA away from making any concrete statements that could subsequently be challenged. In essence, the

heightened use of cognitive verbs allows the industry to do two things: assert their non-factive position with respect to an issue and also to side skirt issues by addressing them from an emotional point of view.

Identifying the Language of Corporate Deception

At this point, the research reviewed has demonstrated that computational approaches to detecting deception are viable and that in addition to interpersonal deceptive communication indicators, there are a number of corporate deception indicator candidates. Our next hurdle is to assess the occurrence of these candidate linguistic indicators of deceptive corporate strategy in tobacco industry documents. Then we can investigate whether these indicators (either alone or working together) can be used to rank internal documents for deceptive corporate strategy in the TDC. In order to determine the presence and relative frequency of these indicators, I have developed computer programs to identify instances of the proposed indicators in text files. These computer programs and the computational strategies used to rank TDC documents for deception are outlined in the next chapter.

CHAPTER 3

METHODS

The last chapters surveyed the topics of deception, corpus linguistics and the history of the tobacco industry. The six potential indicators of corporate deception were introduced as part of a general review of deception and language research. Also, an overview of some fundamental aspects of corpus linguistics laid out some of the boons and shortcomings of corpus linguistic approaches to research. Biber's (1988) extensive suite of features has set the stage for using linguistic features to differentiate between genres, styles and registers. At the same time, Biber's features have been difficult to decipher and are too comprehensive for the current task of detecting deception. Using his suite of features for a task like deception and fraud detection might be like trying to hunt bumble-bees with machine guns. Instead, a precision-laser is needed to attack a precise problem.

The conceptual precision-laser for the task of automatically ranking documents based on deception and fraud is the suite of six proposed indicators of corporate fraud and deception: adversarial language, allness and superlative terms, group mentality, strategically ambiguous intent, deprofiling agency, and image management. These indicators have been extensively researched and defined. They represent a synthesis of the most frequently observed features of corporate deception and fraud (Buller & Burgoon, 1996; Burgoon et al., 2003; Burgoon, Buller, Floyd et al., 1996; B. M. DePaulo, Lindsay et al., 2003; Dyas, 2003; Ekman, 2001; Galasiânski, 2000; Harrison et al., 1978; Lebaron, 1996; Newman et al., 2003; Porter & Yuille, 1996; Shuy, 1998; Todd-Mancillas & Kibler, 1979; Todd, 1977; Vrij & Mann, 2001; Zhou, Burgoon, Nunamaker et al., 2004; Zhou et al., 2003). In order to simplify a study that already involves

potentially confounding factors such as differences in register and document functions, this set of features is small in number but directly derived from deception literature.

This chapter explicates the methods used in this study. Relevant methods include corpus building, vector ranking algorithms, operationalizing the proposed indicators of deception and parameters of the discourse analysis to be used ultimately in qualitatively assessing a selection of TDC documents. It also outlines statistical methods used to determine the distribution of indicators of deception in subsets of TDC documents based on the hypotheses proposed in Chapter 2.

Building a Corpus

As a first order of business for an assessment of deceptive language in the tobacco industry, a corpus of documents needed to be constructed. As of November, 2004, the searchable tobacco documents at the Legacy library website (http://legacy.library.ucsf.edu/) totaled 41,847,214 pages. Even if each page has only ten words on it (approximately just over one full sentence), the total would be half a billion words. But creating a corpus non-systematically from such a universe of discourse poses severe risks. What if, in creating your manageable-sized tobacco industry document corpus, you take the first 2000 documents and inadvertently exclude all documents originating from the American Tobacco company source? In that case, you would be missing language patterns from an entire branch of the industry. In a slightly better scenario, suppose that you choose to create your manageable-sized corpus randomly. This might reduce your risk of excluding all of a certain type of document, for example memos. Despite reducing this risk, statistically speaking, by using a totally random corpus creation method, you very well have a chance (albeit a small one) of not pulling any memos into your corpus. The real risk of an

unbalanced corpus is that it would misrepresent the larger document collection and possibly invalidate your results over the larger collection.

To combat the risks of an unbalanced corpus, a corpus must be built with a plan in mind to balance types of documents and to preserve representation of document classes. This kind of balanced and representative corpus is necessary for accurate and reliable corpus study results (Kennedy, 1998; Kretzschmar, Meyer, & Ingegneri, 1997; Meyer, 2002; Oakes & NetLibrary Inc., 1998). A goal for corpus creation is to have the proportion of any document type in the larger document set match the proportion of that same document type in the final corpus.

A balanced corpus enables a researcher to generalize about specific studies conducted within that corpus (Leech, 1992). In comparison to corpora that may be compiled by sheer convenience, a representative corpus is rigorously designed to choose documents or samples of documents in such a way that distribution of document classes and other metadata features (e.g. which language a document is written in) are preserved. This design helps ensure that the corpus represents the larger document collection. A balanced and representative stratified corpus of tobacco documents enables researchers to generalize about the language of the larger set of documents. The 600,000-word Tobacco Documents Corpus (TDC) is just such a balanced and representative corpus.

Kalton (1983) proposed quota samples as a way to create balanced datasets. In order to create quota samples, a researcher first compiles a random sample, and then replicates the parameters of that sample based on a sample frame (Meyer, 2002). In the case of the TDC relevant sample parameters included such factors as intended audience for the document (industry internal or industry external), author number (singular or corporate), date (by decade)

and industry source (company of origin). The Tobacco Documents Corpus Project (Kretzschmar et al., 2004) used just such a method to create a representative sample of tobacco industry documents. The sample frame and final parameters of the TDC are outlined in the following sections

Since the TDC has been constructed with balanced representation in mind, researchers using it can generalize findings about the corpus out to the entire 3 million+ tobacco industry documents that had been disclosed at the time of its creation (2003). Note that since the development of TDC, there has been a major increase in documents images publicly disclosed and available for inspection online. The more current document archive includes many that were part of discovery in the recently concluded massive US Department of Justice civil racketeering trial, others from recently indexed files of the Tobacco Index, and yet others from the Guildford document depository in the UK. For example, the TDC has proportionally as many 1950s documents as there are in the whole tobacco industry documents as a result of its stratification by decade. The final size of the TDC – over 600,000 words – compares favorably with that of one of the few data sets available for deception communication studies. That alternative data base consists of only 13,243 words (Twitchell et al., 2004).

Distribution and Features of TDC Classes

To represent the entire tobacco industry documents with the TDC, the Tobacco Documents Corpus research team had to first define classes of interest. Classes of interest included audience (internal vs. external), decade, source (American Tobacco, Brown and Williamson, Lorillard, Philip Morris, R. J. Reynolds, Council for Tobacco Research, Tobacco Institute) and audience number or *addressee* (individual vs. corporate). The first step in

establishing this corpus was to determine the frequency of documents by decade so that the TDC could be historically representative and usable in diachronic language studies. Dates could not be ascertained for some of the documents, and the frequency of the "no decade" (referred to within the corpus as *19xx* documents) documents was likewise replicated. In addition, one set of documents was determined to be so key as to warrant treating it as a stratification class along with decade. These were the 36000+ Bliley set documents. Named for United States House Commerce Committee Chairman, Thomas Bliley, these documents were released to the public after the settlement of the Minnesota tobacco litigation, at Bliley's request (Legacy Tobacco Documents Library, 2005). The importance of the Bliley set of documents rests on the fact that these documents were not initially made available to the public, due to their claimed privileged and confidential status. The court decision demanded the release of these documents (under the crime fraud exception to attorney-client privilege - U.S. District Court, District of Columbia ("212 FRD 421 DDC," 2002)). However, the tobacco industry balked and challenged the verdict, which threatened to keep the documents in limbo for years. To counteract this, Bliley subpoenaed the documents and posted them on the House Commerce Committee's website, making them *de facto* public documents. These documents are of particular interest because they represent a selection of documents that the tobacco industry asserted attorney-client privilege over. Other such documents are indexed with the rest of the tobacco industry documents, but are not able to be viewed due to assertions of privilege.

It was determined that no decade classification within this exploratory core sample (including the Bliley set) should include fewer than 10 elements, since to do so would be to essentially exclude that class of documents from further analysis. The following table (adapted

from Kretzschmar et al, 2004, p. 35) illustrates the percentages of the overall tobacco industry documents that were then proportionally replicated in the TDC.

Table 1
Sampling Targets for Exploratory Core Sample

(replicated in part from Kretzschmar, Darwin, Brown, Rubin & Biber, 2004, p.35)

Group Decade	*Total Documents*	*Year/Month*	*Documents*	*Need*	*%/ total Sample*
1900-1959	103,574	1/August	1,193	10	3.055
1960	223,544	0/April	1,136	22	6.593
1970	660,223	12/September	5,895	66	19.473
1980	1,318,823	1/January	7,185	132	33.898
1990	988,793	6/June	1,679	99	29.164
Undated	62,494	n/a	n/a	6	1.843
Bliley	33,003	n/a	n/a	3	0.973
Total	3,390,444		338	349	

This core sample was a random 0.01% of the entire body of tobacco industry documents. This meant that the core sample was large enough to represent each decade and the Bliley set,

but small enough for the team to examine relatively quickly by hand. Each document in this core sample was categorized according to document source, addressee and audience. This helped us determine our initial counts of the different strata that would make up the random stratified sample (see Table 2, below).

Audience and addressee were the two binary document classes, and also two classes of documents which might help researchers investigate deception and corporate fraud. Audience was of interest due to the perceived formality and extensive editing of documents with external audiences versus the potential informality and "off-the-cuff" quality of documents with internal audiences. This document division by audience was intended for use in determining whether internal audience documents were more candid about supporting deceptive and fraudulent agendas. Addressee was also chosen as a class in part to investigate the same question. Ideally, documents addressed to single (*named*) addressees would be personal communications, and might be more candid. By contrast, the theory was that documents addressed to a list of individuals or an unnamed corporate recipient would be more indicative of day-to-day corporate documents, memos and reports.

Lastly, source was of interest as a document class because of the different strategies and purposed for different corporate entities. As outlined in chapter 2, for instance, CTR and TI source documents were expected to differ quite a bit from the other sources. In addition, strategies for dealing with the public differed between the major industry players, and this was expected to be reflected in language in those sources.

The distribution of other classes of interest within that the initial core sample is illustrated in the table below adapted from Kretzschmar et al., 2004, p.37. Eventually, short

documents less than fifty words and non-English documents were excluded from the TDC because they have been determined to be unfit for the type of document-based linguistic analysis that was envisioned.

Not surprisingly, the tobacco industry archives contained a substantial number of documents that were not authored by industry sources at all. These included items such as newspaper clippings, articles from scientific journals, public relations pieces from anti-smoking organizations, and government documents. These documents from industry-external sources needed to be excluded from the final TDC. However the boundary between industry-internal and industry-external sources can be murky in an enterprise so far reaching as the tobacco industry. For example, many research reports were authored by individuals and shadow organizations with substantial financial support from industry. Even newspaper articles may have been published by industry shills. Industry-internal source was eventually operationalized as those documents from entities or individuals who had been paid at some point by the tobacco industry. It was determined that accepting payment amounted to becoming part of the industry and thus industry-internal. Documents from sources external to the tobacco industry, thus defined, were excluded from further analysis.

Table 2

Distribution of Classification Categories of Exploratory Core Sample

	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
Documents judged to have public health significance	9	10	10	22	61	126	96	334
Forms	2	0	0	2	8	18	19	49
Documents that were primarily graphic images	0	0	2	1	1	0	0	4
English	10	9	10	22	63	130	99	343
Documents showing editing marks	3	1	0	2	3	3	5	17
Documents containing marginalia	4	8	5	12	34	73	39	176
Short documents	3	2	2	4	20	37	33	101

The exploratory core sample then served as a template for establishing sampling criteria for the TDC. Approximately 1000 documents were sampled according to the frequencies ascertained from the core exploratory sample That template was augmented, however, with the addition of 100 more documents, likewise stratified by decade, addressed to industry-external audiences. It was deemed important to oversample these documents addressed to industry-external audiences because they were so scarce in the overall population, and the resulting sample-frame. In order for the division of internal vs. eternal audience to possibly ever be statistically significant, the TDC needed to incorporate more industry-external audience documents. These documents were added as a supplement to the original TDC.

In general, audience was defined similarly to internal vs. external source definitions. Documents were considered to have an internal audience if they were a draft, or if they were addressed to an entity or person within the industry. Here again, the broader definition of industry as any entity or individual who accepted funds from the tobacco companies, institutes or centers was used. External audience documents were addressed to entities outside of the tobacco industry. Additionally, named addressee vs. corporate addressee was operationalized by recipient. If the recipient named was a specific person (regardless of the number of people) then the document was considered named addressee. Conversely, if there was no specific person named as recipient, the document was classified as having a corporate (unnamed) addressee.

The final counts for the classes of interest in the TDC were as follows: 939 Industry-Internal Audience Documents, 174 Industry-External Audience Documents; 80 American Tobacco Company documents, 117 Brown and Williamson, 45 Center for Tobacco Research, 97 for Lorillard, 396 Philip Morris documents, 99 Tobacco Institute documents; 492 Named

Addressee Documents, 621 Corporate Unnamed Addressee Documents. The classes (independent variables) focused on for this dissertation project are decade (including Bliley set), audience affiliation (industry-internal and industry-external), addressee number (named and corporate) and source (the six corporate entities). From a deception and language standpoint, differences amongst the various levels of these classes in the distribution of the six indicators of deception are of interest.

Statistical Analysis

The distribution of deception indicators across the TDC's independent variables of audience affiliation, addressee number, source and decade has been examined. Certain basic inferential statistics have been calculated. Levels of the following two independent variables were compared using t-tests (due to their binary nature): audience (internal vs. external) and addressee (named vs. corporate) with respect to the six dependent variables of the indicators of deception. Source (company of origin) and decade were compared with one-way ANOVAs, since each of the independent variables has more than two levels. In each case an *a priori* level of statistical significance was established as $\alpha=.05$.

Document Classification: Vector Analysis

Vector space models have been used in document searching and classification in various ways (Frakes & Baeza-Yates, 1992; Li & Jain, 1998; Markov & Last, 2005; Raghavan & Wong, 1986). As early as 1980, Salton was using vector space models to inform intelligent text searching (Salton, 1980). Additionally, even as recently as 2005, researchers were asserting that "most web classification methods are based on the vector space document representation of

information retrieval" (Markov & Last, 2005, p.1). At this point, many classification methods are more sophisticated than simple vector space document representations, and may include Multidimensional Scaling techniques or other mathematical clustering algorithms. This research does not use additional clustering methods to augment the simple vector space method. Due to the relatively small number of documents and number of indicators, it was not necessary to adopt more sophisticated mathematical models. This choice was made because part of the goal of this dissertation is to evaluate the efficacy of the indicators of deception.

In the context of this dissertation, providing accurate results does not translate to a goal of having this algorithm correctly assess documents for deception. In the context of this dissertation, providing accurate results means creating a straightforward way to analyze the efficacy of these indicators as they behave alone and as they behave together in the population. The vector space method is used without extra augmentation to assess how well the indicators perform when taken together to rank the population. This pared down approach allows for an accurate representation of how the indicators might work together in a professionally developed interface, while at the same time keeping the mathematical model simple enough that it can be easily untangled for further and more in-depth analysis of each of the indicators.

The vector space method applies theories of vector space to document populations as a way to determine the similarity of two documents (Manning & Schütze, 1999). According to a vector space model, any document (or for that matter *anything*) can be represented as an n-dimensional vector with axes that represent each feature of interest. For illustrative purposes, imagine that there are three very short documents to categorize. Document A reads : *the chair*. Document B reads : *another chair*. Document C reads : *hi there*. If we used the vector space

model to predict which document of B and C was topically most similar to Document A, we could use lexical items to produce a vector. Document A's vector would look be (1,1,0,0,0) where the lexical items represented are (*the*, *chair*, *another*, *hi*, *there*). Document B would be (0,1,1,0,0) and Document C, (0,0,0,1,1). Without going through the actual mathematics, in this case we can simply look at the vectors and predict that Document A and Document B have more in common than Document A does with Document C. And indeed, "the chair" lexically has more in common with "another chair" than "hi there."

On the level of topic and lexical similarity, the vector space method has been extremely productive and is a standard for document classification (Frakes & Baeza-Yates, 1992; Li & Jain, 1998; Markov & Last, 2005; Raghavan & Wong, 1986; Salton, Wong, & Yang, 1975). It is unusual, however, to use the vector space method for document classification beyond subject matter and lexical similarity. This research specified that only the six linguistic indicators of corporate deception and fraud are the features of interest. The vector in this study is a six-dimensional vector where each dimension represents the relative occurrence of one of the potential indicators of corporate deception and fraud. In order to rank the TDC documents for degree of deceptiveness, each individual document is represented as a vector in a 6-dimension space where each linguistic indicator is one dimension. The current research uses the vector space model to rank documents according to similar goals of corporate deception and fraud.

Since vector analysis is primarily a heuristic that matches documents to some kind of template, this study needs the "perfect" deceptive document as a benchmark to order the other documents against. In other situations, this benchmark could be a series of "hot" documents that are precise examples of whatever topic is of interest. With this in mind, it was imperative to

determine what constitutes the perfect deceptive document. The most deceptive document in one corpus or one situation could differ from the most deceptive document in a different corpus or situation. Within the realm of tobacco industry documents, however, the class of external audience documents stands out as being quintessential deceptive and misinformative documents. The overwhelming body of tobacco control research supports the position that public audience documents are strongholds of deception and misinformation (Barnoya & Glantz, 2002; Bero, Glantz, & Hong, 2005; Dearlove et al., 2002; Givel & Glantz, 2001; Glantz, 1996; Kessler, 2004; Pollay, 1989, 1990, 1994, 1995; Pollay & Dewhirst, 2002; Ringold, 1986; Wulinger, 1984). However, patterning the ranking on one area of tobacco industry deception could have limited the types of deception to those used in a narrow area of deceptive corporate strategy, such as tobacco industry lobbying (Givel & Glantz, 2001), tobacco industry use of competing interest disclosures to limit free speech (Bero, Glantz, & Hong, 2005), or the "Latin Project" (an overarching plan to prevent regulation of second-hand smoke in Latin America) (Barnoya & Glantz, 2002). Since there really are so many different ways that the tobacco industry enacted deceptive corporate strategies, rather than pattern this ranking system on one area of tobacco industry deception, it was more appropriate to take the average of external-facing (public) documents as a whole as the benchmark of deception. This choice included highly deceptive documents in the group deemed appropriate for the benchmark of deception, and although not a "hot" document approach, represents deception within the industry in a wide variety of forms. The entire program of the tobacco industry with respect to its communication with the public was supporting corporate deception and for this reason, sampling only specific deceptive acts for the purposes of this study would have been shortsighted. The propaganda referred to below

characterizes prototypical corporate deception and fraud (Schudson, 1984, quoted in Pollay, 1995):

> "Not since the days when the vendor of harmful nostrums was swept from our streets, has this country witnessed such an orgy of buncombe (bunk), quackery and downright falsehood and fraud as now marks the current campaign promoted by certain cigarette manufacturers…" (p. 2)

With this in mind the "typical" deceptive document for this study should be an average of external audience documents. The typical deceptive document for this study will be represented by an average vector standing for the average external audience document in the TDC.

For the TDC, each internal document has an ordered set of six numbers that represent the incidence of the indicators of corporate deceptive strategy and fraud. Remember that only internal audience documents from the TDC will be ranked for deceptive corporate strategy, as the external documents are being used as the benchmark for deception. Without internal support, public misinformation and deception could never have happened. Accordingly, the pragmatic intent that may be extrapolated to other industries as well is that using a "most deceptive document," we can find internal documents that were most instrumental in supporting deceptive public endeavors. This process may be useful in monitoring and policing internal industry practices in the future.

Each document can be represented by the vector from the origin (0,0,0,0,0,0) to point (n,m,l,k,j,i). One reason to use this vector representation is that it allows for very easy comparison of documents. Two documents are most similar with respect to deception if the angle

between their vectors is closest to 0. This model is simply a mathematical way to represent documents and their similarity or difference from one another according to whatever indicators are part of the assessment. In large industry endeavors, for example, a vector representing the information from 1000 or more indicators could be used to triangulate more information about the document. In this instance, however, because of the necessity of assessing not only the documents, but also the indicators themselves, this number was greatly reduced, to a very manageable six indicators. Since we have reason to believe that external audience documents are the most deceptive, internal audience documents are ranked for deception based on the angle of their representative vector from the average external audience document vector. The industry-internal audience documents with the smallest angles are most similar to the average external audience document vector. Being most similar to the most deceptive documents, these small angle documents should show more evidence of deceptive corporate strategy than other documents. Conversely, documents that have vectors whose angles from the average external document vector approach 90 degrees should be the most dissimilar to external audience documents and thus have the least deceptive corporate strategy.

In order to actually calculate the angle, I used the 6-indicator-dot-product.py script which is adapted from a python dot product module from the Game Program Wiki (http://gpwiki.org). The six linguistic indicators of deception were calculated for each document (more on those calculations in the next section). These indicators were compiled in a spreadsheet and indexed by an identifying document number. Using this spreadsheet and sorting by audience, I determined what the average external document vector was (0.117737053, 0.133239, 0.220337, 0.00478, 0.001892, 0.034159134). Using this average external document vector and the dot product script,

I calculated the angle between the average external TDC vector and each internal TDC document. Mathematically, for vectors *a* and *b*, the dot product is the length of *a* times the length of *b* times the cosine of the angle (theta) : ($a.b=|a|*|b|*\cos\Theta$). So, in order to find the value of theta, the 6_indicator_dot_product script takes the arccosine of the dot product of *a* and *b* divided by the length of *a* times the length of *b*: ($\Theta = \arccos((a.b)/(|a|*|b|))$).

The computer programs and processes used to establish the indicators that make up the vector elements are outlined in the next section.

Assessing the Linguistic Indicators

The potential linguistic indicators of deceptive corporate strategy and fraud are the following: adversarial language, lack of obligation, group mentality, deprofiling agency constructions, cognitive verbs, and allness terms. For each of the nominated linguistic indicators of deception, modules in the Python scripting language were written to estimate the count of instances of each indicator within a document (see Appendix A for their Python computer programs). The indicators as a whole are intended to be generic enough to resist influence based on genre or style. Though these indicators were developed in part from studies of spoken language, they are anticipated to be general enough to gracefully handle a transition from spoken to written language. Additionally, they should not differ significantly by individual author due to their generic nature. However, in the case that they do differ by individual author, the multi-indicator vector model assessment should enable the indicators to work together to smooth over these differences, and other genre or style difference.

Each of these modules outputs tab or comma-delimited material that is Excel spreadsheet readable. During the statistical analysis, ranking and subsequent discourse analysis phases of this

project, documents were referenced only by their Bates number in order to maintain uniformity and document anonymity.

Cognitive Verbs

Image management verbs were calculated as the total stative verbs plus total relational verbs divided by the total number of base form verbs. Base form verbs were defined as the VB-tagged verbs using the MontyTagger (Liu, 2004). Stative verbs included the following (Kiparsky & Kiparsky, 1970; Lebaron, 1996; Shuy, 2003b): *believe, consider, think, like, mind, recognize, prefer, seem, doubt, abhor, adore, astonish, desire, detest, dislike, feel, forgive, guess, hate, hear, imagine, impress, intend, know, love, perceive, please, presuppose, realize, recall, regard, remember, satisfy, see, smell, suppose, taste, thought, understand, want, wish*. Relational verbs include the following: *concern, consist of, contain, depend on, deserve, involve, lack, matter, need, owe, own, possess, require, resemble, appear, become, appreciate*. See Appendix A for the programming module for stative verbs.

Group Mentality

Group *mentality* was calculated as the total plural first and third person pronouns (*us, our, ours, we, them their, they, theirs*) relative to the total number of personal pronouns, possessive pronouns and plural nouns (see Hess, 1997). Counts for first person plural pronouns and third person plural pronouns were combined in order to reflect the *us v. them* stance of group mentality instead of focusing singularly on one side of this group mentality dynamic. The group mentality pronouns were denominated by the total number of personal pronouns and plural

nouns because these would be candidate constructions that could have been replaced with a plural pronoun. See Appendix A for the programming module for group mentality.

Adversarial Language

Adversarial language is calculated as the number of adversarial words divided by total words in a document. The adversarial words dictionary entries were nominated based on words mentioned in the literature (Lebaron, 1996; Vrij, 2000). The *synset* function of Wordnet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1993) was used to nominate synonymous words for *battle*, *war* and *adversary*. A synset is the set of synonyms in Wordnet for each particular word. Using the synsets of the core adversarial words listed above, the final word list for adversarial language is the following: *war, battle, oppose, defend, crusade, campaign, push, agitate, combat, assault, dogfight, duel, strife, conflict, fight, struggle, attack, assail, hostile, antagonism*. See Appendix A for the programming module for adversarial language.

Deprofiling Agency Constructions

Deprofiling agency constructions – that is, grammatical constructions that reduce the salience of the logical agent of actions – were calculated as the total passive phrases divided by the total number of phrases as defined by a language and grammar tagger – the MontyTagger (Liu, 2004). These calculations are based on integrating a function of MontyTagger (Liu, 2004) that tags phrases for passives. See Appendix A for the programming module for Deprofiling Agency Constructions.

Ambiguity of Intent Index

The Ambiguity of Intent index was calculated by adding the total number of *will*, *shall*, and *going to* divided by the total base form verbs defined as VB tagged verbs from the MontyTagger (Liu, 2004). See Appendix A for the programming module for Ambiguity.

Allness Index

Allness was calculated as the total number of allness words, superlative adjectives and superlative adverbs divided by the total number of words. Allness words include *always, never, nobody, forever, everybody* and *everyone*. These allness terms specifically exclude the "base forms" such as *all*, *every* and *each*. The reason for this exclusion is that these base forms are not necessarily used to hyperbolically discuss everyday items. *All*, *every* and *each* are frequently limited to specific sets of things that are not necessarily extensive, hyperbolic or exaggerations. For example, in the following sentence, *all* only serves as a term indicating a whole-set, and not necessarily as an *allness* term: *Would* all *the left-handed, blond-haired girls wearing red jackets and between the age of five and six please come to the front of the room?*

In addition, using the MontyTagger (Liu, 2004), superlative adjectives (tagged as *JJS*) and superlative adverbs (tagged as *RBS*) were totaled as well. See Appendix A for the programming module for Allness.

Procedure

After running the computer programs to determine the relative instance of each of the six potential indicators of deception in every internal audience document in the TDC, these counts were compiled into one Excel spreadsheet where internal audience documents were indexed by

Bates number. The Bates number of a document is a unique alpha-numerical identifier given to each document as it entered either litigation or the body of publicly available tobacco industry documents. In the spreadsheet, each document has a set of six numbers representing the level of deception indicator in that document. We can think of this set of numbers as a vector of six dimensions. Each internal audience TDC document was represented by its vector.

After the average external document vector was established, the angle between each internal audience document's 6-dimensional vector and the average deceptive document vector was calculated. In order to calculate the angle, a Python-based programming module was used (see Appendix A for the code of this programming module). Then, the internal audience documents were ranked by angle against the average external audience document. Theoretically, this angle should represent how directly supportive of deception a document is: documents with the smallest angle or distance from the vector representing an average of external (deceptive) documents will be the most directly involved in eventual deceptive corporate strategy. Documents with the angle closest to 90 degrees from the average external audience document vector should have the least deceptive corporate strategy and be the furthest removed from public misinformation and deception.

Discourse Analysis of Deception-Ranked Documents

In order to test the efficacy of the ranking based on the linguistic indicators of deception and also to offer more detailed discourse-level descriptions of language characterizing levels of deception, it was imperative to more closely examine some of the internal audience documents from the TDC. To accommodate detailed discourse analyses, documents were elected for further analysis based on their angle/rank: five were the documents with the angles closest to the

deceptive template of average external audience documents (i.e., the five internal-audience documents that would be stylistically closest to external-audience documents and therefore likely to be most deceptive); four were the very middle documents, and five were the documents with the largest angles. A thorough and detailed discourse analysis was conducted on these fourteen documents while the actual ranking of the individual documents remained masked. This discourse analysis is reported in the next chapter and was based on general discourse analysis approaches by Gee (2005), Fairclough (2003) and Shuy (1998).

The goal of this discourse analysis was to determine in part the deceptiveness of the document with respect to public health, and also how integral the document might have been in supporting deceptive communication. Functionally, this discourse analysis approach incorporated micro-analysis on a lexical or syntactic level with more whole-document-level macro-analysis. The macro-analysis in this case was facilitated by historical characterizations not only of the major players in the tobacco industry, but of many of the specific documents in the tobacco industry set. The social and cultural aspects of this analysis (usually associated with critical discourse analysis) were focused specifically on potential deception, public endangerment and business ethics standards.

These discourse analyses are intended to situate the documents and determine whether (or to what degree) these documents are deceptive and/or involved in unethical business practices. At the crux of this analysis, of course, was the question of whether the proposed indicators of corporate deception and fraud had been effectively used to accurately rank documents in order of deceptiveness.

Approaches to discourse analysis methods are at best less than uniform (Gee, 2005). There are, however, some general algorithms proposed in the literature that cover most of the major issues. For the purposes of this study, tobacco document texts identified as the five most deceptive, five least deceptive and four exact middle have been analyzed using complementary approaches outlined in Gee (2005) and Fairclough (2003).

Discourse Analysis – Gee

Gee (2005) identified seven "building tasks" that are of primary importance for a thorough discourse analysis. The term "building tasks" underscores Gee's assertion that language is being used constantly to build things. Situations, relationships, histories and cultures are "built" via language. The following sections outline Gee's seven tasks and Fairclough's algorithms for dismantling assumptions about language. The information in these sections has been used to develop a full discourse analytic picture of selected tobacco documents which is reported later in this dissertation.

Significance

Language is used to make meaning or significance out of everyday objects and occurrences. For instance, if we whisper (this works especially well with children), other people will begin whispering, too. At the least, they may ask why we need to whisper. Here, language gets used to create the meaning of secrecy or some other need for quiet.

Discourse Analysis Question: "How is this piece of language being used to make certain things significant or not and in what ways?" (Gee, 2005, p.13).

Activities

Language is used to create activities. For example, in certain settings, the phrase "Let us pray" will signal that a group should now begin the activity of praying. Language employs activities to make it evident to others what a speaker considers themselves to be doing.

Discourse Analysis Question: What activity or activities is this piece of language being used to enact (i.e., get others to recognize as going on)?

Identities

Language is used to create roles or identities. A person, for instance, who uses language to answer detailed questions or use specialized language, may be creating the identity of "expert" for themselves.

Discourse Analysis Question: What identity or identities is this piece of language being used to enact (i.e., get others to recognize as operative)?

Relationships

Language is used to create relationships we want to have with our hearers. Gee gives the example of a committee chair saying "Prof. Smith I'm very sorry to have to move us on to the next agenda item" versus "Ed, it's time to move on." The former signals a very formal and deferential relationship with this person, while the second demonstrates (and creates) an informal relationship.

Discourse Analysis Question: What sort of relationship or relationships is this piece of language seeking to enact with others (present or not)?

Politics (the distribution of social goods)

Language is used to create perspectives on social goods. Gee names guilt, blame, legal responsibility, correctness, normalcy etc. as types of social goods created through language. This building task is the task most closely related to the question of ethics that I am posing to the selected documents. People use language to demonstrate their perspective on social goods and establish a "political" stance for themselves. "Political" here does not mean "involved in politics" but something more like "persuasion." For example, if I say "Joan hit Billy in the head," I am showing a perspective where Joan is at fault for hitting Billy. However, if I say "Billy was hit in the head" it doesn't at all include Joan as an actor and in doing so absolves her from responsibility in this situation.

Discourse Analysis Question: What perspective on social goods is this piece of language communicating (i.e., what is being communicated as to what is taken to be "normal", "right", "good", "correct", "proper", "appropriate", "valuable", "the ways things are", "the way things ought to be", "high status or low status", "like me or not like me", and so forth)?

Connections

Language is used to build connections or relevance to certain things. This will be of particular interest for the historical situation of these documents. How is this document building connections between players in the industry? How is this document building a connection to an idea? Taken a little more broadly, how is this document connected to other documents? What are the documents surrounding this document, both topically and also sequentially?

Discourse Analysis Question: How does this piece of language connect or disconnect things; how does it make one thing relevant or irrelevant to another?

Sign Systems and Knowledge

Different sign systems include different languages, different varieties, types of specialized language (e.g. air traffic control English), and non-language communication systems.

Discourse Analysis Question: How does this piece of language privilege or disprivilege specific sign systems (e.g., Spanish vs. English, technical language vs. everyday language, words vs. images, words vs. equations, etc.) or different ways of knowing and believing or claims to knowledge and belief?

Critical Discourse Analysis

As a compliment to the type of general discourse analysis inquiry outlined by Gee (2004), Critical Discourse Analysis approaches promulgated by Fairclough (2003) have been used to situate documents in a broader power dynamic. Since this analysis is in essence not only a discursive analysis but also an analysis of ethical stance, an analytic approach was required that could adequately address issues of power and control within a business context. The explicit aims of CDA to expose relations of power and control in discourse (Van Dijk, 1993, 2002) were well suited to this endeavor.

As a very brief introduction, CDA has been at the forefront of interactions between linguists (particularly discourse analysis researchers) and applied attempts to actively change attitudes and power structures in society. As such, it has especially focused on discourse between large powerful entities (governments and companies, mostly) and individuals represented by "the public." For example, a critical discourse analysis of communist Romania's Nicolae Ceauşescu highlights observed links between encoding "deliberate meaning violations" in public speeches and changes in social reality (Ilie, 1998).

Like Gee, Fairclough (2003) asserted that language's primary goal is to help language users create context. The elements of this context are reality, actions and identities (referred to in his work as representation, action and identification) (Fairclough, 2003). Based on these elements, Fairclough proposed a detailed heuristic for approaching CDA questions. Questions of reality, actions and identities dovetail with the attention to historical context that has been adopted for the discourse analyses in this dissertation.

From an overall point of view, the discourse analyses of the selected internal audience TDC documents in this dissertation will focus on creating a context for these documents and assessing the documents within this context. As many contexts as possible will be reviewed to allow for as full a picture as possible. This means considering the documents from the level of language within the documents, from the level of the tobacco industry from a historical perspective, and also in a larger social and ethical context, using principles of CDA.

CHAPTER 4

RESULTS AND ANALYSIS

This chapter reports the findings from statistical, computational and discourse analyses of the TDC. Classes of documents have been tested for occurrence of each of the linguistic indicators of deception and corporate fraud separately. Hypotheses with regard to these outcomes were outlined in chapter 2. Also, the documents ranked as the top, middle and lowest by the vector ranking method have been assessed using discourse analytic principles as well as using historical methods to situate the documents within the history of the tobacco industry. The impact and ramification of findings from the correlation and ranking tests are discussed in Chapter 5: Discussion.

Statistical Comparisons between Groups of Documents

For the TDC, the following independent variables were compared using two-tail t-tests (in order to be able to identify any aberrant patterns): audience affiliation (internal vs. external) and addressee (named vs. corporate) with respect to the six indicators of deception. Source (company of origin) and decade were compared with one-way ANOVAs, since those independent variables have more than two levels. Each group comparison was run for all six dependent variables of interest: adversarial language, allness and superlative terms, group mentality, strategically ambiguous intent, deprofiling agency, and image-managing verbs. To adjust the family-wise error rate for the number of dependant variables (six), the standard significance level of alpha $<.05$ was not used. Instead, to be more conservative, only results

attaining p <0.01 are reported as significant in these group comparisons. Table 3 below reports on simple correlations between the indicators.

Table 3
Correlation Matrix between the Indicators (Pearson)

Linguistic Indicator	*Deprofiling Agency*	*Cognitive Verbs*	*Group mentality*	*Allness and Superlative Language*	*Adversarial Language*	*Strategically ambiguous intent*
Deprofiling agency	**1**	0.022	**-0.082**	**-0.103**	**-0.100**	0.053
Cognitive verbs	0.022	**1**	**0.566**	0.025	0.042	**0.453**
Group mentality	**-0.082**	**0.566**	**1**	0.025	0.053	**0.324**
Allness and Superlative Language	**-0.103**	0.025	0.025	**1**	**0.113**	-0.028
Adversarial Language	**-0.100**	0.042	0.053	**0.113**	**1**	0.058
Strategically ambiguous intent	0.053	**0.453**	**0.324**	-0.028	0.058	**1**

Values in bold are significantly different from 0 with a significance level alpha=0.01

Results of H1

H1: *Compared with documents addressed to audiences internal to tobacco companies, documents addressed to external audiences will exhibit higher relative frequencies of each of the following linguistic indicators of deception: markers of deprofiled agency, strategically*

ambiguous intent markers, adversarial vocabulary, image management verbs, allness markers, group mentality markers.

T-tests were used to assess the impact of audience on linguistic indicators of deception. As a whole, this hypothesis is not supported by the evidence.

The means by audience (internal/ external) of the six potential indicators of deception are presented in Table 4.

Table 4
Means by Audience of the Six Potential Indicators of Deception

Linguistic Indicator	*Internal Audience*	*External Audience*
Adversarial language	0.0017	0.0019
Strategically ambiguous intent	0.0957	0.0342
Group mentality	0.5388	0.2203
Deprofiling Agency	0.1681	0.1177
Image management verbs	0.3406	0.1332
Allness and Superlative Language	0.046	0.048

The results from the t-test of audience (internal / external) effects are presented in Table 5.

Table 5
T-Tests of Audience Effects on Linguistic Indicators of Deception

Linguistic Indicator	*DF*	*Ave Diff*	*SD*	*T Value*	*Pr > |t|*
Adversarial language	1106	0.0002	0.0033	0.63	0.5293
Strategically ambiguous intent	1106	0.062	0.2493	2.98	0. 0029**
Group mentality	1106	0.319	2.0736	1.86	0.0638
Deprofiling Agency	1106	0.05	0.1056	-5.75	<.0001**
Image management verbs	1106	0.207	0.7639	-3.28	0.0011**
Allness and Superlative Language	1106	0.0002	0.0073	0.35	0.7272

**p <0.01

(a) Deprofiling Agency

Counter to the predicted result, internal audience documents had more markers of deprofiling agency (e.g., passive structures) per total phrases than external documents. The mean number of passive structures per total phrases for internal audience documents was 0.1681 (sd ±0.111), while the mean for external audience documents was 0.1177 (sd ±0.0699). The analysis yielded a statistically significant, but unexpected result.

(b) Strategically Ambiguous Intent Markers

Counter to the predicted result, internal audience documents had significantly *more* strategically ambiguous intent markers per total words than external documents. The mean number of strategically ambiguous intent markers per total words for internal audience documents was 0.0957 (sd ±0.2704), while the mean for external audience documents was 0.0342 (sd ±0.0521). The t-test result demonstrated a significant difference between external and internal audience documents for this factor (p<.0001), but not in the anticipated direction.

(c) Adversarial Vocabulary

The t-test analyzing results of adversarial language distribution between internal and external documents did not yield a statistically significant result. Thus, the null hypothesis could not be rejected.

(d) Image-management Verbs

Counter to the predicted result, internal audience documents had significantly *more* stative verbs per total verbs than external documents. The mean number of image management verbs per total verbs for internal audience documents was 0.3406 (sd ±0.8307) while the mean for external audience documents was 0.1332 (sd ±0.0786). Hypothesis H1, in contrast, had predicted that documents directed to external audiences would have a higher incidence of image management verbs relative to documents directed to internal audiences.

(e) Allness and Superlative Language

The t-test analyzing results of allness language distribution between internal and external documents did not yield a statistically significant result. Thus, the null hypothesis could not be rejected.

(f) Group Mentality

The t-test analyzing results of group mentality language distribution between internal and external documents did not yield a statistically significant result. Thus, the null hypothesis could not be rejected.

Inspection of group means indicates that documents addressed to internal audiences exceeded values of documents directed to external audiences for indicators of ambiguous intent,

for deprofiling agency, and for image management verbs. None of the other three linguistic indicators manifested a significant difference between documents directed to internal versus external audiences.

Addressee Status

The means by addressee status of the six potential indicators of deception are presented in Table 6.

Table 6

Means by Addressee of the Six Potential Indicators of Deception

Linguistic Indicator	*Named Addressee*	*Corporate Addressee*
Adversarial language	0.0018	0.0017
Strategically ambiguous intent	0.0762	0.094
Group mentality	0.4762	0.4993
Deprofiling Agency	0.1487	0.1694
Image management verbs	0.2779	0.3323
Allness and Superlative Language	0.0045	0.0047

The results of the six t-tests of addressee status (*named* / *corporate*) effects are presented in Table 7. Inspection of group means indicates that documents with corporate addressees, or recipients, had significantly more deprofiling agency (mainly passive) structures than documents addressed to a specific person (i.e., a *named* addressee). No other linguistic indicator manifested a significant difference between documents directed to *named* versus *corporate* addressees.

Table 7

T-Tests of Addressee Effects on Linguistic Indicators of Deception

Linguistic Indicator	*DF*	*Ave Diff*	*SD*	*T Value*	*Pr > |t|*
Adversarial language	1106	0.0001	0.0033	0.42	0.6766
Strategically ambiguous intent	1106	0.018	0.2501	1.24	0.2382
Group mentality	1106	0.023	2.0768	0.18	0.8546
Deprofiling Agency	1106	0.021	0.1067	3.20	0.0014*
Image management verbs	1106	0.054	0.7671	1.17	0.2414
Allness and Superlative Language	1106	<0.0002	0.0073	0.37	0.7136

* p <0.01

Results for Hypotheses H2

H2: Linguistic indicators of deception should vary significantly by source, and Tobacco Institute and CTR documents in particular should stand out as having higher relative frequencies of the linguistic indicators of deception.

Results of the means of the six linguistic indicators of deception by source are presented in Table 8.

Table 8
Means by Industry Source of the Six Potential Indicators of Deception

Linguistic Indicator:	*Image management verbs*	*Adversarial Language*	*Deprofiling Agency*	*Group Mentality*	*Allness and Superlative Language*	*Strategically Ambiguous Intent*
1. American Tobacco Company	0.22646099	0.00118236	0.1778911	0.62988779	0.00444749	0.05678858
2. Brown and Williamson	0.35659243	0.00131911	0.17024792	0.34797835	0.00468144	0.13780807
3. Center for Tobacco Research	0.37768039	0.00077171	0.16736023	0.57732234	0.00578101	0.0732263
4. Lorillard	0.26809672	0.00144681	0.1614616	0.2312131	0.00430322	0.10166696
5. Philip Morris	0.29588398	0.0017676	0.1568612	0.59770538	0.00448242	0.07644187
6. RJ Reynolds	0.28640067	0.00208623	0.16848753	0.42551209	0.00439379	0.09249172
7. Tobacco Institute	0.43673446	0.00238673	0.11883834	0.50280296	0.00547829	0.05875189

Results of the ANOVA of document source differences among the six linguistic indicators of deception are presented in Table 9.

Table 9
One-way ANOVAs of Effects of Industry Source on Six Linguistic Indicators of Deception

Linguistic Indicator	*DF*	*F Value*	*Pr > F*
Adversarial language	6	2.55	0.018*
Strategically ambiguous intent	6	1.42	0.2034
Group mentality	6	0.64	0.7015
Deprofiling Agency	6	3.36	0.0028**
Image management verbs	6	0.84	0.54
Allness and Superlative Language	6	0.51	0.799

*p<0.05; **p <0.01

Post hoc pair-wise comparisons of industry source cell means utilized the Scheffé procedure with a 95% confidence interval or the LSD procedure with a 95% confidence interval. The Scheffé procedure indicated that for deprofiled agency, the mean for American Tobacco Company (mean =0.178, sd ± 0.110) and RJ Reynolds (mean =0.168, sd ± 0.107) significantly exceeded TI documents (mean = 0.119, sd ± 0.062). The LSD procedure indicated that for adversarial language, the mean for RJ Reynolds (mean =0.0023, sd ± 0.004) and the Tobacco Institute (mean =0.0023, sd ± 0.003) significantly exceeded the mean for the American Tobacco Company (mean =0.0011, sd ± 0.002), Brown and Williamson (mean =0.001, sd ± 0.003), and the Center for Tobacco Research (mean =0.0007, sd ± 0.002). Additionally, the mean for the and the Tobacco Institute (mean =0.0023, sd ± 0.003) significantly exceeded the mean of Lorillard

(mean =0.0014, sd ± 0.003) for adversarial language. No other pair-wise comparisons were statistically significant.

Decade

Results of the means of the six linguistic indicators of deception by decade are presented in Table 10. Decade in the TDC was defined as one of the following for each document: 1950, 1960, 1970, 1980, or 1990. In addition, two non-standard "decades" were recorded. 19xx includes documents without a verifiable date, and Bliley documents were documents from the Bliley collection. These "decades" have been included in analysis because otherwise these documents would have been excluded from this measurement altogether.

Table 10
Means by Decade of the Six Potential Indicators of Deception

Linguistic Indicator:	*Image management verbs*	*Adversarial Language*	*Deprofiling Agency*	*Group Mentality*	*Allness and Superlative Language*	*Strategically Ambiguous Intent*
1950	0.20033071	0.00038267	0.13790005	0.29142092	0.00996326	0.05098744
1960	0.49499187	0.00114256	0.17265267	0.90375549	0.00577834	0.04983807
1970	0.35004667	0.0017189	0.16289894	0.40070035	0.00470565	0.08931661
1980	0.2841245	0.00137931	0.18195378	0.35835314	0.00436593	0.09287508
1990	0.28080435	0.00228628	0.14059376	0.59922695	0.00398197	0.08852265
19xx	0.34801536	0.00357848	0.11794244	0.34547786	0.0039322	0.0806726
Bliley	0.17482488	0.00133829	0.14654971	0.26604017	0.00597922	0.10435859

The results from the ANOVA of decade effects are presented in Table 11.

Table 11
One-way ANOVAs of Effects of Decade on the Six Potential Linguistic Indicators of Deception

Linguistic Indicator	*DF*	*F Value*	*Pr > F*
Adversarial language	6	4.96	<.0001*
Strategically ambiguous intent	6	0.05	0.8077
Group mentality	6	1.20	0.3041
Deprofiling Agency	6	5.78	<.0001*
Image management verbs	6	1.40	0.2126
Allness and Superlative Language	6	3.96	0.0006*

* p <0.01

A post-hoc Scheffé procedure indicated that adversarial language increased significantly from the 1980s (mean = 0.001, sd ± 0.002) to 1990s (mean = 0.002, sd ± 0.004). On the other hand, a Sheffé procedure comparing each decade on relative frequency of allness language indicated that the mean for 1950s documents (mean = 0.010, sd ± 0.018) was significantly higher than that of 1970s (mean = 0.005, sd ± 0.006), 1980s (mean = 0.004, sd ± 0.006) and 1990s documents (mean = 0.004, sd ± 0.007). This procedure also indicated that for language signaling deprofiled agency, the mean for 1980s document (mean =0.182, sd ± 0.119) was significantly higher than 1990s documents (mean =0.141, sd ± 0.091). No other pair-wise comparisons were statistically significant.

Ranking the TDC Internal Documents for Deceptiveness

The averages of the six indicators were calculated for the external TDC documents. For the external TDC documents, the average 6-dimension vector representing linguistic deception was: [0.117737053,0.133239,0.220337,0.00478,0.001892,0.034159134].

I used the average external document vector as the average vector representing quintessential "deceptiveness." Previous research has established that tobacco industry documents with external audiences are deceptive, fraudulent and misinformative (Pollay, 1989, 1990, 1994, 1995; Pollay & Dewhirst, 2002; Ringold, 1986; Rubin et al., 2002; Shuy, 2003a; Wulinger, 1984). The internal documents, by contrast, should range from relatively straightforward communications necessary for day-to-day business to documents that are integral to extensive public deception and fraudulent activities. Where particular documents emerge as potentially highly deceptive from among this set of presumptively overall low deception documents—i.e., from among the internal-audience documents—then such a document is likely to be extraordinary indeed.

The 933 internal documents in the TDC were ranked according to the angle of their vector from the average external document vector. The five documents with the smallest angles (the closest to external documents) were presumably the most engaged in deceptive corporate strategy. These five plus the middle four and furthest five documents (the documents most orthogonal to deceptive corporate strategy) were analyzed more closely using discourse analysis techniques outlined in Chapter 3.

Results of the (Critical) Discourse Analysis

The following sections focus on these industry-internal audience documents in depth and approach them by how they were ranked within the internal documents of the TDC. Within each analysis, references to the indicators of deception (*cognitive verbs, group mentality, adversarial language, deprofiled agency, strategically ambiguous intent* and *allness*), Gee's (2005) list of seven discourse analytic tasks of language (*significance, activities, identities, relationships, politics* or *social goods, connections* and *sign systems*) and Fairclough's critical discourse analysis areas of interest (*power, reality, actions* and *identities*) are italicized for easier reference. Excerpts from the documents have been included in these results. In addition, full images of the fourteen discourse samples are available in Appendix B.

Least Deceptive Documents

The group of presumptively least deceptive documents covers two main document genres: letters and reports. These five documents represent reports and communication about information necessary to the day-to-day business endeavors of the tobacco industry. One (Bates Number *1003390563)*, however, does stand out as having more deceptive corporate strategy.

1) Appeal ISO; Bates number: 2028456398; Angle = 86.88°; (H. W. Gaisch, 1986a).

This is a cover letter for a final draft of an appeal to the Committee on Tobacco and Tobacco Products, which is Technical Committee 126 of the International Standards Organization (ISO/ TC 126). This appeal ("Against the Approval by ISO/TC 126 of the Draft International Standards ISO/DIS 4387 and ISO/DIS 8453") protests ISO standards rulings on the validity of machine-smoking.

This document is only a short letter (one page). It is a straightforward introduction of the Appeal cited above, which is being transmitted for reference or approval to another member of Philip Morris Europe. It has a familiar tone of address: "kind regards and best wishes." This familiar tone, from a discourse analytic point of view, could be considered to be creating a *significance* of familiarity in this document where the *relationship* between addressee and addressor is convivial. This intimated closeness of the salutation is in contrast to the rest of the document that seems otherwise distant and perhaps even cold. There is no direct mention at all of the addressee in the body of the letter. There is only really one explicit *relationship* established from a discourse analytic point of view, which is the group of "Stefano Sandri, Carlo Giarré, Bradley Brooks, Francesco Lopes and myself." Interestingly, though, this document does not include mentions of unspecified groups that might have trigger the *group mentality* indicator. This suggests that although there are group *identities* acknowledged in this document - "Stefano Sandri, Carlo Giarré, Bradley Brooks, Francesco Lopes and myself" and "the Italians" – there is no *group mentality* in the way that might identify this document as participating in deceptive corporate strategy. This *participant structure* includes the *relationship* established between the author and the recipient(s). However, there are other *participants* referenced that are suggested to be involved with the *activity* of reviewing the draft. The groups mentioned, rather than being constructed as part of a deceptive ploy to distribute responsibility, create the *participant structure* of the letter, and perhaps of the activity itself.

Additionally, this letter lacks explicit information. The *connections* established in this document consistently reference items not delineated in the text. For instance, "the meeting with Monital yesterday morning" is not more fully explained; anyone who did not know what the

meeting with Monital was about would be lost in this conversation. The author also uses "the Italians" instead of a more explicitly precise name to refer to the group that the enclosed draft will be presented to. This establishes their *identity* with a mass noun instead of any name or names that would be more specific. If this letter were addressed to someone less "in-the-know," the author might have to specify not only who the Italians are, but even in the instance of a courier perhaps, the exact address where one could find them. In the end, the familiar tone and lack of explicit information demonstrate that the *connections* between the draft of the Appeal and the addressee, the Italians and the author should be known to everyone involved in this communication chain.

The following attachment of this document (which was NOT ranked by this research's algorithms) is a draft of the ISO appeal and is highly technical, using *sign systems* that are more technically specific compared to the rather everyday English used in the letter. Despite differences in *sign systems* where the letter uses everyday standard written English and the appeal uses technical language, both the short letter (our document of interest) and the following attachment use language that excludes the general public. However, this letter was not addressed to the general public. For the audience it was addressed to, the vocabulary (*sign system*) and implicit information (less explicit *connections*) in this letter would have been straightforward and necessarily specific to the subject matter and the primary intentions of straightforward communication.

To situate this document in a larger, *power*-driven context, á la Fairclough, it is necessary to review other documents related to this letter. Interestingly enough, situated historically, this document represents the tip of an industry iceberg. A storm was raging between British

American Tobacco and Philip Morris over “channel ventilated” cigarettes, cigarettes which presumably reduced smokers’ toxic intake by mixing ambient air with tobacco smoke. This conflict was fought over five years and two continents: the debate raged from the ISO 1985 meeting in Paris ("Draft Report of the Thirteenth Plenary Meeting Held in Paris on 850328 and 850329," 1985) to Turkey in 1986 (H. W. Gaisch, 1986b; Lopes, 1986) and then to China in 1988. What was all the fuss about?

A memo from a Philip Morris researcher which outlines the events of the Turkey ISO meeting reports the following about the ISO Standard (ISO/DIS 4387) in question:

> “ISO/DIS 4387 ‘Cigarette Determination of Total and Dry Particulate Matter Using a Routine Analytical Cigarette-Smoking Machine – Glass Fibre Smoke Trap Method’ was developed over some ten years and was approved…subject to the Working Group (WG) 6 finding a preamble concerning the applicability of the testing method.” (Gaisch, 1986b)

The applicability of the inhalation testing method was the base issue. A decision excluding certain types of cigarettes from legitimately being tested for tar using one method would have had huge impacts on that cigarette, as well as ripple effects through the industry and market. Testing for tar and other additives had been the strategy of the tobacco industry to assuage fears surrounding cancer and tobacco use (Bialous & Yach, 2001). Tobacco companies were trying to keep down smoke health risks (or at least to create the illusion that they were reducing risk) by using filters. However, some filters actually had holes in them that allowed

fresh air to pass through into the cigarette-smoking machine. Cigarettes that used these filters, which were *channel* ventilation filters instead of regular ventilation, registered in cigarette-smoking machines as being on-the-whole lower tar than regular ventilation filters. Classifying all cigarettes as being part of the same bucket in this ISO Standard would have the final impact of positioning channel ventilation filtered cigarettes in the market as the best option for a health and tar-conscious smoker (Bialous & Yach, 2001).

Although this document was addressed to an industry-internal audience, it is discussing plans to hash out tobacco industry interests in a supposedly non-tobacco realm. A Philip Morris document describes the ISO as an "[organization] controlled by the industry" (Boder, 1990). The ISO Technical Committee 126 was actually one of two controlled organizations. The other, CORESTA, was an organization of tobacco researchers aiming for a common goal of advancing tobacco industry interests around the world. In tandem, these organizations monopolized standards for the tobacco industry. Tobacco control research on the ISO TC 126 reports:

> "Industry participation in the development of ISO standards is not exclusive to tobacco, but unlike other products, such as screws and credit cards, the determination of standards by the industry, without the participation of other interested parties, has lead to the development of standards that protect the political and commercial interests of the industry rather than those of the consumer. In the case of ISO technical committee 126tobacco and tobacco products standards (ISO TC/126, established in 1968), the standards are developed in fact by the Cooperation Centre for

> Scientific Research Relative to Tobacco (CORESTA).” (Bialous & Yach, 2001)

Accordingly, it is worth noting that this document’s *participant structure* involves not only the groups that are a part of the *activity* of reviewing the draft standards, but that external connections place the document (and its participants) within a context that could extend internationally. The document in question is a one-page cover document that is not deceptive, but sits at the epicenter of an international struggle. This document demonstrates that even absolutely innocuous language use within the tobacco industry can be related to impactful tobacco industry strategies. However, this document does not support the deceptive agenda of the tobacco industry in a direct and meaningful way. Additionally, there are no instances of any of the *indicators of deception* in this document at all. This document’s very short length sets up a situation where it would have been less likely for any one indicator to be present. As a whole, for this document the lack of indicators coincides with a non-deceptive strategy document.

2) Telex: Dear Bob; Bates number: 1003390563; (angle 89.04°) ("--No Title," 1980)

This document is a one-page letter to Dr. Robert Seligmann, “Bob” - vice president, research & development, PM Incorporated, 1975–1980. It discusses changes in the denitration process, the main *activity* referenced in this document. At first blush, this letter is straight-forward and not deceptive nor contributing directly to the deceptive agenda of the tobacco industry. There are inklings, however, of other *activities* that could be of interest. These extra-text *activities* are intimated as part of the web of *connections* established with denitration. Denitration, within this document is *connected* with defensive research, “sidestream smoke,” “loss of share[s],” and finally “embarrassment.” Additionally, the author establishes this

exchange in the context of an ongoing working *relationship* by recalling a previous conversation: "I would like to recall part of our long-term research concept…" The denitration process removes "smoke constituents from mainstream and sidestream" cigarette smoke. The author reminds Dr. Seligmann that the research surrounding this process is "of defensive nature." Darkly, the author alludes to some greater problem that this study may be addressing: "The whole concept…might give rise to acute embarrassment and loss of market share…" This reference to "acute embarrassment" is an overt example of language use being *political* (according to Gee's (2005) definition of *politics* as addressing what is being communicated as "right," "good" or "correct"). The reference to embarrassment suggests some *political* social violation on the part of the tobacco industry here. Why else would the author be concerned about embarrassment other than if he understood that the social norms would make some behavior (or lack of behavior) inappropriate.

The author explicitly acknowledges *not* discussing the specifics of the research in this format: "I kept this telex deliberately less explicit." In doing so, he highlights the social *significance* of the material in the telex, which to untrained eyes might seem extremely uninteresting and technical. In the same way that Gee (2005) talks about speaking in a whisper as establishing the *significance* of information remaining secret, this line establishes the delicacy and sensitivity of topics discussed. This elliptical treatment of the topic again feeds in to the *political* idea that the things being discussed in this telex are things that are potentially detrimental (either for the public, the industry, or both) and would be perceived socially as improper.

Additionally, although this document was intended to be an industry-proprietary document, from a Critical Discourse Analysis point of view, there is a player indirectly referenced in this document. References to being "less explicit" and leakage of certain information causing "acute embarrassment" suggests that there is another phantom addressee, the public. In this instance the language of the document itself bars the public from full access. These inexplicit references serve to establish a *relationship* with the public that is both explicitly and implicitly stilted and impoverished.

In general, there are constant sideways references within the document to amorphous external entities. These references to embarrassment, defense and secrecy creates a *participant structure* that I would suggest is actually *not* centered around the two obvious participants of author and addressee. This document's *participant framework* in some ways actually focuses on an instantiation of the social goods as a silent participant in this document. Is this social goods participant a public audience that has a different moral structure than the author and recipient of this document? Or is the phantom participant a monitor from some governmental agency? Perhaps, even, this phantom participant is an instantiation of an inner morality of the author, which is at odds with his *actions* referenced in this document.

Explicitly, the public (or this phantom *participant*) is cut out of this conversation. Additionally (and perhaps unbeknownst to the author), the public is also implicitly cut out of the conversation by the technical *sign system* which is used throughout the telex. The *sign system* in this document is a mix of standard written British English, and more technical language. The coherent intermix of these two systems suggests that the author did not perceive the technical *sign system* to exclude anyone. The author must have been comfortable enough with this

technical language to have determined that in order to exclude other relationship (such as those with the public) he would have to additionally leave out key information.

Also of interest is the name of the research that the document refers to in this project: "Project Protagoras." Protagoras is a particularly interesting choice here because he was a key proponent of agnosticism. "I don't know" or "we don't believe so" was part of the standard Philip Morris defense for decades. In addition, however, Protagoras is also known as a proponent of strengthening the weakest argument. Using the name Protagoras gives this project an uncertain *identity* as a project determined to support the agnostic stance of the tobacco industry.

From a linguistic indicator of deception standpoint, this document should not be involved in deceptive corporate strategy. There are no cognitive verbs. There is no explicit adversarial language. The agency in this document is direct and profiled. For example, "…the method which we have developed" (italics mine) can be attributed to specific responsible individuals. Also, the author outlines his specific agency in crafting the letter: "I am writing this to you…" There is no *ambiguity of intent*, and no *allness*. However, despite the lack of linguistic indicators in this document, it seems to be supporting a deceptive corporate strategy of non-communication and opacity with the public. This document even seems to be operating in the vein of deception by omission, instead of the other types of more easily accessed deception. This telex indicates a different participation framework than it encodes linguistically. This indicated framework includes a public or other audience that is never directly addressed or mentioned, but that continues to drive the linguistic choices of the document and the cautious choice of words of the author.

3) Ogilvy & Mather Insertion Order; Bates number 689478896/8897;

http://legacy.library.ucsf.edu/tid/gra60f00 (angle 89.04°)

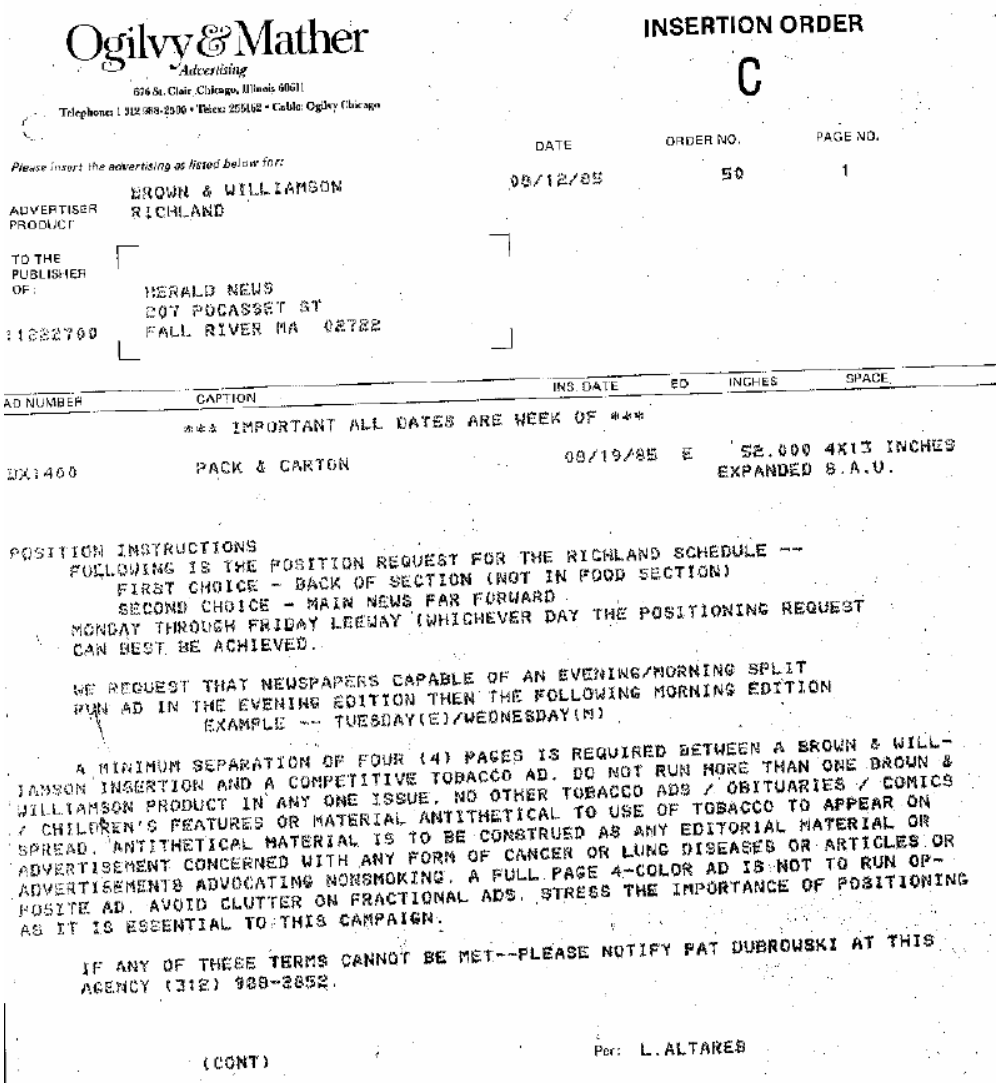
This document is an Insertion Order enumerating specifications for printing a Brown & Williamson advertisement. This document is virtually identical to 3023 other insertion orders in the larger universe of Brown & Williamson's previously secret documents. The change orders range in date from 1979-1998. These forms were part of quotidian business between Ogilvy & Mather (Brown & Williamson's advertising agency) and publication entities running B&W advertisements. In fact, these types of orders are so common that "Advertising Insertion Orders" would be found outside of the context of the tobacco industry altogether, in the business dealings of most large consumer product corporations. Such formulaic advertising insertion orders are used across the print and advertising industry to specify printing orders for advertisements. To illustrate their ubiquity, a quick survey of results of a search for "insertion order" using the Google search engine online produced four insertion order forms within the first ten documents.

As is the case with many formulaic document genres, this insertion order *participates* in a structure that has been pre-tested and pre-formed. The *participant structure* of this document, due to its formulaic content, was codified even before it was actually instantiated as a work-product document. The form-ness of this document effectively solidifies the *participants* in the *structure* into acting as roles instead of acting as individual people. Certain roles within B & W and Ogilvy & Mather use this form to communicate, and by using this document type, shorthand some of the communication and double-checking that might otherwise be necessary.

Part of this need for shorthanded communication may be due to the *activity* initiated by this document. The *activity*, inserting advertisements into print, is very important to the success

of B & W, but is not something that they can do themselves. B & W as a collection of individuals, in order to fulfill their corporate goal of profit, has to "communicate" with Ogilvy & Mather. Because the people handling this communication can change at any time, and because the communication is so very key to the success of the company, this document has been extracted from a *participant framework* that would have supported one-on-one communications between individuals. Instead, this document has become a vehicle for information exchange without the typical interpersonal *participant structure* that another document in a letter-based genre might have.

This insertion order (like all insertion orders) is a list of specifications for printing an advertisement for B&W. This document is not written with deceptive corporate strategy or intent to mislead or defraud the public. Although advertising within the tobacco industry – obviously directed to industry-external audiences (i.e., tobacco consumers or potential consumers) – has been seen to be deceptive and misleading (Cummings, Morley, Horan et al., 2002; Cummings, Morley, & Hyland, 2002; Cummings & Pollay, 2002; Harris et al., 2006; John, 1989; Wakefield, Morley, Horan, & Cummings, 2002), this industry-internal order for creating advertisements is the most rote and least strategically deceptive of the documents so far. This document, shown in figure 3 below, simply contains specific instructions for the placement of an ad. The letter's primary mode of communication is informative, not strategic.



Ogilvy & Mather
Advertising
636 St. Clair, Chicago, Illinois 60611
Telephone: 312/988-2500 • Telex: 255102 • Cable: Ogilvy Chicago

INSERTION ORDER

C

DATE 08/12/85 | ORDER NO. 50 | PAGE NO. 1

Please insert the advertising as listed below for:

ADVERTISER: BROWN & WILLIAMSON
PRODUCT: RICHLAND

TO THE PUBLISHER OF:
HERALD NEWS
207 POCASSET ST
FALL RIVER MA 02722
11032700

AD NUMBER	CAPTION	INS DATE	ED	INCHES	SPACE
	*** IMPORTANT ALL DATES ARE WEEK OF ***				
DX1460	PACK & CARTON	08/19/85	E	52.000	4X13 INCHES EXPANDED S.A.U.

POSITION INSTRUCTIONS
FOLLOWING IS THE POSITION REQUEST FOR THE RICHLAND SCHEDULE --
FIRST CHOICE - BACK OF SECTION (NOT IN FOOD SECTION)
SECOND CHOICE - MAIN NEWS FAR FORWARD
MONDAY THROUGH FRIDAY LEEWAY (WHICHEVER DAY THE POSITIONING REQUEST CAN BEST BE ACHIEVED.

WE REQUEST THAT NEWSPAPERS CAPABLE OF AN EVENING/MORNING SPLIT RUN AD IN THE EVENING EDITION THEN THE FOLLOWING MORNING EDITION
EXAMPLE -- TUESDAY(E)/WEDNESDAY(M)

A MINIMUM SEPARATION OF FOUR (4) PAGES IS REQUIRED BETWEEN A BROWN & WILLIAMSON INSERTION AND A COMPETITIVE TOBACCO AD. DO NOT RUN MORE THAN ONE BROWN & WILLIAMSON PRODUCT IN ANY ONE ISSUE. NO OTHER TOBACCO ADS / OBITUARIES / COMICS / CHILDREN'S FEATURES OR MATERIAL ANTITHETICAL TO USE OF TOBACCO TO APPEAR ON SPREAD. ANTITHETICAL MATERIAL IS TO BE CONSTRUED AS ANY EDITORIAL MATERIAL OR ADVERTISEMENT CONCERNED WITH ANY FORM OF CANCER OR LUNG DISEASES OR ARTICLES OR ADVERTISEMENTS ADVOCATING NONSMOKING. A FULL PAGE 4-COLOR AD IS NOT TO RUN OPPOSITE AD. AVOID CLUTTER ON FRACTIONAL ADS. STRESS THE IMPORTANCE OF POSITIONING AS IT IS ESSENTIAL TO THIS CAMPAIGN.

IF ANY OF THESE TERMS CANNOT BE MET--PLEASE NOTIFY PAT DUBROWSKI AT THIS AGENCY (312) 988-3852.

(CONT)

Per: L.ALTARES

Figure 3. Bates number: *689478896*; Title: Insertion Order C: B&W Richland

4) Monthly Returned Goods Report; ATX03_0025429

http://legacy.library.ucsf.edu/tid/oji05f00 (angle 83.12°) (Leake, 1982)

This document is a Monthly Returned Goods Report from the Assistant Research and Development Director of American Tobacco (P.H. Leake) to their Executive Vice President (R.P. Truitt). Like the insertion order, this document is formulaic, using a formulaic *sign system*. In the entire universe of tobacco industry documents, there are 80 of these reports that vary to some extent, but have very similar information structures.

This report is entirely impersonal. As seen in the last document, due to the formulaic content, this document's *participant structure* at root does not actually involve individuals, but instead is a vehicle for communication between specific *roles* within the company. For example, although the document was addressed to a single individual as is typical of personal communications, its carbon copy (cc) list of people goes on for two pages. This report exists in the form of a memo, but it is in no way a personal correspondence. Instead, the report is mostly rote. From this monthly returned goods report and three other monthly returned goods reports found in the online archive, we get almost all the information in the form of *cigarettes* + **amount returned** + comparison to another period: "… *cigarettes* **returned** during December is approximately 5.0 million more than reported for November" (Leake, 1986), "Year to date results are 4.09% compared to 2.06% in 1992" (Goodridge, 1993), "Monthly **returns** are up slightly compared to the last three periods" (Goodridge, 1994), "… cigarettes had been **returned**. This is about 16.7 million more than… returned during the first five months of 1980" (Leake, 1986).

In addition, there are no references to people (individuals or *groups*) in the discourse of this document. There is no *group mentality* evidenced. The *identities* and *relationships* established from a discourses analyst point of view in this document seem to be those of set roles instead of individuals. The only references to individuals in this document are in the "TO" and "FROM" boxes. Additionally, the individuals themselves are named more fully as their title or role than their given name: "Mr. R. P. Truitt, Executive Vice President" and "P. H. Leake, Assistant Research and Development Director." The *identities* of these individuals in this document are firmly created as formal roles. "Executive Vice President" is not "Executive VP" and "Assistant Research and Development Directors" is not "Assistant R&D Director."

The focus on creating the *identities* of individuals as merely roles ties in to the focus of the entire document. The title of the document ("Monthly Returned Goods Report") first signals that the focus or *significance* of the document is the returned goods. This focus seems pretty obvious, but on the other hand, this information could have been conveyed in a fashion that more clearly focused on retailers and wholesalers as individuals, rather than just on the inanimate product. For instance, instead of reporting on a massive "total returns," the author could have focused more on people, and reported on "returned goods per individuals buying goods" or "returned goods per individuals cataloging goods." Keeping with this non-human, *non-agentive* focus, returned goods are fore-fronted not only in the title, but also in sentence structure throughout the document. Although it was posited in chapter 2 that a preponderance of agentless sentences would be associated with deceptive documents due to *deprofiled agency*, it also seems that this document is an example of agentless sentences being associated with formalized information.

The result is a focus on inanimate entities (the returned goods). For example, "cigarettes had been returned" is a common phrase in these monthly returned goods reports. In effect, using agentless sentences across the board formalizes distance from individuals and people altogether in this document and obscures the responsibility for defective goods. In this instance, avoiding pinning responsibility on an individual or group is at least standard considering the parameters of a report genre. A certain amount of a good is bound to be returned for one reason or another in a large-scale corporate operation. These Monthly Returned Goods Reports would only be of real interest if analysts detected a great and sudden change in goods returned from one month to another.

5) Report of Telephone Conversation with Client 680112152 (Angle 84.35°) (Heller, 1976)

http://legacy.library.ucsf.edu/tid/vvb50f00

This document is a report of a telephone conversation between staff of Ted Bates & Company advertising agency and their client, Brown & Williamson Tobacco Corporation. This conversation in May of 1976 was between Michael J. McCue, Marketing Department Brand Assistant for Viceroy at the time (Source: B&W's Initial Disclosure, State of Texas vs. ATC, et al., 6/5/96), and Arthur Heller, Ted Bates & Company Senior V.P. and Account Director (http://tobaccodocuments.org/profiles/people/heller_art.html). Brown & Williamson had asked Ted Bates & Co. to "determine most efficient manner of including new April 1976 FTC Tar and Nicotine numbers on promotional materials" (Heller, 1976).

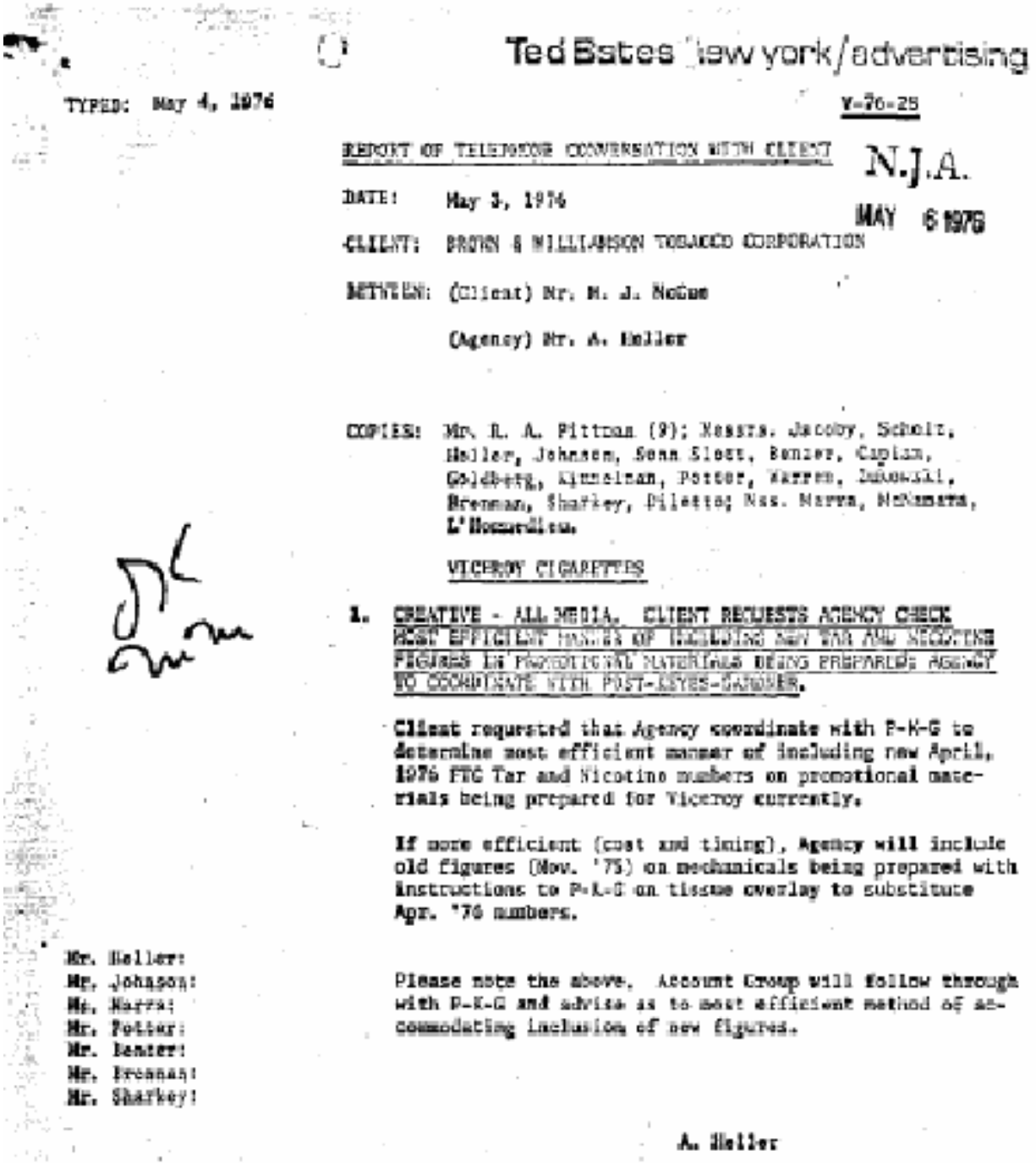
This document is a clear request for the agency to help Brown & Williamson comply with new FTC Tar and Nicotine number regulation. Discursively, it establishes Brown &

Williamson as the client in a *relational* position to make requests. Within Ted Bates & Co. the Account Group "will follow through…" with the requests.

From a historical perspective, regulations on tobacco advertising started to ramp up in the early 1970s. In 1970, a Congressional bill introduced a warning to be attached to cigarette packages: "Warning: Cigarette Smoking is Dangerous to Your Health." Subsequent to that, disclosing amounts of tar was required on cigarette packaging and advertising.

Additionally, this document *connects* the client's request to the new "Apr. '76 [tar and nicotine level] numbers." This establishes a *connection* between requests and the new numbers. By making this connection, the document is thrown into a framework that contains participants far outside the immediate scope of the document. B & W and Ted Bates & Co.'s Account Group form an immediate *participation structure* that situates the activity primarily between the two companies. However, this document is actually situated in a larger *participant structure* that includes the entities that have created the new tar and nicotine level number regulation.

In general, this document does not seem to have deceptive corporate strategy or be supporting corporate fraud. To the contrary, it is asking for help complying with external cigarette advertising regulations. This particular report, displayed in figure 4, is of an interaction that does not support the deceptive agenda of the tobacco industry, but rather actually leads to increased disclosure and transparency.



Ted Bates New york/advertising

TYPED: May 4, 1976

Y-76-25

REPORT OF TELEPHONE CONVERSATION WITH CLIENT

N.J.A.

MAY 6 1976

DATE: May 3, 1976

CLIENT: BROWN & WILLIAMSON TOBACCO CORPORATION

BETWEEN: (Client) Mr. M. J. McCue

(Agency) Mr. A. Heller

COPIES: Mr. R. A. Pittman (9); Messrs. Jacoby, Scholt, Heller, Johnson, Sonn Sloot, Benzer, Caplan, Goldberg, Kimmelman, Potter, Warren, Jakowski, Brennan, Sharkey, Pilette; Mss. Marra, McNamara, L'Hommedieu.

VICEROY CIGARETTES

1. CREATIVE - ALL MEDIA. CLIENT REQUESTS AGENCY CHECK MOST EFFICIENT MANNER OF INCLUDING NEW APRIL AND RECENT FIGURES IN PROMOTIONAL MATERIALS BEING PREPARED; AGENCY TO COORDINATE WITH POST-KEYES-GARDNER.

Client requested that Agency coordinate with P-K-G to determine most efficient manner of including new April, 1976 FTC Tar and Nicotine numbers on promotional materials being prepared for Viceroy currently.

If more efficient (cost and timing), Agency will include old figures (Nov. '75) on mechanicals being prepared with instructions to P-K-G on tissue overlay to substitute Apr. '76 numbers.

Please note the above. Account Group will follow through with P-K-G and advise as to most efficient method of accommodating inclusion of new figures.

Mr. Heller:
Mr. Johnson:
Ms. Marra:
Mr. Potter:
Mr. Benzer:
Mr. Brennan:
Mr. Sharkey:

A. Heller

Figure 4. *Report of Telephone Conversation with Client (Heller, 1976)*

Middle Level Documents

6) DS Scan Results on Several Menthol Cigarettes, Bates number B01281539/1545, http://legacy.library.ucsf.edu/tid/xdb11f00 (angle 41.20°)

This document is a scientific report from Brown & Williamson's R&D department (Moldeveanu & Kulshreshtha, 1991). It reports on results of DS (Direct Silyiation) scans comparing six different menthol cigarette brands. Direct Silylation scans are an *activity* that were used, according to internal Brown & Williamson documents, as "an analytical tool for comparing tobacco" (Moldoveanu, Johnson, & Burch, 1990). DS scans, Tea-Bag, and PY-GC-MS (pyrolysis-gas chromatography-mass spectrometry (Moldoveanu & Alford, 1988)) were techniques used to compare tobacco from different cigarettes for competitive research within Brown & Williamson. DS scan was a standard testing method within Brown & Williamson, but was not used across entire the tobacco industry.

From a text analysis point of view, this document follows expectations for an academic scientific report. It sports a participation framework that is impersonal and role-driven (see document #2 for more about formulaic language and role-driven participant structures). Also, genre and corpus linguistic studies have outlined standard parts of academic scientific reports (Crookes, 1986) that coincide with headings within the current report document: "Abstract," "Introduction," "Results," "Discussion," and even "References." In addition, this report includes a conclusion, which is another standard part of a scientific report. This suggests that the *sign system* of this document is limited to a report system and not a standard narrative or conversational system.

Instead of formally heading a section with a Conclusions title, conclusions are introduced within the Results and Discussion framework more informally. The author reports, "several conclusions can be drawn" (Moldeveanu & Kulshreshtha, 1991). Individuals are either not given *identities* or are only referred to by their last name. This is also part of the formal scientific report genre parameters.

7) Bates Number: 2501205056/5058, http://legacy.library.ucsf.edu/tid/lhu22e00 (angle 41.20°)

This document is a report on how cleaning air conditioning systems can greatly reduce airborne irritants and contaminants. This report talks about the importance of clean air conditioning systems and ducts.

At first look, this document is only a report of the necessity of cleaning duct work. Unclean duct work and other problems with air conditioning systems add to health problems associated with indoor air. However, within the broader context of tobacco interests, focus on "indoor air cleanliness" has been used to keep public attention away from banning smoking indoors. Tobacco control research such as that reported in "Tobacco industry efforts to present ventilation as an alternative to smoke-free environments in North America" (Drope, Bialous, & Glantz, 2004) demonstrates that the tobacco industry tried to diminish concerns about secondhand smoke by saying that second hand smoke was "an insignificant component of a much larger problem of indoor air quality and inadequate ventilation" (p. 41). In doing so, the tobacco industry was able to forestall indoor bans of smoking for many years (Bialous & Glantz, 2002; Drope et al., 2004).

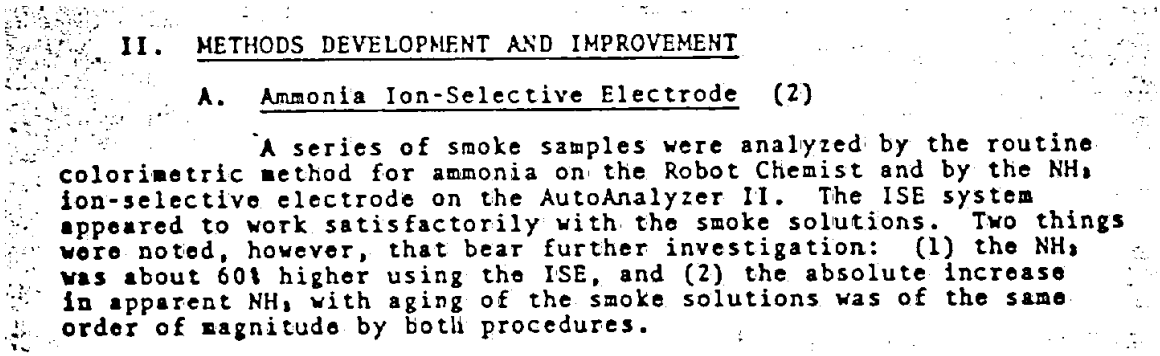
This document is within a range of topics that the tobacco industry has used to "smokescreen" the issue of tobacco and health. A "smokescreen" or "knowledge shield"

(Messick, 1998) includes strategies of deception by obfuscation, including putting out info that will create doubt about an issue, or feigning ignorance about a certain subject. This document is an excellent example of background information that was used by the tobacco industry as a smokescreen. It is also interesting that for background information such as this, there is no discernable *participation structure* within the document. It is as if someone jotted this down and left it on a parkbench, where another completely unrelated person picked up the document. It is only by instantiating a *participation framework* that this document begins to operate as a support to deceptive tobacco strategies. In this research-instantiated participation framework, the audience of the *idea* at least (if not the document itself) is the public, and public officials.

Other tobacco control research demonstrates that in some areas of the world, this smokescreen approach to second hand smoke regulations is still being employed by the tobacco industry (Barnoya & Glantz, 2002). As part of that focus on indoor air quality issues as a smokescreen, this document represents a distant supportive corporate strategy meant to bolster the deceptive rhetorical practices of the tobacco industry. From a discourse and CDA perspective, however, this document does not establish any clear *power* differentials. The only way for us to know that these issues are being used as a smokescreen is for us to have access to other industry documents that can connect this *activity* (air quality) to the tobacco industry, or establish a *participation framework* that identifies the public as audience and the tobacco industry as at minimum working in *relationship* with the author.

8) General Analytical Chemistry Section (Monthly Summary) (Oakley, 1976) (angle 49.25°)

This document is a General Analytic Chemistry Section Monthly Summary for the month of September, 1976. Elizabeth T. Oakley, the facility leader, is the author. It falls in the genre of a scientific or technical report and is specific, detailed and veridical. For example, in the excerpt shown in figure 5 (below), a series of sentences are presented as agentless passives (which should be associated with *deprofiled agency* according to the linguistic indicators): "samples were analyzed," "things were noted" (Oakley, 1976).



II. METHODS DEVELOPMENT AND IMPROVEMENT

A. Ammonia Ion-Selective Electrode (2)

A series of smoke samples were analyzed by the routine colorimetric method for ammonia on the Robot Chemist and by the NH_3 ion-selective electrode on the AutoAnalyzer II. The ISE system appeared to work satisfactorily with the smoke solutions. Two things were noted, however, that bear further investigation: (1) the NH_3 was about 60% higher using the ISE, and (2) the absolute increase in apparent NH_3 with aging of the smoke solutions was of the same order of magnitude by both procedures.

Figure 5. Excerpt from General Analytical Chemistry Section Monthly Summary. (Oakley, 1976).

The agentless passive construction is associated not only with scientific reports (Peters, 1985) but also with news media (Quirk et al., 1985, p.166) and ideological doublespeak (Ilie, 1998). However, in this instance, the passive is used to focus on "the procedures and processes themselves" instead of "personal involvement in experimental routines" (Peters, 1985). In the current document, agentless passives are not being used to distort an image of reality, but are part of the form and convention of technical and scientific reports. This report also resembles the

reports already review above because the *identity* of people involved in the report is reduced to last name only. This discursive reduction of individual *identity* serves in this instance to highlight the primary importance or *significance* of the technical information in this report, as opposed to highlighting who might have been involved in creating the report. This document, like documents #6 and #2, has a *participation framework* that is impersonal and role-driven. This impersonal framework, while reducing emphasis on *identity*, facilitates structured communication and highlights the importance of subject matter over *relationships* within the text.

9) Untitled, MNAT00778295 (angle 49.25)

This April 19, 1967 document is a note or memo about some particulars of a smoking device developed to help test the components of cigarette smoke. The memo discusses whether this device is a candidate for a patent. This document is very honest about the shortcomings of the device and boldly outlines some of the problems with the device in question, including the fact that it allows large particles to collect in its chamber and in doing so, produces "extremely irritating smoke" (see figure 6).

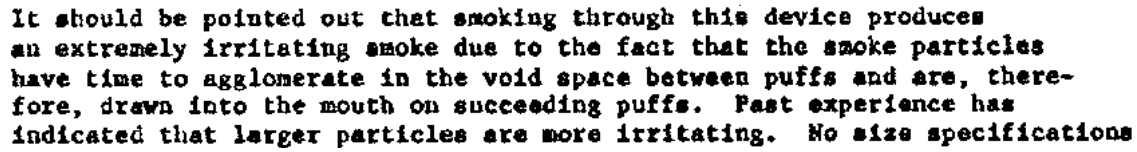

It should be pointed out that smoking through this device produces an extremely irritating smoke due to the fact that the smoke particles have time to agglomerate in the void space between puffs and are, therefore, drawn into the mouth on succeeding puffs. Past experience has indicated that larger particles are more irritating. No size specifications

Figure 6. Untitled (Suggestions: Filter Devices), Bates number: MNAT00778295.

Discursively, the author of this document establishes an *identity* for him/herself as a decision-maker and group member: "We have reviewed the patent [of your memorandum] in some detail…" Reviewing something in this instance seems to indicate that the reviewers have some authority as to what is done with the things that are being reviewed. Indeed, even though the text is inconclusive with respect to the device in questions, after a brief introductory paragraph, all of the rest of the document is a critique of the device, which becomes the single most *significant* item of the text.

It is interesting to note the date of this document. Although by 1967 the dawn of awareness about public health risks had already begun, at that time, the industry was not yet regulated with respect to warning labels and advertising restrictions. Therefore, tobacco company officials did not yet perceive their industry as being under widespread attack.

This document's treatment seems candid and not strategically deceptive or conspiratorial. This is due in part to a *participation structure* that seems to only include the author and the addressee. Not even the creator of the device in question is included as a *participant* in the communication, though he or she is referenced. Since the *participants* are reduced to just a few people, the author seems more open to speaking his or her mind. This openness can also be observed in other documents that resemble this one. For example, documents with Bates numbers in the same range are harshly candid about suggestions made for creating new filters. One document within a range of ten from the document in question reports that in a filter proposal "so many statements [in the proposal] … were contrary to our thoughts." This same document that neighbors the one under examination continues with even more incredulity regarding the proposal: "How anyone could come to the conclusion…is difficult to understand."

Highly Deceptive Documents

As a whole, the presumptively highly deceptive documents are diverse. There are memos, letters and an outline of a strategic plan. Three documents stood out as having elements within the text itself of planning for outright public deception: 1000134454 (H. Gaisch, 1983), 680262216 (Tobacco Industry Research Committee, 1953), TIMN0069106 (Kornegay, 19XX). They discussed a plan to portray the tobacco industry as "impartial," and they promoted strategic plans to buy more time for the industry. Emerging through purely statistical criteria in this most deceptive group of documents is a draft of the Frank Statement (discussed in Chapter 2), which is considered to be the public document that established the deceptive tobacco agenda in the 1950s that carried into the Twenty-First Century. In all, it is noteworthy that the set includes a majority of documents that clearly participate in one way or another in the tobacco industry's deceptive practices.

10) Bates Number: 664063379/3381, http://legacy.library.ucsf.edu/tid/tql83f00 (angle 3.55°)

This document discusses details of KOOL Milds' generic and proposed marketing campaigns in African American communities. It mentions advertisement placement, written copy, and pictures. This letter is a response to individuals at the Ted Bates and Company Advertising agency. This agency worked for Brown & Williamson and other tobacco companies.

This letter does not seem overtly strategic with respect to public health. It is a marketing specification document. The writing is frequently straightforward direction: "…include the addition of…," "recommend a third visual," "proceed to finish material." However, though this particular document appears innocuous, it is part of one of the most cynical and manipulative chapters in the annals of marketing. The difference between the limited *participation structure*

as it is represented in the text itself and the vast *connections* between *activities* in the document and insidious tobacco marketing *activities* gives a clue as to why this document seems so straightforward, and yet was involved in such deceptive strategy. The *participation structure* of the document is limited and includes B & W and Ted Bates & Co. By contrast, other tobacco industry documents unmask overall campaigns to market cigarettes – especially menthol cigarettes like KOOLs – to African Americans as especially exploitative. KOOL has a long history targeting African Americans that still continues today. Marketing campaigns by KOOL as recently as 2004 targeted African American youth (American Lung Association, 2006; Hafez & Ling, 2006). The *connection* in this document via KOOL to African American Marketing Campaigns makes this document a candidate for a function to support deception within the industry. However, the support provided by this document is very distant and cannot be closely enough linked to assert that this document is itself part of that deceptive strategy.

11) Bates Number: 2048908735/8737, http://legacy.library.ucsf.edu/tid/evg48d00 (angle 6.67°)

This letter introduces attached materials intended for use in a Dutch case in which the tobacco industry was suing the Dutch Public Health Foundation – *Stichting Sigaretten Industrie v. Stichting Voksgezondheid en Roken* (roughly translated cigarette industry foundation v. foundation of public health and smoke). James T. Newsom of Shook, Hardy & Bacon (Kansas City, Missouri) is writing to Charles Gielen, part of the Amsterdam counsel for Philip Morris. This document discusses enclosed materials (which are indexed as different documents) that include criticisms to the Tobacco Institute's response to the 1986 Surgeon General's Report, a literature review of environmental tobacco smoke, and a memorandum concerning errors in defendant's claims. In discussing these materials, it creates *connections* between the Dutch court

case, 1986 Surgeon General's Report, criticisms of this report, Infotab's review of Environmental Tobacco Smoke (more commonly known as *secondhand smoke*) and a case where the verdict was not for the tobacco industry, the Cipollone Case.

These events and documents are linked to the Dutch court case at hand in part because it seems they have been requested by the Dutch counsel. These items have been provided in response to a *relationship* that seems to have been built for the purpose of this type of information exchange. For example, these items are provided "pursuant to our telephone conversation." The document also mentions inclusion of additional material relating to "the EC [European Commission] proposal to restrict the use of trademarks, etc." Interestingly, this additional material was not requested. The *relationship* that Newsom alludes to (e.g. "in *response* to your letter," "pursuant to *our telephone conversation*," "*your earlier request* for materials" – italics mine) seems to have been limited to discussions of the specific items Newsom has attached to this letter.

This *relationship* creates an *identity* for Newsom that is lodged between expert and helper. Newsom is eager to please Gielen: "we hope these materials are helpful," "please let us know if there is anything more we can do." These niceties could be seen as part of a value of professional politeness within the letter. This helpful politeness underscores the *social good* in this letter of providing information and informative help. In fact, this letter exists entirely to outline which informational items will be exchanging hands. The *activity* of information exchange is the forefront reason for this letter. But this *activity* goes beyond what has been requested and actually includes materials that were not requested, but that Newsom has deemed potentially helpful for the Dutch counsel. "In addition," Newsom states, "I have enclosed some

additional materials…which may be helpful." Moving out from the direct scope of this letter, it is interesting to consider the difference in the *social good* of free and helpful information exchange in this letter in contrast to the *political* policy of the tobacco industry to withhold information to the public.

James T. Newsom worked for Shook, Hardy and Bacon, a law firm representing Philip Morris. This semi-formal legal letter is addressed to Mr. Charles Gielen of Nauta Van Haersolte, Philip Morris counsel in Amsterdam. This document presents a prime example of a letter that is supporting tobacco machinations. For example, the attached materials, which are listed in the letter, are not entirely comprised of sensitive materials. Three of the four mentioned items were potentially publicly available materials even before the forced disclosure of industry documents. These nonsensitive documents included the Tobacco Institute's reply to the Surgeon general, the criticisms of the Surgeon General's report, and the international research unit of the tobacco industry's annual review of ETS (Environmental Tobacco Smoke or. second-hand smoke).

These items enclosed with the document in question closely resemble other items that are available to the public as part of the tobacco industry documents. However, even after the court mandate to make all documents available to the public, Philip Morris has continued to shield these specific documents. Instead of releasing them, these documents are classified as "Confidential communications from Philip Morris USA, Inc. counsel to Philip Morris USA, Inc. employees," and with that label, they are unable to be viewed.

This *confidential* label has raised red flags for tobacco control researchers. Tobacco control researchers have noticed a startling trend with regard to confidential material (Muggli, 2000). Namely, it appears that very frequently, material marked "confidential" (and by virtue of

that label, withheld from the public) has simply been passed over a lawyer's desk with the specific intent of protecting it as a confidential attorney-client communication, despite the fact that the subject matter may have nothing to do with legal matters. In this way, the goal of making tobacco industry information public is undermined.

From a CDA point of view, this document specifically forecloses the possibility of texts/voices that could oppose tobacco positions. This sequestering of information could be perceived as a generally expected exclusion within the context of a series of legal texts. Lawyers, after all, are advocates for their clients, and not employed to encourage opposing voices. Despite the complicating factor of legal status, this exclusion of other anti-tobacco voices does contribute to a larger historical pattern of aggressive suppression of opposing voices.

Within the adversarial framework of a trial, this document might not be considered abnormal or deceptive. The document constructs the *identity* of the anti-tobacco defendants as wrong-headed and erroneous: "[D]efendant's assertions contain many errors." In contrast, as was outlined above, the author of this letter constructs his own *identity* as helpful and "on their side." In general, from a public health and endangerment position, this document constitutes deceptive practices because it participates in a climate of information suppression. To be sure, in the context of the legal profession, acting on behalf of a client – regardless of how nefarious that client may be – these practices are normative and in accordance with is its own kind of ethical imperative or network of *social goods*.

This document hints at a rather standard *participation structure* within law firms and legal support. Though there are two main individuals participating in this conversation, there are hints within the document that a network of lawyers and legal aides has been involved in

supporting this exchange. References to "we" on behalf of Newsom must be attributed to a staff that assists him or works with him. These individuals are not active participants within this *structure*, but participate as silent witnesses and information providers.

From a linguistic indicator standpoint, this document shows little evidence of *cognitive verbs* (with the exception of *hope*). There is definitely a group mentality evidenced here. Newsom and Gielen are working together with the tobacco industry as a group, incorporating information from Infotab and the Tobacco institute (their reviews of ETS literature, and responses to the Surgeon General's report, respectively). Technically, however, only one instance of a pronoun is used in a way that would feed in to this study's automated assessment: "*our* conversation." Other pronouns that would have been counted for this indicator were showing group mentality as well, but definitely not an *us v. them* mentality. For instance, the following instances of group mentality are well within the bounds of a corporate department or unit working as a whole to produce effective work products: "*we* hope these materials are helpful;" "*we* would also appreciate;" "let *us* know if there is anything more *we* can do." In contrast to Newsom's off-the-cuff addition ("*I* have enclosed some additional materials"), these instances of group mentality seem to support a standard corporate setting whereby many people have worked together to provide these items.

Additionally, there is no *deprofiling of agency* with the use of passive constructions. There is little *strategic ambiguity*, and no *adversarial language* that would technically have been caught with the current versions of the computer program for adversarial language. The overall context of this document is a litigious context. And yet, due to the professionally congenial *relationship* evidenced between Newsom and Gielen, there is no adversarial language despite the

context. Here, the *political* goodwill expressed toward the recipient of the letter outweighs the adversarial context of the Dutch litigation.

12) –No Title. 1000134454 (angle 7.44°)

This document suggests that the industry use longitudinal studies of the epidemiological impacts of smoking to "buy time for the industry." Studies will be run at "arm's length," presumably to allow the industry to direct studies while having the added benefit of employing researchers that will do legitimate research. Scientists working for the major tobacco companies (within this document, it is impossible to tell which specific company scientists worked for) were concerned that epidemiological studies might be conducted without taking into account "all possibly associated risk factors." That is, from the point of view of the tobacco industry, the intent was to delay studies of adverse tobacco outcomes, since all potential non-tobacco factors contributing to health risks had not been identified.

This document discusses the International Committee on Smoking Issues (ISCOSI) plan to forestall industry evidence about smoking and health for *ten years*, from 1977 to 1986 when the first of the proposed epidemiological data would be available. ISCOSI is the international equivalent of CTR and TIRC which were both dissolved as part of the master settlement agreement due to their position as "front groups set up by industry to block public health policies" (Yach & Bettcher, 2000). This document outlines one of the tobacco industry's most insidious techniques for public endangerment: funding studies and research with the explicit purpose of buying time and "proving" that no link between tobacco use and public health could be conclusively established.

The obvious *activities* discussed in this document are a hypothetical epidemiological study and an introductory telephone call on the subject of epidemiology that precipitated this letter. In talking about the hypothetical *activity* of an epidemiological study, there is evidence in the document of *deprofiled agency*. Instead of stating that the tobacco industry would have to run any epidemiological study, this letter sets up a series of demands on how to do an epidemiological study that in the end would exclude any other entity from doing this kind of a study and having it be considered valid. For instance, the *agency* of the tobacco industry is deprofiled by remarks such as the following "such a study would have to be run [by the industry]."

In general, the hypothetical nature of the subject creates some interesting linguistic patterns that in some cases correspond to the linguistic indicators of deception. For instance, the items that indicate *strategic ambiguity* occur throughout this document, but do not in this instance mark a strategy of behaving in one way while intentionally reporting your behavior ambiguously. In this instance, the lack of futurity associated with *strategic ambiguity* is linked to the fact that this document is discussing a hypothetical and not yet real event or series of events. This document is not masking its intentions with lack of futurity; it is simply positing a possible future that as yet is so indefinite as to be ambiguous.

There are no *cognitive verbs* or instances of *adversarial language* in this document. On the other hand, the group mentality is pervasive through this document: "our group," "the group," "the industry." The group mentality in this document seems linked to an issue of power with respect to the hypothesized epidemiological study. There is an undercurrent of this document suggesting that the tobacco industry is the "only" industry "that could initiate such a

unified and coordinated study." The participant structure of the document, though entirely tobacco-industry internal, makes these arguments and explains reasons why the tobacco industry is the best candidate to do this type of study. This argument structure suggests a participant framework that includes a silent adversary who might block the proposal for the tobacco industry to run an epidemiological study. This industry knows that they have the *power* to direct just such a study. They have the financial and political resources to direct an epidemiological study, but "run it at arm's length via consultants."

13) A Frank Statement to the Public, 680262216/2218, http://legacy.library.ucsf.edu/tid/hfz70f00 (angle 5.07)

As explained in Chapter 2 of this dissertation, the Frank Statement was the first collective public statement from the entire tobacco industry (under the guise of the Tobacco Industry Research Council). It promised straightforward communication with the public about health facts and findings with regards to tobacco and cancer. It was widely circulated and originally printed in 1954 in over 400 venues. Its expansive circulation and uniqueness as a statement to the public from the entire tobacco industry has made this document a key document in tobacco industry legal proceedings. This document is a best example of a publicly disseminated manipulative rhetoric was supported by deceptive tobacco industry strategy. As discussed in Chapter 2, this is a quintessential deceptive tobacco document.

The Frank Statement is a primarily strategic text (Brown & Rubin, 2005). That is, the *activity* (Gee, 2005) created by this document is not information exchange, but persuasion. For example, unneeded adjectives are used to legitimize what would already be presumed to be legitimate constructs. Thus readers are presented with adjective-noun constructions such as,

"eminent scientists." In the context of the time when this document was written, with internal research already linking cancer to cigarette smoking, it is no wonder that the industry might have felt the need to doubly legitimize the credibility of their researchers.

Indirect reporting and references to external legitimators serve to bolster the credibility of the claims made in the Frank Statement. By referencing these external sources, they bring them in to the *participation framework*, effectively making any of these external sources complicit with the tobacco industry. The linguistic strategy of deferring to external legitimators – that is, those "eminent scientists"- effectively removes the locus of the topic (in this instance the controversy over smoking and cancer) to outside of the author's purview. Despite these appeals to external legitimacies, in the parts of this document where the tobacco industry does linguistically create an *agentive* identity for itself, it makes rather promises that have been shown to be hollow: "We are pledging," "We always have and always will cooperate," "We accept an interest in people's health as a basic responsibility." This last assertion would later become a prominent focus of legal actions against the tobacco industry. These outlined *social goods* of honesty and cooperation are a stark contrast to the *activities* that occurred within the tobacco industry which effectively cut honest and cooperative communications with the public to a minimum.

In this document, we see the full effect of an attempted *relationship* creation between the authors (the tobacco industry) and the intended recipients of this message, the public. In this instance, the recipient is extremely important. Would we have been as interested in this message had the tobacco industry never published it, and thus never attempted a discursive relationship between the industry and the public? I contend that the industry still would have been held

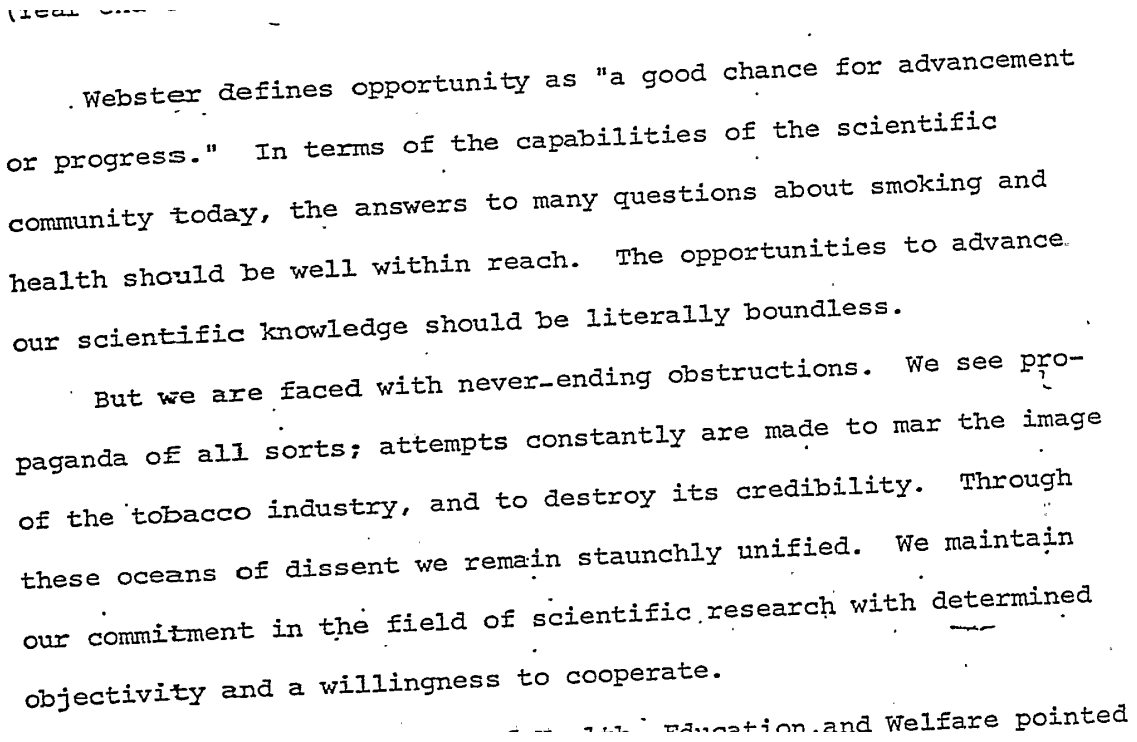
responsible for deaths associated with tobacco and cigarette smoking. On the other hand, there are court cases underway at present that never address the question of scientific linkages between tobacco smoking and health. Instead, they leverage the difference in message between what the tobacco industry delivered to the public for the last fifty years versus what the industry was discussing internally. So what if the industry had never communicated with the public at all; would they have escaped the scathing indictments that they are now facing? I think not. They were still (and are still) producing products that are detrimental to public health, but I do think it would have made a difference in public opinion for them to not have consistently misrepresented and lied about their knowledge of health risks.

14) – No Title [End of Year Statement from the Chairman of the Tobacco Institute to the United States Tobacco Journal], TIMN0069106 (angle 6.79°)

This document is a fragment of correspondence between the United States Tobacco Journal and The Tobacco Institute. This document is actually the year-end statement from the Tobacco Institute Chairman at the time, Horace Kornegay (Vice President, Leaf Operations, RJ Reynolds). Although the actual document is undated, Kornegay was only Chairman of the Tobacco Institute during 1985, so this must have been written during 1985.

Assertions made in this draft public statement are supportive of the public disinformation agenda of the tobacco industry. Some of the assertions are clearly lies. For example, statements shown in figure 7 (below) assert, "we maintain our commitment in the field of scientific research with *determined objectivity* and a *willingness to cooperate*" (italics mine, TIMN0069106). Strangely, this document establishes two core values, two *social goods*, as being of primary importance: scientific rigor and truth. References to "determined objectivity" and statements that

"scientific objectivity will prevail" make clear the political importance for the author of scientific rigor. Casting scientific rigor as an imperative *social good* is one of the ways that the tobacco industry actually undercut other scientific studies and attempted to invalidate information about the health risks of tobacco use.



. Webster defines opportunity as "a good chance for advancement or progress." In terms of the capabilities of the scientific community today, the answers to many questions about smoking and health should be well within reach. The opportunities to advance our scientific knowledge should be literally boundless.

But we are faced with never-ending obstructions. We see propaganda of all sorts; attempts constantly are made to mar the image of the tobacco industry, and to destroy its credibility. Through these oceans of dissent we remain staunchly unified. We maintain our commitment in the field of scientific research with determined objectivity and a willingness to cooperate.

Education and Welfare pointed

Figure 7. Bates number: TIMN0069106.

Additionally, cooperation and objectivity in scientific endeavors that could potentially link smoking and health risks was absolutely *not* a *social good* or commitment of the tobacco industry. Quite the contrary, document research has demonstrated that the tobacco industry acted to undermine and stymie scientific research that was not directly supportive to their viewpoint (Carter & Chapman, 2003). The tobacco industry was also definitively *not* objective, forcing

scientists within their grant structure to report only positive or ambiguous findings about links between tobacco and public health risks (Bero et al., 2005).

This document almost has an over-the-top defensive quality about it as well. The *allness* language in this document is pervasive; opportunities should be "boundless;" obstructions are "never-ending," and there are "oceans of dissent." Massive *group mentality* is evident as well. Kornegay uses a repetitive sentence structure that fronts the pronoun "we" which adds to the *group mentality* expressed in the document. He repeats: "we will discover truth;" "we remain staunchly unified;" "we maintain our commitment;" "we believe scientific objectivity will prevail." Additionally, he uses *cognitive verbs* to underscore the *identity* of the tobacco industry as one that is unified, committed, and objective. The tobacco industry receives an *identity* that resonates with *adversarial language* and war imagery; it remains "staunchly unified" despite attempts by other parties to "mar" the tobacco industry image.

In addition to lexical items and phrases that resonate with *adversarial language*, Kornegay also outlines two longer narrative *events* that firmly create adversarial *relationships* between the tobacco industry and the American Cancer Society and the Surgeon General. Kornegay relates that once the American Cancer Society's "smoking-dog" results were released, the Tobacco Industry asked for the underlying data and was denied. Though I have not been able to verify that this incident actually happened, it is interesting to note that in this address, Kornegay represents the tobacco industry as outraged over the fact that information was not freely shared with them. This is particularly interesting considering that the industry firmly refused to distribute in-house reports that would have been detrimental to the tobacco industry.

In calling for truth to prevail at the end of his speech, Kornegay is not actually calling for truth, but for his and the tobacco industry's version of the truth to prevail.

Throughout the letter, Kornegay sets up adversarial positions that culminate in *participation framework* that is binary in nature. On one side, Kornegay, the Tobacco Industry, Scientific Reason and even Humanity speak with one voice, while on the other side, a cacophony of adversaries work against them. Kornegay ends his address with a rallying cry that appeals to human nature as validation for his actions:

> "One of mankind's inherent instincts is self preservation. And if we persevere and continue to search for the facts, we <u>will</u> discover what Mark Twain once called 'the most valuable thing we have:' Truth." (Kornegay, 19XX)

This appeal the nature of humanity is an appeal to even greater legitimization than references to the doctors and scientists in other documents. Kornegay introduces the idea of the nature of humanity to underscore the *political* rightness of the actions of the tobacco industry. This appeal underscores deceptive and fraudulent actions of the tobacco industry in attempting to solidify false links between the tobacco industry and truth, and the tobacco industry and "inherent" human nature.

CHAPTER 5: DISCUSSION

Discussion of General Results

This study used the Tobacco Documents Corpus (TDC) to investigate deceptive corporate strategy in a body of corporate documents. Extensive background research in the field of deception studies established the basis for six proposed linguistic indicators of corporate deception and fraud. Using corpus linguistic and computational methods, these indicators were operationalized using computer programs designed for this purpose and then investigated one-by-one and in a multivariate vector model-based ranking system. These indicators were used to investigate differences between classes of TDC documents and finally to rank internal audience TDC documents for deceptive corporate strategy. Some of the general research questions of interest in this research included the question of whether interpersonal traits of deception would be transferable to a corporate context, how the linguistic indicators of deception would perform in ranking documents for deceptive corporate strategy, and whether hypotheses about the nature of the indicators of deception would be substantiated.

This chapter discusses results of the statistical tests using the indicators of deception and corporate fraud. Additionally, it discusses results from the discourse analyses of fourteen ranked documents (full analyses are in Chapter 4). Finally, this chapter addresses implications of these results in the fields of linguistics and deception research including future research avenues and research caveats.

The Hypotheses

Internal- Versus External-Audience Documents

Hypothesis 1 predicted that internal audience documents would demonstrate lower instances of the proposed six indicators of corporate deception. Hypothesis 2 predicted that there would be observable and significant differences in the instance of the indicators of corporate deception between sources. For Hypothesis 2, how CTR and TI documents performed were of particular interest. Neither hypothesis was supported. Contrary to theory-based expectations incorporated into Hypothesis 1, documents directed toward industry-internal audiences exceeded those directed toward external audiences on deprofiling agency, strategically ambiguous intent markers, and cognitive verbs. There was no significant difference between the two sets of documents for adversarial vocabulary, allness language, and group mentality. By and large, this pattern of results demonstrates that the fundamental assumptions about differences between the internal-audience and external-audience documents were not on target. This may be in part due to differences between spoken and written data. However, I also contend that some of the indicators, while performing well in their original studies, are not robust enough to withstand a shift into corporate discourse.

One possible explanation for the failure to confirm Hypothesis 1 is that industry-external audience documents are *not* any more deceptive than are industry-internal audience documents. That possibility seems unlikely, in that the motivation to manipulation and dissembling in external communications is naturally much higher than the motivation in internal communication. However, the constitution of the two classes of documents is complicated by variation in genres both within and between those classes. The internal-audience documents are made up of a vast array of different styles and genres including technical reports, memos, emails,

in-house speeches, marketing reports, and so forth. In comparison, external-audience documents are limited mostly to press releases to the public. Thus any effort to demonstrate differences between documents directed to the different audience targets will be confounded by these genre differences, and genre/situation is most likely the strongest determinant of stylistic traits of all (Hymes, 1974). Additionally, the impact of intended recipient may have interfered with the indicators.

Another possible explanation for this unexpected research outcome is that no single indicator by itself can serve as a reliable indicator of deception. Rather, given genre-specific stylistic formulations, only a weighted *constellation* of language markers can even begin to distinguish misleading communication from truthful.

Finally, a third possible explanation for the failure to confirm predicted differences between internal audience and external audience is that the six linguistic indices selected for testing in this dissertation were simply the wrong ones. To be sure, the existing literature on deception in language offers more promise in the case of spontaneous spoken discourse issuing from a single speaker than for deliberate, edited written discourse from corporate authors. The more promising approaches to identifying deceptive language in speech arise from indicators of cognitive burden. Thus, for example, fluctuations in type-token ratios may indicate where in the stream of speech a speaker is grasping to overcome the burden of disinformation, or has over-rehearsed a response (Carpenter, 1981). However the variables selected for examination here are the best available given the current state of research on deception in written language. This list was intentionally limited to offer each variable a full examination, both theoretically and experimentally. Accordingly, the six indicators of corporate deceptive corporate strategy were distilled from a wide range of different approaches to deception research including case studies

observations (Lebaron, 1996; Vrij & Mann, 2001), experimental deception studies (M. Burgoon, Callister, & Hunsaker, 1994; Newman et al., 2003; Qin & Burgoon, 2005; Vrij & Mann, 2001), deception detection studies (including those studies testing both how well people can detect deception, as well as those that are testing methods for detecting deception) (Burgoon et al., 2003; B. M. DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997; B. M. DePaulo, Jordan, Irvine, & Laser, 1982; B. M. DePaulo & Pfeifer, 1986; B. M. DePaulo, Tang, & Stone, 1987; P. J. DePaulo & DePaulo, 1989; Ekman & Osullivan, 1991; Frank & Feeley, 2003; Frank, Paolantonio, Feeley, & Servoss, 2004; Harrison et al., 1978; Porter & Yuille, 1996; Vrij, Edward, Roberts, & Bull, 2000), and theoretical accounts of deception constructs such as Interpersonal Deception Theory (Buller & Burgoon, 1996; Buller, Burgoon, Buslig, & Roiger, 1994; Buller, Burgoon et al., 1996; Buller, Burgoon, White, & Ebesu, 1994; Buller, Stiff, & Burgoon, 1996; Burgoon & Buller, 1994; Burgoon, Buller, Dillman, & Walther, 1995; Burgoon, Buller, Ebesu, & Rockwell, 1994; Burgoon, Buller, Ebesu, White, & Rockwell, 1996; Burgoon, Buller, & Floyd, 2001; Burgoon, Buller, & Guerrero, 1995; Burgoon, Buller, Guerrero et al., 1996; Burgoon, Buller, White, Afifi, & Buslig, 1999). This rich literature, though somewhat scattered, seemed to have themes that translated into the six indicators of deception.

Difference between sources for TDC documents

Some of Hypothesis 2 was supported by the data. For deprofiled agency, instance in documents from the American Tobacco Company and RJ Reynolds exceeded that in the TI documents. Tobacco Institute documents significantly exceeded the instance of adversarial language compared to the American Tobacco Company, Brown and Williamson, CTR and Lorillard. Additionally for adversarial language, RJ Reynolds documents exceeded American Tobacco Company, Brown and Williamson and CTR documents. There was no significant

difference between the two sets of documents for adversarial vocabulary, allness language, group mentality and image management verbs. The results of H2 indicate that a corpus of single-source documents could be used in future research as possible collections of "most-deceptive" documents. Differences between the two industry front organizations – The Center for Tobacco Research (CTR) and the Tobacco Institute (TI) – and the rest of the sources were not statistically significant in the way predicted (where *both* would have higher instances of the indicators than the rest of the sources). However, the Tobacco Institute was significantly higher than four of the other sources for deprofiled agency. In addition, between the two, either TI or CTR had the highest average representation of a linguistic indicator of deception for four out of the six indicators: adversarial language, image management verbs and allness and superlative language.

The Center for Tobacco Research (CTR) and the Tobacco Institute (TI) were originally chartered in the 1940s to deal with public attitudes towards tobacco and health. The deceptiveness and anti public-health stance of both the CTR and the TI was such an issue in the proceedings against the tobacco industry in the late 1990s that both were dissolved as part of the Master Settlement Agreement (Glantz, 1996). Court decisions to disband CTR and TI confirm that these were organizations only developed to distract and mislead the public. As just such institutions, CTR and TI produced documents that were either highly deceptive or supportive of a deceptive tobacco industry agenda. Despite this fact, however, their internal documents are not necessarily especially deceptive.

Proposed Indicators of Deception and Corporate Fraud with Other Classes of TDC Documents

Decade

No specific hypotheses had been posed for decade. Differences among decade categories (which included the Bliley set as one category) for the six indicators of deception did not produce statistically significant results.

Addressee

No specific hypothesis had been posed for Addressee. Documents with corporate addressees exceeded those with named addressees for deprofiling agency. Deprofiled agency was operationalized as the number of passive structures per total phrases in a document. Anecdotally, considering the documents reviewed as part of the discourse analysis of ranked documents, it is not surprising that corporate addressed documents would have higher deprofiled agency. One of the striking features of many of the report-like documents is their use of passive, null-agent structures.

There were no other significant differences between documents to named addressees and documents addressed to corporate addressees.

Using Vector Analysis to Rank Documents for Deception

Considering the poor match between the hypotheses and findings, we might expect that the results of the corporate fraud and deception ranking would also be poor, or in a best case scenario, unclear. On the contrary, the discourse analyses of the 14 documents (five most deceptive, four middle and five least deceptive) demonstrate the start of some promising trends. It seems that while the distribution of deceptive indicators in the different classes of the TDC do

not map out as expected, the effect of them used together to rank deceptive corporate strategy is interesting and worth further investigation.

Deception, in general, seems to be much too complex a phenomena to be adequately triangulated with only one indicator. A multivariate approach, as the discourse analysis results show, gives a better estimate than does a single indicator approach. However, even with a multivariate (all six indicators used together) approach, there are lingering problem points that result directly from the complexity of language use.

The theoretically least deceptive documents actually showed some very interesting potential deception by omission. In one document, a phantom audience – perhaps the public – kept an employee from being as specific as he might have wanted: "I kept this telex less explicit." But by keeping the message less explicit, this document could not be diagnosed as actually deceptive. This document and the other least deceptive ranking documents did not function as smoking gun or "hot" documents. At the same time, however, they addressed topics that warranted further research and investigation. These least deceptive ranking documents led me in the larger document population (the entire 7 million tobacco documents) to discover that denitration most likely was connected to health issues focusing on pesticide residue on cigarettes. These least deceptive ranked documents also led me to investigate more about the workings of the ISO – International Standards Organization – which in turn led me to research that demonstrated how the tobacco industry has used the ISO for nefarious purposes.

The middle documents, by contrast, were a selection of duplicate or near-duplicate documents in the entire 7 million tobacco document population. These forms and reports are very much a part of instantiated day-to-day business language. However, it is interesting to consider what these quotidian documents really mean for the TDC. On one hand, having

documents in the TDC that are part of sets of everyday business language is important from a representative standpoint. In order to get the best picture of what was happening in the tobacco industry, we must at least represent everything. But reports and forms are from a linguistic standpoint very different than conversational, epistle, or narrative language. For example, the linguistic variability between two letters in the tobacco industry documents dataset can be quite high while the linguistic variability of a DS Scan report is very low in comparison. One way to address this in the future might be to spend some time in the document population cutting out duplicate or near-duplicate sets of documents. Definitely in a study looking at something as linguistically variable as deception, this type of preprocessing could be beneficial.

Finally, focusing on the most deceptive (ranked closest to deceptive) documents, it is striking to me that for each document in this group, there seems to be a different trend and distribution of the six linguistic indicators of deception. Despite this variability along the axes of the linguistic indicators of deceptive corporate strategy, I do find that three of the five most deceptive documents are smoking guns from not only a discourse analytic perspective, but also a historical one. The internal version of "A Frank Statement to Cigarette Smokers," the plan to "buy the industry time" with a 10-year epidemiological study and the unpublished letter to the tobacco industry are all highly deceptive.

The Frank Statement is unquestionably a deceptive document. This is a document that has been proven in court to be not only misinformative, but full of lies. There is not a document that I know of within the tobacco industry that has been as scrutinized and vilified. Despite the difficulty of linking the individual instances of the indicators of deception to deception itself, the indicators of deceptive corporate strategy worked extremely well in this case to rank the most deceptive document in the industry exactly where it should be, with the most deceptive

documents. The appearance of the Frank Statement—perhaps the most notorious example of tobacco industry duplicity—as one of the most deceptive documents is a strikingly confirmation of a correct direction for the results. It lends validity to the algorithm by which it was selected despite the fact that some of the indicators of deceptive corporate strategy did not instantiate as hypothesized in every instance.

This dissertation research did not substantiate hypothesized claims linking single indicators to deception. However, by using the indicators of deception in a multivariate approach, this research has automatically created sets of documents that appear to be stratified with respect to their deceptive function. The documents selected as most highly deceptive area contained deceptive statements themselves. More importantly, however, the documents in the highly deceptive ranking outlined and promoted public disinformation programs that contributed to the pervasive deceptive *information climate*. This corporate climate of deception characterized tobacco industry corporate communication practices within it. In addition, a link between the documents selected as least deceptive and deception by omission is interesting and worth further investigation. Overall, the discourse analyses confirmed that documents in each area – highly deceptive, mid-level and least deceptive – were qualitatively different from each other in informative ways.

Conclusions

A number of methodological, practical and theoretical questions guided this research. These were questions in general about the extendibility of previous corpus linguistic research linking goals and intents to language usage patterns and questions more specifically about corporate deception and fraud.

The results of this dissertation indicate that goals and intents can be deciphered through language usage. However, results of this dissertation also show that this coding between usage patterns and the specific goal or intent to deceive is multifaceted, and requires more sophisticated and comprehensive accounts of language beyond that of individual lexicon, syntactic structures or even thematic roles. In this research, I have not proposed a comprehensive account of language use in settings of corporate deception, but I have begun to move in that direction.

Application of the Discussion for Corpus Linguistics

This research also demonstrates that linguists need not use opaque methods of computation to have fruitful results. The vector matching method, though computationally intense, is straightforward and has no opaque techniques such as clustering or neural networks. Use of more sophisticated computational or analytic techniques such as neural networks or multi dimensional scaling analysis could produce results, but this research demonstrates that good results are achievable with relatively basic mathematical modeling. These results suggest that in order of complexity, simple correlation should be assessed first (e.g. correlation between linguistic indicators and TDC document classes), followed by more transparent mathematical approaches (e.g. vector model matching). As a next step, verifying these results with the use of more sophisticated computational modeling (e.g. neural networks or clustering techniques) could produce results that are even more efficacious. However, the clearly productive results of the current research are strong affirmation that transparent computational techniques supported by linguistic theory and observation-driven research are not just sufficient, but are additionally a successful combination.

Instead of “kitchen sink”-ing the proposed “genre” of deceptive corporate strategy in this dissertation, I strategically chose a small number of indicators to work with. Linguists can learn

from this discussion that a theory- and observation-based approaches to genre and corpus linguistics can be productive now that the field of corpus linguistics has already established general parameters for exploring differences and similarities between one type of language use and another.

There is still plenty to be done exploring language use from a corpus linguistic perspective. Deception, deceptive corporate strategy, corporate fraud and conspiracy are obvious areas that could continue to provide research possibilities. In addition, with the TDC and the Enron email corpus, corpus linguistics has brand new opportunities to delve into usage patterns and corporate culture. For instance, one of the things that struck me in reviewing the fourteen documents was the prevalence of formulaic documents in both the least and mid-range deceptive corporate strategy samples. A simple study to catalogue formulaic documents could be an interesting next step. In addition, although I did not choose to use a full battery of features and indicators, trying the same method with Biber's suite of features would be a possible follow-up for this dissertation.

The Future of Linguistic Deception Research

Despite the fact that this study does little to provide a full theoretical account of deception, I have tried to cover as much of the field of deception research as possible. More attention could be paid to corporate deception, not just in the realm of business ethics, but also in areas that typically model interpersonal communication. On the whole, observations taken into account from interpersonal models of deception in this dissertation were valuable jumping-off points for practical work in a corporate linguistic setting. A quick word of warning should also accompany any study in deception research, however. Automated deception detection is not only impractical, but teeters dangerously close to being unethical. Postulating that you can

automatically detect deception is risky. People are all too ready to trust in an easy lie detector, to the detriment of science and innocent parties.

On the other hand, the impact of distributing documented evidence of deception has immediate results. Attempts to counteract the tobacco industry's deceptive public materials have resulted in localized and immediate action. For instance, the quit rate after the release of the Surgeon General's report on smoking was higher in college educated-populations, due to their access to the report (Cutler & Kadiyala, 2003). Additionally, in 1997, the tobacco industry settled a lawsuit with the state of Florida and as part of that settlement released internal documents to the public. The next year, money from that settlement had funded the Florida anti-tobacco "Truth" public health campaign that widely distributed information found in the previously internal documents. Two years after the settlement and the release of these documents, smoking had decreased 11% in Florida (Sly, Hopkins, Trapido, & Ray, 2001). These examples demonstrate that distributing document evidence of tobacco industry deception can make an impact on public health. The impact of truthful revelations after so many years of misinformation and deception cannot be overestimated.

The "Truth" campaign was focused on exposing tobacco industry advertising and marketing practices (Farrelly et al., 2002), and was developed in part based on information garnered from the previously unreleased tobacco documents. These documents are tobacco industry-internal documents that prove previous public tobacco statements to be false and deceptive. While there is no way to prove that decreased smoking rates result even indirectly from the release of this material, it is striking that anti-tobacco campaigns that make extensive use of the released documents are associated so strongly with decreased smoking (Sly et al., 2001). This link between releasing documents and change in public behavior makes this

dissertation research imperative. Results of research exploring linguistic correlates of internal planning for public corporate deception could be used to facilitate faster and more accurate expert assessment of deceptive documents. Better and quicker assessments would result in more extensive access to documents. In general, locating and appropriately distributing documents that support publicly released false or misleading documents are urgent tasks that can have formidable impacts on public health, quality of life and even mortality.

Limitations

The dissertation in general builds on the prior literature of linguistic features and deception, corpus research and tobacco control. Several venues for future research should be considered. First of all, this research should be repeated on a different corpus. Would the same observations hold? Additionally, what other types of "average deceptive documents" could be considered for use in this template? The documents available from CTR and TI are candidates for average deceptive documents based on results of tests verifying hypothesis H2. These questions should be pursued as part of further research with different corpora in both the corporate and even potentially interpersonal realm.

In addition, although the proposed indicators of deception in this dissertation were thoroughly researched and attested in purely observational as well as experimental setting, to round out this research, these indicators could be revisited. What might happen if only five of the six indicators were used to rank documents?

The method of ranking documents proposed in this dissertation should be tested in a document population where document genre does not vary. For instance, the emails from the Enron document collection are perhaps the best data to use for continued research in this area.

Within the context of one style of document, how would the six indicators of deception perform? Would the same positive result be demonstrated?

I am not entirely convinced that by using the six linguistic indicators of deception, instead of a broad swath of linguistic indicators, I was able to overcome automatic grouping of the documents by genre. For example, three of the five "least-deceptive" documents were short letters, whereas only one of the ten mid- and most-deceptive documents was short letter style. This could have occurred for a number of reasons. First off, perhaps the genre of short letters is simply not as inherently deceptive in a strategic corporate setting as is longer epistles or plans for marketing or strategy. Alternatively, this distribution of short letters could indicate that there are aspects of the linguistic indicators of deception (e.g. group-mentality) that simply do not occur as frequently in the context of written letters.

Another potentially confounding genre could be the formulaic orders and provisions scattered throughout the lowest and mid-range documents. This pattern points to a broader strategy: exclude documents with multiple duplicate or near-duplicate copies from consideration in a corporate deception and fraud setting.

Lastly, having used the external document as my "hot" documents in this research was non-standard for the vector method model (or for that matter any document classification or matching techniques). Though this choice seemed the best at the time, a selection of "hot" documents as a benchmark could be used in the future to produce better ranking results. From the results, documents originating from the Tobacco Institute source may be good candidates for this type of more focused deceptive benchmark. Averaging all external documents meant that I was sure to get some non-deceptive documents in the mix that would serves as the benchmark. Moreover, using "hot" documents in the future would be more consistent with any efforts to

identify deceptive documents on a specific topic, rather than indicting an entire industry as deceptive.

A Future Research Direction

The six indicators of deception were used instead of a larger number of corpus counts in an attempt to avoid confounding genre with deception and corporate fraud. I was worried originally that using a large number of document indicators for the document vector would replicate the style or genre of the document and not so much its deceptiveness. The high frequency of letter documents at the non-deceptive side of the spectrum raises the question of whether the use of the six linguistic indicators of deception was not specific enough to extricate deception from document genre. Are the results of this research showing a genre-based ranking that happens to coincide with deceptiveness, or do these results demonstrate a deception-based automatic ranking of documents? The answer is probably some of both. A resolution to this question points to future research possibilities.

As a whole, though, these results are promising. Written documents have clear potential indicators of deception and corporate fraud. The proposed indicators of deception and corporate fraud can be used to rank documents for potential deception and corporate fraud. Although this kind of methodology will never be able to replace a detailed examination of documents by an expert, when faced with a huge number of documents like the 7 million tobacco industry documents, the methods demonstrated in this dissertation could be of use in whittling down a population to a target number of most likely deceptive documents. The crucial link that this research never intends to provide is the final analysis of whether something is deceptive or not. That step is for experts and investigators.

These results support the value of pursuing research that integrates cross-field methods and observations (e.g. discourse analysis, corpus linguistics and deception research areas). The linguistic indicators of deception were developed from the deception research literature (composed diversely of case studies as well as experimental psychological and linguistic research). They are not comprehensive and have been shown to be related to situations outside of deception. For instance, in the case of deprofiled agency, while this may be a trait of deception, it is also a trait of typical scientific report writing. In this case in particular, using a much larger corpus to investigate deception and language helped me determine that deprofiling agency with high use of passive structures might indeed be linked to deception, but only in limited contexts that do not include scientific report writing.

This is really all to say that iterative linguistic testing – moving experimental or observational results through corpora and discourse analytic frameworks – can only give us more information about the results we are getting. This is particularly important when you consider the potential impact of a "lie-detector" algorithm. As linguists are more often looked to for expert advice on everything from marketing to forensics, it becomes increasingly important that researchers use all the methods available to them to test and retest hypotheses and observations. Linguists must continue to make reporting not only good results, but also disabused hypotheses, a top priority.

REFERENCES

212 FRD 421 DDC (U.S. District Court, District of Columbia 2002).

American Lung Association. (2006). Smoking and African Americans fact sheet [Electronic Version], March. Retrieved Oct 8 from http://www.lungusa.org/site/pp.asp?c=dvLUK9O0E&b=35988.

American Nonsmoker's Rights Foundation. (2006). Municipalities with local 100% smokefree laws. Retrieved September 9, 2006, from http://www.no-smoke.org/pdf/100ordlisttabs.pdf

Argenti, P. A., & Druckenmiller, B. (2004). Reputation and the corporate brand. *Corporate Reputation Review, 6*(4).

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, B. T. S., & Levin, B. (1995). Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography, 8*(2), 85-114.

Austin, J. L. (1962). *How to do things with words*. Cambridge: Harvard University Press.

Axelsson, M. W., & Berglund, Y. (1999, 22-23 April, 1999). *The Uppsala Student English Corpus (USE): A multi-faceted resource for research and course development.* Paper presented at the Language and Computers, Parallel corpora, Parallel worlds: Uppsala University, Sweden.

Barnoya, J., & Glantz, S. (2002). Tobacco industry success in preventing regulation of secondhand smoke in Latin America: the "Latin Project". *Tobacco Control, 11*(4), 305-314.

Barrionuevo, A. (2006, October 2). Fastow a key for plaintiffs in bank suits. *The New York Times*.

Baskett, G. D., & Freedle, R. O. (1974). Aspects of language pragmatics and the social perception of lying. *Journal of Psycholinguistic Research, 3*(2), 117-131.

Berger, C. R., Karol, S. H., & Jordan, J. M. (1989). When a lot of knowledge is a dangerous thing - the Debilitating effects of plan complexity on verbal fluency. *Human Communication Research, 16*(1), 91-119.

Bero, L. A., Glantz, S., & Hong, M. K. (2005). The limits of competing interest disclosures. *Tobacco Control, 14*(2), 118-126.

Bialous, S. A., & Glantz, S. A. (2002). ASHRAE Standard 62: Tobacco industry's influence over national ventilation standards. *Tobacco Control, 11*(4), 315-328.

Bialous, S. A., & Yach, D. (2001). Whose standard is it, anyway? How the tobacco industry determines the International Organization for Standardization (ISO) standards for tobacco and tobacco products. *Tobacco Control, 10*(2), 96-104.

Biber, D. (1988). *Variation Across Speech and Writing*. Unpublished manuscript, Cambridge, England: Cambridge University Press.

Biber, D., Finegan, E., & Atkinson, D. (1994). ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, G. Tottie

& P. Scheider (Eds.), *Creating andUsing English Language Corpora* (pp. 1-14). Amsterdam: Rodopi.

Bible, G. C. (January, 1996). From: Philip Morris (source). Remarks Geoffrey C. Bible Chairman and Chief Executive Officer Philip Morris Companies, Inc. 960409 Meeting with Employees. http://legacy.library.ucsf.edu/tid/bbx18d00.

Breitborde, N. J., Lopes, S. R., Ochs, E., & Nuechterlein, K. H. (2005). Emotional over-involvement, self-efficacy, and recovery from schizophrenia: an understanding of an index of expressed emotion through narrative. *Schizophrenia Bulletin, 31*, 196.

Brown, C., & Rubin, D. L. (2005). Being Frank: Corporate identity construction and the Tobacco Industry's "A Frank Statement to Cigarette Smokers". Paper presented at *SALSA: Symposium about language and society, Austin*, Austin, Texas.

Bruner, J. S. (1990). *Acts of meaning*. Cambridge, Mass.: Harvard University Press.

Bruner, J. S., Postman, L., & Rodrigues, J. (1951). Expectation and the perception of color. *American Journal of Psychology, 64*(2), 216-227.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 6*(3), 203-242.

Buller, D. B., Burgoon, J. K., Buslig, A. L. S., & Roiger, J. F. (1994). Interpersonal deception .8. Further analysis of nonverbal and verbal correlates of equivocation from the Bavelas Et-Al (1990) Research. *Journal of Language and Social Psychology, 13*(4), 396-417.

Buller, D. B., Burgoon, J. K., Buslig, A. L. S., & Roiger, J. F. (1996). Testing interpersonal deception theory: The language of interpersonal deception. *Communication Theory, 6*(3), 268-288.

Buller, D. B., Burgoon, J. K., White, C. H., & Ebesu, A. S. (1994). Interpersonal deception .7. Behavioral profiles of falsification, equivocation, and concealment. *Journal of Language and Social Psychology, 13*(4), 366-395.

Buller, D. B., Stiff, J. B., & Burgoon, J. K. (1996). Behavioral adaptation in deceptive transactions - Fact or fiction: Reply. *Human Communication Research, 22*(4), 589-603.

Burgoon, J. K., Blair, J. P., Qin, T. T., & Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. In *Intelligence and Security Informatics, Proceedings* (Vol. 2665, pp. 91-101).

Burgoon, J. K., & Buller, D. B. (1994). Interpersonal deception .3. Effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior, 18*(2), 155-184.

Burgoon, J. K., Buller, D. B., Dillman, L., & Walther, J. B. (1995). Interpersonal deception .4. Effects of suspicion on perceived communication and nonverbal behavior dynamics. *Human Communication Research, 22*(2), 163-196.

Burgoon, J. K., Buller, D. B., Ebesu, A. S., & Rockwell, P. (1994). Interpersonal deception .5. Accuracy in deception detection. *Communication Monographs, 61*(4), 303-325.

Burgoon, J. K., Buller, D. B., Ebesu, A. S., White, C. H., & Rockwell, P. A. (1996). Testing interpersonal deception theory: Effects of suspicion on communication behaviors and perceptions. *Communication Theory, 6*(3), 243-267.

Burgoon, J. K., Buller, D. B., & Floyd, K. (2001). Does participation affect deception success? A test of the interactivity principle. *Human Communication Research, 27*(4), 503-534.

Burgoon, J. K., Buller, D. B., Floyd, K., & Grandpre, J. (1996). Deceptive realities - Sender, receiver, and observer perspectives in deceptive conversations. *Communication Research, 23*(6), 724-748.

Burgoon, J. K., Buller, D. B., & Guerrero, L. K. (1995). Interpersonal deception .9. Effects of social skill and nonverbal-communication on deception success and detection accuracy. *Journal of Language and Social Psychology, 14*(3), 289-311.

Burgoon, J. K., Buller, D. B., Guerrero, L. K., Afifi, W. A., & Feldman, C. M. (1996). Interpersonal deception .12. Information management dimensions underlying deceptive and truthful messages. *Communication Monographs, 63*(1), 50-69.

Burgoon, J. K., Buller, D. B., White, C. H., Afifi, W., & Buslig, A. L. S. (1999). The role of conversational involvement in deceptive interpersonal interactions. *Personality and Social Psychology Bulletin, 25*(6), 669-685.

Burgoon, J. K., & Qin, T. T. (2006). The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology, 25*(1), 76-96.

Burgoon, J. K., Stern, L. A., & Dillman, L. (1995). *Interpersonal adaptation : Dyadic interaction patterns*. Cambridge; New York: Cambridge University Press.

Burgoon, M., Callister, M., & Hunsaker, F.G. (1994). Patients who deceive: An empirical investigation of patient-physician communication. *Journal of Language and Social Psychology, 13*, 443–468.

Carpenter, R. H. (1981). Stylistic analysis for law enforcement purposes: A case study of a suspect's caution in phrasing answers. *Communication Quarterly, 29*(1), 32-39.

Carroll, A. B., & Buchholtz, A. K. (2006). *Business & society: Ethics and stakeholder management* (6th ed.). Mason, Ohio: Thomson/South-Western.

Carter, S. M., & Chapman, S. (2003). Smoking, disease, and obdurate denial: The Australian tobacco industry in the 1980s. *Tobacco Control, 12*(90003), 23iii-30.

Center for Disease Control. (2002). *Morbidity and mortality weekly report* (No. 51).

Center for Disease Control. (2005). *Morbidity and mortality weekly report* (No. 26).

Chambers, K. L., & Zaragoza, M. S. (2001). Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Memory & Cognition, 29*(8), 1120-1129.

Chambers, S. M. (1995). Age, prior opinions, and peer interactions in opinion restructuring. *Child Development, 66*(1), 178-192.

Chung, C. K., & Pennebaker, J. W. (2005). Assessing quality of life through natural language use: Implications of computerized text analysis. In W. R. Lenderking & D. A. Revicki (Eds.), *Advancing health outcomes research methods and clinical applications* (pp. 79-94). Washington, DC: Degnon Associates.

Cicres i Bosch, J. (2003). *Estudi pilot d'identificació de parlants amb finalitat forense*.Unpublished manuscript, Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada, Barcelona.

Connor, U., Precht, K., & Upton, T. A. (2002). Business English: Learner data from Belgium, Finland and the U.S. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 175-194). Amsterdam: John Benjamins.

Connor, U., & Upton, T. A. (2004). *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam ; Philadelphia: J. Benjamins.

Crookes, G. (1986). Towards a validated analysis of scientific text structure. *Applied Linguistics, 7*(1), 57-70.

Cummings, K. M. (2003). A promise is a promise. *Tobacco Control, 12*(2), 117-118.

Cummings, K. M., Morley, C. P., Horan, J. K., Steger, C., & Leavell, N.-R. (2002). Marketing to America's youth: Evidence from corporate documents. *Tobacco Control, 11 Suppl 1*, I5-17.

Cummings, K. M., Morley, C. P., & Hyland, A. (2002). Failed promises of the cigarette industry and its effect on consumer misperceptions about the health risks of smoking. *Tobacco Control, 11*(90001), 110i-117.

Cummings, K. M., & Pollay, R. W. (2002). Exposing Mr. Butts' tricks of the trade. *Tobacco Control, 11*(90001), 1i-4.

Cutler, D., & Kadiyala, S. (2003). The return to biomedical research: Treatment and behavioral effects. In K. Murphy & R. Topel (Eds.), *Measuring the gains from medical research: An economic approach* (pp. 110-162). Chicago: The University of Chicago Press.

Dearlove, J. V., Bialous, S. A., & Glantz, S. A. (2002). Tobacco industry manipulation of the hospitality industry to maintain smoking in public places. *Tobacco Control, 11*(2), 94-104.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality & Social Psychology Review, 1*, 346-357.

DePaulo, B. M., & Jordan, A. (1982). Age changes in deceiving and detecting deceit. In R. S. Feldman (Ed.), *Development of nonverbal behavior in children* (pp. 151-180). New York: Springer-Verlag.

DePaulo, B. M., Jordan, A., Irvine, A., & Laser, P. S. (1982). Age-Changes in the Detection of Deception. *Child Development, 53*(3), 701-709.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology, 16*(3), 249-267.

DePaulo, B. M., Tang, J., & Stone, J. I. (1987). Physical attractiveness and skill at detecting deception. *Personality and Social Psychology Bulletin, 13*(2), 177-187.

DePaulo, B. M., Wetzel, C., Sternglanz, R. W., & Wilson, M. J. W. (2003). Verbal and nonverbal dynamics of privacy, secrecy, and deceit. *Journal of Social Issues, 59*(2), 391-410.

DePaulo, P. J., & DePaulo, B. M. (1989). Can deception by salespersons and customers be detected through nonverbal behavioral cues. *Journal of Applied Social Psychology, 19*(18), 1552-1577.

Draft report of the thirteenth plenary meeting held in Paris on 850328 and 850329. (Jun 1985). From: International Standards Organization (source). Bates No. 2028456530/2028456557. http://legacy.library.ucsf.edu/tid/dzz14e00

Drope, J., Bialous, S. A., & Glantz, S. A. (2004). Tobacco industry efforts to present ventilation as an alternative to smoke-free environments in North America. *Tobacco Control, 13*(90001), 41i-47.

Drozhashchikh, A. V. (1991). The pragmatic parameters of english passive constructions in scientific prose. *Vestnik Leningradskogo Universiteta, Istoriya-yazykoznanie-literaturovedenie, 46*(3(2)), 107-109.

Dyas, J. D. (2003). Linguistic features of lying under oath: An experimental study of English and French. *Dissertation Abstracts International, A: The Humanities and Social Sciences, 64*(1), 127-A.

Eisenberg, E., & Riley, P. (1988). Organizational symbols and sense making. In G. Goldhaber & G. Barnett (Ed.), *Handbook of organizational communication*. Norwood, NJ: Ablex.

Ekman, P. (1992). Facial expressions of emotion - An old controversy and new findings. *Philosophical transactions of the Royal Society of London series B-Biological sciences, 335*(1273), 63-69.

Ekman, P. (2001). *Telling lies: clues to deceit in the marketplace, politics, and marriage* (3rd ed.). New York: W.W. Norton.

Ekman, P., & Osullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*(9), 913-920.

Eriksen, M. P. (2000). Best practices for comprehensive tobacco control programs: opportunities for managed care organizations. *Tobacco Control, 9*(90001), 11i-14.

Fairclough, N. (1989). *Language and power*. London ; New York: Longman.

Fairclough, N. (2003). *Analysing discourse : textual analysis for social research*. London ; New York: Routledge.

Farrelly, M. C., Healton, C. G., Davis, K. C., Messeri, P., Hersey, J. C., & Haviland, M. L. (2002). Getting to the truth: Evaluating national tobacco countermarketing campaigns. *American Journal of Public Health, 92*(6), 901-907.

Forsberg, B. (2003). Anti-smoking treaty [Electronic Version]. *Radio Netherlands* from http://www.radionetherlands.nl/features/development/030527tobacco.html.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval : data structures & algorithms*. Englewood Cliffs, N.J.: Prentice Hall.

Francis, W. N. (1964). *A standard sample of present-day English for use with digital computers*. Providence, RI: Brown University.

Francis, W. N., & Kučera, H. (1979). *Brown corpus manual. Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers*: Department of Linguistics, Brown University.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*: Houghton Mifflin.

Frank, M. G., & Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research, 31*(1), 58-75.

Frank, M. G., Paolantonio, N., Feeley, T. H., & Servoss, T. J. (2004). Individual and small group accuracy in judging truthful and deceptive communication. *Group Decision and Negotiation, 13*(1), 45-59.

Fuller, C., Biros, D. P., Adkins, M., Burgoon, J. K., Nunamaker, J. F., & Coulon, S. (2006). Detecting deception in person-of-interest statements. In *Intelligence and security informatics, Proceedings* (Vol. 3975, pp. 504-509).

Gaisch, H. (20 Jul 1983). From: Fabriques De Tabac Reunies S. A. (source). Epidemology. Bates No. 1000134454/1000134457. http://legacy.library.ucsf.edu/tid/fht87e00

Gaisch, H. W. (6 Nov 1986a). From: Fabriques De Tabac Reunies S. A. (source). Appeal Iso. Bates No. 2028456398. http://legacy.library.ucsf.edu/tid/jcn56e00

Gaisch, H. W. (1 Oct 1986b). From: Fabriques De Tabac Reunies S. A. (source). The Iso Meeting in Turkey. Bates No. 2028457027/2028457033. http://legacy.library.ucsf.edu/tid/szm56e00

Galasiânski, D. (2000). *The language of deception : a discourse analytical study*. Thousand Oaks, Calif.: Sage Publications.

Gee, J. P. (2005). *An introduction to discourse analysis : theory and method* (2nd ed.). New York: Routledge.

George, J. F., Marett, K., Burgoon, J. K., Crews, J. M., Cao, J., Lin, M., et al. (2004). *Training to detect deception: An experimental investigation.* Paper presented at the 37th Annual Hawaii International Conference on System Sciences (HICSS 2004), Big Island, Hawaii.

Givel, M. S., & Glantz, S. A. (2001). Tobacco lobby political influence on US state legislatures in the 1990s. *Tobacco Control, 10*(2), 124-134.

Glantz, S. A. (1996). *The cigarette papers*. Berkeley: University of California Press.

Goodridge. (3 Jun 1993). From: American Tobacco Company (source). Monthly Returned Goods Report - May 1993. Bates No. 970242296. http://legacy.library.ucsf.edu/tid/zoq35f00

Goodridge. (4 Nov 1994). From: American Tobacco Company (source). Monthly Returned Goods Report - October 1994. Bates No. 970216518/970216521. http://legacy.library.ucsf.edu/tid/hdu35f00

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.

Habermas, J. (1985). *Der philosophische Diskurs der Moderne : zwèolf Vorlesungen* (1. Aufl. ed.). Frankfurt am Main: Suhrkamp.

Hafez, N., & Ling, P. M. (2006). Finding the Kool Mixx: How Brown & Williamson used music marketing to sell cigarettes. *Tobacco Control, 15*(5), 359-366.

Harris, F., MacKintosh, A. M., Anderson, S., Hastings, G., Borland, R., Fong, G. T., et al. (2006). Effects of the 2003 advertising/promotion ban in the United Kingdom on awareness of tobacco marketing: findings from the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control, 15*(suppl_3), iii26-33.

Harrison, A. A., Hwalek, M., Raney, D. F., & Fritz, J. G. (1978). Cues to deception in an interview situation. *Social Psychology, 41*(2), 156-161.

Heller, A. (4 May 1976). From: Brown and Williamson (source). Report of Telephone Conversation with Client 760503. Bates No. 680112152. http://legacy.library.ucsf.edu/tid/vvb50f00

Hess, J. E. (1997). *Interviewing and interrogation for law enforcement*. Cincinnati, OH: Anderson Pub. Co.

Hirschhorn, N. (2004). Corporate social responsibility and the tobacco industry: hope or hype? *Tobacco Control, 13*(4), 447-453.

Hyland, K. (2000). Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts. *Language Awareness, 9*(4), 179-197.

Hyland, K., & Milton, J. (1997). Qualification and certaintly in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183-205.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics, 25*(2), 156-177.

Hymes, D. (1974). *Foundations of sociolinguistics: An ethnographic cell research*. Philadelphia: University of Pennsylvania.

Ilie, C. (1998). The ideological remapping of semantic roles in totalitarian discourse, or, How to paint white roses red. *Discourse & Society, 9*(1), 57-80.

Jespersen, O. (1909-49). *A modern English grammar on historical principles*. Copenhagen: Munksgaard.

Johansson, S., Leech, G. N., & Goodluck, H. (1978). *Manual of information to accompany The Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo, Norway: Department of English, University of Oslo.

John, F. Q. (1989). Moral theory and defective tobacco advertising and warnings (the business ethics of Cipollone v. Liggett Group). *Journal of Business Ethics, V8*(11), 831-840.

Kahn, N. (2004). *Will/Shall* and b*e going to* and future in English. *Journal of Indian Education*(May), 89-97.

Katz, J. E. (2005). Individual rights advocacy in tobacco control policies: an assessment and recommendation. *Tobacco Control, 14*(suppl_2), ii31-37.

Kennedy, G. D. (1998). *An introduction to corpus linguistics*. London ; New York: Longman.

United States' Final Proposed Findings of Fact, (2004).

Kiparsky, P., & Kiparsky, C. (1970). Fact. In E. M. Bierwisch & K. Heidolph (Eds.), *Progress in linguistics* (pp. 143-173). The Hague: Mouton.

Klimt, B., & Yang, Y. M. (2004). The Enron corpus: A new dataset for email classification research. In *Machine Learning: Ecml 2004, Proceedings* (Vol. 3201, pp. 217-226).

Knapp, M., Hart, R. P., & Dennis, H. S. (1974). An exploration of deception as a communication construct. *Human Communication Research, 1*(1), 15-29.

Kohut, G., & Segars, A. (1992). The president's letter to stockholders: An examination of corporate communication strategy. *Journal of Business Communication, 29*(1), 7-21.

Koller, V. (2004). Businesswomen and war metaphors: 'Possessive, jealous and pugnacious'? *Journal of Sociolinguistics, 8*(1), 3-22.

Koop, C. E. (1998). The tobacco scandal: where is the outrage? *Tobacco Control, 7*(4), 393-396.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51*, 481-537.

Kornegay, H. (Jan 19XX). From: Tobacco Industry Research Committee (source). --No Title [End of Year Statement from the Chairman of the Tobacco Institute to the United States Tobacco Journal]. Bates No. TIMN0069106/0069108. http://legacy.library.ucsf.edu/tid/sbx92f00

Kozlowski, L. T. (2003). First, tell the truth: a dialogue on human rights, deception, and the use of smokeless tobacco as a substitute for cigarettes. *Tobacco Control, 12*(1), 34-36.

Kozlowski, L. T., & Edwards, B. Q. (2005). "Not safe" is not enough: smokers have a right to know more than there is no safe tobacco product. *Tobacco Control, 14*(suppl_2), ii3-7.

Kretzschmar, W. A., Darwin, C., Brown, C., Rubin, D. L., & Biber, D. (2004). Looking for the smoking gun: Principled sampling in creating the Tobacco Industry Documents Corpus. *Journal of English Linguistics, 32*(1), 31-47.

Kretzschmar, W. A., Meyer, C. F., & Ingegneri, D. (1997). Uses of inferential statistics in corpus studies. In *Corpus-based studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17) Stockholm, May 15-19, 1996.* Ljung, Magnus [Ed], Amsterdam: Rodopi, 1997, pp 167-177.

Labov, W. (1964). Phonological correlates of socials-stratification. *American Anthropologist, 66*(6), 164-176.

Lachowicz, D. (1981). On the use of the passive voice for objectivity, author responsibility and hedging in EST. *Science of Science, 2*(2(6)), 105-115.

Landman, A. (2000). Push or be punished: tobacco industry documents reveal aggression against businesses that discourage tobacco use. *Tobacco Control, 9*(3), 339-346.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Leake. (19 Jul 1982). From: American Tobacco Company (source). Monthly Returned Goods Report - June 1982. Bates No. 990942620/990942622. http://legacy.library.ucsf.edu/tid/oji05f00

Leake. (4 Feb 1986). From: American Tobacco Company (source). Monthly Returned Goods Report - December 1985 - Corrected Copy. Bates No. 950637123/950637124. http://legacy.library.ucsf.edu/tid/cxk80a00

Leavell, N.-R. (1999). The low tar lie. *Tobacco Control, 8*(4), 433-437.

Lebaron, C. D. (1996). Looking for verbal deception in Clarence Thomas's testimony. In S. L. Ragan, D. Bystrom & L. L. Kaid (Eds.), *The lynching of language : gender, politics, and power in the Hill-Thomas hearings* (pp. 113-132). Urbana: University of Illinois Press.

Lee, K., Gilmore, A. B., & Collin, J. (2004). Breaking and re-entering: British American Tobacco in China 1979-2000. *Tobacco Control, 13 Suppl 2*, ii88-95.

Leech, G. N. (1992). 100 million words of English: The British National Corpus (BNC). *Ohak Yonku/Language Research, 28*(1), 1-13.

Legacy Tobacco Documents Library. (2005). Announcements: LTDL adds "Bliley" documents. Retrieved Nov 11, 2006, from http://legacy.library.ucsf.edu/announcements_2005.html

Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *The Computer Journal, 41*(8), 537-546.

Lindsay, D. S., Allen, B. P., Chan, J. C. K., & Dahl, L. C. (2004). Eyewitness suggestibility and source similarity: Intrusions of details from one event into memory reports of another event. *Journal of Memory and Language, 50*(1), 96-111.

Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions - A problem for the Match-to-Description Lineup Foil selection strategy. *Law and Human Behavior, 18*(5), 527-541.

Liu, H. (2004). MonyLingua: An end-to-end natural language processor with common sense. Available at: http://web.media.mit.edu/~hugo/montylingua.

Lopes, F. (19 Sep 1986). From: Fabriques De Tabac Reunies S. A. (source). Iso Meeting. Bates No. 2022220257/2022220260. http://legacy.library.ucsf.edu/tid/cux74e00

Lopipero, P., & Bero, L. A. (2006). Tobacco interests or the public interest: 20 years of industry strategies to undermine airline smoking restrictions. *Tobacco Control, 15*(4), 323-332.

MacWhinney, B. & Snow, C. (1990). The child language data exchange system. *ICAME Journal 14*:3-25.

Malone, R. E., & Balbach, E. D. (2000). Tobacco industry documents: treasure trove or quagmire? *Tobacco Control, 9*(3), 334-338.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Manning, M., Rymes, B., Weninger, C., & Brown, C. (2004). *Negotiating discourse flows by telling success stories.* Paper presented at the Graduate Association of Multicultural Studies Conference, Athens, Georgia.

Markov, A., & Last, M. (2005). A simple, structure-sensitive approach for web document classification. In *Lecture notes in computer science : Advances in web intelligence* (pp. 293-298). Berlin / Heidelberg: Springer.

Master, P. (1991). Active verbs with inanimate subjects in scientific prose. *English for Specific Purposes, 10*, 15-33.

McDaniel, P. A., & Malone, R. E. (2005). Understanding Philip Morris's pursuit of US government regulation of tobacco. *Tobacco Control, 14*(3), 193-200.

McDaniel, P. A., Smith, E. A., & Malone, R. E. (2006). Philip Morris's Project Sunrise: weakening tobacco control by working with it. *Tobacco Control, 15*(3), 215-223.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London and NewYork: Routledge.

Mehrabian, A. (1971a). Nonverbal betrayal of feelings. *Journal of Experimental Research in Personality, 5*, 64-73.

Mehrabian, A. (1971b). Verbal and nonverbal interaction of strangers in a waiting situation. *Journal of Experimental Research in Personality, 5*, 127-138.

Meyer, C. F. (2002). *English corpus linguistics : an introduction*. Cambridge, U.K. ; New York: Cambridge University Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1993). Introduction to WordNet: An on-line lexical database.

Moldeveanu, S., & Kulshreshtha, N. (24 Sep 1991). From: Brown and Williamson (source). File Note. DS Scan Results on Several Menthol Cigarettes Note 8/329. Bates No. B01281539/01281545. http://legacy.library.ucsf.edu/tid/xdb11f00

Moldoveanu, S., & Alford, E. (13 Jan 1988). From: Brown and Williamson (source). Comparison of the Paper Recons From Marlboro KS and Eli Cutter/460. Paper recon samples separated from Marlboro KS and Eli Cutter cigarettes, and B&W s production

EBR and PJS paper recons were analyzed by an improved pyrolysis-gas chromatography-mass spectrometry (PY-GC_MS) method. Bates No. 599005829/599005836. http://legacy.library.ucsf.edu/tid/hmi41f00

Moldoveanu, S., Johnson, R., & Burch, T. (14 Nov 1990). From: Brown and Williamson (source). B&W Research & Development File Note. Comparison of Marlboro Full Flavor Cigarettes by Manufacturing Site and by Style - DS Scan Results/327. Bates No. B12913577/12913583. http://legacy.library.ucsf.edu/tid/ecy01f00

Morley, C. P., Cummings, K. M., Hyland, A., Giovino, G. A., & Horan, J. K. (2002). Tobacco Institute lobbying at the state and local levels of government in the 1990s. *Tobacco Control, 11*(90001), 102i-109.

Nevalainen, T., & Raumolin-Brunberg, H. (1996). *Sociolinguistics and langauge history: Studies based on the corpus of Early English correspondence*. Amsterdam: Rodopi.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality & Social Psychology Bulletin, 29*(5), 665-675.

--No Title. (12 Jan 1980). From: Philip Morris (source). Bates No. 1003390563. http://legacy.library.ucsf.edu/tid/aeb64e00

Novotny, T. E., & Carlin, D. (2005). Ethical and legal aspects of global tobacco control. *Tobacco Control, 14*(suppl_2), ii26-30.

Nyklâiécek, I., Temoshok, L., & Vingerhoets, A. J. J. M. (2004). *Emotional expression and health : advances in theory, assessment and clinical applications*. Hove, East Sussex ; New York: Brunner-Routledge.

O'Sullivan, M., Ekman, P., Friesen, W., & Scherer, K. (1985). What you say and how you say it - The contribution of speech content and voice quality to judgments of others. *Journal of Personality and Social Psychology, 48*(1), 54-62.

Oakes, M. P., & NetLibrary Inc. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oakley, E. T. (27 Sep 1976). From: Philip Morris (source). General Analytical Chemistry Section (Monthy Summary) 760830 to 760920. Bates No. 1001511601/1001511604. http://legacy.library.ucsf.edu/tid/lax54e00

Olson, D. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review, 47*, 257-281.

Ong, W. J. (1982). *Orality and Literacy*. London: Methuen.

Partington, A. (1998). *Patterns and meanings : using corpora for English language research and teaching*. Amsterdam ; Philadelphia: J. Benjamins Pub.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC (Version 2nd edition). Mahwah, NJ: Erlbaum.

Pilkington, P., & Gilmore, A. B. (2004). The Living Tomorrow Project: how Philip Morris has used a Belgian tourist attraction to promote ventilation approaches to the control of second hand smoke. *Tobacco Control, 13*(4), 375-378.

Pollay, R. W. (1989). Filter, flavor . . . flim-flam, too!: Cigarette advertising content and its regulation. *Journal of Public Policy & Marketing, 8*, 30-39.

Pollay, R. W. (1990). Propaganda, puffing and the public interest: The scientific smoke screen for cigarettes. *Public Relations Review, 16*(3), 27-42.

Pollay, R. W. (1994). Exposure of US youth to cigarette television advertising in the 1960s. *Tobacco Control, 3*(2), 130-133.

Pollay, R. W. (1995). Targeting tactics in selling smoke: Youthful aspects of 20th century cigarette advertising. *Journal of Marketing Theory and Practice, 3*(1), 1-22.

Pollay, R. W. (2000). Targeting youth and concerned smokers: evidence from Canadian tobacco industry documents. *Tobacco Control, 9*(2), 136-147.

Pollay, R. W., & Dewhirst, T. (2002). The dark side of marketing seemingly "Light" cigarettes: successful images and failed fact. *Tobacco Control, 11* (Suppl 1), I18-31.

Pollay, R. W., & Dewhirst, T. (2003). A Premiere example of the illusion of harm reduction cigarettes in the 1990s. *Tobacco Control, 12*(3), 322-332.

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior, 20*(4), 443-458.

Pryke, S., Lindsay, R. C. L., Dysart, J. E., & Dupuis, P. (2004). Multiple independent identification decisions: A method of calibrating eyewitness identifications. *Journal of Applied Psychology, 89*(1), 73-84.

Qi-bo, Z. 1989. A quantitative look at the Guangzhou Petroleum English Corpus. *ICAME Journal 13*: 28–38

Qin, T. T., & Burgoon, J. K. (2005). An empirical study on dynamic effects on deception detection. In *Intelligence and Security Informatics, Proceedings* (Vol. 3495, pp. 597-599).

Raghavan, V. V., & Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 37*(5), 279-287.

Ringold, D. J. (1986). A preliminary investigation of the information content of cigarette advertising: A longitudinal analysis. *Advances in Consumer Research, 14*, 269-273.

Rubin, D. L., Hirschhorn, N., & Turner, L. J. (2002). *Manipulative language strategies in tobacco industry media statements.* Paper presented at the National Conference on Tobacco or Health, Boston, MA.

Rymes, B. (1995). The construction of moral agency in the narratives of high-school drop-outs. *Discourse & Society, 6*(4), 495-516.

Salton, G. (1980). Automatic term class construction using relevance - Summary of work in automatic pseudocClassification. *Information Processing & Management, 16*(1), 1-15.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*, 613-620.

Schiappa, E., Gregg, P. B., & Hewes, D. E. (2004). Can a television series change attitudes about death? A study of college students and *Six Feet Under*. *Death Studies, 28*, 459-474.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, England: Cambridge University.

Seeman, M., Seeman, T., & Sayles, M. (1985). Social networks and health status: A longitudinal analysis. *Social Psychology Quarterly, 48*, 237-248.

Shuy, R. W. (1998). *The language of confession, interrogation and deception*. Thousand Oaks: Sage Publications.

Shuy, R. W. (2003a). *Tobaccospeak: Image repair as a variety of American English.* Paper presented at the American Dialect Society, Atlanta, Georgia.

Sinclair, J. M. (1987). *Looking up : An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary*. London: Collins ELT.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (2004). *How to use corpora in language teaching*. Amsterdam ; Philadelphia: J. Benjamins.

Sloan, F. A., Mathews, C. A., & Trogdon, J. G. (2004). Impacts of the Master Settlement Agreement on the tobacco industry. *Tobacco Control, 13*(4), 356-361.

Sly, D. F., Hopkins, R. S., Trapido, E., & Ray, S. (2001). Influence of a counteradvertising media campaign on initiation of smoking: The Florida "truth" campaign. *American Journal of Public Health, 91*(2), 233-238.

Spence, S. A., Hunter, M. D., Farrow, T. F. D., Green, R. D., Leung, D. H., Hughes, C. J., et al. (2004). A cognitive neurobiological account of deception: evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 359*(1451), 1755-1762.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology, 20*(4), 421-446.

Stamatatos, E., Kokkinakis, G., & Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471-495.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 27*(5), 523-540.

Stubbs, M. (1996). *Text and corpus analysis : Computer-assisted studies of language and culture*. Oxford, UK ; Cambridge, Mass.: Blackwell.

Stubbs, M. (2001). *Words and phrases : Corpus studies of lexical semantics*. Oxford ; Malden, Mass.: Blackwell Publishers.

Subramanian, R., Insley, R. G., & Blackwell, R. D. (1993). Performance and readability: A comparison of annual reports of profitable and unprofitable companies. *Journal of Business Communication, 30*(1), 49-62.

Szymanczyk, M. E. (Jan 1996). From: Philip Morris (source). Remarks by Michael E. Szymanczyk Executive Vice President, Sales and Marketing Philip Morris U.S.A. Action against Access. Bates No. 2062906065/2062906069. http://legacy.library.ucsf.edu/tid/hct49e00

Teague, C. A. (2 Feb 1953). From: Council for Tobacco Research (source). Survey of Cancer Research with Emphasis on Possible Carcinogens from Tobacco. Bates No. CTRMN043168/043190. http://legacy.library.ucsf.edu/tid/dpt30a00

Thomas, G., & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics, 12*(4), 365-382.

Thomas, J. (1997). Discourse in the marketplace: The making of meaning in annual reports. *Journal of Business Communication, 34*(1), 47-67.

Thomas, S., & Hawes, T. P. (1995). Reporting verbs in medical journal articles. *English for Specific Purposes, 13*(2), 129-148.

Tobacco Industry Research Committee. (28 Dec 1953). From: Tobacco Industry Research Committee (source). A Frank Statement to the Public by the Makers of Cigarettes. Bates No. 680262216/680262218. http://legacy.library.ucsf.edu/tid/hfz70f00

Todd-Mancillas, W. R., & Kibler, R. J. (1979). A test of concurrent validity for linguistic indices of deception. *Western Journal of Speech Communication, 43*(2), 108-122.

Todd, W. R. (1977). *Linguistic indices of deception as manifested by women: A content analytic study.* Unpublished Ph.D. dissertation, Florida State University.

Tofler, A., & Chapman, S. (2003). "Some convincing arguments to pass back to nervous customers": The role of the tobacco retailer in the Australian tobacco industry's smoker reassurance campaign 1950-1978. *Tobacco Control, 12*(90003), 7iii-12.

Twitchell, D. P., Forsgren, N., Wiers, K., Burgoon, J. K., & Nunamaker, J. F. (2005). Detecting deception in synchronous computer-mediated communication using speech act profiling. In *Intelligence and Security Informatics, Proceedings* (Vol. 3495, pp. 471-478).

Twitchell, D. P., Nunamaker, J. F., & Burgoon, J. K. (2004). Using speech act profiling for deception detection. In *Intelligence and Security Informatics, Proceedings* (Vol. 3073, pp. 403-410).

Van Dijk, T. A. (1993). Principles of Critical Discourse Analysis. *Discourse & Society, 4*(2), 249-283.

Van Dijk, T. A. (2002). Political discourse and political cognition. In *Politics as text and talk: Analytic approaches to political discourse, Chilton, Paul Anthony, & Schaffner, Christina [Eds], Amsterdam: John Benjamins, pp 203-237*.

Verbeke, W., & Bagozzi, R. P. (2002). A situational analysis on how salespeople experience and cope with shame and embarrassment. *Psychology and Marketing, 19*(9), 713-742.

Vrij, A. (2000). *Detecting lies and deceit : The psychology of lying and the implications for professional practice*. Chichester, England ; New York: John Wiley.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*(4), 239-263.

Vrij, A., & Mann, S. (2001). Telling and detecting lies in a high-stake situation: The case of a convicted murderer. *Applied Cognitive Psychology, 15*(2), 187-203.

Wakefield, M., Morley, C., Horan, J. K., & Cummings, K. M. (2002). The cigarette pack as image: new evidence from tobacco industry documents. *Tobacco Control, 11*(90001), 73i-80.

Wekker, H. C. (1976). *The expression of future time in Contemporary British English*. Amsterdam: North Holland.

White, C. H., & Burgoon, J. K. (2001). *Adaptation and communicative design: Patterns of interaction in truthful and deceptive conversations*.

Wong, P. T. P. (2002). Lessons from the Enron debacle: Corporate culture matters! *International Network on Personal Meaning* (February 22, 2002).

Wulinger, G. T. (1984). The constitutional rights of puffery: Commercial speech and the cigarette broadcast advertising ban. *Federal Communications Law Journal, 36*(1).

Yach, D., & Bettcher, D. (2000). Globalisation of tobacco industry influence and new global responses. *Tobacco Control, 9*(2), 206-216.

Yang, H. 1985. The JDEST Computer Corpus of texts in English for science and technology. *ICAME News 9*: 24–25.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*(1), 81-106.

Zhou, L., Burgoon, J. K., & Twitchell, D. P. (2003). A longitudinal analysis of language behavior of deception in e-mail. In *Intelligence and Security Informatics, Proceedings* (Vol. 2665, pp. 102-110).

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T. T., & Nunamaker, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139-165.

Zhou, L., Burgoon, J. K., Zhang, D., & Nunamaker, J. F. (2004). Language dominance in interpersonal deception in computer-mediated communication. *Computers in Human Behavior 20*(3), 381-402.

APPENDIX A

PYTHON COMPUTER PROGRAMS

Adversarial Language Python Script

```
## Rates documents for adversarial language
## Put documents in text format in the same folder as this program.
## This produces a comma delimited text, suitable for transferring into
## an excel spreadsheet.
## draft 2 1.16.05 Cati Brown
## works
## This program counts the instances of specifically chosen adversarial
## language lexemes.

import glob

filelist = glob.glob("*.txt")

for thing in filelist:
          f=open(thing,'r')
          s=f.read()
          war = s.count('war')
          battle = s.count('battl')
          oppose = s.count('oppose')
          defend = s.count('defend ')
          crusade = s.count('crusade')
          campaign = s.count('campaign')
          push = s.count('push')
          agitate = s.count('agitat')
          combat = s.count('combat')
          assault = s.count('assault')
          dogfight = s.count('dogfight')
          duel = s.count('duel')
          war = s.count('war')
          strife = s.count('strife')
          conflict = s.count('conflict ')
          fight = s.count('fight')
          struggle = s.count('struggl')
          attack = s.count('attack')
          assail = s.count('assail')
          hostile = s.count('hostil')
          antag = s.count('antagonism')
          adverselang = battle + war + oppose + defend + crusade + campaign +
push + agitate + combat + assault + dogfight
            print thing + ','+'%2d' % (adverselang)
```

Passive Counter Python Script

```
## Passive_counter
## Cati Brown
## 6.10.05
## In a folder with all the MontyTagger accoutrements plus all your non-
tagged
## text files, this program will open each file and tag it for part of
speech.
## Then, using MontyTagger's extract_info function, it will count the total
## phrases in the text and the total passive phrases.
## The output is a comma-delimited file that lists each file and the total
## number of phrases, plus the total number of passive phrases.
##

from MontyLingua import MontyLingua
m=MontyLingua(0)

import glob
import os
path = os.getcwd()

newfile= open(path+'\\Passive.csv','w')

newfile.write("file,passive,total_phrases")

filelist = glob.glob("*.txt")
for file in filelist:
            print file
          a=open(file,'r')
          b=a.read()
          count = 0
          total = 0
          c=m.tokenize(b)
          d=m.tag_tokenized(c)
          e=m.chunk_tagged(d)
          f=m.extract_info(e)
          g=f['parameterized_predicates']
          for x in g:
                if len(x[0][1])==0:
                    total = total+1
                    count=count

                else:
                    if x[0][1][-1]=='passive_voice':
                        count=count+1
                        total = total +1
                    else:
                        count=count
                        total =total+1
            a.close
            length = len(g)
          towrite= '\n'+file+ ','+ '%2d' % (count)+','+'%3d' % (length)
            newfile.write(towrite)
```

```
newfile.close
```

Cognitive Verb Counter Python Script

```
## Cognitive Verb Counter
## Cati Brown
## June 10, 2005
## Counts the instances of lexical cognitive verbs.
## Put documents in text format in the same folder as this program.
## This produces a comma delimited text, suitable for transferring into
## an excel spreadsheet.

import glob

filelist = glob.glob("*.txt")

print('file, stative, relational')

for thing in filelist:
          f=open(thing,'r')
          s=f.read()
          f.close
          a = s.count('believe')
          #stative verbs
          b = s.count('consider')
          # I accidently left c out
          d = s.count('think')
          e= s.count('like')
          ff= s.count('mind')
          g= s.count('recognize')
          h= s.count('prefer')
          i= s.count('seem')
          j= s.count('doubt')
          k= s.count('abhor')
          l= s.count('adore')
          m= s.count('astonish')
          n= s.count('desire')
          o= s.count('detest')
          p= s.count('dislike')
          q= s.count('feel')
          r= s.count('felt')
          t= s.count('forgive')
          u= s.count('forgave')
          v= s.count('guess')
          w= s.count('hate')
          x= s.count('hear')
          y= s.count('imagine')
          z= s.count('impress')
          aa= s.count('intend')
          bb= s.count('know')
          cc= s.count('knew')
          dd= s.count('love')
          ee= s.count('perceive')
          fff= s.count('please')
          gg= s.count('presuppose')
          hh= s.count('realize')
          ii= s.count('recall')
```

```
            jj= s.count('regard')
            kk= s.count('remember')
            ll= s.count('satisf')
            mm= s.count('see')
            nn= s.count('saw')
            oo= s.count('smell')
            pp= s.count('suppose')
            qq= s.count('taste')
            rr= s.count('thought')
            ss= s.count('understand')
            tt= s.count('understood')
            uu= s.count('want')
            vv= s.count('wish')
            ##relational verbs
            ww= s.count('concern')
            xx= s.count('consist of')
            yy= s.count('consisted of')
            zz= s.count('contain')
            aaa= s.count('depend on')
            bbb= s.count('depended on')
            ccc= s.count('deserve')
            ddd= s.count('involve')
            eee= s.count('lack')
            ffff= s.count('matter')
            ggg= s.count('need')
            hhh= s.count('owe')
            jjj= s.count('own')
            kkk= s.count('possess')
            lll= s.count('require')
            mmm= s.count('resemble')

            nnn= s.count('appear')
            ooo= s.count('become')
            ppp= s.count('appreciate')

stative1=a+b+d+e+ff+g+h+i+j+k+l+m+n+o+p+q+r+t+u+v+w+x+y+z+nnn+ooo+ppp

stative2=aa+bb+cc+dd+ee+ff+gg+hh+ii+jj+kk+ll+mm+nn+oo+pp+qq+rr+ss+tt+uu+vv

relational=ww+xx+yy+zz+aaa+bbb+ccc+ddd+eee+ffff+ggg+hhh+jjj+kkk+lll+mmm
            stative=stative1+stative2
              print thing + ','+'%2d '% (stative)+','+'%3d'%(relational)
```

Group Mentality Python Script

```
## Group Mentality
## Cati Brown
## Draft June 7, 2005
## Counts first person plural pronouns (us, we, our, ours, ourselves)
## and also third person plural pronouns (they, them, their, theirs,
themselves).

## Instructions:
## Put documents in text format in the same folder as this program.
## This produces a comma delimited text, suitable for transferring into
## an excel spreadsheet.
##
## works

import glob

filelist = glob.glob("*.txt")

for thing in filelist:
          f=open(thing,'r')
          s=f.read()
          us = s.count(' us ')
          we = s.count(' we ')
          our = s.count(' our ')
          ours = s.count(' ours ')
          ourse = s.count(' ourselves ')
          they = s.count(' they ')
          them = s.count(' them ')
          their = s.count(' their ')
          theirs = s.count(' theirs ')
          thems = s.count(' themselves ')
          uss = us+we+our+ours+ourse
          themm = they+them+their+theirs+thems
            print thing + ','+'%2d'% (uss) +','+'%3d' % (themm)
```

Ambiguity Python Script

```
## Ambiguity
## Cati Brown
## 6.10.05
## In a folder with all the MontyTagger accoutrements plus all your non-
tagged
## text files, this program will open each file and tag it for part of
speech.
## Then, using MontyTagger's chunk_tagged function, it will count the total
## verb phrases in the text as well as the future verbs
## The output is a comma-delimited file that lists each file.
##

from MontyLingua import MontyLingua
m=MontyLingua(0)

import glob
import os
path = os.getcwd()

newfile= open(path+'\\ambiguity.csv','w')

newfile.write("file,verb_phrase, will, shall, going_to")

filelist = glob.glob("*.txt")
for file in filelist:
            print file
          a=open(file,'r')
          b=a.read()
          a.close
          count = 0
          c=m.tokenize(b)
          d=m.tag_tokenized(c)
          e=m.chunk_tagged(d)
          mvphr = e.count('VX') # counts number of verb phrase tags in
MontyTagger
                                # verb phrases are tagged at the beginning
                                # and the end, so divide this tag
                                # count by 2:
            vphr = mvphr/2
            will = c.count('will')
            shall = c.count ('shall')
            am = c.count('am going to')
            iss = c.count('is going to')
            are = c.count('are going to')
            going = am+iss+are
            towrite= '\n'+file+ ','+ '%2d' % (vphr)+','+'%3d'% (will)+','+
'%4d' % (shall)+','+'%5d'% (going)
            newfile.write(towrite)

newfile.close
```

Allness Python Script

```
## Cati Brown
## May 13, 2005 (Friday)
##
## INSTRUCTIONS:
## Run this file in the same folder as a group of text files (.txt)
## and with all of the MontyTagger files.
## This script will tag each file for part of speech, then count
## the number of superlative adjectives and adverbs, plus the number
## of hyperbolic words mentioned below.
## This script will produce a comma-delimited text file with each filename
## and it's corresponding number of hyperbolic terms.

import glob

import os

from MontyTagger import MontyTagger

m=MontyTagger(0)

path = os.getcwd()
newfile = open(path+'\\hyperbole.csv','w')
newfile.write('file,hyperbole')

filelist = glob.glob("*.txt")

for file in filelist:
    f=open(file,'r')
    t=f.read()
    regular=t.lower()
    s=m.tag(regular)
    a=s.count('always')
    b=s.count('never')
    c=s.count('nobody')
    d=s.count('forever')
    e=s.count('everybody')
    ff=s.count('everyone')
    gg=s.count('JJS')
    hh=s.count('RBS')
    hyperbole = a+b+c+d+e+ff+gg+hh

    towrite= '\n'+file+ ','+ '%2d' % (hyperbole)
    newfile.write(towrite)

newfile.close
```

Six Indicator Dot Product Python Script

```
# Cati Brown
# six_indicator_dot_product
# and web
# Parts adapted from
# http://gpwiki.org/index.php/MathGem:Vector_Operations
# are marked as WEB.
# June 14, 2005
# This one works!
# Notes: you must use this with a comma-delimited file saved in .txt format
# and called "test.txt" which is in the same folder as this program.
# ALSO: this only works for the following format:
# 'file_identifier,num1,num2,num3,num4,num5,num6 /n' etc...
# AND: Don't forget to change w (the vector against which you're measuring
all the other vectors)
# to the appropriate vector.

# WEB v and w are tuples representing 3d vectors
def dot(v, w):
    return v[0]*w[0] + v[1]*w[1] + v[2]*w[2] + v[3]*w[3] + v[4]*w[4] +
v[5]*w[5]

# WEB v is a tuple representing a 3d vector
def normalize(v):
    len = length(v);
    return (v[0] / len, v[1] / len, v[2] / len,v[3] / len, v[4] / len, v[5] /
len)

# WEB v is a tuple representing a 3d vector
def length(v):
    return (v[0]**2 + v[1]**2 + v[2]**2+v[3]**2 + v[4]**2 + v[5]**2) ** 0.5


w=[0.117737053,0.133239,0.220337,0.00478,0.001892,0.034159134]
lenw=length(w)

import os
import math
path = os.getcwd()

newfile= open(path+'\\test.txt','r')
c=newfile.read()
newfile.close

while len(c)> 0:
    first=c.find(',')
    ll=c[first+1:]

    second=ll.find(',')

    alphanum1=ll[:second]
```

```
mm=ll[second+1:]
third=mm.find(',')
alphanum2=mm[:third]


nn=mm[third+1:]
fourth=nn.find(',')
alphanum3=nn[:fourth]


oo=nn[fourth+1:]
fifth=oo.find(',')
alphanum4=oo[:fifth]


pp=oo[fifth+1:]
sixth=pp.find(',')
alphanum5=pp[:sixth]


qq=pp[sixth+1:]
end=qq.find('\n')
alphanum6=qq[:end]


new=c.find('\n')

c=c[new+1:]
num1=float(alphanum1)
num2=float(alphanum2)
num3=float(alphanum3)
num4=float(alphanum4)
num5=float(alphanum5)
num6=float(alphanum6)
tupled=[num1,num2,num3,num4,num5,num6]
lent=length(tupled)
prod =dot(tupled,w)
between= prod/(lent*lenw)
theta=math.acos(between)

degrees = theta*57.2957795
print degrees
```

APPENDIX B

TOBACCO DOCUMENTS

Appeal ISO; Bates number: 2028456398; Angle = 86.88°; (H. W. Gaisch, 1986a)

06-11-1986

APPEAL ISO

The enclosed second draft is the final version which is going to be submitted by Carlo Giarré to the Italians today towards midday. It takes into account the the input based on the meeting with Monital yesterday morning and developed yesterday afternoon in Rome by Stefano Sandri, Carlo Giarré, Bradley Brooks, Francesco Lopes and myself.

Kind regards and best wishes,

Helmut W. Gaisch
SCIENCE and TECHNOLOGY
FTR/PME NEUCHATEL

Copies for information :

EEC : Messrs. Darrah and Brooks

EEMA : Messrs. Dulles, Ware and Robinson

R&D : Messrs. Häusermann and Lopes

Telex: Dear Bob; Bates number: 1003390563; (angle 89.04°) ("--No Title," 1980)

PMASIN RS34662VVVV
RCA DEC 01 0415+
PM RD RCH
ETOEU FTR CH

DR. ROBERT B. SELIGMANN

1.12.1980

DEAR BOB,
AS A POSSIBLE INPUT TO THE DISCUSSION ON TUESDAY - SHOULD THE SUBJECT COME UP - I WOULD LIKE TO RECALL PART OF OUR LONG-TERM RESEARCH CONCEPT WHICH YOU AND DR. HAEUSERMANN HAVE APPROVED (IN YOUR MEETINGS ON COORDINATION WITH THE RICHMOND AND NEUCHATEL ACTIVITIES).

DENITRATION, BE IT ON LAMINA, SHEET OR STEN, IS PART OF A PRODUCT RESERACH CONCEPT OF DEFENSIVE NATURE GEARED PRIMARILY TO POSSIBLE NEEDS OF THE REGION PMENA. THE WHOLE CONCEPT CONCERNS THE REMOVAL OF SMOKE CONSTITUENTS FROM MAINSTREAM AND SIDESTREAM THAT MIGHT GIVE RISE TO ACUTE EMBARRASSMENT AND LOSS OF MARKET SHARE IN SOME MARKETS OF THE REGION. MOST OF THE COMPOUNDS IN QUESTION HAPPEN TO BE NITROGEN DERIVATIVES WHICH ARE EITHER NITRATE OR PROTEIN DERIVED. THEREFORE THE LOGIC EXTENSION OF DENITRATION IS DEPROTEINATION.

THE METHOD WHICH WE HAVE DEVELOPED IN PROJECT PROTAGORAS FOR WHICH WE WHICH WE ARE IN THE PROCESS OF OBTAINING PATENT COVERAGE IS IN PRATICAL TERMS A SIMPLE EXTENSION OF THE NINO PRINCIPLE, USING THE SAME EQUIPMENT. THE MAIN MODIFICATION BEING A COMBINED DIGESTION / EXTRACTION STEP INSTEAD OF THE SIMPLE EXTRACTION IN NINO.

INTEGRATED INTO THE NEW PROCESS IS A NINOMASS DERIVED RE ADDITION STEP OF REACTION FLAVOURS. THIS IS IN ORDER TO MAKE UP FOR THE LOSS OF THOSE AMINO ACIDS WHICH ARE FLAVOUR PRECURSOR.

I AM WRITING THIS TO YOU IN ORDER TO AVOID THAT NINO IS BEING LOOKED AT IN ISOLATION. DR. HAEUSERMANN WILL CERTAINLY BE WILLING TO GIVE YOU FURTHER DETAILS IF SO REQUIRED. I KEPT THE TELEX DELIBERATELY LESS EXPLICIT.

YOUR QUERY SURGEON GENERAL-INDEPENDENT SCIENTIFIC COMMITTE.- THERE HAD BEEN Y VISIT BY DR. STEINFELD AND SOME OTHER AMERICANS TO DISCUSS DIFFERENCES OF OPINION WITH THE INDEPENDENT SCIENTIFIC COMMITTEE. US ATTITUDE IS ALLEGEDLY: THERE ARE NO LESS HAZARDOUS CIGARETTES. UK (ISC) ATTITUDE: LESS HAZARDOUS PRODUCTS ARE INDEED POSSIBLE.

BY THE WAY : RUSSEL AND WALD ARE GOING TO HELP DRAFT THE FORTHCOMING US REPORT. SHOULD BE READY FOR JUNE AND IS SUPPOSED TO CONCENTRATE ON COMPENSATION.

[illegible] I WOULD NOT [illegible]

PLEASE GIVE A COPY TO DR. HAEUSERMANN AND MR. RESNIK ON MY BEHALF.

Ogilvy & Mather Insertion Order; Bates number 689478896/8897; (angle 89.04°)

Ogilvy & Mather
Advertising
[illegible]

INSERTION ORDER

C

Please insert the advertising as listed below for:

		DATE	ORDER NO.	PAGE NO.
ADVERTISER PRODUCT	BROWN & WILLIAMSON RICHLAND	09/12/85	50	1

TO THE PUBLISHER OF:

HERALD NEWS
207 PUCASSET ST
11222700 FALL RIVER MA 02722

AD NUMBER	CAPTION	INS. DATE	ED.	INCHES	SPACE
	*** IMPORTANT ALL DATES ARE WEEK OF ***				
DX1409	PACK & CARTON	09/19/85	E	52.000 4X13 INCHES EXPANDED S.A.U.	

POSITION INSTRUCTIONS
FOLLOWING IS THE POSITION REQUEST FOR THE RICHLAND SCHEDULE --
FIRST CHOICE - BACK OF SECTION (NOT IN FOOD SECTION)
SECOND CHOICE - MAIN NEWS FAR FORWARD
MONDAY THROUGH FRIDAY LEEWAY (WHICHEVER DAY THE POSITIONING REQUEST CAN BEST BE ACHIEVED.

WE REQUEST THAT NEWSPAPERS CAPABLE OF AN EVENING/MORNING SPLIT RUN AD IN THE EVENING EDITION THEN THE FOLLOWING MORNING EDITION
EXAMPLE -- TUESDAY(E)/WEDNESDAY(M)

A MINIMUM SEPARATION OF FOUR (4) PAGES IS REQUIRED BETWEEN A BROWN & WILLIAMSON INSERTION AND A COMPETITIVE TOBACCO AD. DO NOT RUN MORE THAN ONE BROWN & WILLIAMSON PRODUCT IN ANY ONE ISSUE. NO OTHER TOBACCO ADS / OBITUARIES / COMICS / CHILDREN'S FEATURES OR MATERIAL ANTITHETICAL TO USE OF TOBACCO TO APPEAR ON SPREAD. ANTITHETICAL MATERIAL IS TO BE CONSTRUED AS ANY EDITORIAL MATERIAL OR ADVERTISEMENT CONCERNED WITH ANY FORM OF CANCER OR LUNG DISEASES OR ARTICLES OR ADVERTISEMENTS ADVOCATING NONSMOKING. A FULL PAGE 4-COLOR AD IS NOT TO RUN OPPOSITE AD. AVOID CLUTTER ON FRACTIONAL ADS. STRESS THE IMPORTANCE OF POSITIONING AS IT IS ESSENTIAL TO THIS CAMPAIGN.

IF ANY OF THESE TERMS CANNOT BE MET--PLEASE NOTIFY PAT DUBROUSKI AT THIS AGENCY (312) 908-2852.

(CONT)

POI: L.ALTARES

Ogilvy & Mather

- Please check reproduction material against proof to make sure it corresponds in every respect.
- If instruction proof and material for reproduction differ, advise us immediately.
- This insertion order is subject to standard conditions on reverse side of this order.

Member of
AMERICAN ASSOCIATION OF ADVERTISING AGENCIES
SUBJECT TO STANDARD CONDITIONS ON BACK HEREOF

689478896

4043.0785

Printed in U.S.A.

Ogilvy & Mather
Advertising

[illegible] Chicago, Illinois 60611
Telephone: [illegible] • Telex: [illegible] • Cable: Ogilvy Chicago

INSERTION ORDER

C

Please enter / the advertising as listed below for:

		DATE	ORDER NO.	PAGE NO.
ADVERTISER PRODUCT	BROWN & WILLIAMSON RICHLAND	08/12/85	50	2

TO THE PUBLISHER OF:

HERALD NEWS
207 POCASSET ST
FALL RIVER MA 02722

1362700

P NUMBER	CAPTION	INS. DATE	SD	INCHES	SPACE

MATERIAL INSTRUCTIONS

(√) MATERIAL HEREWITH.

( ) THIS IS A REPEAT OF AD WHICH APPEARED IN THE WEEK OF ___/___/___. PICK UP MATERIALS.

(√) HOLD MATERIALS FOR POSSIBLE REPEAT.

SPECIAL INSTRUCTIONS

1-NON-ACB SUBSCRIBERS PLEASE MAIL COMPLETE ISSUE FOR CHECKING PURPOSES TO OGILVY & MATHER / CHECKING DESK / 676 ST. CLAIR / CHICAGO IL 60611.

2-ACB SUBSCRIBERS PLEASE MAIL TEARSHEETS/CHECKING COPIES TO ACB OFFICE IN YOUR GEOGRAPHICAL AREA.

3-PLEASE SEND 6 COPIES OF COMPLETE SECTION THAT OUR AD APPEARED IN FOR CHECKING OF REPRODUCTION TO NANCY ALEJANDRO / OGILVY & MATHER / CHICAGO.

4-MAIL INVOICES AS INSTRUCTED ABOVE OR TO ATTN. B&W MEDIA PAYMENT / OGILVY & MATHER / 676 ST. CLAIR / CHICAGO IL 60611.

SEND CHECKING COPIES TO MR. PAUL WESSEL / BROWN & WILLIAMSON TOBACCO CORP / 1500 BROWN & WILLIAMSON TOWER / LOUISVILLE GALLERIA / LOUISVILLE KY 40202.

ON COUPON ADS ALSO SEND 1 COMPLETE COPY TO MS. NANCY SPRIGGS / BROWN & WILLIAMSON TOBACCO CORP. / LOUISVILLE, KY (ADDRESSED AS ABOVE).

SEND 2 COMPLETE CHECKING COPIES TO MR. BRIAN SULLIVAN / GEER DU BOIS ADVERTISING / 114 FIFTH AVENUE / NEW YORK NY 10011.

Per: L. ALTARES

Ogilvy & Mather
Advertising

• Please [illegible] age for proof to make sure it corresponds in every respect.
• If there [illegible] material for reproduction differ, advise us immediately.
• This insertion order is subject to standard conditions on reverse side of this order.

MEMBER OF
AMERICAN ASSOCIATION OF ADVERTISING AGENCIES
SUBJECT TO STANDARD CONDITIONS ON BACK HEREOF

6894788S7

1045-8788

PRINTED IN U.S.A.

Monthly Returned Goods Report; ATX03_0025429 (angle 83.12°) (Leake, 1982)

MEMORANDUM

The American Tobacco Company

DATE: July 19, 1982

TO: Mr. R. P. Truitt, Executive Vice President

FROM: P. H. Leake, Assistant Research and Development Director

REF: Monthly Returned Goods Report - June 1982

The monthly summary of cigarettes returned from the market during June 1982 and the preceding twelve months is attached.

The total quantity of 59,910,340 cigarettes returned during June is approximately 11.1 million more than reported for May 1982 and 4.0 million less than for June 1981.

Through June 1982, a total of 311,141,480 cigarettes had been returned. This is about 1.9 million less than the 313,024,240 returned during the first six months of 1981.

As of May 31, 1982, there were 18,922,100 cigarettes on hand at the Service Department at the Durham Branch for which credits had not been issued. On June 30, 1982, there were 17,059,040 cigarettes on hand for which credits had not been issued.

The DERINGER Cigars and the major brands of smoking tobacco returned during June are listed below:

Little Cigars	Total Returns	Major Smoking Tobaccos	Total Returns (lbs)
DERINGER	60,100		
		BOURBON BLEND	333
		CUTTY PIPE	56
		FIVE BROTHERS	197
		GENUINE BULL DURHAM	64
		HALF AND HALF	479
		PALADIN BLACKCHERRY	255

The quantity of smoking tobacco returned during June is 1,384 pounds, about 247 pounds less than for May 1982 and 671 pounds less than for June 1981.

Total smoking tobacco returned during 1982 amounted to 8,307 pounds, and for 1981, to 9,651 pounds.

P. H. Leake

AIF:ODL
Attachment

cc: (Please see page 2.)

cc: Mr. V. H. Lougee, III
President and Chief Executive Officer

Mr. T. C. Hays
Executive Vice President

Mr. C. C. Kern, Jr.
Vice President - Manufacture and Leaf

Mr. C. H. Mullen
Vice President - Sales

Mr. T. F. McGuire
Manager - Sales Operations

Mr. W. J. Moore
Advertising Director

Mr. J. H. Wells
General Sales Director

Mr. J. H. Hager
Coordinator - Leaf

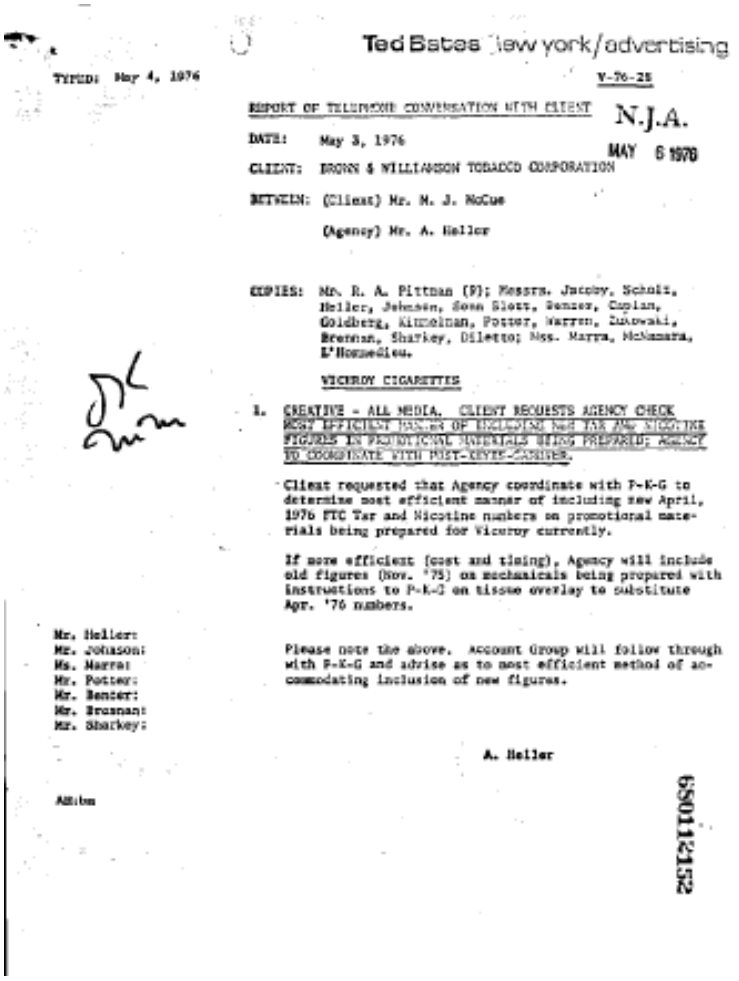
Mr. E. A. Martin
Operations Planning Manager
Department of Manufacture and Leaf

Mr. R. S. Sprinkle, III
Vice President - Research and Development

RETURNED GOODS REPORT
THOUSANDS OF CIGARETTES RETURNED FOR INDICATED MONTH

	1981							1982						Returned Goods Divided By Avg. Sales
	JUNE	JULY	AUG	SEPT	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUNE	%
AMERICAN LIGHTS FILTER 120	26	9	88	23	11	40	35	27	7	35	27	18	18	-
AMERICAN LIGHTS MENTHOL 120	15	25	67	34	23	38	42	78	37	56	49	20	46	-
BULL DURHAM	59	58	115	64	30	69	81	91	56	58	43	50	42	5.46
CARLTON FILTER	4344	3187	1976	2934	5351	5342	4123	3990	3245	4584	3655	2952	3412	1.05
CARLTON FILTER BOX	1708	1101	1391	722	1258	1332	1034	1065	993	1610	1636	1367	1374	29.32
CARLTON 100 FILTER	2979	2032	2365	1504	2384	2632	1997	2035	1315	2181	2252	2123	2283	0.61
CARLTON 100 FILTER BOX	1162	874	985	557	841	900	988	1392	911	1513	1469	853	423	0.57
CARLTON MENTHOL	4830	2801	2904	1705	3065	3291	3772	4580	3600	5289	5030	3888	4749	3.78
CARLTON 100 MENTHOL	2103	1986	2354	1746	3479	3363	1812	1838	1423	2065	2205	1823	2431	1.50
CARLTON 100 MENTHOL BOX	-	-	-	-	-	-	1	95	235	45	130	143	127	39.64
CARLTON 120 FILTER	-	-	-	7	15	41	159	198	437	639	1147	1500	2455	6.31
CARLTON 120 MENTHOL	-	-	-	6	5	38	89	174	415	637	751	759	1577	5.74
HALF AND HALF	177	125	187	122	159	91	125	75	68	56	72	62	64	1.38
HERBERT TAREYTON	843	537	743	531	680	550	564	601	460	743	691	533	624	4.04
ICEBERG 100	264	183	0	154	203	177	129	112	138	354	275	238	226	6.23
LONG JOHNS FILTER 120	60	43	34	61	53	61	66	43	41	35	45	55	33	-
LONG JOHNS FILTER MENTHOL 120	55	35	41	32	45	72	87	121	97	125	125	174	85	-
LUCKY STRIKE	2472	1887	1	1738	2535	2449	2131	2045	1403	2591	2523	3246	2615	0.58
LUCKY STRIKE LOW TAR	-	-	-	-	-	-	0	0	5	0	5	22	30	0.59
LUCKY STRIKE LOW TAR BOX	-	-	-	-	-	-	0	0	5	0	12	7	8	0.17
LUCKY 100	581	696	1064	923	1162	1183	661	491	144	120	608	622	0	0.00
LUCKY 100	863	474	774	475	677	585	516	525	455	712	831	494	591	16.38
MONTCLAIR	196	102	187	113	249	153	189	295	165	184	158	94	133	7.44
PALL MALL	2992	2030	2187	1764	2554	2368	1840	1732	1357	2288	2394	2153	2461	0.13
PALL MALL FILTER KING	2506	1843	2252	1831	2526	2424	1924	1885	1469	2291	2314	1984	2419	12.39
PALL MALL FILTER 100	2180	1552	1728	1277	1847	1727	1248	1386	948	1674	1573	1411	1764	0.27
PALL MALL EXTRA LIGHT	2542	1739	2139	1386	2055	2030	1469	1529	1151	1932	1823	1336	1665	18.50
PALL MALL LIGHT 100	3723	2227	3445	2381	3698	3122	2068	2121	1941	3234	3374	2627	3258	8.83
PALL MALL LIGHT 100 LTM	2610	1807	2119	1751	3054	2860	1892	1863	1681	2711	3088	2448	2876	12.49
SILVA THINS FILTER	2379	1634	1733	1319	1924	1750	1181	1355	1161	1701	1836	1603	2046	1.61
SILVA THINS MENTHOL	2491	1617	1655	1192	2113	2303	1526	1782	1295	2241	2207	1892	2402	0.72
TALL FILTER 120	473	339	532	209	494	466	478	531	299	535	567	463	734	2.60
TALL MENTHOL 120	891	659	786	565	921	860	626	726	496	708	750	553	529	6.25
TAREYTON 85	4739	3148	3099	2304	3427	3396	2632	2488	1948	3348	3384	2919	3430	1.04
TAREYTON 100	2582	1806	1997	1432	2342	2205	1814	1908	1256	2115	1982	1810	2358	0.88
TAREYTON LONG LIGHTS	3531	2452	2569	1760	3378	3471	2636	2652	2305	3336	3167	2611	3832	5.35
TAREYTON LIGHTS	6102	4566	5097	3271	5901	5685	4026	4457	3696	6257	6242	4796	6097	11.87
TAREYTON ULTRA LOW TAR MENTHOL	1261	722	740	600	970	982	814	781	399	881	639	471	633	219.07
RETURNED GOODS - CREDITED	43908	44240	52112	36691	59845	58077	45211	47614	36872	59057	58852	48837	58910	1.34
RETURNED GOODS - NOT CREDITED	7205	14539	9284	17835	12790	11898	5163	13031	17334	18172	14512	18922	17059	
ADJUSTED RETURNED GOODS	48359	51374	47057	45222	54831	57184	38475	55462	41195	59894	55192	51248	58047	

Report of Telephone Conversation with Client 680112152 (Angle 84.35°) (Heller, 1976)



Ted Bates New York/advertising

TYPED: May 4, 1976

V-76-25

REPORT OF TELEPHONE CONVERSATION WITH CLIENT

N.J.A.

MAY 6 1976

DATE: May 3, 1976

CLIENT: BROWN & WILLIAMSON TOBACCO CORPORATION

BETWEEN: (Client) Mr. M. J. McCue

(Agency) Mr. A. Heller

COPIES: Mr. R. A. Pittman (9); Messrs. Jacoby, Scholz, Heller, Johnson, Rose Slott, Benzer, Caplan, Goldberg, Kimmelman, Potter, Warren, Zukowski, Brennan, Sharkey, Diletto; Mss. Marra, McNamara, L'Hommedieu.

VICEROY CIGARETTES

1. CREATIVE - ALL MEDIA. CLIENT REQUESTS AGENCY CHECK MOST EFFICIENT MANNER OF INCLUDING NEW TAR AND NICOTINE FIGURES IN PROMOTIONAL MATERIALS BEING PREPARED; AGENCY TO COORDINATE WITH PUST-KEYES-GARDNER.

Client requested that Agency coordinate with P-K-G to determine most efficient manner of including new April, 1976 FTC Tar and Nicotine numbers on promotional materials being prepared for Viceroy currently.

If more efficient (cost and timing), Agency will include old figures (Nov. '75) on mechanicals being prepared with instructions to P-K-G on tissue overlay to substitute Apr. '76 numbers.

Mr. Heller:
Mr. Johnson:
Ms. Marra:
Mr. Potter:
Mr. Benzer:
Mr. Brennan:
Mr. Sharkey:

Please note the above. Account Group will follow through with P-K-G and advise as to most efficient method of accommodating inclusion of new figures.

A. Heller

AHlbm

680112152

DS Scan Results on Several Menthol Cigarettes, Bates no. B01281539/1545, (angle 41.20°)

BROWN & WILLIAMSON TOBACCO CORPORATION

RESEARCH & DEVELOPMENT

FILE NOTE

TITLE: DS SCAN RESULTS ON SEVERAL MENTHOL CIGARETTES (NOTE [illegible]/339.

AUTHORS: S. C. Moldoveanu, M. F. Kulshreshtha.

DATE: September 24, 1991

CC: A. McMurtrie, B. B. Chakraborty, D. V. Cantrell, R. F. Bender, D. L. Scholcen, R. H. Honeycutt, J. H. Lauterbach, E. F. Litzinger, K. H. Walker, E. A. Conterman, W. E. Conway, J. Y. Tang, Library.

ABSTRACT

Tobacco (total blend) from six cigarette brands: KOOL KS, KOOL MILDS (KS), KOOL XL, Newport Box, Salem KS, and Salem Box was analyzed by DS scan.

The results indicated that Newport Box is not similar to the other menthol brands, except some similarities with KOOL XL. At least part of the Newport peculiarities are due to the paper resin used in these cigarettes.

Salem KS and Salem Box showed a series of differences between them. In the whole set of samples, these two RJR products are somehow closer to KOOL KS or KOOL MILDS.

INTRODUCTION

The purpose of this study was to find differences and similarities between the total blend from six menthol cigarettes: KOOL KS, KOOL MILDS KS, KOOL XL, Newport Box, Salem KS, and Salem Box. (Newport Box and Newport soft cup are supposed to have the same blend). The analyses applied were DS scan and several routine methods. The results of the study are described in this report.

This DS scan study will be continued with additional analyses on the blend components of the same cigarette brands.

RESULTS AND DISCUSSION

Six menthol cigarettes: KOOL KS, KOOL MILDS KS (KOOL MLD), KOOL XL, Newport Box, Salem KS, and Salem Box were analyzed by DS scan, and by several routine procedures.

812 81539

By directly inspecting the DS scan chromatograms there were several findings:

- Most differences in chromatograms are only quantitative.

- The chromatogram for Newport showed certain small peaks which were different from KOOL family or Salem family. The identification of these small peaks (which is still tentative) indicated that they are carbohydrate related compounds. Further work is needed to find if these differences may reveal anything special.

- The only obvious qualitative difference was found in Salem KS which uses hydrogenated corn syrup (sorbitol and maltitol identified in the DS scan).

The quantitative comparison of the DS scan data for the total blend tobacco of the six samples, obtained by using the selected similarity index (SSI) is given in Table 1. The diagonal values of the SSI in Table 1 show the comparison between duplicates of the same resin sample (values above 92% - 93% of the SSI show no practical difference between the samples, or some differences in only very few compounds).

Table 1

Selected Similarity Index for several Menthol Brands

	KOOL KS	KOOL MLD	KOOL XL	Newport	Salem KS	Salem Box
KOOL KS	97.5	91.6	85.8	80.5	80.8	83.6
KOOL MLD		92.8	84.6	82.0	81.7	86.3
KOOL XL			97.6	87.3	75.1	78.6
Newport				95.6	76.5	80.9
Salem KS					92.9	83.4
Salem Box						95.4

Table 1 showed that there are certain overall differences between the six analysed cigarettes (regarding the compounds seen by DS scan). For the KOOL family Table 1 shows as expected good similarity between KOOL KS and KOOL MLD, and lesser similarity with KOOL XL. KOOL XL is more similar to Newport. The RJR cigarettes are closer to KOOL KS than to KOOL XL or Newport. The two Salem cigarettes (KS and Box) showed also significant differences between them.

A more detailed evaluation of the data was required to identify the type of compounds where the differences are more pronounced. The normalized area count values for a series of compounds in the DS scan are given in Table 2. Some routine analytical data which were available are given in Table 3.

Inspecting the data from Table 1, Table 2 and Table 3, several conclusions can be drawn:

- Only KOOL XL and Newport showed signs of synthetic sugar ammonia chemistry. The levels of deoxyfructosazines (DF"s) in these two products was about equal, and double the levels in the other brands. Although KOOL XL contains ESE which has elevated levels of DF's while the Newport paper recon has lower levels of DF's. The presence of higher phosphate level in the two brands (coming from the paper recon) probably determined the higher level of DF's in the total blend.

- The level of reducing sugars in all samples was rather similar except in Newport which was slightly higher.

- The level of sucrose showed a set of differences. KOOL XL had the highest levels, followed by KOOL KS and KOOL MLD (which were close together). Newport had a rather low level of sucrose, but not as low as RJR products.

- Citric, and malic acid levels were shown by DS scan slightly higher in Newport. This may come from its paper recon which contains higher levels of these compounds".

- Lactic acid was not found elevated by DS scan in any brand. However, KOOL MLD has 250 ppm lactic acid added in the top flavor (by use of MOSATE), which was not identified by DS scan. The difference made by this addition, from the levels of lactic acid commonly found in cigarettes, was too small to be measured by DS scan. Possibly formation of dimers of lactic acid may also have occurred, and these are not measured in DS scan.

- Nicotine level in Newport is higher than in other brands.

- The level of menthol (in tobacco) is about equal in KOOL KS, KOOL MLD, and Salem KS. In the same time, KOOL XL, Newport and Salem Box are lower in menthol.

In conclusion, DS scan results (confirmed by routine analyses), showed the following:

- Newport KS tobacco is not similar with other menthol brands, except some similarities with KOOL KL.
- It appears that part of Newport's peculiarities are due to the paper recon used.
- Salem KS and Salem Box showed a series of differences between them. In the whole set of samples, the two RJR products are somehow closer to KOOL KS or KOOL MLD.
- Additional analyses on the blend components of the same cigarette brands is needed for more detailed informations.

N. P. Kulshreshtha

S. C. Moldoveanu

p3f518.wp

Table 2

The Normalised Area Count for a Series of Compounds in DS Scan.

COMPOUNDS	KOOL KS	KOOL MLD	KOOL XL	Newport	Salem KS	Salem Box
1) Propylene Glycol	1.66	1.62	1.15	1.89	1.53	1.51
2) Lactic Acid	0.04	0.04	0.04	0.04	0.04	0.05
3) Glycolic Acid	0.03	0.03	0.04	0.05	0.05	0.04
4) Proline + Sorbic Acid	0.11	0.12	0.13	0.08	0.04	0.00
5) Phosphate	0.38	0.35	0.55	0.46	0.31	0.12
6) Glycerine	6.41	6.74	4.87	9.05	4.53	5.55
7) Phenyl Acetic + Hiscinic Acid	0.28	0.27	0.20	0.25	0.20	0.11
8) Succinic Acid	0.09	0.10	0.09	0.09	0.07	0.07
9) Nicotine	1.70	1.56	1.57	2.14	1.76	1.77
10) Glyceric Acid	0.26	0.24	0.26	0.29	0.25	0.24
11) Menthol	0.64	0.73	0.49	0.51	0.72	0.42
12) Free Malic Acid	1.85	1.93	1.98	2.38	1.78	1.89
13) Pyroglutamic Acid	0.35	0.32	0.41	0.36	0.25	0.28
14) Trihydroxybutanoic Acid(1)	0.17	0.17	0.21	0.19	0.16	0.17
15) Trihydroxybutanoic Acid(2)	0.53	0.52	0.56	0.61	0.46	0.47
16) Arabitol	0.07	0.07	0.08	0.11	0.07	0.06
17) Free Citric Acid	0.14	0.15	0.17	0.18	0.13	0.13
18) Fructose(1)	4.22	4.29	4.28	4.86	3.95	4.45
19) Fructose(2)	2.43	2.65	2.74	2.85	2.42	2.43
20) Fructose(3)	0.12	0.14	0.13	0.16	0.13	0.16
21) Glucose(1)	0.68	0.76	0.73	0.89	0.72	0.79
22) Quinic Acid	1.56	1.45	1.64	1.90	1.30	1.36
23) Glucose + Fructose	2.62	2.90	2.70	2.64	2.48	2.84
24) Iso-Inositol	0.49	0.45	0.39	0.39	0.43	0.81
25) Beta-D-Glucose	3.64	4.07	3.78	3.62	3.50	4.09
26) Palmitic Acid	0.09	0.09	0.09	0.09	0.10	0.10
27) Glucaric Acid	0.23	0.22	0.27	0.19	0.20	0.11
28) Inositol	1.37	1.43	1.48	1.35	1.39	1.51
29) Unknown	0.11	0.11	0.11	0.10	0.11	0.11
30) Unknown	0.14	0.16	0.16	0.17	0.17	0.19
31) al-Cembratrienediol	0.04	0.04	0.04	0.04	0.04	0.05
32) be-Cembratrienediol	0.06	0.06	0.07	0.06	0.06	0.06
33) Sac I.S.	0.34	0.36	0.36	0.35	0.35	0.34
34) Sucrose	2.20	1.81	3.44	1.17	0.96	0.85
35) 2,3-Deoxyfructosazine	0.11	0.09	0.16	0.18	0.07	0.08
36) 2,6-Deoxyfructosazine	0.07	0.06	0.14	0.14	0.05	0.06
37) 2,5-Fructosazine	0.02	0.01	0.04	0.05	0.00	0.01
38) 2,6-Fructosazine	0.03	0.03	0.06	0.04	0.00	0.05
39) Iso-Chlorogenic Acid	0.10	0.10	0.11	0.11	0.12	0.11
40) Chlorogenic Acid	1.01	1.09	0.93	0.94	0.89	1.04

p3ft16.wp

Table 3

Routine Analysis Results

	KOOL [illegible]	KOOL [illegible]	KOOL [illegible]	Newport	Salem [illegible]	Salem [illegible]
Material pH	5.42	5.52	5.49	5.46	5.21	5.40
ANALYTE	X̄ S.D.	X̄ S.D.	X̄ S.D.	X̄ S.D.	X̄ S.D.	X̄ S.D.
Alkaloids	2.25 ± 0.01	2.07 ± 0.05	1.96 ± 0.00	2.49 ± 0.04	2.37 ± 0.00	2.16 ± 0.04
Nitrates	1.22 ± 0.02	1.26 ± 0.00	1.22 ± 0.02	1.26 ± 0.05	1.30 ± 0.06	1.24
Chlorides	0.61 ± 0.00	0.57 ± 0.01	0.56 ± 0.01	0.38 ± 0.00	0.59 ± 0.00	0.58
Ammonia	0.17 ± 0.00	0.18 ± 0.00	0.22 ± 0.00	0.20 ± 0.00	0.17	0.17
Fructose	2.65	3.20	2.79	3.61	3.51	2.88 ± 0.02
Glucose	2.05	3.10	2.31	2.70	2.57	3.05 ± 0.26
Sucrose	1.08	1.75	2.40	0.94	0.79	0.76 ± 0.01
Reducing Sugars	4.70 ± 0.41	5.10 ± 0.17	7.41 ± 0.07	9.10	6.8	7.26 ± 0.13
Total Sugars	10.68 ± 0.15	11.02 ± 0.09	11.76 ± 0.02	11.29	9.8	9.86 ± 0.21
Phosphates	0.68 ± 0.01	0.50 ± 0.02	0.83 ± 0.01	0.76	0.51	0.66
Propylene Glycol	0.97 ± 0.02	0.90 ± 0.02	1.08 ± 0.02	1.14	0.85	0.88 ± 0.01
Glycerin	2.83 ± 0.04	2.16 ± 0.07	1.96 ± 0.06	2.41	1.70	1.58 ± 0.03
Cocoa	0.26 ± 0.00	0.20 ± 0.00	0.59 ± 0.15	-	0.77	0.71
Licorice	-	-	1.28	0.48	0.62	0.59
Menthol	0.62 ± 0.10	0.47 ± 0.02	0.53 ± 0.01	0.36	0.45	0.36 ± 0.01

p0f518.wp

Table 3

Routine Analysis Results

	KOOL FS		KOOL MLD		KOOL M		Newport		Salem KS		Salem Box	
Material pH	5.42		5.52		5.49		5.46		5.31		5.40	
ANALYTE	X	S.D.	X	S.D.	X	S.D.	X	S.D.	X	S.D.	X	S.D.
Alkaloids	2.25 ± 0.01		2.07 ± 0.03		1.96 ± 0.00		2.49 ± 0.04		2.37 ± 0.00		2.18 ± 0.04	
Nitrates	1.33 ± 0.03		1.26 ± 0.00		1.22 ± 0.02		1.26 ± 0.05		1.30 ± 0.04		1.24	
Chlorides	0.61 ± 0.03		0.57 ± 0.01		0.56 ± 0.01		0.38 ± 0.00		0.59 ± 0.00		0.58	
Ammonia	0.17 ± 0.03		0.18 ± 0.00		0.23 ± 0.00		0.29 ± 0.00		0.17		0.17	
Fructose	3.05		3.20		2.79		5.41		3.51		2.80 ± 0.02	
Glucose	2.05		3.10		2.31		2.70		2.57		3.05 ± 0.26	
Sucrose	1.08		1.75		2.40		0.94		0.79		0.76 ± 0.01	
Reducing Sugars	4.75 ± 0.41		5.13 ± 0.17		7.41 ± 0.07		9.15		6.8		7.26 ± 0.13	
Total Sugars	10.68 ± 0.15		11.02 ± 0.09		11.76 ± 0.02		11.29		9.8		9.86 ± 0.21	
Phosphates	0.68 ± 0.01		0.50 ± 0.02		0.83 ± 0.01		0.76		0.51		0.46	
Propylene Glycol	0.97 ± 0.03		0.75 ± 0.02		1.08 ± 0.02		1.14		0.85		0.80 ± 0.01	
Glycerin	2.83 ± 0.04		2.16 ± 0.07		1.96 ± 0.06		2.41		1.70		1.58 ± 0.03	
Cocoa	0.26 ± 0.00		0.20 ± 0.00		0.58 ± 0.15		-		0.77		0.71	
Licorice	-		-		1.28		0.48		0.62		0.59	
Menthol	0.42 ± 0.10		0.47 ± 0.02		0.33 ± 0.01		0.36		0.45		0.36 ± 0.01	

p01518.wp

Burson-marsteller
Press Clipping

Client
Subject
Publication/Date Nieuwsblad Page 2/9 Circulation 298,000
Standaard Circulation 75,000
Gentenaar Page 2/9 Circulation 30,000
24/6/89

KEEP THE AIR INDOORS HEALTHY
Regular Cleaning of Air Conditioning is a Necessity

"Problems with air within buildings resulting from inefficient operation of air conditioning systems can be laid at the door of bad design in 20% of all cases and lack of maintenance in 80% of all cases", state Emilienne Clement and Ray De Koster of De Kobra, a company that specialises in hygiene audits and maintainance of air-conditioning systems.
"Bad design results from failing to adapt the system to alterations within the office, incorrect installation, a filter system whose performance is inadequate and an inappropriate supply of fresh air, by for example obtaining such so-called 'fresh air' from underground car-parks."

Picture Caption:
A correctly cleaned ventilation duct.

A bad interior climate can usually be blamed on the air conditioning system. The first part of of the system consists of a filtering system which is intended to remove dust particles from freshly supplied air and recycled air, in as far as this is possible. This filtering can never be 100% effective with the result that there are always dust particles which penetrate to the inside. When the filters are contaminated there is an increase in the difference in pressure across the filter which can result in bursting and the distribution of high concentration of micro-organisms throughout the system and the premises. Periodic replacement of the filter is a solution. A damp filter adjacent to the grill which sucks in air from the outside then proves an ideal source of food which permit micro-organisms to grow. The temperatures in the heating and cooling pipework provide good conditions for the development of heat-loving and spore-forming micro-organisms. Checking and cleaning of pipework as and when required reduces these risks.
A third harmful area is the humidification section, because water and humidifiers are excellent habitats for algae and bacteria. The water within the cask is rich in oxygen and is the same as the ambient temperature of the nearby rooms. When this water is sprayed onto the air which is being driven along, such micro organisms present are transferred to the air. Any additional pollution can be prevented by regularly treating the humidification water with disinfectants or by the use of steam humidifiers.

DUCTWORK SYSTEM
A fourth source of emission is the ductwork system, where as a result of poor maintainence dirt and dust can gather. Problems effecting the interior air may be significantly reduced by means of preventive maintainence of the system. Emilienne Clement has a loathing for systems that can only be reached with difficulty. "To ensure good maintainence it is important to take easy access to the various elements of the system into account during the design stage. Most frequently, maintainence is only considered when problems arise. Designers, managers, employers and doctors need to be better informed about the problems and preventative measures. As regards this it is important to distinguish between checking the technical aspects of the system and maintainence for the purpose of interior airway hygiene."
When examining complaints regarding a building, Ray De Koster advises a systematic analysis of the various symptoms.
"During the first phase we gather together as much data as possible about the complaints on health and comfort submitted by the workforce or residents, taking into account their activities out of the workplace, their motivation and psychological factors etc. The questionnaire provides useful information and permits one to pick out the critical areas within the building."
Simultaneously, information is gathered about the situation of the system and its working, the construction materials used for the building, the quality of the illumination and the noise levels. The third phase consists of the measurement and analysis of various air hygiene and comfort figures: carbon monoxide, carbon dioxide, micro-organisms, temperature and humidity. After assessing the various factors an action plan is drawn up. The possible solutions include adapting the ventilation system or a programme of cleansing and a disinfection or maintainence programme."

NEW PROCEDURE
At the end of June Ray De Koster introduced a new procedure for cleaning air conditioning systems. As cleaning ductwork was until recently a highly labour intensive and therefore expensive exercise, a license in an American system was purchased, enabling more rapid and efficient working.
"In the past cleaning required a great deal of preparatory work", Emilienne Clement explains. "Frequently the false ceiling has to be broken open. Every 4 metres there is a cleaning aperture measuring 300mm by 300mm (1 foot x 1 foot). Revolving brushes loosen the dirt and dust which is sucked up by a vacuum cleaner. The area in which our staff are working must be separated by means of a plastic cover from the surrounding area because when the brushes are removed from the ducting a large quantity of the contamination is introduced into the rooms. By this method 32 metres per day may be cleaned and this was very difficult to justify in financial terms. To check for contamination in the ductwork after the passage of time the false ceiling and cleaning apertures must be reopened."
With the new cleaning system, small round 25mm (1 inch) diameter holes are drilled into the ductwork every 8 to 10 metres. These

holes are hermetically sealed when work is concluded and connect directly to the false ceiling. Before cleaning commences the ventilation duct-work is visually inspected by means of endoscopic techniques. This checking permits an estimation of the quantity of accumulated dirt and if required dust samples may be analysed by micro-biological methods.

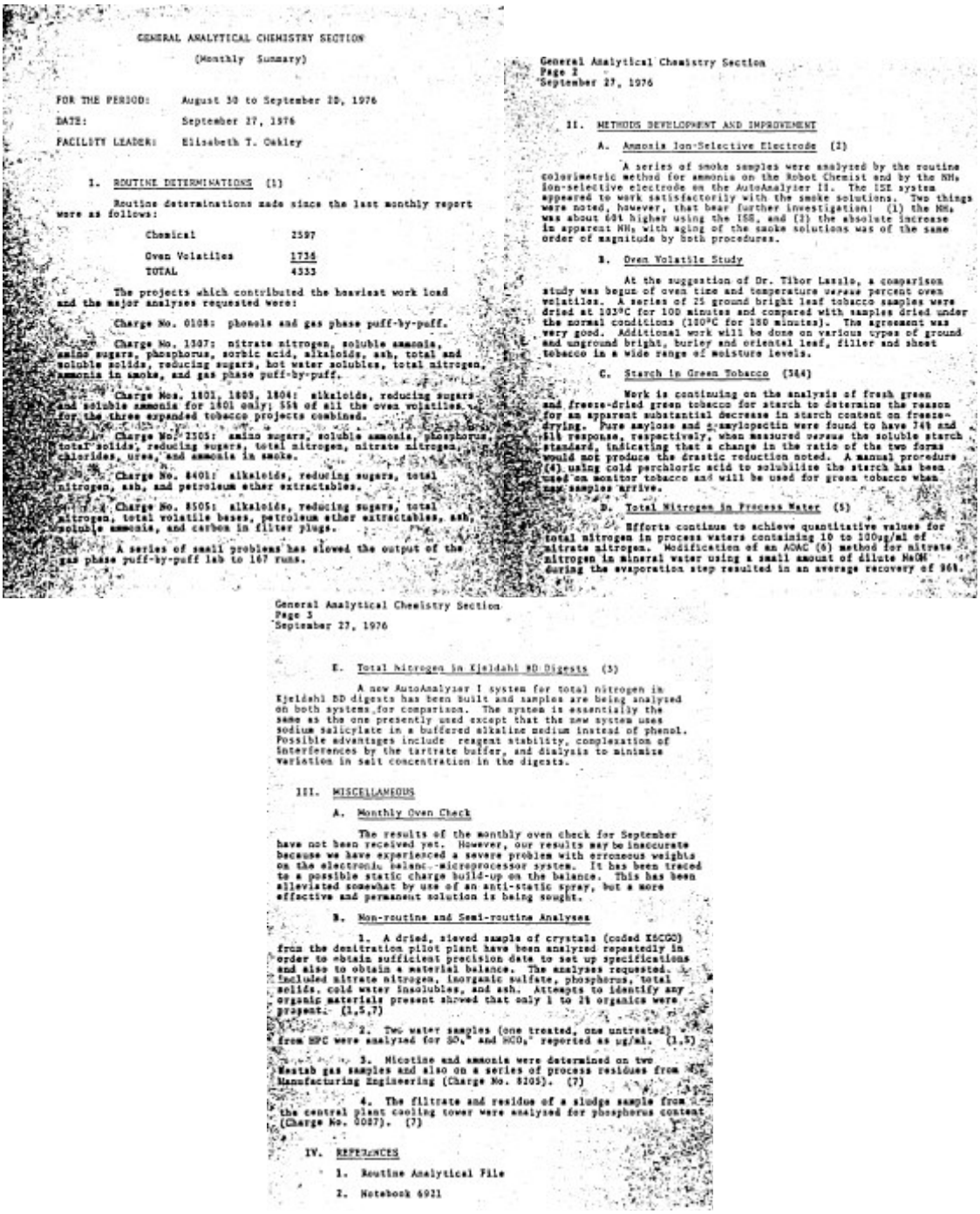
Picture Caption:
The new American cleansing system uses equipment which makes the danger of contamination of surrounding areas almost impossible.

PHASES
Following this the ventilation plant; ductwork system, plenums batteries and convectors are cleansed. The ductwork is cleaned in phases. After a length of ductwork of between 30 and 50 metres is shut off, the vacuum equipment is connected to a grill. A piece of equipment driven by compressed air is introduced into the ductwork by way of the access opening and this blows and sweeps the dirt away, the dirt is collected and filtered out in the vacuum equipment. The extraction air is returned to the atmosphere following filtration (99.67% return on filtration). During cleaning the ductwork is under negative pressure and as a result, Emilienne Clement states, there is no danger of contaminating the surrounding areas. Where heating and cooling batteries are involved a choice is made between various concentrations of grease removing and disinfecting products, according to the degree of contamination. After permitting the chemicals to work in the batteries are cleaned with a high pressure cleaner.
"This new cleansing method was developed in the United States after many years of research, and when compared to previous methods it could be called a revolutionary method. When work is in progress the building does not have to be taken out of service. The cleansing equipment works by itself and reaches all valves, ductwork surfaces and sides, irrespective of size or type. Insulated or fibreglass ductwork is just as easily cleaned as non-insulated ductwork. The most important factor is the elimination of the danger of contamination from the environment."

Ventilation systems are usually easily accessible so that cleaning does not pose any problems. With regular checking, cleaning and disinfecting contamination and pollution of ductwork avoided.

Reporter: Colette Demil

General Analytical Chemistry Section (Monthly Summary) (Oakley, 1976) (angle 49.25°)



GENERAL ANALYTICAL CHEMISTRY SECTION

(Monthly Summary)

FOR THE PERIOD: August 30 to September 25, 1976

DATE: September 27, 1976

FACILITY LEADER: Elisabeth T. Oakley

I. ROUTINE DETERMINATIONS (1)

Routine determinations made since the last monthly report were as follows:

Chemical	2597
Oven Volatiles	1736
TOTAL	4333

The projects which contributed the heaviest work load and the major analyses requested were:

Charge No. 0108: phenols and gas phase puff-by-puff.

Charge No. 1307: nitrate nitrogen, soluble ammonia, amino sugars, phosphorus, sorbic acid, alkaloids, ash, total and soluble solids, reducing sugars, hot water solubles, total nitrogen, ammonia in smoke, and gas phase puff-by-puff.

Charge Nos. 1801, 1805, 1804: alkaloids, reducing sugars and soluble ammonia for 1801 only; 55% of all the oven volatiles for the three expanded tobacco projects combined.

Charge No. 2105: amino sugars, soluble ammonia, phosphorus, total solids, reducing sugars, total nitrogen, nitrate nitrogen, chlorides, urea, and ammonia in smoke.

Charge No. 8401: alkaloids, reducing sugars, total nitrogen, ash, and petroleum ether extractables.

Charge No. 8505: alkaloids, reducing sugars, total nitrogen, total volatile bases, petroleum ether extractables, ash, soluble ammonia, and carbon in filter plugs.

A series of small problems has slowed the output of the gas phase puff-by-puff lab to 167 runs.

General Analytical Chemistry Section
Page 2
September 27, 1976

II. METHODS DEVELOPMENT AND IMPROVEMENT

A. Ammonia Ion-Selective Electrode (2)

A series of smoke samples were analyzed by the routine colorimetric method for ammonia on the Robot Chemist and by the NH_4 ion-selective electrode on the AutoAnalyzer II. The ISE system appeared to work satisfactorily with the smoke solutions. Two things were noted, however, that bear further investigation: (1) the NH_4 was about 6% higher using the ISE, and (2) the absolute increase in apparent NH_4 with aging of the smoke solutions was of the same order of magnitude by both procedures.

B. Oven Volatile Study

At the suggestion of Dr. Tibor Lassiu, a comparison study was begun of oven time and temperature versus percent oven volatiles. A series of 2% ground bright leaf tobacco samples were dried at 103°C for 100 minutes and compared with samples dried under the normal conditions (100°C for 180 minutes). The agreement was very good. Additional work will be done on various types of ground and unground bright, burley and oriental leaf, filler and sheet tobacco in a wide range of moisture levels.

C. Starch in Green Tobacco (344)

Work is continuing on the analysis of fresh green and freeze-dried green tobacco for starch to determine the reason for an apparent substantial decrease in starch content on freeze-drying. Pure amylose and amylopectin were found to have 74% and 41% response, respectively, when measured versus the soluble starch standard, indicating that a change in the ratio of the two forms would not produce the drastic reduction noted. A manual procedure (4) using cold perchloric acid to solubilize the starch has been used to monitor tobacco and will be used for green tobacco when new samples arrive.

D. Total Nitrogen in Process Water (5)

Efforts continue to achieve quantitative values for total nitrogen in process waters containing 10 to 100µg/ml of nitrate nitrogen. Modification of an AOAC (6) method for nitrate nitrogen in mineral water using a small amount of dilute NaOH during the evaporation step resulted in an average recovery of 86%.

General Analytical Chemistry Section
Page 3
September 27, 1976

E. Total Nitrogen in Kjeldahl BD Digests (3)

A new AutoAnalyzer I system for total nitrogen in Kjeldahl BD digests has been built and samples are being analyzed on both systems for comparison. The system is essentially the same as the one presently used except that the new system uses sodium salicylate in a buffered alkaline medium instead of phenol. Possible advantages include reagent stability, complexation of interferences by the tartrate buffer, and dialysis to minimize variation in salt concentration in the digests.

III. MISCELLANEOUS

A. Monthly Oven Check

The results of the monthly oven check for September have not been received yet. However, our results may be inaccurate because we have experienced a severe problem with erroneous weights on the electronic Selectomicroprocessor system. It has been traced to a possible static charge build-up on the balance. This has been alleviated somewhat by use of an anti-static spray, but a more effective and permanent solution is being sought.

B. Non-routine and Semi-routine Analyses

1. A dried, sieved sample of crystals (coded K6C00) from the denitration pilot plant have been analyzed repeatedly in order to obtain sufficient precision data to set up specifications and also to obtain a material balance. The analyses requested included nitrate nitrogen, inorganic sulfate, phosphorus, total solids, cold water insolubles, and ash. Attempts to identify any organic materials present showed that only 1 to 2% organics were present. (1,5,7)

2. Two water samples (one treated, one untreated) from HPC were analyzed for $SO_4^=$ and HCO_3^- reported as µg/ml. (1,5)

3. Nicotine and ammonia were determined on two Westab gas samples and also on a series of process residues from Manufacturing Engineering (Charge No. 8305). (7)

4. The filtrate and residue of a sludge sample from the central plant cooling tower were analyzed for phosphorus content (Charge No. 0087). (7)

IV. REFERENCES

1. Routine Analytical File

2. Notebook 6921

Untitled, MNAT00778295 (angle 49.25)

Pub Suggestions
Re Filter Devices - Misc

Mr. L. P. Runyon, Jr.
Assistant Manufacturing and Leaf Director

April 19, 1967

R. M. Irby, Jr., Manager - New Products Division
Research and Development

Your memorandum and attachments of April 6 have been referred to the undersigned. We have reviewed the patent in some detail and have fabricated the device outlined in the patent. This device is based on the principle of the hookah pipe. We are sending to you under separate cover our mock-up model for your information.

It should be pointed out that smoking through this device produces an extremely irritating smoke due to the fact that the smoke particles have time to agglomerate in the void space between puffs and are, therefore, drawn into the mouth on succeeding puffs. Past experience has indicated that larger particles are more irritating. No size specifications were given in the patent, but obviously this device could be made much smaller if so desired. You may recall that Mr. E. S. Harlow demonstrated to Marketing a similar device in the early development stages of WATERFORD.

kad

Bates Number: 664063379/3381, (angle 3.55°)

August 7, 1979

Ms. Dianne Neral
Ted Bates & Co.

Dear Dianne:

This confirms next steps and approvals for KOOL Milds generic and Black campaigns, per our conversation yesterday afternoon.

Generic

1. The second execution "New KOOL Milds 100's" was approved by Mr. Johnston as submitted, with the addition of "TM" and "available in King Size" positioned at the bottom and to the right of the 100's package.

2. The third ad should be a revised version of ad two, using the same visual. Modifications to be investigated should include the addition of body copy, similar to ad one, positioned to the right of the 100's pack; and a revised subhead (e.g., "Taste Menthol Mist").

3. The third execution "Couple on Rock" has been rejected. We would prefer to introduce people into this campaign when they are shot both on location and with recognizable product involvement. Now that we are committed to incorporating people into this campaign, we would rather wait to do it in conjunction with the upcoming shoot. In this way, executions can benefit from thoughtful preproduction planning regarding the attitude and involvement of the smokers.

4. Please re-examine the chromes which were presented in Louisville. Recommend a third visual which emphasizes the waterfall and mist in a tone and manner similar to the two approved visuals. Frank will be sending you a set of guidelines for determining the appropriate amount of emphasis. Yes, this is very judgmental, but for now the rushing water and mist should be the two dominant elements in the visual. This third visual will be held in reserve, pending the outcome of the upcoming shoot and depending on our timing requirements.

Black

Magazine

1. "Tuxedo" was approved as submitted. Please recommend whether black or white type will read better in the body copy.

2. Ad three should be "Yellow Dress Sitting." It was approved as is.

3. Ad four will be "Hand on Sweater," approved as is. Proceed to finished materials on these three ads.

4. Other comps were approved with one exception. However, no work should proceed until directed. "Edge Sweater, Mustached Man Staring at Camera" has been held up. Review chromes of this couple/situation to see if there is one with the man not staring at the camera. Let us know what you find and give us a recommendation.

5. Recommend an allocation schedule for all ads assuming that we will proceed with those which have been comped and approved.

Outdoor

1. Criterion executions should generally follow the magazine schedule. Please be aware of the difficulties in translating magazine to out-of-home, particularly with respect to copy legibility and to registration of complex visuals. Recommend changes in layout, if any, to make these executions read.

2. For all out-of-home recommend a color for the copy. Day-glo, yellow, reverse white, other?

3. 6-Sheet executions approved were:

   - "Night People" as next execution
   - "Winged Hat" as third execution

4. The 6-Sheet version of "Umbrella" has been held up pending our ability to prove the legibility of copy and visual elements.

5. Send us art for all out-of-home as soon as it becomes available so we may proceed with production.

6. Send criterion comps and dupes as soon as they are available.

If this direction is unclear, give me a call.

Sincerely,

I. R. Holmes, Jr.

IRH/tc

cc: F. E. McKeown
G. T. Reid
P. C. Weinsteiner (Bates)

Bates Number: 2048908735/8737, (angle 6.67°)

LAW OFFICES
SHOOK, HARDY & BACON
ONE KANSAS CITY PLACE
1200 MAIN STREET
KANSAS CITY, MISSOURI 64105
(816) 474-6550

September 13, 1989

FEDERAL EXPRESS

Mr. Charles Gielen
Nauta Van Haersolte
1016 GB Amsterdam
Keizersgracht 384
1001 EW Amsterdam
Postbus 10880
Netherlands

Re: Stichting Sigaretten Industrie v. Stichting Volksgezondheid en Roken

Dear Charles:

In response to your letter of August 22, 1989 and pursuant to our telephone conversation on September 8, 1989, I have enclosed the following materials:

1. The Tobacco Institute's reply to the 1986 Surgeon General's Report.

2. Other criticisms of the 1986 Surgeon General's Report.

3. INFOTAB's annual review of the Environmental Tobacco Smoke literature, as of October, 1988. The updated 1989 version will be available shortly and I will provide you with a copy when it is published.

4. A memorandum prepared in our office with respect to the claims made by defendants concerning what happened in the Cipollone case. As indicated in the memorandum, defendant's assertions contain many errors which you could point out in a further brief or at an oral hearing.

In addition, I have enclosed some additional materials from INFOTAB which may be helpful with respect to your earlier request for materials in connection with the EC proposal to restrict the use of trademarks, etc.

2048908735

We hope these materials are helpful and please let us know if there is anything more we can do. We would also appreciate your keeping us advised of the progress of the case, including when a hearing may be scheduled.

Sincerely,

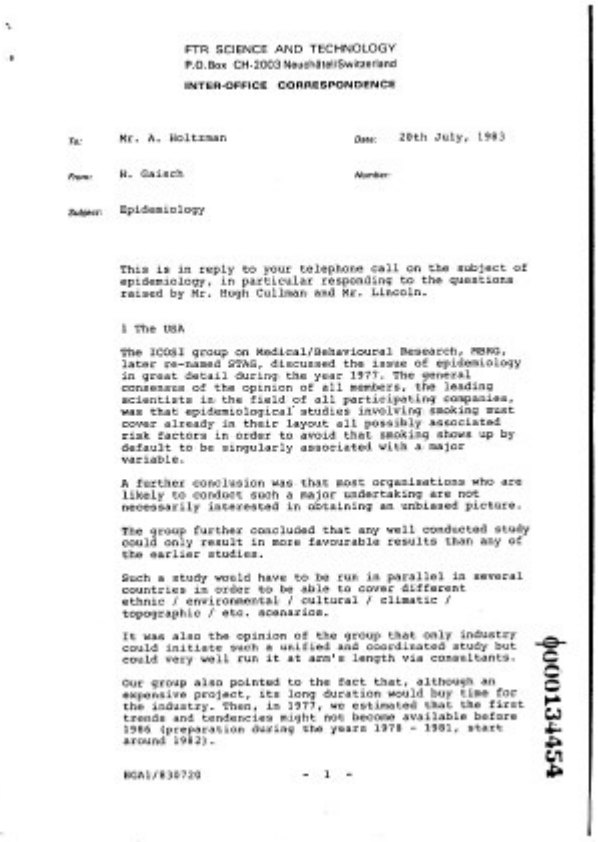
James T. Newsom

JTN/tlb
Enclosures

cc: Donald K. Hoel, Esq. (w/encls.)
London Office

bcc: Steven C. Parrish (w/encls.)

–No Title. 1000134454 (angle 7.44°)



FTR SCIENCE AND TECHNOLOGY
P.O.Box CH-2003 Neuchâtel/Switzerland
INTER-OFFICE CORRESPONDENCE

To: Mr. A. Holtzman Date: 20th July, 1983

From: H. Gaisch Number:

Subject: Epidemiology

This is in reply to your telephone call on the subject of epidemiology, in particular responding to the questions raised by Mr. Hugh Cullman and Mr. Lincoln.

1 The USA

The ICOSI group on Medical/Behavioural Research, MBRG, later re-named SSAS, discussed the issue of epidemiology in great detail during the year 1977. The general consensus of the opinion of all members, the leading scientists in the field of all participating companies, was that epidemiological studies involving smoking must cover already in their layout all possibly associated risk factors in order to avoid that smoking shows up by default to be singularly associated with a major variable.

A further conclusion was that most organisations who are likely to conduct such a major undertaking are not necessarily interested in obtaining an unbiased picture.

The group further concluded that any well conducted study could only result in more favourable results than any of the earlier studies.

Such a study would have to be run in parallel in several countries in order to be able to cover different ethnic / environmental / cultural / climatic / topographic / etc. scenarios.

It was also the opinion of the group that only industry could initiate such a unified and coordinated study but could very well run it at arm's length via consultants.

Our group also pointed to the fact that, although an expensive project, its long duration would buy time for the industry. Then, in 1977, we estimated that the first trends and tendencies might not become available before 1986 (preparation during the years 1978 - 1981, start around 1982).

HGAl/830720 - 1 -

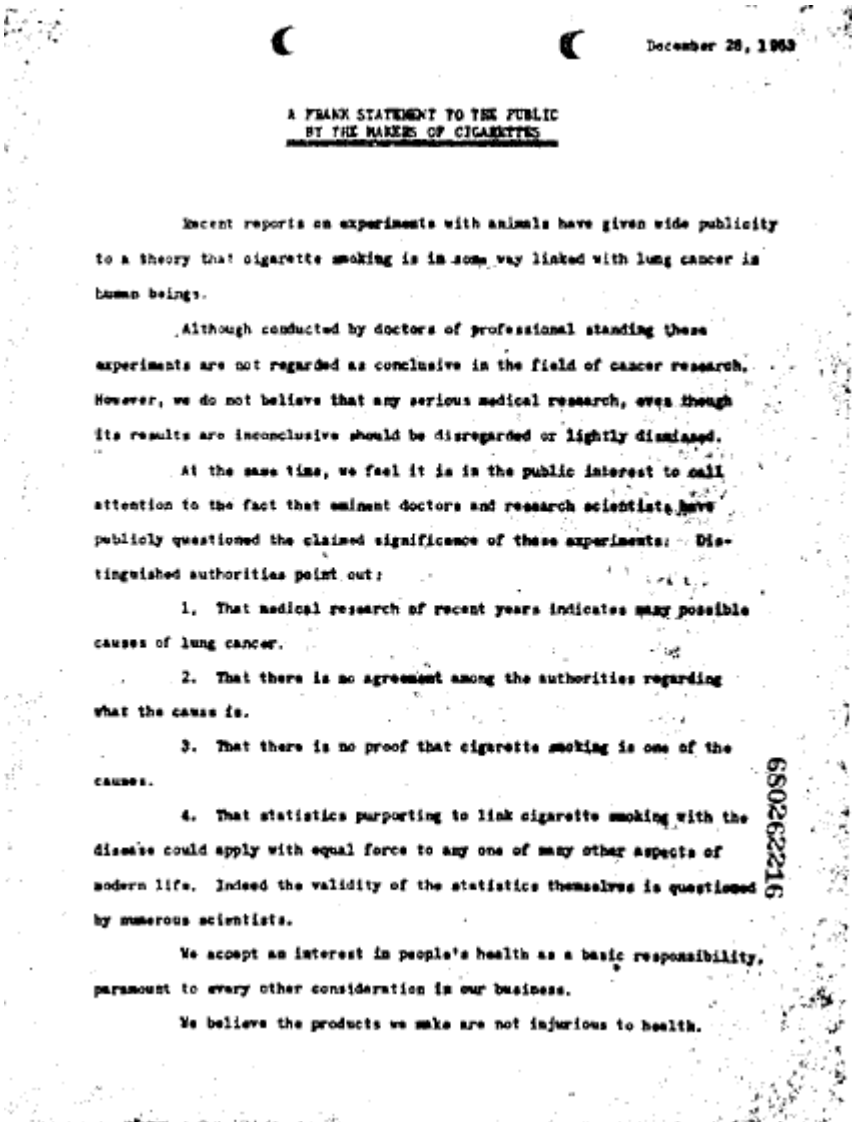
1000134454

Enclosures as mentioned

Distribution list:
(Without enclosures)

cc.:
Mr. A. G. Buzzi
Mr. M. Serrano,
Dr. T. S. Osdene
MCS

A Frank Statement to the Public, 680262216/2218, (angle 5.07)



December 28, 1953

A FRANK STATEMENT TO THE PUBLIC
BY THE MAKERS OF CIGARETTES

Recent reports on experiments with animals have given wide publicity to a theory that cigarette smoking is in some way linked with lung cancer in human beings.

Although conducted by doctors of professional standing these experiments are not regarded as conclusive in the field of cancer research. However, we do not believe that any serious medical research, even though its results are inconclusive should be disregarded or lightly dismissed.

At the same time, we feel it is in the public interest to call attention to the fact that eminent doctors and research scientists have publicly questioned the claimed significance of these experiments. Distinguished authorities point out:

1. That medical research of recent years indicates many possible causes of lung cancer.

2. That there is no agreement among the authorities regarding what the cause is.

3. That there is no proof that cigarette smoking is one of the causes.

4. That statistics purporting to link cigarette smoking with the disease could apply with equal force to any one of many other aspects of modern life. Indeed the validity of the statistics themselves is questioned by numerous scientists.

We accept an interest in people's health as a basic responsibility, paramount to every other consideration in our business.

We believe the products we make are not injurious to health.

680262216

We always have and always will cooperate closely with those whose task it is to safeguard the public health.

For more than 300 years tobacco has given solace, relaxation and enjoyment to mankind. At one time or another during those years critics have held it responsible for practically every disease of the human body. One by one these charges have been abandoned for lack of evidence.

Regardless of the record of the past, the fact that cigarette smoking today should even be suspected as a cause of a serious disease is a matter of deep concern for us.

Many people have asked us what we are doing to meet the public's concern aroused by the recent reports. Here is the answer:

1. We are pledging aid and assistance to the research effort into all phases of tobacco use and health. This joint financial aid will of course be in addition to what is already being contributed by individual companies.

2. For this purpose we are establishing a joint industry group consisting initially of the undersigned. This group will be known as Tobacco Industry Research Committee.

3. In charge of the research activities of the Committee will be a scientist of unimpeachable integrity and national repute. In addition there will be an Advisory Board of scientists disinterested in the cigarette industry. A group of distinguished men from medicine, science and education will be invited to serve

on this Board. These scientists will advise the Committee on its research activities.

This statement is being issued because we believe the people are entitled to know where we stand on this matter and what we intend to do about it.

TOBACCO INDUSTRY RESEARCH COMMITTEE

Sponsors:

American Tobacco Company
Paul M. Hahn, President

Benson & Hedges
Joseph F. Cullman, Jr., Chairman of the Board and President

Brown & Williamson Tobacco Corporation
Timothy V. Hartnett, President

P. Lorillard Company
Herbert A. Kent, Chairman

Philip Morris & Co., Ltd., Inc.
O. Parker McComas, President

R. J. Reynolds Tobacco Company
E. A. Darr, President

Tobacco Associates, Inc.
John B. Matson, President

U. S. Tobacco Company
J. E. Peterson, President

No Title [End of Year Statement from the Chairman of the Tobacco Institute to the United States Tobacco Journal], TIMN0069106 (angle 6.79°)

011110

(Year end edition--Kornegay) about 500 words

Webster defines opportunity as "a good chance for advancement or progress." In terms of the capabilities of the scientific community today, the answers to many questions about smoking and health should be well within reach. The opportunities to advance our scientific knowledge should be literally boundless.

But we are faced with never-ending obstructions. We see propaganda of all sorts; attempts constantly are made to mar the image of the tobacco industry, and to destroy its credibility. Through these oceans of dissent we remain staunchly unified. We maintain our commitment in the field of scientific research with determined objectivity and a willingness to cooperate.

In 1969, the Secretary of Health, Education,and Welfare pointed out the great need for cooperation when he said, "I believe the ((tobacco)) industry and government working together offers great promise of finding the answers we need. I am confident our joint effort will yield a cooperative research program which strongly promotes the public interest."

Said a late 1970 Advertising Age editorial:

"The consumer's legitimate need for information...is one of the major irritants operating between business and the public. If it is resolved only on the basis of lawsuits and legislation, the infection will persist. Fortunately, as in the case of the tobacco people, we

TIMN 0069106

011111

are beginning to see affirmative efforts by business which suggest that the outlook is changing."

The outlook is indeed changing. Avoiding what may very well have been years of unnecessary litigation, the industry volunteered to take cigarette commercials off the air. The industry volunteered to list the Federal Trade Commission's "tar"-nicotine test figures in cigarette advertising. And the industry voluntarily put more money into smoking-health research than any other individual agency, government or private.

There has been little cooperation. When the American Cancer Society announced the results of the "smoking-dog" experiment, The Tobacco Institute asked that the detailed underlying data be made available to an expert panel for impartial review, and was refused.

The Surgeon General of the U. S. Public Health Service, an "impartial" observer, stated in public that cigarette smoking is "a dirty, smelly, foul, chronic form of suicide."

The odds sometimes seem stacked against the industry, but, as many seem to agree, the industry is far from its way out. A research analyst for New York based Oppenheimer & Company recently projected that "the cigarette industry is on the eve of fresh growth."

The New Year may introduce many new problems; it may underline many of the old nemeses. But in the end, we believe scientific objectivity will prevail.

TIMN 0069107

One of mankind's inherent instincts is self perpetuation. And if we persevere and continue to search for the facts, we will discover what Mark Twain once called "the most valuable thing we have:" Truth.

###