THE ‘COMPANY’ WORDS KEEP: A CORPUS-BASED ANALYSIS OF COLLOCATIONAL PATTERNING IN BUSINESS TERMINOLOGY

by

MARLENE A. KEMP-DYNIN

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

ABSTRACT

This dissertation focuses on twenty-five business terms and the words with which they co-occur within a span of four words to the left of the term and four to the right. Two corpora were utilized in this study: the Tobacco-Documents Corpus, a corpus of over 500,000 words compiled at the University of Georgia from tobacco industry documents, and a reference corpus of general business language, consisting of over one million words and compiled specifically for the study. The strength of attraction between the terms and their primary collocates was analyzed per corpus using the mutual information statistic. The validity of each collocation was measured by means of the t-score. The results indicate that within the span of 4:4, industry and/or company terms will collocate significantly with general business terms. Furthermore, it was shown that while some common terminology will be shared across industries or companies, there are major differences between general business language and that of specific industries or companies.

INDEX WORDS: Corpus Linguistics, Lexicography, Business Terminology, Collocation, Tobacco-Documents Corpus, Language Variation, English for Specific Purposes, Mutual Information
THE ‘COMPANY’ WORDS KEEP: A CORPUS-BASED ANALYSIS OF COLLOCATIONAL PATTERNING IN BUSINESS TERMINOLOGY

by

MARLENE A. KEMP-DYNIN

B.A., St. Bonaventure University, 1975
M.B.A., Monterey Institute of International Studies, 1992

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

2005
THE ‘COMPANY’ WORDS KEEP: A CORPUS-BASED ANALYSIS OF COLLOCATIONAL
PATTERNING IN BUSINESS TERMINOLOGY

by

MARLENE A. KEMP-DYNIN

Major Professor: William A. Kretzschmar, Jr.
Committee: Marlyse Baptista
Peter A. Jorgensen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005
DEDICATION

To my husband, George Dynin

A truly extraordinary human being

Words cannot express my feelings, not even for a linguist.
ACKNOWLEDGMENTS

There are several people whom I would like to acknowledge for their support during my pursuit of this doctoral degree.

I would like to thank Dr. William Kretzschmar for his invaluable contributions to my education in linguistics and this dissertation. It was he who introduced me to the fields of corpus linguistics and lexicography. In addition, I would like to thank my committee members, Dr. Peter Jorgensen and Dr. Marlyse Baptista, not only for being on my committee, but also for expanding my knowledge of linguistics. I admire all three professors greatly, and I know how fortunate I was to have had them as my committee. They were encouraging, respectful and kind, and I shall always be grateful.

I would like to thank Dr. Douglas Biber of Northern Arizona University for providing me with a copy of his article, ‘Representativeness in Corpus Design.’ I would also like to acknowledge the late Dr. William G. Provost, who was the director of the linguistics program during the majority of my studies. As the first person I met at the University of Georgia, he impacted my decision to enroll at UGA.

On the personal side, I would like to thank some dear friends: Helen Bykat, Lynnette Lang, and Elle Wendelberger. Helen has known me for several years and has been so supportive and encouraging, always calling to check on me and tell me how proud of me she is. I hope she knows how much it has meant to me. My first contact with Lynnette, who is the administrative assistant in the Linguistics Program, was on-line as I was applying for admission. I appreciate
her technical help, but even more her friendship. I met Elle in Dr. Kretzschmar’s lexicography course, and we became colleagues and friends. I am truly grateful for her friendship and encouragement.

Finally, I would like to acknowledge my brother, Stephen, and my sister, Cheryl, who were the original computer experts. I’m not sure they ever thought that I would be so entrenched in computer-related endeavors. I wish to acknowledge my late parents, Clyde H. and Victoria Walsh Kemp, who not only provided me with intelligence, but also taught me never to do a half-way job.
# Table of Contents

| Acknowledgments                                                                 | v |
| List of Tables                                                                 | xi |

## Chapter

1. **Introduction**
   1.1 **Background of the Study**
   1.2 **Problem Statement**
   1.3 **Significance of the Study**
   1.4 **Overview of the Methodology**
   1.5 **Delimitations of the Study**
   1.6 **Organization of the Dissertation**

2. **Corpus Linguistics**
   2.1 **Early Corpus-based Research**
   2.2 **Computerized Corpora**
   2.3 **Specialized Corpora**
   2.4 **Corpus-based Studies**
   2.5 **Summary**

3. **Words and Collocations**
   3.1 **The Word**
   3.2 **The Collocation**
3.3 TERMINOLOGY ...................................................................................................52
3.4 SUMMARY ..........................................................................................................53

4 METHODOLOGY ............................................................................................................54
4.1 METHODOLOGY ..................................................................................................56

5 ANALYSIS OF THE TERMS ..............................................................................................69
5.1 PRODUCT ............................................................................................................69
5.2 BRAND ...............................................................................................................74
5.3 DATA ..................................................................................................................77
5.4 PRODUCTS ..........................................................................................................81
5.5 BRANDS ..............................................................................................................85
5.6 ADVERTISING .....................................................................................................88
5.7 SHARE ................................................................................................................91
5.8 MARKET .............................................................................................................95
5.9 SALES .................................................................................................................98
5.10 WORK .............................................................................................................101
5.11 COMPANY .......................................................................................................105
5.12 LEVELS ...........................................................................................................108
5.13 INCREASE .......................................................................................................111
5.14 ANALYSIS .......................................................................................................115
5.15 INDUSTRY .......................................................................................................118
5.16 DEVELOPMENT ...............................................................................................122
5.17 PROJECT .........................................................................................................125
5.18 SYSTEM ..........................................................................................................128
LIST OF TABLES

Table 4.1: composition of the reference corpus .................................................................61
Table 4.2: top twenty-five business terms and their rank order per corpus...............................62
Table 4.3: occurrences of terms by business field in the reference corpus...............................63
Table 5.1: mutual information and t-scores from the reference corpus .....................................71
Table 5.2: mutual information and t-scores from the Tobacco-Documents Corpus...............72
Table 5.3: span positions of tobacco-related collocates by frequency and rate of occurrence ......74
Table 5.4: mutual information and t-scores from the reference corpus .....................................75
Table 5.5: mutual information and t-scores from the Tobacco-Documents Corpus................76
Table 5.6: span positions of tobacco-related collocates by frequency and rate of occurrence ......77
Table 5.7: mutual information and t-scores from the reference corpus .....................................78
Table 5.8: mutual information and t-scores from the Tobacco-Documents Corpus................78
Table 5.9: span positions of tobacco-related collocates by frequency and rate of occurrence......81
Table 5.10: mutual information and t-scores from the reference corpus .................................82
Table 5.11: mutual information and t-scores from the Tobacco-Documents Corpus...............84
Table 5.12: span positions of tobacco-related collocates by frequency and rate of occurrence....84
Table 5.13: mutual information and t-scores from the reference corpus .................................86
Table 5.14: mutual information and t-scores from the Tobacco-Documents Corpus.................87
Table 5.15: span positions of tobacco-related collocates by frequency and rate of occurrence....88
Table 5.16: mutual information and t-scores from the reference corpus .................................89
Table 5.17: mutual information and $t$-scores from the Tobacco-Documents Corpus..........................90
Table 5.18: span positions of tobacco-related collocates by frequency and rate of occurrence...91
Table 5.19: mutual information and $t$-scores from the reference corpus...........................................92
Table 5.20: mutual information and $t$-scores from the Tobacco-Documents Corpus.........................93
Table 5.21: span positions of tobacco-related collocates by frequency and rate of occurrence...94
Table 5.22: mutual information and $t$-scores from the reference corpus...........................................96
Table 5.23: mutual information and $t$-scores from the Tobacco-Documents Corpus.........................97
Table 5.24: span positions of tobacco-related collocates by frequency and rate of occurrence...98
Table 5.25: mutual information and $t$-scores from the reference corpus...........................................99
Table 5.26: mutual information and $t$-scores from the Tobacco-Documents Corpus.......................100
Table 5.27: span positions of tobacco-related collocates by frequency and rate of occurrence..101
Table 5.28: mutual information and $t$-scores from the reference corpus.........................................103
Table 5.29: mutual information and $t$-scores from the Tobacco-Documents Corpus.......................104
Table 5.30: span positions of tobacco-related collocates by frequency and rate of occurrence..104
Table 5.31: mutual information and $t$-scores from the reference corpus.........................................105
Table 5.32: mutual information and $t$-scores from the Tobacco-Documents Corpus.......................107
Table 5.33: span positions of tobacco-related collocates by frequency and rate of occurrence..107
Table 5.34: mutual information and $t$-scores from the reference corpus.........................................109
Table 5.35: mutual information and $t$-scores from the Tobacco-Documents Corpus.......................110
Table 5.36: span positions of tobacco-related collocates by frequency and rate of occurrence..111
Table 5.37: mutual information and $t$-scores from the reference corpus.........................................112
Table 5.38: mutual information and $t$-scores from the Tobacco-Documents Corpus.......................114
Table 5.39: span positions of tobacco-related collocates by frequency and rate of occurrence..114
Table 5.40: mutual information and $t$-scores from the reference corpus.....................................116
Table 5.41: mutual information and $t$-scores from the Tobacco-Documents Corpus..................117
Table 5.42: span positions of tobacco-related collocates by frequency and rate of occurrence..118
Table 5.43: mutual information and $t$-scores from the reference corpus.....................................119
Table 5.44: mutual information and $t$-scores from the Tobacco-Documents Corpus...............121
Table 5.45: span positions of tobacco-related collocates by frequency and rate of occurrence..121
Table 5.46: mutual information and $t$-scores from the reference corpus.....................................123
Table 5.47: mutual information and $t$-scores from the Tobacco-Documents Corpus...............124
Table 5.48: span positions of tobacco-related collocates by frequency and rate of occurrence..125
Table 5.49: mutual information and $t$-scores from the reference corpus.....................................126
Table 5.50: mutual information and $t$-scores from the Tobacco-Documents Corpus...............127
Table 5.51: span positions of tobacco-related collocates by frequency and rate of occurrence..128
Table 5.52: mutual information and $t$-scores from the reference corpus.....................................130
Table 5.53: mutual information and $t$-scores from the Tobacco-Documents Corpus...............131
Table 5.54: span positions of tobacco-related collocates by frequency and rate of occurrence..131
Table 5.55: mutual information and $t$-scores from the reference corpus.....................................133
Table 5.56: mutual information and $t$-scores from the Tobacco-Documents Corpus...............134
Table 5.57: span positions of tobacco-related collocates by frequency and rate of occurrence..134
Table 5.58: mutual information and $t$-scores from the reference corpus.....................................135
Table 5.59: mutual information and $t$-scores from the Tobacco-Documents Corpus...............137
Table 5.60: span positions of tobacco-related collocates by frequency and rate of occurrence..137
Table 5.61: mutual information and $t$-scores from the reference corpus.....................................138
Table 5.62: mutual information and $t$-scores from the Tobacco-Documents Corpus...............139
Table 5.63: span positions of tobacco-related collocates by frequency and rate of occurrence..140
Table 5.64: mutual information and $t$-scores from the reference corpus..................................142
Table 5.65: mutual information and $t$-scores from the Tobacco-Documents Corpus...............143
Table 5.66: span positions of tobacco-related collocates by frequency and rate of occurrence..144
Table 5.67: mutual information and $t$-scores from the reference corpus.................................145
Table 5.68: mutual information and $t$-scores from the Tobacco-Documents Corpus...............146
Table 5.69: span positions of tobacco-related collocates by frequency and rate of occurrence..147
Table 5.70: mutual information and $t$-scores from the reference corpus.................................149
Table 5.71: mutual information and $t$-scores from the Tobacco-Documents Corpus...............150
Table 5.72: span positions of tobacco-related collocates by frequency and rate of occurrence..150
Table 5.73: mutual information and $t$-scores from the reference corpus.................................152
Table 5.74: mutual information and $t$-scores from the Tobacco-Documents Corpus...............153
Table 5.75: span positions of tobacco-related collocates by frequency and rate of occurrence..154
Table 7.1: comparison of statistical values for associations with the tobacco-related terms. .....176
Table 7.2: shared collocates among the five most frequent collocates in both corpora. ............177
Table 7.3: shared collocates among the five most frequent collocates in only one corpus. 178-179
Table 7.4: position of majority of the collocates in relation to the node. .................................194
Table 7.5: most likely span position(s) for industry/company-specific collocate per term........195
CHAPTER 1

INTRODUCTION

This dissertation is the result of a corpus-based study of collocational patterning in business terminology. The study is based upon two corpora: a corpus of general business language, which is composed of approximately one million words, and the Tobacco-Documents Corpus, which is composed of over 500,000 words from documents internal to the tobacco industry. This chapter presents the background of the study, states the problem of the study, describes its importance, gives an overview of the methodology, and points out the delimitations of the study. The chapter concludes by providing a summary of the following chapters.

1.1 BACKGROUND OF THE STUDY

There are a few factors which influenced this study. The first factor is that the field of corpus linguistics is constantly developing. Since the advent of computerized corpora in 1949, corpora have evolved steadily in size, potential, and accessibility. The first electronic corpus compiled for linguistic research, the Brown Corpus, consisted of approximately one million words when it was created in the 1960s. Many corpora no longer have finite sizes, but rather are expanding constantly as more and more material is added to them. Furthermore, numerous research projects which were not possible in the past can be undertaken on recent corpora. Some corpora are tagged for parts-of-speech, making them more searchable and suitable for grammatical investigations. There is also an increasing number of software programs capable of searching
electronic corpora. Whereas previous research on a corpus was often restricted to its compilers and, perhaps, to other scientists, many corpora can now be accessed online, publicly or through institutions such as universities. Additionally, it is now possible to compile a corpus either from electronic sources or by scanning material into a computer using optical character recognition (OCR) software. All of these developments facilitate linguistic research, as well as open up new possibilities.

A second influential factor for this study is the increasing importance of English as the language of international business. While this trend makes Business English an important variety of English, very little linguistic research has been performed on business language. Most previous work in this area has focused on lexicography and communication. With regard to the former, there are monolingual dictionaries, which serve merely to define terms, and bilingual/multilingual ones, which give an equivalent or near-equivalent translation for the key term. The purpose of the latter is to promote proper and effective communication in the business environment. This growth in the number of speakers, and potential speakers, of Business English has precipitated the need for linguistic research on this particular variety of English.

A third factor impacting the decision for this study is the growing acceptance that word use is highly patterned, combined with the technology to reveal such patterning. While patterns of word use were studied prior to the introduction of electronic corpora, they are now more easily and accurately identified. Lists of co-occurring words for a specified term can be generated using concordance software, eliminating the reliance on observation and intuition.

Therefore, the combination of these three factors — advancements in corpus linguistics, the rise in the importance of Business English, and the recognition of the vast amount of patterning in word use — led to this study.
1.2 PROBLEM STATEMENT

This study is based on the premise that word use is highly patterned and that these patterns vary according to context and text type. This leads to the hypothesis that collocates of most business terms will vary according to the specific business practice, industry, or company. That is, the language used in different industries or companies will share some common business terminology, but the common terms will be embedded in other language that is specific to the industry or company.

1.3 SIGNIFICANCE OF THE STUDY

This study extends the knowledge of the nature of collocations in particular. It provides further insight into the manner in which words co-occur by examining terms within a specific context, the Tobacco-Documents Corpus, and within a specific register, business language. Moreover, it extends linguistic knowledge about the nature of business language in general. By comparing the two corpora, it is possible to determine not only the most likely collocates of a particular business term, but also to what extent the collocates of general business terms will reflect the industry or company in which such terms are used. In addition, one can identify the positions where it is most likely for a collocate of a certain term to occur. These results can be used in the field of lexicography, both for separate collocational dictionaries of business language and for incorporation in bilingual, multilingual, and ESL dictionaries. In the field of business, such knowledge can be applied in areas such as marketing, advertising, management, and communications, where concise, effective language is crucial.
1.4 OVERVIEW OF THE METHODOLOGY

Two corpora were used in this study: the Tobacco-Documents Corpus and a reference corpus. The Tobacco-Documents Corpus is a corpus of more than 500,000 words from documents internal to the tobacco industry. The reference corpus, which I compiled, contains over one million words of running text. The sources for this corpus were randomly selected from electronic versions of books and articles (newspapers, journals, and magazines) in the areas of marketing, management, accounting, finance, and economics.

Once I had compiled the reference corpus, I used WordSmith Tools to generate frequency lists from each corpus. I then chose the twenty-five most frequently occurring business terms which the two corpora had in common. Next, I ran lists of collocates per term/per corpus within a span of four words to the right and four to the left of the term under consideration. Following this procedure, I compared the collocates of each term between the corpora and used quantitative and qualitative methods to assess the significance of the results. Finally, I wrote dictionary entries for each term utilizing the collocates. A detailed discussion of the methodology is covered in Chapter 4.

1.5 DELIMITATIONS OF THE STUDY

There are a few boundaries which have a bearing on this study. First, the size of the reference corpus was kept to approximately one million words, as this size is considered appropriate for specialized corpora (Kjellmer 1991, 115; Fang 1993, 74; Teubert and Čermáková 2004, 119). Second, since both corpora contained only written text, this study is not representative of spoken Business English. Third, the reference corpus was restricted to the areas of marketing,
management, economics, finance, and accounting. Other business-related areas were not considered.

1.6 ORGANIZATION OF THE DISSERTATION

Chapter 2 presents the development and applications of corpus linguistics. It progresses from early research to computerized corpora. Specialized corpora and specific studies are discussed in order to place this study within a context.

In Chapter 3 the underlying concepts of this study — the word, the collocation, and terminology — are addressed. Varying terms and definitions as to what constitutes the term *word* are described. Next the composition and characteristics of collocations are detailed. The chapter concludes with a brief discussion of the notion of terminology.

Chapter 4 covers the methodology used in this study. It includes sampling techniques, analytical tools, and quantitative methods for measuring significance. In addition to explaining the two statistical measures which were utilized in this study, it highlights other applications for determining the strength of attraction between words.

Chapter 5 contains the analysis of the twenty-five business terms. It offers both quantitative and qualitative assessments of the associations formed per term and per corpus. Additionally, it indicates tendencies in the placement of the collocates per term/per corpus.

Chapter 6 focuses on the area of lexicography. It begins with a discussion of the role of collocations in the dictionary. The next section describes the methodology I used to create collocational dictionary entries with each of the twenty-five business terms. The actual entries are then presented.
The dissertation concludes with Chapter 7. In this chapter, the problem of the study is restated, a brief review of the methodology is provided, and the overall results are detailed. Finally, the significance of the study and the implications for future research are discussed.
Corpus-based research is not a recent development. What is recent is the technique for the compilation and analysis of corpora: through the utilization of computers. In order to perform a computer analysis on a corpus, it must be in machine-readable format. The easiest source for such a corpus is text which is in electronic form. The number of books, newspapers, magazines, journals, and so forth, which are available in this format is increasing rapidly. Text which is not already in electronic form must be keyboarded or converted into such form by optical character recognition. The resulting corpus can then be analyzed by using the appropriate software. The following chapter discusses corpus linguistics from its beginnings to recent developments. Individual corpora and corpus-based studies are described, in particular, those relevant to this study.

2.1 Early corpus-based research

As early as the mid-eighteenth century, linguistic research was undertaken on a corpus, when Alexander Cruden (1769) produced his Concordance based on the King James Version of the Bible in 1736. In addition to making concordances of the major content words in the Bible, Cruden constructed concordances of some function words (e.g., he, you, how) and specific collocations (e.g., how much, how much more, how many, how often). The purpose of this
research was to verify factual consistency between parts of the Bible. Forty-two editions of this scholarly work were published (Kennedy 1998, 13-14).

Around the same time, Samuel Johnson was compiling his *Dictionary of the English Language*. In striving to base his examples on actual usage, Johnson amassed a corpus of over 150,000 citations from literature to be used in the entries of approximately 40,000 headwords. However, while Johnson is usually credited with being the first to use illustrative quotations, the practice had already been utilized for about 150 years (Kennedy 1998, 14; Landau 2001, 190).

In 1897 Käding, in an effort to establish spelling conventions in German, compared the frequency distributions of letters, as well as sequences of letters, in that language. According to McEnery and Wilson (2004, 3), his research was based on a corpus of eleven million words, making it not only sizable for that time, but also comparable to more recent corpora.

During the first half of the twentieth century, important research on corpora in the field of language pedagogy was undertaken by both Thorndike (1921) and Palmer (1933). Thorndike created a word frequency list from a 4.5 million-word corpus which he had created from forty-one sources, 75% of which were from the Bible and 19th century English fiction. The remaining 25% was composed of newspaper articles, letters, and school readers. His objective in constructing such a list was to develop more effective materials for teaching native speakers of English in the U.S. to read. Along with Lorge, Thorndike (1944) increased the size of the corpus to eighteen million words in the 1930s. Their research impacted English language instruction for thirty years.

Around the same time, H. E. Palmer, working in Japan, conducted corpus-based research in which he not only devised word counts, but also included collocations. In the publication based on his work, he incorporated more than 6,000 English collocations. Palmer proposed that
collocations were more prevalent in everyday vocabulary than had been estimated previously (1933, 7).

In addition to their contribution to linguistic scholarship, these early studies were significant due to the magnitude of the effort and time required to undertake them. However, such manual research was susceptible not only to human error, but also to oversight. Therefore, a faster, more reliable, and more efficient method of conducting corpus-based research became a priority. The following section provides an overview of major computerized corpora.

2.2 COMPUTERIZED CORPORA

While there were numerous studies which utilized corpora, the advent of corpus linguistics coincides with the compilation and analysis of corpora stored in computerized databases. In 1949 Father Roberto Busa (1992) began work on the production of a machine-readable corpus which he could search by means of a machine-aided concordance. Called the Index Thomisticus, it was a concordance to the works of St. Thomas Aquinas and similar authors. His corpus consisted of 10,000 sentences, with each sentence stored on a different card. As he was investigating the word in, he devised a manual index which indicated those sentences in which the word in appeared. Busa knew, however, that, if he wanted to do any further studies, he would need to expand his search mechanism. He managed to secure support for his research from Thomas J. Watson, Sr. of IBM, and his corpus was transferred onto punch cards. IBM computers were then able to perform search and retrieval of words on a word-by-word basis. By the project’s end in 1967, Father Busa had compiled a corpus of 10,600,000 words from medieval philosophy and 5,000,000 words in various languages. His most significant contribution was that he initiated
computational methods in the study of language in the humanities (Hockey 2000, 5; McEnery and Wilson 2003, 20).

Although Father Busa had already developed an electronic corpus, an extremely significant corpus which was not electronic would follow. In 1959 Randolph Quirk began work on his Survey of English Usage (SEU) Corpus. The original corpus consisted of thousands of slips of paper, complete with detailed grammatical annotations. The corpus is now in a computerized format with each lexical item tagged for word class. The SEU Corpus is composed of one million words of written and spoken British English, which was produced between 1955 and 1985 (University College London). This Survey Corpus was important in empirical language studies in several ways. First, it was a source of numerous corpus-based studies, in particular grammatical ones such as Quirk’s *A Comprehensive Grammar of the English Language*, which became a standard for English grammar (Quirk et al. 1985). Through the marking of features, it instituted new principles in the area of corpus design. Additionally, some leading corpus linguists began their research on the SEU (Kennedy 1998, 19; McEnery and Wilson 2003, 22; Teubert 2004, 107).

In 1975 the spoken portion of the SEU Corpus was computerized by Jan Svartvik of Lund University in Sweden and became known as the London-Lund Corpus. Thirteen texts were added to the original eighty-seven texts of transcribed speech, creating a corpus of approximately 500,000 words (Svartvik 1990) and, according to Kennedy (1998, 32), making it not only the largest electronic corpus of spoken English, but also the most widely used one until the mid-1990s.

The first computer corpus compiled for linguistic research was the Brown University Standard Corpus of Present-Day English, commonly known as the Brown Corpus. Created in
1963-64 by W. Nelson Francis and Henry Kučera of Brown University, the 1,014,312-word corpus is composed of 500 samples of text, with each sample containing approximately 2,000 continuous words. The samples were taken from material printed in the year 1961 in the United States. Additional requirements for inclusion were that, as far as could be determined, the author should be a native-speaker of American English, the material should be in prose, and such prose should not contain more than 50 percent dialogue (Francis and Kučera 1964). According to Kučera and Francis (1967, xvii), the word *standard* in the title of the corpus did not refer to ‘standard English,’ but rather to their hope that the corpus might “serve as a standard of comparison for a variety of studies and analyses of present-day English.”

The British equivalent of the Brown Corpus, the Lancaster-Oslo/Bergen (LOB) Corpus, was compiled through the collaborative efforts of the University of Lancaster, the University of Oslo, and the Norwegian Computing Centre for the Humanities at Bergen. From 1970 until 1976, the project was directed by Geoffrey Leech at the University of Lancaster and financed by grants from the Longman Group and the British Academy. Stig Johansson became the director in 1977 when the project was moved to the University of Oslo. With financial support contributed by the Norwegian Research Council for Science and the Humanities (NAVF), the corpus was completed in 1978. The goal of the project was to compile a corpus which would be a general representation of text types, thus allowing for research in a variety of aspects of language. The corpus consists of approximately one million words from 500 texts of around 2,000 words each. The LOB duplicates the Brown Corpus in respect to both date of publication (1961) and sampling techniques. It has been the basis for word frequency, as well as grammatical, studies (Johansson et al. 1978; Hofland and Johansson 1982).
In 1980 the University of Birmingham and Collins, a British publishing company, set up a joint venture to undertake a project which would lead to a description of the English language based on naturally-occurring texts rather than on observations. The focus of the study would be lexical, and the findings would be published in various works, with the initial publication being that of a dictionary of current English for advanced learners. Therefore, a corpus was to be compiled and processed so that lexicographers could utilize the data. Later the same year, work on the corpus began under the direction of John Sinclair. The Birmingham Corpus, also known as the Cobuild Corpus (Collins Birmingham University International Language Database), consists of a ‘Main’ Corpus and a ‘Reserve’ Corpus. The Main Corpus contains 7.3 million words in the form of six batches of concordances, six million words from written text and 1.3 million from transcribed speech. After this corpus was completed, the English Department at the University of Birmingham continued to gather data, with the objective being variety rather than balance. The resulting thirteen million-word corpus became known as the Reserve Corpus, and the concordances from this corpus were collated with those from the Main Corpus (Renouf 1987). The lexicographers working on the Cobuild Dictionary project were initially provided with the Main Corpus; however, in the later stages of compilation, they made use of the entire twenty million-word corpus to analyze forms which appeared fewer than fifty times in the smaller corpus (Krishnamurthy 1987, 63). The Cobuild Project was important for several reasons. As Kennedy (1998, 46) points out, it was the first significant computerized corpus-based project in lexicography since the American Heritage School Dictionary project in the 1970s, as well as being the first significant mega-corpus project. Additionally, in the compilation process the project made use of Zgusta’s (1971) four main lexicographical tasks of material collection,
entry selection, entry collection, and entry arrangement (Krishnamurthy 1987, 62). The work from this project culminated in the Collins Cobuild English Language Dictionary.

In 1991 Cobuild decided to embark on a new initiative which would expand the Birmingham Corpus. As of January 2002, the corpus, now known as the Bank of English, contained 450 million words, including millions of words of transcribed speech from broadcasts of the BBC and National Public Radio. There are constant reviews of the texts in the Bank of English in order to maintain a variety, and new sources are added to preserve the balance in order to reflect contemporary mainstream English. The Collins Cobuild Advanced Learner’s English Dictionary is based on this corpus (Cobuild Collins).

The compilation of a large corpus for comparative studies was proposed by Sydney Greenbaum, Director of the Survey of English Usage at University College London, in 1988. According to Greenbaum (1991, 83), the original aim was to compile corpora of British and American spoken English in order to compare speech across the two varieties. However, it was decided to add written material, including both printed and non-printed texts. Work was begun on the International Corpus of English (ICE) in 1990 at twenty centers around the world (e.g., Great Britain, the U.S., New Zealand, the Philippines), with each center compiling a corpus of its own national or regional variety of English. Each ICE corpus is adhering to a common design and common annotation scheme. Each corpus contains 500 texts of roughly 2,000 words per text, totaling approximately one million words in size. Additionally, each text category must have a minimum of ten texts (20,000 words). In general the texts range in date from 1990 until 1994; however, some corpora include more recent material, especially spoken text. In order for a text to be included in the corpus, the author or speaker of the text must be eighteen years of age or older and either formally educated up to the completion of secondary school through the medium of
English or of a status which warrants their inclusion (e.g., broadcasters, politicians). In addition, the author or speaker must have been born in the country in whose corpus they are to be included or else have moved there at a young age and been schooled through the medium of English in that country. Since the corpora aim to be representative of the English in the respective country, the text selection is not random: deliberate attention is paid to social variables, such as sex, age, region, occupation (Greenbaum 1991; Greenbaum 1996; Nelson 1996). The British component is complete and is available on CD-ROM (International Corpus of English). According to Meyer (2002, 146), who is the project’s international coordinator, the East African and New Zealand components are also complete.

Between 1991 and 1994, the British National Corpus (BNC) was compiled by a consortium consisting of Oxford University Press, publishers Addison-Wesley-Longman and Larousse Kingfisher Chambers, Oxford University Computing Services, the Centre for Computer Research on the English Language at the University of Lancaster, and the Research and Innovation Centre of the British Library. The project was funded by the Science and Engineering Council and the Department of Trade and Industry under the Joint Framework for Information Technology program, with additional funding by the British Library and the British Academy. The 100,106,008-word corpus consists of 4,124 texts, of which 90 percent are written and 10 percent are transcribed speech. Each word in the corpus is automatically tagged for part of speech. In order that the corpus might be used in a broad range of research, the corpus was designed to be as representative of British English as was possible. Additionally the corpus is encoded in compliance with the Guidelines of the Text Encoding Initiative (TEI), using SGML—Standard Generalized Markup Language (Oxford University Computing Services).
Kennedy (1998, 50) observes that, due to the BNC’s finiteness, it can be used as a point of reference for British English.

This section has described some of the major general-purpose corpora, which were compiled for various language research. In the following section, the focus shifts to special-purpose corpora.

2.3 SPECIALIZED CORPORA

According to Kennedy (1998, 19-20), what distinguishes general corpora from specialized corpora is that the former is compiled to provide answers to specific questions about language, while the latter is assembled according to specific research projects. Sinclair (1991a, 17) states that a general reference corpus is “not a collection of material from different specialist areas—technical, dialectal, juvenile, etc. It is a collection of material which is broadly homogeneous, but which is gathered from a variety of sources so that the individuality of a source is obscured, unless the researcher isolates a particular text.” The corpora discussed in the previous section fall into the category of general corpora. The following discussion outlines some examples of specialized corpora. The focus of the first three corpora is in the area of language learning, while that of the fourth deals with registers. Both of these topics are relevant to this study.

When the Birmingham Corpus was completed, Cobuild decided to produce a series of smaller corpora. Between 1982 and 1985, several specialized corpora were compiled. The first such corpus was the TEFL Corpus, which was created from twenty-six EFL texts. According to Renouf (1987, 18), the corpus is unique because it “reveals hitherto inaccessible facts about the language which occurs in the major course books, both as a whole and individually.” Besides its importance as a research resource, the corpus assisted lexicographers on the Cobuild project on
topics such as which linguistic choices would be appropriate for use in a definition and which meanings and usages would be less familiar to the learner.

An example of a corpus created for research in the field of language acquisition is the CHILDES (Child Language Data Exchange System) child language database. Begun in the mid-1980s, this corpus contains over twenty million words of predominantly child language acquisition data from more than 500 children (MacWhinney and Snow 1990). As Kennedy (1998, 41) points out, the CHILDES Database can be used in a variety of research projects such as comparing different languages, studies involving first and second language acquisition, and studies on the acquisition of normal and impaired language.

The International Corpus of Learner English (ICLE) project was begun in October 1990 at the Catholic University of Louvain, Belgium, under the directorship of Sylviane Granger. The project, which was initiated in conjunction with the International Corpus of English project, is the result of a collaborative effort between several international universities. The ICLE is defined as a corpus of English as a Foreign Language and is composed of over two million words from essays written by advanced learners of English as a foreign language from nineteen different mother-tongue backgrounds, with each language variety forming the basis of a sub-corpus (e.g., E2German, E2French, E2Russian). Each national corpus is to consist of at least 200,000 words of argumentative essays of 500-1,000 words each from advanced EFL learners, who are usually university students in the 3rd or 4th year of studying English. Additionally, there is a control corpus of argumentative essays, the Louvain Corpus of Native English Essays (LOCNESS) which were written by university students of the same age who are native speakers of American and British English. A spoken corpus, the Louvain International Database of Spoken English Interlanguage (LINDSEI), is being compiled by researchers from a number of ICLE sites. This
corpus consists of informal interviews with native speakers and EFL students. The purpose of the ICLE is to serve as a resource for extensive comparative studies of advanced EFL learners with vastly different L1 backgrounds (Granger 1993; Centre for English Corpus Linguistics; Lund University).

There have been several so-called ‘historical’ corpora of English compiled. Perhaps the best known is the Helsinki Corpus of English Texts, assembled at the University of Helsinki. One such corpus which was developed in the United States is ARCHER, ‘A Representative Corpus of Historical English Registers.’ ARCHER is a 1.7 million-word corpus which is composed of 1,037 texts of ten registers, including American English and British English. Three of the registers are speech based; the written registers contain both formal and informal texts. The texts were sampled from seven historical periods between 1650 and 1990. There are approximately 20,000 words per register for each period, and, with the exception of some of the earlier period of the news register, texts generally are at least 2,000 words in length. Potential uses of the corpus include comparative studies between American and British English over time, stylistic examinations of individual registers and authors, and the drift of registers over time (Biber, Finegan, and Atkinson 1994).

2.3.1 DOMAIN-SPECIFIC CORPORA

In compiling specialized corpora, the focus can be narrowed even further. Leech (1991, 11) refers to corpora which are designed for a specific function, such as representing the language of a particular industry, as domain-specific corpora. Since the present study is based on such a corpus, some corpora which fall into this category will be discussed.
The JDEST Computer Corpus of Texts in English for Science and Technology was compiled at Jiao Tong University (Jiaotong Daxue) in Shanghai. The one million-word corpus, which was completed in June of 1983, is composed of 2,000 texts of approximately 500 words each. Texts were taken from ten fields, which were chosen according to need and availability: civil engineering, electrical engineering, chemical engineering, computers, physics, metallurgy, atomic energy, naval architecture, aircraft manufacturing, and machine building. Each text has been coded to indicate the date of publication, the genre classification, and the form of English (e.g., American, British). The majority of texts were published after 1975. The primary aim for compiling such a corpus was to have a means for verifying that the language used in English for Science and Technology courses was an accurate reflection of the disciplines involved (Yang 1985).

Another corpus based on technical language is the Guangzhou Petroleum English Corpus (GPEC), a corpus of the English used in the petroleum industry. Compiled in 1987 by Zhu Qi-bo at the Guangzhou Training College of the Chinese Petroleum University, it contains 411,612 words from 700 texts written in Petroleum English from 1975 to 1986. The texts, which are between 500 and 600 words in length, were randomly sampled. Goals for developing such a corpus include acquiring more knowledge about Petroleum English and making available vocabulary lists and other authentic information to educators and students of Petroleum English (Qi-bo 1989).

A one million-word corpus of the English of computer science was compiled at the Hong Kong University of Science and Technology. The HKUST Corpus is composed of text taken from 166 textbooks which were used in computer science courses in three universities, two polytechnic schools, and two post-secondary colleges in Hong Kong. Most of the text is from
American publications published after 1985 or revised editions later than 1985. Three 2,000-word samples were randomly extracted from each textbook. The objective of the corpus is twofold. First, the corpus will serve as a source for pedagogical materials. Second, a dictionary of computer science terminology will be created with entries under each lemma listed according to their frequency of occurrence in the corpus (Fang 1993).

The Student Engineering English Corpus (SEEC) consists of approximately two million words taken from English-language textbooks in general areas of engineering. Compiled from textbooks which were compulsory for all engineering students at Walailak University in Thailand, the corpus is composed of entire texts rather than extracts. According to Moudraia (2003, 28), the word lists which are generated from this representative corpus of student engineering can be used as a lexical source of Engineering English. Additionally research into this specialized form of English can be performed on the corpus in areas such as lexis, syntax, morphology, and discourse.

In the field of business, TERMINEC is an English–Norwegian text and term database which is being developed by the Department of Languages at the Norwegian School of Economics and Business Administration under the direction of Magnar Brekke. TERMINEC will be composed of two parallel corpora, one in English and one in Norwegian, and a parallel term database. The parallel corpora will form a textbank, which will be a representative corpus of whole texts from approximately thirty sub-domains of economics and administration sampled from textbooks, newspaper and journal articles, and research articles. The term database will be linked to the text corpora and draw its contents from them. While the initial size of each corpus is being set at 500,000 words, there are plans to turn TERMINEC into a monitor corpus. As of 2002, the textbank contained roughly 100,000 words from thirty-two texts in eight sub-domains.
TERMINEC is important not only as a research tool, but also as a resource for bilingual specialized dictionaries and genre-related instructional materials. Brekke also proposes that the corpus will establish a point of reference and perhaps standardization, thereby assisting Norwegian export companies and the Norwegian government in conducting business, both of whom depend upon multilingual databases and some manner of translation (Brekke 2002).

While the previous sections discussed particular corpora, the following sections concentrate on various studies which make use of corpora. These studies were selected due to relevance to the current study, e.g., studies based on language taken from newspapers, studies based on text types, and collocational studies.

2.4 Corpus-based studies

Two names which are associated with corpus linguistics are M. A. K. Halliday and John Sinclair. Halliday compiled his first corpus in 1949, a collection of Cantonese sentences with which he intended to study the grammar of the dialects of the Pearl River Delta (1992, 61). In the early 1960s, he and Sinclair designed a project for the collection of spontaneous English conversation. The resulting corpus, the first ever electronic corpus of speech, contained 135,000 words recorded at the University of Edinburgh and University College, London. This corpus formed the basis for a study of English collocations, entitled English Lexical Studies (or the OSTI Report), in which Sinclair et al. (1970/2004) illustrated the collocational patterning of grammatical words. In addition to Sinclair’s renowned work on the Cobuild Corpus, he has undertaken numerous other corpus-based studies. Using the Cobuild corpus, Sinclair (1991a) investigates the word back in order to demonstrate that what people often perceive to be the ‘core’ meaning of a word, that which initially comes to mind, is often not the meaning which is most common. In the case
of back, the most frequent usage is as an adverb indicating location, although the core meaning provided by most people and dictionaries is that of the posterior part of the body (Sinclair 1991a, 112).

In 1984, a project was initiated at the University of Birmingham under John Sinclair and Tim Johns. This project was based on a corpus of scientific English which had been assembled by Dr. Peter Roe and was to establish a method of automatically identifying scientific/technical terms and to characterize text types. The study of two-word combinations from this corpus has demonstrated that there exists a significant difference between scientific texts and general English ones. These results can be utilized in the characterization of text types. According to Yang (1986, 101), “This suggests that it is possible to characterize text types on the basis of collocational behavior.”

Fillmore (1992) used corpora in his work on a lexical description of the words risk and home. Using all 1,743 sentences containing risk from a twenty-five million-word corpus, he set up a framework to characterize the word. For example, he found all instances of the verb risk to be transitive, with noun-phrase objects signifying deed, harm, or valued possession. Even though most sentences fit into his framework for components and adjuncts, a few did not. According to Fillmore, adjunct prepositional phrases with on, in, and to did not fit, and without the corpus he probably would have overlooked them. Fillmore used the Wall Street Journal section of the DCI (Data Collection Initiative) corpus to research the word home. This eight to nine million-word section consists of the text of the 1989 Wall Street Journal. Fillmore was interested in the word’s usage without an article, particularly when it functions as a locative or adverb of direction, and he extracted approximately 450 sentences from the WSJ corpus in which home was not preceded by a determiner. From this research Fillmore concludes that corpus evidence provides word
usages and meanings which are lacking in dictionaries and grammars. At the same time, he points out that other judgments need to be employed as well because while corpora illustrate what is possible, they do not indicate what is not possible.

Robert Gephart (1997) of the Faculty of Business at the University of Alberta compiled a small corpus to analyze the discourse in a public inquiry relating to a major industrial accident. The corpus was composed of 217 pages of official proceedings from the Alberta Energy Resources Conservation Board and a sixteen-page final report on the incident by the same board. Gephart then applied quantitative methods to illustrate the roles that different vocabularies and logic played in sensemaking for the two groups involved.

Several corpora have been compiled from newspaper language. Scott (1997b) assembled two reference corpora, one English and one Portuguese, for the purpose of investigating the concept of ‘key key words’ and their ‘associates’ in the same text and then the potential to match not only the key key words, but also the associates between the two languages. The English corpus was comprised of ninety-five million words taken from 4,645 Home News stories from the Guardian newspaper covering the years 1991-1994. The eight million-word Portuguese corpus consisted of 9,912 texts from the Folha de São Paulo from 1994 and 1995. The length of the texts ranged from 200 to 5,000 words, averaging about 1,000 words per text. Due to the low number of words per text, Scott set his minimum frequency of occurrence at two. Word lists were processed for the corpora, with a key word list being generated for each text. From these key word lists, a keywords database was created for each language. Scott’s aim was to determine whether the proper placement of an English word is similar to the proper placement of a word in another language. From this study, he proved such to be the case, leading him to conclude that his findings will benefit both language learning and translation.
Davies (2001) compiled two web-based corpora from online newspapers. One corpus contained thirty-five million words of Modern Spanish, with fifteen million words from eight newspapers in Spain and twenty million words from Latin American newspapers. For the Latin American portion, one million words from at least two newspapers from each of the twenty Spanish-speaking countries was selected. The other corpus consisted of twenty-six million words of Modern Portuguese, with fifteen million words from regional Brazilian newspapers and ten million words from eight newspapers in Portugal. Also included in the Portuguese corpus were one million words from interviews which were extracted from the same newspapers. In this study, Davies examined syntactic variation and change and showed, by comparison to smaller corpora, that emerging constructions such as object-to-subject raising are evidenced only in large corpora.

In an analysis employing the *Le Soir* Corpus, a sixty million-word Belgian francophone newspaper corpus, Degand and Bestgen (2003) developed a procedure aimed at automatically retrieving idiomatic expressions from large text corpora. From this corpus, Degand and Bestgen extracted only content articles, omitting stock exchange reports, weather forecasts, television schedules, and so forth. After lemmatization and the elimination of items such as punctuation, digits, and function words, a corpus of 5,500,000 words remained. Then they considered only word sequences which occurred in articles with a minimum of fifty words. This restriction reduced the corpus to 5.4 million words. While the authors acknowledge that their procedure is not completely automatic at this point, they assert that it is a major improvement over manual retrieval (2003, 258).
2.4.1 REGISTER STUDIES

Biber has undertaken several corpus-based register studies in order to investigate linguistic variations among varieties of English. In one such study, Biber (1987) compared nine written genres from American and British English with regard to three dimensions which he had previously classified. The dimensions were: Dimension 1: interactive versus edited text; Dimension 2: abstract versus situated content; Dimension 3: reported versus immediate style. For his analysis, Biber used the Brown Corpus for the American texts and the Lancaster-Oslo/Bergen Corpus for the British texts. The results of his study demonstrated that systemic linguistic differences do exist between American and British written genres across the first two dimensions.

In another study, Biber built upon his previous research to develop “a comprehensive model of textual relations among spoken and written genres in English” (1988, 58). He used the Lancaster-Oslo/Bergen Corpus and the London-Lund Corpus of Spoken English, as well as a collection of personal and professional letters (included to compensate for the lack of unpublished material in the corpora) to compile a one million-word corpus of 481 texts covering twenty-three genres. Using sixty-seven previously identified linguistic features, Biber performed a factor analysis of the distribution of these features to identify six dimensions: Dimension 1: ‘Involved versus Informational Production’; Dimension 2: ‘Narrative versus Non-Narrative Concerns’; Dimension 3: ‘Explicit versus Situation-Dependent Reference’; Dimension 4: ‘Overt Expression of Persuasion’; Dimension 5: ‘Abstract versus Non-Abstract Information’; and Dimension 6: ‘On-Line Informational Elaboration’. According to Biber (1990, 263), each dimension “represents a distinct grouping of linguistic features that co-occur frequently in texts.” Biber states that his multi-feature/multi-dimension approach to textual variation is not merely
applicable to variation between speech and writing, but that it can be used in areas such as dialect comparison; discourse, stylistic, and historical comparisons; composition research; comparison of stance types; cross-linguistic textual comparisons; and a typology of English texts.

Using a five-dimensional model of variation, Biber (1989) formulated a classification of texts in English. This categorization consists of eight text types representing groups of texts with similar sets of features. In order to identify these groupings of texts, Biber analyzed the same 481 texts as in his previous study using cluster analysis. The resulting text types are: type 1: intimate interpersonal interaction; type 2: informational interaction; type 3: scientific exposition; type 4: learned exposition; type 5: imaginative narrative; type 6: general narrative exposition; type 7: situated reportage; and type 8: involved persuasion. Biber states that such a typology of texts is important as a theoretical basis for the study of both discourse in English and register studies.

In order to investigate the types and distribution of referential expressions in text, Biber (1992) extracted fifty-eight texts from nine written and spoken genres of both the Lancaster-Oslo/Bergen Corpus and the London-Lund Corpus. Using the term ‘anaphoric’ to mean those expressions which had a common referent with a prior referring expression, Biber analyzed the distribution of the anaphor in the initial 200 words from each text. In addition, he used factor analysis to explore the existence of underlying referential dimensions, as well as their associations to textual dimensions which he had classified previously.

2.4.2 Collocational Studies

Rydén (1975) made a corpus from six British newspapers, The Observer, The Times, the Guardian, the Daily Telegraph, the Daily Mail, and the Daily Express, from June and July 1971. The purpose of his research was to study the collocations of non-titular nouns with personal
names, and whether or not the non-titular noun took a definite/indefinite article. Rydén found that the collocation without an article was a stylistic feature which differentiated popular journalistic language from the more conventional journalistic prose.

Kennedy (1991, 1992) researched the usages of the words *between* and *through* in the Lancaster-Oslo/Bergen corpus and then used the Oxford Concordance Program to collect information about their collocational patterns. Kennedy points out that the difficulty in distinguishing the usage of one word over the other comes in part from the fact that both words share some functions. However, he adds, there is a significant difference between the functions each word serves and the words, and word class, with which it collocates.

Renouf and Sinclair (1991) studied discontinuous pairs (those separated by one word) of high-frequency grammatical words and their intermediate collocates. For this study they used a one million-word corpus of spoken British English and a ten million-word corpus of written British English, both of which were taken from the Birmingham Collection of English Text. According to Renouf and Sinclair, these pairings or ‘frameworks’, such as *a* + ? + *of*, demonstrate that the “choice of word class and collocate is specific, and governed by both elements in the framework” (1991, 143). Additionally, they assert that the significance of these frameworks is exhibited by their high type-token ratio.

Following along the same lines as Renouf and Sinclair, Eeg-Olofsson and Altenberg (1994) researched recurrent discontinuous word combinations in the London-Lund Corpus. Their framework was represented as *X* ? *Y*, with ? being an intermediate word or string of words; however, they limited their study to two intermediate words. In order to describe these frames, they used the type/token ratio, entropy, and mutual information, statistical measures which calculate internal variability, filler predictability, and statistical uniqueness respectively.
According to Eeg-Olofsson and Altenberg, mutual information and entropy are effective tools for the detection and description of collocations, and both measures work well for discovering collocations in frames which occur as little as two times.

Smadja (1991, 1993) developed a set of statistical-based techniques for the identification and retrieval of collocations from large corpora. To demonstrate this set of techniques, known as Xtract, Smadja used three corpora: an eight to nine million-word corpus of stock market news reports taken from the Associated Press newswire, a twelve million-word corpus of articles from the New York Times published in 1987 and 1988, and a four million-word corpus from assorted domains (e.g., health, weather reports, politics) from the Associated Press newswire. Smadja found that the results were corpus-size dependent; if the corpus is too small or if a word has a low frequency in the corpus, the distribution of the collocates will be too small for the statistical methods to be effective. However, Smadja observes that Xtract does allow for the automatic retrieval of both lexical associations and the syntactic connection of the two words under consideration.

Renouf (1992) studied the top 150 words in the Birmingham Corpus to identify how these words combined into phrases. The aim of her research was threefold: how native-speakers use the most common words; what is produced from the constant use of high-frequency words; and how a learner of English can manipulate these core words. From this study, Renouf points out that while the recurrent sequences provide the way for various types of discourse, they are too restrictive to allow for full expression. They do, nonetheless, indicate the formation of collocational sets and give basic information regarding phrasal colligation.

Using an 11.6 million-word sub-sample of the Longman/Lancaster Corpus, Biber (1993a) analyzed collocations for the words certain and right. In his analysis, Biber selected only those
collocations which occurred more than thirty times and only those texts which were longer than 20,000 words. He then applied factor analysis to identify the groups of collocational pairs. The aim of his study was to demonstrate the extent to which factor analysis can be used to identify underlying collocational patterns which reveal major uses and senses of words.

Stubbs (1995a, 1995c, 1996, 2001) investigated the lemma *cause* (noun and verb) based on 38,000 instances from 120 million words of running text from the Cobuild corpus. Using a span of 3:3 (three words to the right of *cause* and three words to the left), he determined that the lemma *cause* collocated only with words having unpleasant connotations. Stubbs points out that such collocational behavior is evidence of the co-selection of lexis and that, once a word has been selected, it is possible to predict the semantic features of the words which accompany it.

2.5 **Summary**

In order to place this study in the proper context, this chapter has provided a history of the field of corpus linguistics. It has described the composition, uses, and significance not only of generalized corpora, but also of specialized ones. Additionally, corpus-based studies were discussed, with emphasis on those relevant to this project. The following chapter details essential concepts: words, collocations, and terminology.
Chapter 3

Words and Collocations

Prior to describing the methodology, analysis, and results of this study, it is useful, and perhaps essential, to include a discussion of fundamental concepts which are involved in it. This chapter focuses on key concepts of the study: the word, the collocation, and terminology.

3.1 The Word

According to the Oxford English Dictionary 2nd Edition, one of the definitions of the term word is ‘an element of speech.’ This is perhaps the meaning which comes to mind most readily, but it is, nonetheless, a rather vague description. In addition there are several other terms which apply to the term word depending upon how the notion of the word is defined. These terms include: word form, lemma, lexeme, lexical unit, lexical item, lexical word, grammatical word, keywords, type, and token.

While it might be argued that the word is a central constituent of language, the concept of what actually constitutes a word is rather nebulous. According to Sapir (1921), it is not possible to construct a definition of the term word. He states:

The word is merely a form, a definitely molded entity that takes in as much or as little of the conceptual material of the whole thought as the genius of the language cares to allow (1921, 32).
Sapir continues:

The best that we can do is to say that the word is one of the smallest, completely satisfying bits of isolated “meaning” into which the sentence resolves itself. It cannot be cut into without a disturbance of meaning, one or the other or both of the severed parts remaining as a helpless waif on our hands (1921, 34).

Akmajian et al. (1995, 15) also cite the difficulty in establishing a definition of word. They state that even though native speakers know intuitively what does and does not constitute a word in their own language, there exists no satisfactory definition of the term itself.

Bloomfield (1933, 178) observes that “for the purposes of ordinary life, the word is the smallest unit of speech.” He points out that since words are separated by spaces in the written language, people are used to dividing linguistic forms into words. The difficulty in discerning word boundaries arises, however, for those people who do not read or write. This latter point illustrates the definition of a word formulated by Sinclair et al. (1970/2004, 5) as “a somewhat arbitrary string of signs.” In order to segment these strings into words, one must have a means of identifying them, that is, through the written language.

Some linguists attempt to assign a more concrete definition to the term word. For example, Trask (1993, 304) defines a word in two ways:

1. A lexical item; a single item belonging to some lexical category, having an identifiable meaning or grammatical function and typically a fairly consistent phonological shape, though possibly exhibiting a certain amount of inflectional variation reflecting its grammatical environment in particular sentences. In this sense of the term ‘word’, such forms as go, goes, went, going and gone are all forms of a single word (lexical item) go.
2. A word form; a particular morphosyntactic form of a lexical item occurring in a particular grammatical environment. In this sense of the term ‘word’, the forms go, goes, went, going and gone are all different words…

Halliday (2004, 2-3) ascribes three senses for the term word. The first sense is that of a lexical item, that element which carries the meaning in a language. The second sense applies to the written language and is that which is written between two spaces. A third sense is that unit which receives its part of speech according to the grammar of a language. Halliday states that the complication in pinpointing exactly what constitutes the term word arises from the fact that there are three senses which actually define it.

According to Crystal (1997, 419), words can be defined physically both in writing and in speech. The written, or orthographic, word is easily identifiable as it occurs between spaces. The spoken, or phonological, word is more difficult to determine; however, phonological cues (e.g., pauses) can help to establish word boundaries. Matthews (1991, 28) defines the term word in a similar manner. For him the written word is a collection of syllables and letters, the spoken word a collection of syllables and phonemes.

3.1.1 Word form

Closely related to the word is the word form. In fact, Stubbs (2001, 26) uses the term to express the same concept as that of the word. He states that the only lexical unit which can be observed directly is the word form and that it is a linear string of these word forms which comprise a text. A spoken text is composed of a sequence of word forms in time, while a written text is composed of a sequence of word forms in space. Crystal (1997, 419) uses the term word form to refer to the orthographic and phonological word collectively, and Trask (1993, 304), as previously
mentioned, uses it as one of the senses for \textit{word}. The term is, however, generally considered to have a slightly different meaning than that of \textit{word}.

Sinclair (1985, 84; 1991a, 28) defines a \textit{word form} as “an unbroken succession of letters,” a definition which one might argue could also be used to define a \textit{word}. However, Sinclair differentiates the two terms by stating that one word has many different word forms. Using the word \textit{drive} to illustrate this point, Sinclair labels \textit{drive}, \textit{drives}, \textit{drove}, \textit{driven}, and \textit{driving} as word forms of \textit{drive}. Additionally he points out that no distinction is made between lower and upper case when identifying a word form, i.e., \textit{drive}, \textit{Drive}, and \textit{DRIVE} would all be occurrences of the same word form. Matthews (1991, 30) observes that the word form can be analyzed in regard to phonemes and letters, thus indicating a definition of \textit{word form} similar to that of Sinclair’s.

3.1.2 \textbf{LEMMA AND LEXEME}

The terms \textit{lemma} and \textit{lexeme} are often used interchangeably to refer to the canonical form of a word. As opposed to word forms, they are abstract and thus not directly observable. This abstract form is used to represent a set of morphological/grammatical variants. As a rule, lemmas are determined by tradition and tend to be the singular form of a noun and the base form of a verb, e.g., the verb \textit{write} would represent \textit{writes}, \textit{wrote}, \textit{written}, and \textit{writing} (Zgusta 1971, 120; Stubbs 1996, xv; Stubbs 2001, 26; Crystal 1997, 220, 400). According to Zgusta (1971, 120), “the guiding principle for the choice of the canonical form is that it should be as good a starting point for the construction of the other forms of the paradigm as possible.”

The \textit{lemma}, which is also called the \textit{headword}, the \textit{entry form}, and the \textit{basic form}, is an element of vocabulary. It is the form listed at the beginning of a dictionary entry (Stubbs 1996,
3.1.3 Lexical Unit and Lexical Item

Zgusta (1971) applies the term *lexical unit* to refer not only to single words, but also to more complex units. He uses the terms *lexical unit* and *word* interchangeably provided that the words referred to are designative ones.

Malmkjær (1991, 291) states that a lexical unit is a component of the vocabulary of a language, with the primary example of a lexical unit being the word. She labels as multiword lexical units those multiword items whose meanings are more than simply the sum of the meanings of the individual words. She illustrates this point with *night owl* and *nocturnal owl*, suggesting that the former would be considered a lexical item, whereas the latter would not (1991, 293).

In a similar manner, Stubbs (2001, 29-31) uses the term *lexical item* to signify both individual words and phrases, e.g., *near*, *near-sighted*, and *Near East*. However, while Malmkjær limits the use of the term *multiword lexical unit* to those phrases whose meaning extends beyond that of the individual words, Stubbs places no such condition on phrases which he considers as lexical units.

For Halliday et al. (2004, 169), a lexical item is not a written or spoken word, but rather one that can be considered as a unit of meaning. Sinclair et al. (1970/2004, 9) expand upon this idea by stating that a lexical item is not necessarily synonymous with an orthographic word. They argue that a lexical item is a “unit of language representing a particular area of meaning which has a unique pattern of co-occurrence with other lexical items.” According to them,
morphemes and homographs also qualify as lexical items. They contrast the term *lexical item* with that of *grammatical item*. As opposed to the lexical item, which is based on meaning, the grammatical item fulfills a grammatical function in the text. For example, in the word *going*, Sinclair et al. call the morpheme *go* a lexical item and the morpheme *ing* a grammatical item.

### 3.1.4 Lexical Words

When used to indicate an individual item, the term *word* can be segmented. One such category is that of the *lexical word*. *Lexical words*, which are also referred to as *content, full, major, or open-class words*, include nouns, main verbs, adjectives, and adverbs. These words, which comprise the majority of the language, due in large part to the open-ended nature of the category, have meanings which are expressible and relatively clear (Halliday et al. 2004, 168; Crystal 1997, 87; Stubbs 1996, 71; Stubbs 2001, 40; Akmajian et al. 1995, 21). According to Stubbs (2001, 39-40), lexical words are what indicate what a text is about and make reference external to the language. They contain actual semantic content (Trask 1993, 159).

### 3.1.5 Grammatical Words

A second category of words is that of the *grammatical word*. *Grammatical words*, which are also referred to as *function, empty, structural, minor, or closed-class words*, perform a primarily grammatical role. Examples of grammatical words include pronouns, articles, and conjunctions. As opposed to lexical words, grammatical words tend to consist of a limited number of fixed elements and do not take inflections. Intrinsically they are mostly or completely void of semantic content, serving the purpose of relating lexical words to one another (Akmajian et al. 1995, 21; Stubbs 2001, 39; Trask 1993, 123; Halliday et al. 2004, 168; Crystal 1997, 162).
Stubbs (2001, 40) asserts that the function of grammatical words is “internal to the language, for example in making explicit the relation of lexical words to each other.” While a language could conceivably contain only lexical words, such as telegraphic forms of language, a language consisting of only grammatical words is unlikely to exist (Akmajian et al. 1995, 21; Stubbs 2001, 40). Halliday (2004, 3) suggests that there is no sharp division between lexical words and grammatical words.

While the previous sections have dealt with defining the term word, the following section concerns perspective.

3.1.6 Keyword
According to Scott (1997a, 236; 1998b, 10), a word is classified as a keyword if its frequency of occurrence in a text can be deemed unusual. Unusual frequency refers to a frequency of occurrence which is either higher or lower than would normally be expected when compared with some type of reference corpus. If a word is key in more than one text of a related type, then it is considered a key key word. Topic homogeneity of the corpus is a determining factor in the establishing of key key words (Scott 1997a, 237; 1998b, 78).

Williams (1976, 13) uses the term keywords in two ways. One definition that he offers is “significant, binding words in certain activities and their interpretation.” The second usage is for “significant, indicative words in certain forms of thought.”

3.1.7 Types and Tokens
A term which coincides with Sinclair’s (1985, 84; 1991a, 28) definition of word form is that of type. The term types refers to the different words in a text. Each occurrence of a word in a text
constitutes one token (Stubbs 2001, 133; Kennedy 1998, 207; Matthews 1991, 32). In the sentence The boy and the girl went to the store to buy candy, there are nine types and twelve tokens. The represents three tokens, but only one type; to represents two tokens, but only one type.

By calculating the type-token ratio, the ratio of the number of different words to the number of running words, one can determine the lexical density of the text. Taking into account the topic of the text and the author’s vocabulary, the number of lexical types increases along with the length of the text, while the number of grammatical types remains limited (Stubbs 2001, 133-34).

3.2 The collocation

Collocation is the co-occurrence of words within a limited span in a text. The literature suggests that words and context are inseparable, that collocations contribute to lexical cohesion, that prosody is a feature which extends over more than one unit, and that collocations vary across text types. In the next section the concept of collocation will be considered before going on to consider some of its aspects.

3.2.1 The co-occurrence of words

Palmer (1933, 10) defined the term collocation as a sequence of words which speakers should learn as “integral wholes rather than pieced together from their counterparts.” However, Firth is generally credited with being the first to use the term collocation in its modern linguistic sense. He referred to collocation as a word “in familiar and habitual company…You shall know a word by the company it keeps!” (1957a, 11). He continued:
...collocation is not to be interpreted as *context*, by which the whole meaning is implied. Nor is it to be confused with *citation*...The habitual collocations in which words under study appear are quite simply the mere word accompaniment, the other word-material in which they are most commonly or most characteristically embedded (1957a, 11-12)...Collocations are actual words in habitual company (1957a, 14).

According to Lyons (1966, 295), it was Firth who made a technical term from the word *collocation*.

Sinclair, who studied under Firth, initially defined collocation as “the co-occurrence of two items in a text within a specified environment” (Sinclair et al. 1970/2004, 10; Jones and Sinclair 1974, 19). He later amended this definition by extending collocation to two or more words which co-occur within a short span of one another in a text and by indicating that the words did not need to be adjacent (1991a, 170). He does, however, limit this span to four words on either side of the term under investigation, the *node* (Jones and Sinclair 1974, 21).

For Greenbaum (1974, 80), two words can be classified as collocates regardless of the distance they are apart, provided they belong to what he labels a “single remembered set.” Kjellmer (1982, 27; 1984, 163; 1987, 133) lists two criteria for determining whether a sequence of words is a collocation. First, there must be more than one occurrence of the sequence in identical form. Second, the sequence of words must be well-structured grammatically. Stubbs (1995a, 381; 1995b, 23; 1995c, 245; 1996, 176; 2001, 29) defines collocation as the frequent and habitual co-occurrence of two or more words within a few words of one another. Furthermore, he states that collocation is a lexical relation between these co-occurring words (2004, 24). Partington (1998, 16-17) points out that patterning in text is not restricted to words. Rather there
exist several types of collocation: word with word, word with phrase, phrase with phrase, phrase with clause, and clause with clause.

According to Altenberg and Eeg-Olofsson (1990, 3-4), the term *collocation* in its most general sense is rather synonymous with “recurrent word combinations,” with the focus being on “word sequences in texts.” Under this definition fall compound and complex words, as well as idioms. The more restrictive meaning, which is that of Firth, refers to “habitually co-occurring lexical items.” As Altenberg and Eeg-Olofsson indicate, this sense focuses on the connection between “lexical items in language,” making it applicable to discontinuous words. This definition goes beyond that of “recurrent word combination” to render a more inclusive lexical notion (Altenberg and Eeg-Olofsson 1990, 4).

Co-occurrence is not a random process; words do not collocate freely. Allerton (1984, 18-39) claims that there are four levels of co-occurrence restrictions which determine the choice of one word over another: syntactic, semantic, locutional, and pragmatic restrictions. Syntactic restrictions involve acceptable grammatical patterns of the language. Allerton states that semantic restrictions on co-occurrence are an automatic consequence of word meaning. Locutional restrictions are arbitrary and are not caused by semantic considerations. An example of such a restriction is the preference for one preposition over others, e.g., *at home*, not *in* or *by*. Pragmatic restriction concerns those cases where one phrase or sentence is chosen over others which are nearly synonymous. Moon (1998a, 28) seems to concur on Allerton’s restrictions when she states,

> Even if a lexical selection is motivated by semantics, it carries with it various collocational and syntactic constraints: the selection of the co-text is not free. There are preferred or typical locutions or structures, preferred ways of saying things.
Greenbaum (1970, 84) notes that collocations which breach lexical, syntactic, and semantic restrictions should be differentiated from those which are possible, even if they have never occurred. This will result in a cline with acceptable collocations at one end and unacceptable ones at the opposite end.

A further issue regarding collocations is how to ascertain if two items are collocates because as Sinclair (1991a, 111) observes, “many phrases have an indeterminate extent.” Since even non-adjacent items can be classified as collocates, two items in different sentences might be considered as collocating, and the problem is in establishing a limit on the distance allowed between the two items. Greenbaum (1970, 11) labels this the “item-oriented approach” and suggests that an arbitrary distance be set on an ad hoc basis. However, he acknowledges that the item-oriented approach has several other drawbacks. One negative aspect is that such an approach tends to mask syntactic restrictions. Greenbaum gives as an example the verb like, which collocates with a following much in negative sentences, e.g., I don’t like him much. In affirmative sentences, much must be pre-modified in order to be a post-verbal collocate of like, e.g., I like him very much. Another problem with this approach, that of homonyms, occurs when semantic considerations are overlooked. Greenbaum (1970, 12) exemplifies this point with the word badly, whose frequent collocates are treat and need. When badly collocates with treat, it is a manner adjunct; with need, it functions as an intensifier.

Even when collocations fulfill the restrictions, it does not mean that they necessarily occur in a text; some collocations are more likely to occur in the language than others (Stubbs 1995a, 383). As Greenbaum points out (1970, 10), the occurrence of collocations is based on probabilities not absolutes. There are many factors which will influence the collocations which appear in a certain text, including topic and level of formality.
Sinclair et al. (1970/2004, 10) distinguish between significant collocation and casual collocation. They define significant collocation as:

regular collocation between two items, such that they co-occur more often than their respective frequencies, and the length of text in which they appear, would predict.

Casual collocation refers to collocation which is not significant. An important component in the behavior of collocations is that of position-dependence (Sinclair et al. 1970/2004, 81-83; Jones and Sinclair 1974, 43-44). Not only do significant collocations exhibit position-dependence, but most significant collocation occurs between the node and the adjacent positions, both to the left and to the right. The amount of significant collocation between the node and the other positions of the span are minimal. Jones and Sinclair (1974, 31) maintain that strong patterns, such as grammatical structures and clichés, tend to make the identification of more lexical relationships between words more difficult. However, when lexical patterns of significant collocation are found, they can provide insight into the lexical organization of the language (Jones and Sinclair 1974, 42).

Greenbaum (1970, 84-85) uses the collocation juvenile delinquency to exemplify the linguistic relevance of strong collocations. He observes that the frequency of this collocation is “relevant to a synchronic statement of the meaning of” the collocations middle-aged delinquency and senile delinquency. Strength of collocation can be used to differentiate between varieties of English (Sinclair et al. 1970/2004, 133). Additionally, collocations can be composed of two words of different frequencies. In such a case, the collocation will be more significant for the less frequently occurring word than for the more frequently occurring one (Sinclair 1991a, 109-15).

According to Kjellmer (1984, 164), collocations exhibit some degree of distinctiveness, which is somewhat context-dependent. Such distinctiveness can be considered as forming a
continuum, with sequences which are highly predictable, yet neither considered nor remembered as units, at one endpoint, and collocations which are listed in dictionaries at the other endpoint. Examples of the former would be *the house* and *to write*, while those of the latter would be *lose one’s place, fall into place*, and *in the first place*. Between the two end points fall occurrences which cannot be clearly categorized. In many ways the behavior of highly distinctive collocations parallels that of single-word lexemes. Not only are such collocations often synonymous or nearly synonymous with single words (e.g., *the windy city* = *Chicago, have a look* = *look*), but they also follow restrictions such as those on word order (e.g., *spic and span*, but not *span and spic*) and syntactic variation (e.g., *off course*, but not *off courses*). Kjellmer (1984, 165-70; 1987, 136) lists several factors which can indicate collocational distinctiveness, provided other things remain equal. A sequence which is a minimum of two units in length is likely to increase in distinctiveness the more frequently it occurs, the more frequently it occurs with regard to its expected frequency of occurrence, the longer it is, the more texts over which it is distributed, the more text categories over which it is distributed, and the more complex it is structurally.

Sinclair (1987a; 1991a, 109-21) has formulated two diametrically opposed principles for interpreting meaning in text: the open-choice principle and the idiom principle. The open-choice principle states that language is composed of an immense number of complex choices and that as each word, phrase, or clause becomes complete, new choices become available, subject only to grammatical restraints. The idiom principle, on the other hand, operates on the assumption that words do not occur randomly. Rather language is composed of a large quantity of semi-preconstructed phrases, and these units are tantamount to single choices. Sinclair proposes a few possible reasons for the existence and use of such phrases: the repetition of like situations in life;
the economization of effort; the constraints of real-time conversation (1987a, 320; 1991a, 110).

Characteristics of the idiom principle include that the boundaries of many phrases cannot be
determined clearly; internal variation, as well as internal syntactic variation is often permitted;
there is often word-order variation; many uses of words and phrases collocate strongly with other
words, have particular grammatical preferences, and/or tend to select specific semantic
environments. Collocation falls under the idiom principle.

Firth distinguished between collocation and colligation. He used collocation to refer to
lexical patterning and colligation to refer to grammatical patterning. In his words:

Collocations are actual words in habitual company. A word in a usual collocation stares
you in the face just as it is. Colligations cannot be of words as such. Colligations of
grammatical categories do not necessarily follow word divisions or even sub-divisions of
words (1957a, 14).

Benson (1985) and Benson et al. (1986) classify collocations as lexical collocations and
grammatical collocations. According to Benson (1985, 61), lexical collocations are usually
composed of two lexical words (nouns, verbs, adjectives, and adverbs) which can be considered
as ‘equals’. The main types of these collocations are combinations of adjective + noun, noun +
verb, and verb + noun. Grammatical collocations are a combination of a noun, verb, or adjective
and a grammatical word, characteristically a preposition, or a grammatical structure, e.g., an
infinitive or a clause. Cowie (1978, 132) uses the term collocation to mean “the co-occurrence of
two or more lexical items as realizations of structural elements within a given syntactic pattern.”
This definition combines the notions of lexical collocation and grammatical collocation.
Likewise, Jones and Sinclair (1974, 22) point out that every word in a text functions
grammatically, and, therefore, lexical patterning cannot be considered as discrete from
grammatical patterning. Kjellmer (1987, 137-38) asserts that the structural complexity of a collocation can be gauged according to the number of structural categories which are relevant to it.

Bahns (1993, 57) claims that one of the primary attributes of collocations is that “their meanings reflect the meanings of their constituent parts.” For him, this is what sets collocations apart from idioms. Sinclair (1994, 23), on the other hand, argues that the result of such co-selection is that the words become somewhat delexicalized, in that the collective meaning differs from the individual meanings. Similarly, Zgusta (1967, 582) describes multiword units from which the meaning cannot be ascertained from the individual constituents.

Collocation can also be thought of as working on meaning in a reciprocal manner. Bublitz (1998, 23) exemplifies this point with the collocations *physical body*, *completely forget*, and *entirely agree*, stating that the meanings of both words in the pairs are affected by their co-selecting one another. He suggests that a shared feature is the minimum semantic connection between nodes and collocates and, as exhibited in the three examples, there is a reduction to their single shared feature. According to him, “Co-selection leads to a re-arrangement of both items’ semantic features which are then different from the meaning they display when not co-selected” (1998, 23). He sees shared meaning in collocations as a “matter of degree,” based upon the extent to which the meanings of the nodes and collocates match (1998, 26).

Another notion of meaning is discussed by Leech (1974, 20), who lists collocative meaning as one of his ‘seven types of meaning.’ He defines collocative meaning as “the associations a word acquires on account of the meanings of words which tend to occur in its environment.” Leech illustrates this point with the noun *woman* and collocating adjectives *pretty* and *handsome*. Both adjectives are acceptable collocates of *woman*; however, due to their
collocative associations, they imply different types of attractiveness. Leech goes on to state that collocative meaning is a special category, “an idiosyncratic property of individual words,” to be assigned meaning only when other explanations are not sufficient.

“Semantic tailoring” is the term given by Allerton (1984, 22-23) to describe situations where the meaning of one word is adjusted (“tailored”) to fit with the meaning of the word (with the more fixed meaning) with which it co-occurs, although the contribution to the total meaning which the collocate makes stays constant. As an example of this phenomenon, Allerton cites the adjective *good* in *good pianist* and *good piano*, stating that the *good* with *pianist* implies a somewhat different meaning than that of the *good* with *piano*. While there are several combinations which can produce this semantic tailoring (e.g., noun and modifying adjective; object noun and transitive verb; noun and preposition), nouns tend to exert the most influence (“tailor”) on the meaning of their collocates, while prepositions are the most likely to be “tailored.” According to Allerton, semantic tailoring should not be confused with modification. While semantic tailoring is a relationship between two words, in which one word restricts its meaning when it occurs with another word, modification refers to a relationship between words and the phrase in which they occur; that is, in a phrase one word modifies another—the traditional usage of *modification* (Allerton 1984, 23).

Cowie (1978, 133-34) discusses meaning as a determining factor in regard to the amount and “semantic variety” of co-occurring words. For him there exists a continuum of co-occurrence, with open collocations at one end of the scale and restricted collocations at the other. At the center of the continuum occur collocations in which the collocates are limited by the meaning of the node. Cowie illustrates this situation with the collocation *explode a claim*. In this case, the verb has taken on the meaning of ‘refute,’ a rather specialized sense, thereby limiting
the number and diversity of nouns with which it can combine. Those nouns with which it co-
occurs—claim, case, theory, fallacy—are semantically related (Cowie 1978, 133).

On the open end of the scale of co-occurrence are those words whose acceptable
collocates are not only open-ended, but also semantically diversified. For instance, Cowie (1978,
133) states that the verb run can co-occur with animate, inanimate, concrete, and abstract nouns,
e.g., machine, army, business, scheme. At the other end of the scale lie restricted collocations.
Here one word takes on a particular meaning due to its co-occurrence with another word. This
situation is similar to what Allerton (1978, 22) calls “semantic tailoring.” Cowie (1978, 134)
uses the example of foot the bill, in which foot assumes the specialized meaning of ‘settle’ under
the influence of bill. This sense is valid only with the noun bill, no substitution is acceptable,
e.g., bill cannot be replaced with account.

McIntosh (1961, 327-28) claims that disregarding any purely grammatical patterning,
there exists only a certain amount of compatibility between words, what he labels as the
“potential of collocability.” He asserts that language users have numerous choices, both
grammatically and collocationally. While collocations which have never been encountered
should not be dismissed, neither should all grammatically-correct sequences be accepted. In
order to determine which collocations are acceptable, McIntosh suggests relying on the familiar
as well as on range, that is, the set of possible alternatives. In the final analysis as to what does or
does not constitute a collocation, McIntosh argues that various factors are relevant, including
grammaticality, both within the collocation and between the collocation and external
constituents; meaning; and situational context (1961, 329-31). It does not necessarily follow,
however, that collocations cannot occur which seem to be unacceptable. Rather, such deviations
are deliberate, used to elicit humor and irony (Bublitz 1998, 29).
3.2.2 The inseparability of words and context

According to Firth (1935, 37), the only way to establish the full meaning of a word is by looking at the entire context. Stubbs (1996, 113) stresses that, in addition to individual words, patterns of words express meaning. It is from this contextual environment that a word receives a part of its meaning. Sinclair (1994, 22) continues with this idea of connectedness between words and context. He states:

One way of describing collocation is to say that the choice of one word conditions the choice of the next, and of the next again. The item and the environment are ultimately not separable, or certainly not separable by present techniques.

Stubbs (2001) views Sinclair’s notion of this connection as a reciprocal indicator. From the words, one can project the environment in which they are likely to occur; from the environment, one can expect that certain words will appear in the text. Additionally, collocation can be seen as a linear relationship, in which a node predicts what precedes or follows it (Stubbs 2001, 30). He sums it up as, “Any choice of words creates a mini-world or universe of discourse, and makes it likely that other words will be co-selected in the same context” (2001, 7).

McIntosh (1961, 333) sees word meaning as directly related to context. A word obtains meaning according to how that word is encountered in different contexts, which Cruse (1986, 16) refers to as the “contextual relations” of a word. Halliday and Hasan (1976, 289) also underline the relationship between word meaning and context:

Without our being aware of it, each occurrence of a lexical item carries with it its own textual history, a particular collocational environment that has been built up in the course of the creation of the text and that will provide the context within which the item will be incarnated on this particular occasion. This environment determines the ‘instantial
meaning’, or text meaning, of the item, a meaning which is unique to each specific instance.

Stubbs (1996, 89) maintains that word meaning originates in texts, and that as a word’s usage changes in texts, so does its meaning.

According to Kjellmer (1990, 172), context is a very important factor in determining collocation. He proposes a continuum based on the contextual company of a word, with ‘entirely predictable’ at one end (e.g., Angeles, Fidel) and ‘entirely unpredictable’ (e.g., therefore) at the opposite end. In his opinion, most words fall towards the predictable end.

3.2.3 LEXICAL COHESION

Hoey (1991, 3) gives a basic definition of cohesion as the manner in which words and grammatical constructions link sentences in a text. Halliday and Hasan (1976, 4-8) elaborate on this definition, stating that cohesion arises when an element in a text relies on another element for its meaning and can traverse sentence boundaries. They emphasize that cohesion relates to how a text is structured semantically, not what it means (1976, 26). Since cohesion is achieved in part through grammar and in part through lexis, they refer to grammatical cohesion and lexical cohesion (1976, 6). Under lexical cohesion falls collocation, or collocational cohesion (1976, 287). Collocational cohesion occurs through the mutual expectancy of which words frequently co-occur in texts (Firth 1957b, 195; Firth 1957c, 12); Halliday and Hasan 1976, 320).

Greenbaum (1974, 80) also discusses expectancy and co-occurrence. In the case of mutual expectancy, he states that the occurrence of turn on predicts light or TV will occur nearby; while the occurrence of light or TV predicts that turn on will occur nearby, although most likely not as strongly. However, he argues that the expectancy in one direction might be
much stronger, such as in the case of *rancid* which predicts *butter* to a much greater extent than vice versa. Additionally, there are those collocations in which one word predicts another in all instances, e.g., *kith and kin, to and fro.*

While whole sentences are unlikely to occur more than once in running text, chunks of sentences do recur. The recurrence of these lexical combinations (e.g., collocations, fixed phrases, idioms) contributes to lexical cohesion (Kjellmer 1991; Hunston and Francis 1998, 2000). For Hoey (1991, 219), collocation is a textual phenomenon. It results directly from relations such as bonding which occur between sentences in texts, as well as across texts. He also notes the connection between recurrence and cohesion:

> Words only have collocations because they appear in recurring combinations in text; indeed our ability to recognize collocations supports the belief that the connections made between sentences across long distances of text are subconsciously recognized by the reader.

Cruse (1986, 40) points out that collocations possess a form of semantic cohesion: their individual constituents are mutually selective to some extent. When the meaning of one or more of the collocational constituents is greatly limited by the context, its semantic cohesion is marked to a greater degree. He illustrates this concept with the adjective *heavy* when it collocates with nouns such as *drinker* and *smoker* or with the prepositional phrase *on petrol,* and he states that in order for this narrow sense of *heavy* to be chosen, the idea of ‘consumption’ must occur in close proximity. In the case in which the meaning of one of the collocational constituents requires a specific lexical item to be nearby, the semantic cohesion becomes even stronger. This condition can be exhibited by the collocation *foot the bill,* in which *foot* in the sense employed here requires the presence of *bill* (Cruse 1986, 41).
According to Stubbs (1996, 89; 2001), the occurrence of certain words leads to the expectations of what other words will occur. However, he asserts that lexical cohesion involves more than merely this expectation of co-occurrence. The content of the text, as well as the text itself, will determine which collocations occur and whether such collocations are retained or not.

3.2.4 Prosody

Firth (1957c, 193-94) used the term prosody in relation to phonological coloring which extended beyond segmental boundaries. Sinclair (1991a; 1996b) and Louw (1993) apply this term to collocations and label it semantic prosody. Sinclair (1994, 22) explains his choice of the term prosody as:

I find myself more and more drawn to Firth’s notion of prosody in phonology to apply to the kind of distribution of meaning that is observed in text when there is a large quantity of organized evidence. Successive meanings can be discerned in the text, and you can associate a meaning or a component of meaning or a shade of meaning with this or that word or phrase that is present in the text.

Sinclair (1996b, 92) supports his decision to employ this term by constructing two examples:

1. I will not make any promises
2. I will make no promises.

He claims that the second example is the more forceful statement, not only rhythmically (c.f. Firth’s notion of prosody, 1957c), but also structurally. It is not possible to establish beforehand the scope of a prosody or how it will be carried out. However, once a word or phrase has acquired a prosody, it will be reflected in the meaning whether or not it is distinctly stated. Bublitz (1998, 25) sums up the phenomenon of semantic prosody as follows:
Apparently words can have a halo or profile, which may be positive, pleasant and good, or else negative, unpleasant and bad. Whenever some such word is uttered, it prompts hearers to expect the following word to have a clear pleasant or unpleasant sense. It prospects ahead and ‘paves the way’ for a particular type of subsequent item. It opens not just any slot but one with a definite semantic set-up. Consequently, a word with such a specific semantic ‘radiation’ demands not a specific word but a semantically defined class of words.

Additionally, Louw (1993) claims that semantic prosodies “hunt in packs.” That is, the occurrence of one semantic prosody prompts another, and one builds upon the other.

Using the 7.3 million-word Cobuild Main Corpus, Sinclair (1987b, 155-56; 1991a, 74-75) performed an analysis of the phrasal verb set in. Regarding the semantic prosody exhibited by set in, Sinclair notes:

The most striking feature of this phrasal verb is the nature of the subjects. In general they refer to unpleasant states of affairs. Only three refer to the weather; a few are neutral, such as reaction, and trend. The main vocabulary is rot (3), decay, malaise, despair, ill-will, decadence, impoverishment, infection, prejudice, vicious (circle), rigor mortis, numbness, bitterness, mannerism, anticlimax, anarchy, disillusion, disillusionment, slump. Not one of these is desirable or attractive.

The Collins COBUILD English Language Dictionary lists three examples of set in which illustrate its semantic prosody: A feeling of anticlimax set in; It must be treated quickly before infection sets in; The bad weather has set in for the winter (Sinclair 1987b, 156; 1991a, 75).

According to Stubbs (1995c, 247-49; 1996, 176; 2001, 45), semantic prosodies can be very general, or it may be evident that the collocates of the node will be from a limited lexical
set. As an example of the general type of semantic prosody, Stubbs gives the word *cause*, in both its nominal and verbal forms, which characteristically collates with words sharing the semantic feature ‘unpleasant’ and whose semantic prosody is strongly negative. Words collating with *cause* include *accident, cancer, damage, death, disease, pain, problems,* and *trouble.* Stubbs illustrates the more restricted type of prosody with the word *provide,* whose collocates are from the semantic field of ‘care,’ ‘food,’ ‘money,’ and ‘shelter.’ He remarks that what is being provided can be positive and valuable, or it can be neutral; however, it cannot be negative or undesirable. Among positive collocates are *assistance, care, employment, food, housing, money, protection,* and *support.* Neutral collocates include *an answer, a copy, evidence, information,* and *a summary* (Stubbs 1995c, 247-48).

As Louw (1993) points out, semantic prosodies are established and refined over a long period of time, exhibiting both pejorative and ameliorative changes. There are both negative and positive prosodies, but negative ones prevail. From his investigation of the connection between irony and semantic prosody, Louw (1993, 164) notes that a prosody which is not well-established cannot be used as a basis for irony. Just as words can be used contrary to their accepted meanings, prosodies can be manipulated for many purposes, among which are to express irony, to express the attitude of a speaker or writer, and to persuade. Louw proposes that it is possible to compare promotional and advertising copy, as well as propaganda, with semantic prosodies in language in general, thereby greatly assisting those in the persuasion industry (Louw 1993).

3.2.5 Collocational variation across text types

Firth (1957a, 12-13) indicated that collocations varied from one text type to another. In this regard, he grouped collocations into “general or more usual collocations” and “more restricted
technical or personal collocations” (1957b, 195). Likewise, Sinclair (1966, 429) and Biber et al. (1998, 234) emphasize the relationship between collocation and register. Biber et al. (1998, 135) define registers as “varieties defined by their situational characteristics.” They are classified by topic, purpose, and so forth. Registers can be general, such as business sections of newspapers, or specific, as in reports on mergers and acquisitions. According to Sinclair (1991a, 110), registers involve large-scale conditioning choices. The choice of a particular register drastically reduces, or precludes altogether, the amount of what Sinclair terms “slot-by-slot” choices.

Hoey (1991, 219) maintains that strength of collocation is text-dependent, while Kjellmer (1987, 133-36) holds that some types of texts are more collocationally-prone than others. He states that collocations are characteristic of official reports and documents. In addition, he suggests that the length of collocations is reflective of text type, with informal text characterized by longer collocations. Kjellmer (1984, 170) argues that when the frequency of collocations is high in a number of texts within one text category, it might be indicative of technical or specialized language.

Partington (1998, 20) lists legal documents, scientific and medical papers, and business reports as genres in which collocations are important elements. He ascribes this to two reasons. First, a writer often needs to designate the register of the text. Second, new information needs to be balanced with old information, thereby reducing the amount of processing which is necessary. This is particularly important with lengthy texts.

3.3 TERMINOLOGY

Sinclair (1996a, 101) distinguishes between the meanings of words and the meanings of terms. The meanings of words are generally subject to change, while the meanings of terms are
intentionally kept from varying. When changes do occur in terminology, they are the result of advances in the various fields or the need for clarification.

According to Sinclair (1996b), new units of meaning are created through the pairing of lexis and grammar. Therefore it is necessary to control rigidly both the vocabulary and the syntax in fields which depend on terminology in order that all members of the language community can understand the texts. However, as Sinclair points out (1996a, 101-2), there is no clear line between words and terms. He states that words such as the names of flora and fauna behave similarly to terms. At the other end of the scale is the highly-restricted language of science, terminology, and the law, as well as some bureaucratic language. In between exists what Sinclair labels as “quasi-terms.” These are words which have specific meanings related to the domain in which they are used, but are not considered as terms. Sinclair lists mouse and window in their senses in the field of computing as examples of quasi-terms.

3.4 **Summary**

This chapter focused on concepts around which this study is based: the word, the collocation, and terminology. First, various views as to what constitutes a word were described. Next, the notion of the collocation was investigated in respect to the co-occurrence of words, the inseparability of words and context, lexical cohesion, prosody, and variation across text types. Finally, a brief discussion of what can be labeled terminology was provided. Chapter 4 explains the methodology involved in the study.
Neither corpus-based research nor the study of collocations is a recent development. Rather it is the use of computers which is recent. Since people tend to notice what is unusual in language, it is only through corpus analysis that the typical patterning in language can be observed (Sinclair 1991a, 17; Biber et al. 1998, 3; Stubbs 1996, 21). By using this method, researchers are able to recognize and analyze what Biber et al. (1998, 5) refer to as “complex association patterns.” They caution, however, that merely stating quantitative results is not sufficient. One must investigate the importance of such results.

Different factors may affect what collocations are found in the corpus. According to Kjellmer (1991, 115), the size of the corpus will have an influence on the recurrence of a particular sequence of words. Kennedy (1998, 117) states that size is not the only factor which has a bearing on the occurrence of collocations; they are also affected by topic diversity in the corpus. This leads us to the matter of corpus construction.

The initial and perhaps most important stage of building a corpus is the planning process. In this phase the corpus builder must decide on the size of the corpus, the length of the samples included in it, the types of text, the number of texts, the sources of the texts, and the time frame. The projected use of the corpus will be the determining factor as to the composition of the corpus (Meyer 2002, 30). This does not mean, however, that changes will not, or cannot, be made
beyond this point. As Biber (1993b, 256) indicates, compiling a corpus should be cyclical, and one should constantly reassess its composition.

For general usage, Sinclair (1991a, 18-19) advocates corpora which are open-ended and comprised of whole texts. In the case of register studies, Biber et al. (1998) state that corpora should include a large number of texts. However, corpora can also be constructed for specialized purposes (cf. Kennedy 1998, 20). Such corpora do not purport to represent the language as a whole and, therefore, the text samples can be collected from subpopulations, e.g., according to genres, topics. The size of such corpora usually ranges between 100,000 and two million words of running text. Meyer (2002, 38) maintains that when text fragments are included in the corpus, they should not be randomly selected, but rather should constitute a coherent unit. The problem with using partial texts, as opposed to full texts, is that rare features, such as relative clauses (Biber 1993b; Biber et al. 1998), cannot be studied. Therefore, it is important to have a clear understanding of the linguistic feature to be analyzed prior to constructing the corpus.

Biber (1993b) and Biber et al. (1998) argue that matters of representativeness are vital in corpus construction. According to Biber (1993b, 378), “Representativeness refers to the extent to which a sample includes the full range of variability in a population.” This representativeness is dependent upon the range of text types and the range of linguistic distributions included in the corpus, as well as other matters, e.g., the number of words in each sample of text and the number of samples which comprise a text.

Biber (1993b) concludes that the representativeness in a corpus cannot be assessed fully until the corpus is finished. In order to insure that a corpus is representative, it must be constantly evaluated and adjusted during the compilation process. Kennedy (1998) considers representation
of a language to be a valid goal of corpus design, but states that the final decision on representativeness and balance will be questions of judgment and merely approximate.

The following sections describe the methodology I used in this study.

4.1 METHODOLOGY

In order to research collocational patterning in business terminology, I made use of two corpora: the Tobacco-Documents Corpus, compiled at the University of Georgia, and a reference corpus of general business language, which I compiled. The compilation processes of the two corpora will be described below.

4.1.1 THE TOBACCO-DOCUMENTS CORPUS

The Tobacco-Documents Corpus was compiled from over four million documents both internal and external to the tobacco industry, rendered by the industry in various litigations. From these documents, a random set of documents was extracted for the purpose of creating an exploratory core sample, which would be used in determining text types and proportions for the creation of a reference corpus. This set contains documents prior to January 1999 from R. J. Reynolds, Philip Morris, Brown and Williamson, Lorillard, The American Tobacco Company, The Tobacco Institute, and The Council for Tobacco Research, as well as 33,000 documents known as the Bliley Collection. The proportion of documents to be taken from each decade was determined by searching their online collections. Seven categories were then established (1) 1900-1959, (2) 1960, (3) 1970, (4) 1980, (5) 1990, (6) Undated, and (7) Bliley. For each decade, one year was randomly chosen, and then one month within that year. Using the established proportions, documents were selected from the applicable month by taking every $n$th document, with $n$
equaling the last two digits of the year representing the decade. The total number of documents in the exploratory core sample is 349 (Kretzschmar et al. 2004).

Following this compilation process, certain criteria were applied to ensure the eligibility of the documents for inclusion in the reference corpus:

1. texts had to be internal to the tobacco industry, which included any individual or organization which received financing from a tobacco source;
2. texts had to relate to public health;
3. texts had to be in English;
4. documents had to contain a minimum of fifty words.

Additional document coding categories included:

1. each document was coded according to the target audience as either industry-internal or industry-external;
2. each document was coded according to whether the audience was named or unnamed.

Quotas were then established based on statistics obtained from the exploratory core sample. The corpus needed to be large enough so that low frequency features would be represented, yet small enough that the available resources would not be exhausted. After applying the criteria, only 202 documents remained from those in the exploratory core sample. Next the proportions in the sampling plan were applied, and the reference corpus was created. This corpus consists of 808 documents and 529,000 words, with the average length of 655 words per document (Kretzschmar et al., 2004).

Since only twenty-eight documents in the reference corpus dealt with external audiences, an ancillary corpus of solely external documents was compiled. This corpus was created using the same sampling techniques as were used in the reference corpus, e.g., the same decade and
named/unnamed quotas. This corpus contains 49,146 words from 100 documents. The average length of the external documents is 422 words (Kretzschmar et al. 2004). It is the combination of the tobacco reference corpus and the supplemental corpus that I refer to as the Tobacco-Documents Corpus.

In the following section, I describe the methodology I used in constructing my reference corpus.

4.1.2 THE REFERENCE CORPUS

Prior to compiling the reference corpus, I established the following criteria:

1. the corpus should consist of approximately one million words of running text;
2. the sources had to be in electronic versions;
3. the sources had to be general in application, “see below”;
4. the sources had to be in full-text version;
5. the sources had to be published in the U.S. or the U.K.;
6. the sources had to be published in 1950 or later;
7. articles had to contain a minimum of 500 words.

In order to determine which fields of business I should incorporate into the corpus, I decided to use the top-ranked graduate schools of business in the U.S. as a guide. Based on rankings in BusinessWeek (2002), Forbes (2003), and U.S. News and World Report (2005), I chose the following schools: Harvard, Stanford, University of Pennsylvania, Massachusetts Institute of Technology, Northwestern, Columbia, University of Chicago, University of California at Berkeley, Dartmouth, University of Michigan, Yale, University of North Carolina at Chapel Hill,
University of Virginia, and Duke. I then checked each school’s website to establish the core courses and selected the top five subjects which the schools had as common requirements:

1. finance — required by all 14 schools;
2. marketing — required by all 14 schools;
3. accounting — required by 13 schools;
4. economics — required by 13 schools;
5. management — required by 13 schools.

I decided to limit the corpus to these five areas because there was too much discrepancy between other required courses at the schools.

Having chosen the five areas of business, I decided to employ a combination of quota sampling and random sampling. Initially I utilized quota sampling. I made the decision that the corpus would be composed of approximately 200,000 words in each of the five categories. Additionally, 90% of the text would be taken from electronic versions of books, and the remaining 10% would be taken from online newspaper and journal articles. I selected these percentages for two reasons. First, due to the length of the majority of articles, it was not feasible to use the same percentage of articles as of books. The number of articles needed in such a case would have been astronomical. Second, books which provided a general variety of business language could be chosen more easily than articles of the same nature.

In the next portion of the selection process, I employed a form of random sampling. Using the website http://www.irony.com/webdice.html, I was able to select both the number of dice and the number of sides per die. Then I could roll the dice electronically.

Beginning with books, I accessed NetLibrary, a collection of electronic books available online through the University of Georgia Libraries. I searched in each of the five categories to
ascertain the number of books available, as this would determine the composition of the dice. I then set the ‘webdice’ for a single roll of four six-sided dice. The results were assigned as follows: accounting, 13; economics, 16; finance, 9; management, 17; marketing, 6. This meant, for example, that starting with the most recent book, I would scroll down to the thirteenth book on the accounting list. If that book was of a general nature, it was selected for inclusion in the corpus. If it covered specific accounting issues, e.g., accounting for certain industries, I took the next applicable book on the list and then continued thirteen places down the list. From each selected book, I blocked and copied the first 2,000 words from each chapter into MS Word, removing all formatting and tables and saving as a text file. For the materials with which I worked, 2,000 words often constituted the bulk of the chapter and not just the beginning. This was partially due to the fact that chapters tended to contain many tables and were, therefore, not so discursive. Since I took samples from each chapter, the results are not biased by this decision. I continued selecting books until I had amassed approximately 180,000 words in the category. I followed the same procedure in the other four categories, varying only in their respective dice-results.

The procedure for selecting articles was somewhat different than that employed for books. For this portion, I used three databases — *Factiva, ProQuest*, and *LexisNexis* — from GALILEO, the interconnected libraries of the University of Georgia system, and decided that each database would contribute approximately 1/3 of the words per category. Once again I used ‘webdice’ to randomly select articles. Since the number of articles was enormous, I set the ‘webdice’ at a single roll of twenty six-sided dice, which gave the following results: accounting, 79; economics, 66; finance, 55; management, 71; marketing, 80. Because most of the electronic books were relatively recent publications, I opted to begin with the earliest available, provided it
was no older than 1950. *Factiva* provided the earliest articles of the three databases, and, therefore, I began with it for all five categories. From this database, I collected articles from 1969 to 1983. Articles prior to 1969 contained less than my 500-word criterion. The second database I employed was *ProQuest*, and I took articles from it beginning with the year 1984. Due to the huge number of articles in *ProQuest* and *LexisNexis* and my desire to cover as large a span of time as possible, I rolled the dice two additional times. First, I rolled two six-sided dice to determine the number of months to skip between articles. Second, I rolled one three-sided die to establish the maximum number of articles per chosen month. The results of these rolls of the dice were: accounting, 3 articles, 10 months; economics, 2 articles, 4 months; finance, 3 articles, 7 months; management, 2 articles, 8 months. Once I had accumulated the number of words I needed from *ProQuest*, I proceeded to collect articles from *LexisNexis*, starting with the date at which I had stopped in *ProQuest* and using the same procedure. Table 4.1 lists the composition of the reference corpus.

Table 4.1: composition of the reference corpus

<table>
<thead>
<tr>
<th>business field</th>
<th>books</th>
<th>articles</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>accounting</td>
<td>179,600</td>
<td>21,972</td>
<td>201,572</td>
</tr>
<tr>
<td>economics</td>
<td>180,183</td>
<td>20,374</td>
<td>200,557</td>
</tr>
<tr>
<td>finance</td>
<td>180,544</td>
<td>19,488</td>
<td>200,032</td>
</tr>
<tr>
<td>management</td>
<td>181,273</td>
<td>19,223</td>
<td>200,496</td>
</tr>
<tr>
<td>marketing</td>
<td>181,277</td>
<td>19,126</td>
<td>200,403</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>902,877</td>
<td>100,183</td>
<td>1,003,060</td>
</tr>
</tbody>
</table>
4.1.3 Analytical Tools

When I had finished compiling the reference corpus, I used WordSmith Tools (Scott 1998a), a concordancing program, to generate a frequency list from the Tobacco-Documents Corpus and one from the reference corpus. Beginning with the most frequently-occurring word on each list, I compared the two lists. Based on my own knowledge of business terms and aided by business language dictionaries (Friedman 1987; Siegel and Shim 1987; Downes and Goodman 1991; Imber 2000), I selected the top twenty-five business terms which the two lists had in common (Table 4.2).

<table>
<thead>
<tr>
<th>term</th>
<th>reference corpus</th>
<th>Tobacco-Documents Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>brand</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>data</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>products</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>brands</td>
<td>117</td>
<td>6</td>
</tr>
<tr>
<td>advertising</td>
<td>92</td>
<td>7</td>
</tr>
<tr>
<td>share</td>
<td>90</td>
<td>8</td>
</tr>
<tr>
<td>market</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>sales</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>work</td>
<td>33</td>
<td>12</td>
</tr>
<tr>
<td>company</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>levels</td>
<td>105</td>
<td>14</td>
</tr>
<tr>
<td>increase</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>analysis</td>
<td>60</td>
<td>16</td>
</tr>
<tr>
<td>industry</td>
<td>73</td>
<td>17</td>
</tr>
<tr>
<td>development</td>
<td>80</td>
<td>19</td>
</tr>
<tr>
<td>project</td>
<td>96</td>
<td>21</td>
</tr>
<tr>
<td>system</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>order</td>
<td>83</td>
<td>23</td>
</tr>
<tr>
<td>production</td>
<td>37</td>
<td>25</td>
</tr>
<tr>
<td>quality</td>
<td>78</td>
<td>27</td>
</tr>
<tr>
<td>marketing</td>
<td>11</td>
<td>29</td>
</tr>
<tr>
<td>rate</td>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>consumer</td>
<td>85</td>
<td>34</td>
</tr>
<tr>
<td>business</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>
By *business term*, I am referring not only to words which are clearly used in the field of business, such as *consumer*, but also to those which take on a ‘business sense’ in the corpora, e.g., *increase* as in *increase in sales*. This is what Sinclair (1996a, 102) refers to as “quasi-terms.” I established these usages by examining the concordance lines in *WordSmith Tools*.

Table 4.3: occurrences of terms by business field in the reference corpus

<table>
<thead>
<tr>
<th>term</th>
<th>accounting</th>
<th>economics</th>
<th>finance</th>
<th>management</th>
<th>marketing</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>261</td>
<td>136</td>
<td>190</td>
<td>477</td>
<td>1,525</td>
<td>2,589</td>
</tr>
<tr>
<td>brand</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>67</td>
<td>718</td>
<td>801</td>
</tr>
<tr>
<td>data</td>
<td>253</td>
<td>28</td>
<td>91</td>
<td>495</td>
<td>211</td>
<td>1,078</td>
</tr>
<tr>
<td>products</td>
<td>221</td>
<td>75</td>
<td>129</td>
<td>271</td>
<td>742</td>
<td>1,438</td>
</tr>
<tr>
<td>brands</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>17</td>
<td>364</td>
<td>386</td>
</tr>
<tr>
<td>advertising</td>
<td>20</td>
<td>10</td>
<td>54</td>
<td>14</td>
<td>392</td>
<td>490</td>
</tr>
<tr>
<td>share</td>
<td>85</td>
<td>49</td>
<td>139</td>
<td>42</td>
<td>199</td>
<td>514</td>
</tr>
<tr>
<td>market</td>
<td>126</td>
<td>765</td>
<td>655</td>
<td>167</td>
<td>993</td>
<td>2,706</td>
</tr>
<tr>
<td>sales</td>
<td>611</td>
<td>56</td>
<td>428</td>
<td>555</td>
<td>651</td>
<td>2,301</td>
</tr>
<tr>
<td>work</td>
<td>174</td>
<td>160</td>
<td>135</td>
<td>345</td>
<td>134</td>
<td>948</td>
</tr>
<tr>
<td>company</td>
<td>1,317</td>
<td>66</td>
<td>875</td>
<td>495</td>
<td>574</td>
<td>3,327</td>
</tr>
<tr>
<td>levels</td>
<td>46</td>
<td>61</td>
<td>60</td>
<td>151</td>
<td>95</td>
<td>413</td>
</tr>
<tr>
<td>increase</td>
<td>140</td>
<td>380</td>
<td>140</td>
<td>65</td>
<td>115</td>
<td>840</td>
</tr>
<tr>
<td>analysis</td>
<td>147</td>
<td>98</td>
<td>242</td>
<td>65</td>
<td>171</td>
<td>723</td>
</tr>
<tr>
<td>industry</td>
<td>70</td>
<td>135</td>
<td>135</td>
<td>100</td>
<td>170</td>
<td>610</td>
</tr>
<tr>
<td>development</td>
<td>38</td>
<td>61</td>
<td>77</td>
<td>211</td>
<td>161</td>
<td>548</td>
</tr>
<tr>
<td>project</td>
<td>58</td>
<td>27</td>
<td>293</td>
<td>70</td>
<td>31</td>
<td>479</td>
</tr>
<tr>
<td>system</td>
<td>445</td>
<td>309</td>
<td>147</td>
<td>149</td>
<td>179</td>
<td>1,229</td>
</tr>
<tr>
<td>order</td>
<td>140</td>
<td>69</td>
<td>100</td>
<td>189</td>
<td>79</td>
<td>577</td>
</tr>
<tr>
<td>production</td>
<td>179</td>
<td>288</td>
<td>136</td>
<td>130</td>
<td>146</td>
<td>879</td>
</tr>
<tr>
<td>quality</td>
<td>68</td>
<td>33</td>
<td>83</td>
<td>277</td>
<td>138</td>
<td>599</td>
</tr>
<tr>
<td>marketing</td>
<td>39</td>
<td>11</td>
<td>51</td>
<td>182</td>
<td>1,476</td>
<td>1,759</td>
</tr>
<tr>
<td>rate</td>
<td>168</td>
<td>762</td>
<td>909</td>
<td>42</td>
<td>62</td>
<td>1,943</td>
</tr>
<tr>
<td>consumer</td>
<td>25</td>
<td>86</td>
<td>51</td>
<td>127</td>
<td>299</td>
<td>588</td>
</tr>
<tr>
<td>business</td>
<td>850</td>
<td>216</td>
<td>336</td>
<td>1,052</td>
<td>361</td>
<td>2,815</td>
</tr>
</tbody>
</table>

Table 4.3 displays the arrangement of the twenty-five terms according to business fields in the reference corpus. Eleven of the terms (44%) occur with the greatest frequency in the field of marketing, although for some of these terms, there are other fields with comparable frequencies, e.g., *sales* and *industry*. Additionally, this preference for the terms to be from the
field of marketing can be considered to be driven by the Tobacco-Documents Corpus. Among the top twenty-five terms which the two corpora had in common, *business* was in thirty-fifth place in the Tobacco-Documents Corpus, whereas the lowest ranking term on the list for the reference corpus was *brands*, which was in 117th place. On the other hand, terms which occurred in a higher position in the reference corpus, placed quite low in the Tobacco-Documents Corpus. For example, *costs*, which appeared in ninth place in the reference corpus, was in 584th place in the Tobacco-Documents Corpus. While I was able to isolate the terms according to the field of business in my reference corpus and determine exactly where they occurred, I was not able to do so with the Tobacco-Documents Corpus. I would assert, however, that my selection process was successful, as exhibited by the fact that the terms are adequately represented in the other four business fields.

At this point it is necessary to make a clarification concerning the terms and the collocates. As discussed in Chapter 3, there are various ways of describing the concept of the word. In selecting the business terms, I used Sinclair’s (1985, 84; 1991a, 28) notion of the *word form*. For example, when a term appeared among the top twenty-five terms in both its singular and plural forms, I took both terms. The reason for this is that different word forms have different collocates. For example, the most frequent collocate of *brand* in the reference corpus is *name*, which forms the collocation *brand name*. The most frequent collocate of *brands* in the same corpus is *new*, and its collocation is *new brands*. Furthermore, the majority of the top-five collocates of *brand* in the reference corpus (73%) occur to its right, while the majority of those of *brands* (83%) appear to its left. With regard to the collocates, I accepted only lexical words.

The next task I undertook was to run two concordances on each of the terms: one for the term in the Tobacco-Documents Corpus and one for the same term in the reference corpus.
created a ‘stop list’ of approximately 125 function words which I did not want to be returned as collocates. Furthermore, I set the frequency limit for collocates at a minimum of five occurrences. Then I generated a table of collocates per term/per corpus with a span of four words to the right and four to the left of the node, in accordance with the optimal span size as proposed by Jones and Sinclair (1974, 21). Starting with the reference corpus, I recorded the collocate which occurred most frequently in each slot to the left of the node (L1, L2, L3, L4) and to the right of the node (R1, R2, R3, R4) for every term, as well as the five most frequent collocates overall. I used this method because some collocates did not occur so frequently as others, but where they did occur, e.g., in the L2 position (two places to the left of the node), they were the primary collocate. I did not consider self-collocates, as it was too time-consuming to ascertain whether they were the result of true self-collocation or merely due to boundary crossing, that is, the same term appearing in two separate sentences but falling within the allotted span.

With regard to the Tobacco-Documents Corpus, I recorded the collocates in the same manner as I had with the reference corpus. However, I was interested in the manner in which terms related to the tobacco industry collocated with the nodes. While some of these terms were the primary collocate in various positions within the span, I decided that their importance was due to the fact that they occurred at all. Therefore, I listed each tobacco-related collocate per node, as well as a ‘total tobacco-collocate’ frequency per node. I designated as tobacco-related terms: obvious terms, such as tobacco, cigarette, smokers, and nicotine; the names of tobacco companies, such as R. J. Reynolds and Lorillard; brand names, such as Marlboro, Winston, Salem, and Virginia Slims, checking the concordance lines to exclude instances referring to Winston-Salem, North Carolina, or to the state of Virginia.
4.1.4 STATISTICAL MEASURES OF SIGNIFICANCE

There are various ways of measuring the strength of collocation between words. Some of these statistical applications include:

1. the log-likelihood coefficient, which can be applied to large and small samples of text in order to compare directly the significance of frequent and infrequent occurrences (Dunning 1993);
2. the C-score, which takes into consideration both co-frequency and proximity (Geffroy et al. 1973);
3. the z-score, which designates in standard deviations how far a value is from the mean — it is the difference between the observed and expected frequencies divided by the standard deviation;
4. Berry-Rogghe’s (1973) z-score, which is the z-score with the span size factored into the formula — a major flaw of this formula is that it does not allow for a node to collocate with itself;
5. the t-score, which is similar to the z-score, but, whereas the z-score is used for large samples, the t-score is used for small ones;
6. mutual information, which measures how strongly two words are associated.

The two statistical measures which I chose for my analysis, mutual information and the t-score, are described below.
4.1.4.1 MUTUAL INFORMATION

Mutual Information (I) measures how strongly two words are associated by comparing their joint probabilities with the probability of their occurring independently, that is, due to chance. The formula for mutual information is:

\[
I(n,c) = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}
\]

where: \( n = \) node; \( c = \) collocate; \( N = \) size of the corpus; \( f(n,c) = \) frequency of co-occurrence; \( f(n) = \) frequency of the node, \( f(c) = \) frequency of the collocate.

According to Church et al. (1991, 119-20), mutual information is a measure of similarity. If the association between two words is real, the probability of their co-occurrence will be much greater than chance. Therefore, \( I(n,c) \) will be greater than 0. Church and Hanks (1990, 24) state that I-values which are greater than 3 indicate interesting relationships between pairs of words. Additionally, Church et al. (1994, 159) point out that such associations are interesting for lexicographic work.

4.1.4.2 T-SCORE

While mutual information measures the strength of association between two words, the t-score is a measure of the confidence that the association is genuine. The formula for the t-score is:

\[
t = \frac{f(n,c) - f(n)f(c)}{\sqrt{f(n,c) / N}}
\]

Clear (1993, 281) maintains that “In general, the collocates highlighted by the t-score statistic will be frequently recurring items (often grammatical words such as prepositions, personal pronouns, determiners and particles) along with those fully lexical words which are
clearly associated with the node word.” Associations with a $t$-score of 2 or greater tend to be both significant and interesting (Barnbrook 1996, 98; Hunston and Francis 2000, 231).

As stated above, I chose these two statistical measures because I found them to be the most appropriate ones for my research purposes. Because mutual information overstates the significance of collocating words whose frequency is low (Barnbrook 1996, 101; Clear 1993, 281), it is a reliable indicator of technical terms. The $t$-score is useful since it focuses on high frequency collocations.

I wrote spreadsheet programs to calculate the mutual information values and the $t$-scores. I then used these templates to calculate the respective values for each node and its collocates per corpus. In the following chapter, I will analyze my findings for each of the twenty-five nodes.
CHAPTER 5

ANALYSIS OF THE TERMS

In this chapter I will perform an in-depth analysis of the twenty-five business terms, comparing and contrasting their collocates in the reference corpus and the Tobacco-Documents Corpus. To reiterate briefly, the statistical measures which I used are mutual information and the $t$-score. For mutual information, a value (I) greater than 3 is considered as significant; for the $t$-score, a value ($t$) of 2 or more indicates significance. Mutual information tends to focus on less frequent collocations, e.g., technical terms. $T$-scores draw attention to collocations which occur frequently. Span positions are L1, L2, L3, L4, R1, R2, R3, and R4. L1 indicates the position one word to the left of the node (the term under consideration), R1 indicates the position one word to the right of the node, and so forth. The following analysis will proceed in order of frequency of occurrence based on the Tobacco-Documents Corpus, beginning with the most frequent node.

5.1 PRODUCT

The five major collocates of product in the reference corpus are new, manager, development, managers, and management. Of the 173 times that new co-occurs with product, 82% of the occurrences are in the L1 position, making it the main collocate in that position. Manager, which collates with product a total of 153 times, appears 83% of the time in the R1 position. While the other three collocates — development, managers, management — are in the top five collocates, they are not the most frequent collocate in any of the positions. Seventy percent of the instances
of these five terms occur to the right of *product*. Other primary collocates to the left with their overall frequency of co-occurrence with the node and their rate of occurrence in that position are:

L2 = *gross* 44, 93%; L3 = *more* 58, 57%; L4 = *sales* 81, 49%.

Those in the corresponding positions to the right are:

R2 = *service* 110, 57%; R3 = *marketing* 70, 26%; R4 = *market* 87, 15%.

When evaluated by means of mutual information, all of the above words form significant collocations with *product* with the exception of the collocation *more product*, whose I-value of 2.99 falls just short of the 3.0 cut-off point. Interestingly, the collocation exhibiting the highest I-value (6.6) is *product development*, although *development* is not the most frequent collocate of *product* in any of the eight span positions. This I-value means that the probability of the collocation *product development* is approximately ninety-seven times\(^1\) greater than chance. Of its 137 occurrences with *product*, 101 of them (74%) are in the R1 position, which is not surprising since *product development* is a combination which comes readily to mind.

In keeping with its tendency to highlight frequently recurring combinations, the collocation *new product* shows the highest *t*-score, 12.78. All eleven collocations have *t*-scores which indicate significant co-occurrence, with the lowest value being 6.66 for the combination of *more* and *product*.

When considering the collocation of *gross* with *product*, it is important to consult the actual concordance lines containing *gross* to discover the expanded collocation. Out of a total of forty-four co-occurrences of *gross* and *product*, *gross* is immediately followed by *domestic* twenty-five times and by *national* seventeen times, resulting in the expanded collocations *gross domestic product* in 57% of the cases and *gross national product* in 39%. As indicated in Table 5.1, both *domestic* and *national* also collocate significantly with *product*; however, the close

\[ 2^{6.6} \approx 97 \]
association between \textit{gross} and the other two collocates indicates that the meaningful collocation is the expanded one.

Table 5.1: mutual information and \( t \)-scores from the reference corpus

\begin{center}
\begin{tabular}{cccccc}
\hline
\textbf{I (n,c)} & \textbf{t (n,c)} & \textbf{f(n,c)} & \textbf{f(c)} & \textbf{c} \\
\hline
5.15 & 12.78 & 173 & 1,900 & new \\
6.22 & 12.20 & 153 & 800 & manager \\
6.60 & 11.58 & 137 & 548 & development \\
5.81 & 11.24 & 131 & 910 & managers \\
4.94 & 10.28 & 113 & 1,427 & management \\
5.74 & 10.29 & 110 & 802 & service \\
3.64 & 8.58 & 87 & 2,706 & market \\
3.77 & 8.34 & 81 & 2,301 & sales \\
3.95 & 7.83 & 70 & 1,759 & marketing \\
2.99 & 6.66 & 58 & 2,829 & more \\
6.40 & 6.55 & 44 & 202 & gross \\
5.58 & 5.18 & 28 & 228 & domestic \\
4.66 & 4.50 & 22 & 339 & national \\
\hline
\end{tabular}
\end{center}

The five most frequent collocates of \textit{product} in the Tobacco-Documents Corpus are \textit{new}, \textit{current}, \textit{development}, \textit{test}, and \textit{consumer}. As with the reference corpus, \textit{new} is not only the most frequent collocate overall — seventy-four co-occurrences — but also the most frequent collocate in the L1 position, fifty-one times (69%). \textit{Development}, which is also a primary collocate of \textit{product} in the reference corpus, is the top collocate in the R1 position, appearing there in nineteen out of thirty co-occurrences (63%).

In this corpus, there are some span positions with only one main collocate and other positions with multiple top collocates. The remaining span positions with their primary collocate(s), overall frequency of occurrence with the node, and rate of occurrence in that position are:
L2 = type 14, 64%; L3 = new 74, 7%; one 16, 31%; packaging 16, 31%; L4 = new 74, 7%; testing 19, 26%; cigarette 17, 29%; tar 17, 29%; brand 6, 83%; R2 = smoker 10, 70%; R3 = image 20, 20%; available 11, 36%; market 8, 50%; R4 = tobacco 16, 31%.

Table 5.2: mutual information and t-score from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>I (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.30</td>
<td>8.38</td>
<td>74</td>
<td>1,224</td>
<td>new</td>
</tr>
<tr>
<td>5.75</td>
<td>5.55</td>
<td>32</td>
<td>387</td>
<td>current</td>
</tr>
<tr>
<td>5.71</td>
<td>5.37</td>
<td>30</td>
<td>373</td>
<td>development</td>
</tr>
<tr>
<td>4.47</td>
<td>5.14</td>
<td>29</td>
<td>854</td>
<td>test</td>
</tr>
<tr>
<td>5.57</td>
<td>4.49</td>
<td>21</td>
<td>288</td>
<td>consumer</td>
</tr>
<tr>
<td>4.60</td>
<td>4.18</td>
<td>19</td>
<td>511</td>
<td>Lights</td>
</tr>
<tr>
<td>2.85</td>
<td>3.55</td>
<td>17</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>4.34</td>
<td>3.92</td>
<td>17</td>
<td>548</td>
<td>tar</td>
</tr>
<tr>
<td>4.27</td>
<td>3.79</td>
<td>16</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>2.12</td>
<td>3.08</td>
<td>16</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>4.09</td>
<td>3.52</td>
<td>14</td>
<td>537</td>
<td>Camel</td>
</tr>
<tr>
<td>3.94</td>
<td>3.37</td>
<td>13</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>4.48</td>
<td>3.02</td>
<td>10</td>
<td>293</td>
<td>smoker</td>
</tr>
<tr>
<td>1.94</td>
<td>2.34</td>
<td>10</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>4.53</td>
<td>3.03</td>
<td>10</td>
<td>282</td>
<td>Winston</td>
</tr>
<tr>
<td>2.52</td>
<td>2.48</td>
<td>9</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>2.22</td>
<td>2.36</td>
<td>9</td>
<td>1,260</td>
<td>smoke</td>
</tr>
<tr>
<td>2.16</td>
<td>2.33</td>
<td>9</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>6.49</td>
<td>2.80</td>
<td>8</td>
<td>58</td>
<td>King</td>
</tr>
<tr>
<td>4.48</td>
<td>2.70</td>
<td>8</td>
<td>234</td>
<td>Kool</td>
</tr>
<tr>
<td>5.07</td>
<td>2.74</td>
<td>8</td>
<td>155</td>
<td>Ultra</td>
</tr>
<tr>
<td>4.77</td>
<td>2.55</td>
<td>7</td>
<td>167</td>
<td>Kent</td>
</tr>
<tr>
<td>4.91</td>
<td>2.56</td>
<td>7</td>
<td>152</td>
<td>KS</td>
</tr>
<tr>
<td>2.35</td>
<td>2.13</td>
<td>7</td>
<td>895</td>
<td>nicotine</td>
</tr>
<tr>
<td>6.05</td>
<td>2.61</td>
<td>7</td>
<td>69</td>
<td>non-menthol</td>
</tr>
<tr>
<td>7.29</td>
<td>2.43</td>
<td>6</td>
<td>25</td>
<td>Golden</td>
</tr>
<tr>
<td>6.03</td>
<td>2.20</td>
<td>5</td>
<td>50</td>
<td>Capri</td>
</tr>
<tr>
<td>3.40</td>
<td>13.79</td>
<td>232</td>
<td>14,314</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

As exhibited in Table 5.2, the mutual information values for collocations made up of the node and one of the top five collocates range from 4.47 to 5.75, indicating that all of these
collocations are associated strongly. The $t$-scores for these same collocations indicate that we can be confident that the associations are valid. In addition to the top five collocates of *product* in the Tobacco-Documents Corpus, Table 5.2 lists all collocates which relate to tobacco and the tobacco industry, including brand names for cigarettes. The figures on the bottom line are for the lexical set of tobacco-related terms and are determined by totaling the frequencies of those combinations of the node and a ‘tobacco collocate,’ totaling the frequencies of the ‘tobacco collocates,’ and then using these sums as the $f(n,c)$ and $f(c)$ respectively in the statistical formulas. When regarded on an individual basis, seven of the twenty-two tobacco collocations (32%) fall below the significance level for mutual information. The highest I-values among these collocations, e.g., those with *King* (6.49), *Ultra* (5.07), *non-menthol* (6.05), *Golden* (7.29), and *Capri* (6.03), result from the low frequencies of occurrence of the collocates in relation to their frequencies of co-occurrence with the node. On the other hand, the $t$-scores for all twenty-two collocations indicate that the associations are significant. The values for collocations with the lexical set of tobacco-related terms are far more noteworthy. The I-value of 3.4 reveals that there is a strong association, and the $t$-score indicates a high level of confidence that there is an association between the node and these terms. This is very important as it verifies that for the term *product* we can be confident that within the span of 4:4 industry and/or company-related terms will occur significantly.

In addition to the fact that the tobacco-related terms form significant collocations with *product* both as a set and generally on an individual basis, it is interesting to consider their positions within the span. In other words, will certain span positions be more likely to contain these types of terms than other positions? Table 5.3 shows the frequencies and rates of occurrence of the tobacco-related collocates of *product* according to span positions.
Table 5.3: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>34 (16%)</td>
<td>40 (17%)</td>
<td>37 (16%)</td>
<td>56 (24%)</td>
<td>*</td>
<td>3 (1%)</td>
<td>16 (7%)</td>
<td>15 (7%)</td>
<td>31 (13%)</td>
</tr>
</tbody>
</table>

As indicated, 72 percent of the tobacco-related collocates occur to the left of the node, *product*, with approximately a quarter of them immediately preceding it.

5.2 Brand

In the reference corpus, *brand* collocates most frequently with *name, product, manager, names,* and *new*. Additionally, *name* is the top collocate in the R1 position, occurring there in 83% of its seventy co-occurrences with *brand*. *Product*, which forms a collocation with *brand* fifty times, is the main collocate in the L2 and L4 positions, 30% and 16% respectively. The other three terms are not top collocates in any of the positions. It is worth noting that, while *name* appears predominantly to the right of *brand* (88% of the instances), it does appear to the left as well, especially in the L4 position. On the other hand, *names* occurs only to the right of the node, with twenty-nine out of thirty-one instances (94%) in the R1 position and one occurrence each in the R3 and R4 positions. *Manager* exhibits a similar pattern, with only one instance to the left of *brand*. The vast majority (73%) of these five terms appear to the right of the node. The other leading collocates according to span positions with their overall frequency of co-occurrence with the node and their rate of occurrence in that position are:

L1 = **strong** 20, 70%; L3 = **consumers** 18, 45%; **meaning** 15, 54%;
R2 = **strategies** 19, 43%; R2 = **membership** 11, 64%; **market** 21, 38%.

Table 5.4 indicates that all the collocations of *brand* with these terms are significant. There are, however, some discrepancies in the level of significance. *Name* has both a high I-value and a high t-score, indicating that it is a typical collocation. On the other hand, *names*
receives a high I-value, but its t-score is somewhat lower. This could be due to the fact that the collocation can be considered a technical one, thus giving it the high I-value (an observed frequency that is approximately 362 times greater than its expected frequency), while the absolute frequency of names is low, resulting in the lower t-score.

Table 5.4: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.50</td>
<td>8.34</td>
<td>70</td>
<td>243</td>
<td>name</td>
</tr>
<tr>
<td>4.60</td>
<td>6.78</td>
<td>50</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>5.78</td>
<td>5.81</td>
<td>35</td>
<td>800</td>
<td>manager</td>
</tr>
<tr>
<td>8.81</td>
<td>5.56</td>
<td>31</td>
<td>87</td>
<td>names</td>
</tr>
<tr>
<td>4.10</td>
<td>4.80</td>
<td>26</td>
<td>1,900</td>
<td>new</td>
</tr>
<tr>
<td>3.29</td>
<td>4.11</td>
<td>21</td>
<td>2,706</td>
<td>market</td>
</tr>
<tr>
<td>6.91</td>
<td>4.43</td>
<td>20</td>
<td>209</td>
<td>strong</td>
</tr>
<tr>
<td>5.89</td>
<td>4.29</td>
<td>19</td>
<td>404</td>
<td>strategies</td>
</tr>
<tr>
<td>5.52</td>
<td>4.15</td>
<td>18</td>
<td>494</td>
<td>consumers</td>
</tr>
<tr>
<td>7.76</td>
<td>3.86</td>
<td>15</td>
<td>87</td>
<td>meaning</td>
</tr>
<tr>
<td>8.11</td>
<td>3.30</td>
<td>11</td>
<td>50</td>
<td>membership</td>
</tr>
</tbody>
</table>

Three-word collocations with the node, with their frequency of occurrence, include brand’s category membership, 6; and brand name strategies, 3. Extended collocations with four terms are product or brand manager, with six occurrences, and the company as brand, with five occurrences.

The top five collocates of brand in the Tobacco-Documents Corpus include new, smokers, styles, awareness, and competitive. Of these collocates, both new and smokers co-occur with brand fifty times. New is the main collocate in the L1 position, occurring there in 62% of its co-occurrences. Smokers is the most frequent collocate in the L3 position, with eleven instances (22%), and in the L4 position, with twelve instances (24%). Styles is the chief collocate in the R1 position where it occurs sixteen times (67%). The primary collocates in the remaining span
positions, with their overall frequency of occurrence with the node and rate of occurrence in that position, are:

\[ L2 = \text{low} \ 14, 86\%; \ R2 = \text{people} \ 9, 89\%; \ R3 = \text{smoked} \ 13, 38\%; \ R4 = \text{market} \ 18, 50\%. \]

Table 5.5: mutual information and \( t \)-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>( I(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.91</td>
<td>6.84</td>
<td>50</td>
<td>1,224</td>
<td>new</td>
</tr>
<tr>
<td>4.80</td>
<td>6.82</td>
<td>50</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>7.64</td>
<td>4.87</td>
<td>24</td>
<td>88</td>
<td>styles</td>
</tr>
<tr>
<td>7.00</td>
<td>4.76</td>
<td>23</td>
<td>132</td>
<td>awareness</td>
</tr>
<tr>
<td>6.09</td>
<td>4.73</td>
<td>23</td>
<td>247</td>
<td>competitive</td>
</tr>
<tr>
<td>4.80</td>
<td>4.42</td>
<td>21</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>3.25</td>
<td>4.00</td>
<td>20</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>4.70</td>
<td>4.19</td>
<td>19</td>
<td>537</td>
<td>Camel</td>
</tr>
<tr>
<td>3.69</td>
<td>3.91</td>
<td>18</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>6.11</td>
<td>3.82</td>
<td>15</td>
<td>159</td>
<td>Slims</td>
</tr>
<tr>
<td>4.33</td>
<td>3.68</td>
<td>15</td>
<td>548</td>
<td>tar</td>
</tr>
<tr>
<td>4.24</td>
<td>3.54</td>
<td>14</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>3.03</td>
<td>3.28</td>
<td>14</td>
<td>1,260</td>
<td>smoke</td>
</tr>
<tr>
<td>6.04</td>
<td>3.55</td>
<td>13</td>
<td>145</td>
<td>smoked</td>
</tr>
<tr>
<td>5.08</td>
<td>3.50</td>
<td>13</td>
<td>282</td>
<td>Winston</td>
</tr>
<tr>
<td>3.98</td>
<td>3.11</td>
<td>11</td>
<td>511</td>
<td>Lights</td>
</tr>
<tr>
<td>5.74</td>
<td>3.25</td>
<td>11</td>
<td>151</td>
<td>Virginia</td>
</tr>
<tr>
<td>5.28</td>
<td>2.92</td>
<td>9</td>
<td>170</td>
<td>Viceroy</td>
</tr>
<tr>
<td>6.41</td>
<td>2.80</td>
<td>8</td>
<td>69</td>
<td>non-menthol</td>
</tr>
<tr>
<td>4.00</td>
<td>2.65</td>
<td>8</td>
<td>366</td>
<td>Philip</td>
</tr>
<tr>
<td>1.29</td>
<td>1.67</td>
<td>8</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>4.46</td>
<td>2.53</td>
<td>7</td>
<td>234</td>
<td>Kool</td>
</tr>
<tr>
<td>3.75</td>
<td>2.45</td>
<td>7</td>
<td>381</td>
<td>Morris</td>
</tr>
<tr>
<td>4.00</td>
<td>2.48</td>
<td>7</td>
<td>320</td>
<td>RJR</td>
</tr>
<tr>
<td>1.37</td>
<td>1.50</td>
<td>6</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>3.65</td>
<td>2.06</td>
<td>5</td>
<td>293</td>
<td>smoker</td>
</tr>
<tr>
<td>3.92</td>
<td>16.15</td>
<td>299</td>
<td>14,504</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

With the exception of \textit{tobacco} and \textit{smoking}, all of the terms form significant associations with \textit{brand}, as can be seen in Table 5.5. It is not unexpected that these two terms have both a low I-value and a low \( t \)-score because, although they each have a high absolute frequency, their frequencies of co-occurrence with the node are low. The I-value for collocations with the lexical
set of tobacco-related terms indicates that the probability of these co-occurrences is fifteen times greater than chance. The high t-score shows that we can be confident that there is a strong association between *brand* and the lexical set. From this, we can expect that industry and/or company-specific terms will occur significantly within a span of 4:4 of *brand*.

In considering the collocates listed in Table 5.5, a couple of points must be made concerning extended collocations. First, every instance of *Virginia* is immediately followed by *Slims*. Although there are four more occurrences of *Slims*, it is due to the fact that the accompanying *Virginia* lies outside the 4:4 span. Second, the same situation applies to *Philip* and *Morris*, with *Morris* outside the span in one instance.

The frequencies and rates of occurrence of the tobacco-related collocates of *brand* according to span positions are presented in Table 5.6.

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>49 (16%)</td>
<td>45 (15%)</td>
<td>35 (12%)</td>
<td>50 (17%)</td>
<td>*</td>
<td>15 (5%)</td>
<td>29 (10%)</td>
<td>36 (12%)</td>
<td>40 (13%)</td>
</tr>
</tbody>
</table>

In the case of *brand*, 60% of the tobacco-related collocates occur to the left of the node. However, with the exception of the R1 position, these terms are rather equally distributed across the span positions.

5.3 Data

*Data* collocates most frequently with *warehouse, operational, entry, systems, and information* in the reference corpus. Of these five terms, three are key collocates positionally as well. *Warehouse* is the top collocate in the R1 position, with sixty-three of its 102 collocations with *data* (62%). *Operational*, which co-occurs with *data* forty-eight times, is the chief collocate in
both the L4 position, where it appears nine times (19%), and the R3 position, where it occurs eleven times (23%). *Systems*, which forms forty-one collocations with *data*, prevails in both the L3 position, with nine occurrences (22%), and the R4 position, with thirteen occurrences (32%). Sixty-seven percent of the occurrences of the top five collocates are to the right of the node. In the remaining positions, the main collocates per position, with their overall frequency of occurrence with the node and rate of occurrence, are:

L1 = *financial* 29, 83%; L2 = *using* 22, 45%; R2 = *process* 31, 55%.

Of the top five collocates, *operational* and *systems* can be considered as forming an extended collocation with *data*. *Systems* immediately follows *operational* in fourteen of the collocations, eight to the left of the node and six to the right. In five of the occurrences to the left, *operational systems* is followed by one of two prepositional phrases: *to the data* or *into the data*. All six instances of *operational systems* to the right are preceded by the prepositional phrase *from the data*. Another extended collocation which is worth mentioning is that of *data acquisition process*. In twelve of the seventeen occurrences of *process* in the R2 position (71%), it is immediately preceded by *acquisition*.

Table 5.7: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>n = data, f(n) = 1,078</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Table" /></td>
</tr>
</tbody>
</table>
All the collocations listed in Table 5.7 fall clearly in the range of significant associations. Even the collocation composed of *financial* and *data* receives an I-value signifying that the probability of its occurrence is greater than eighteen times that due to chance. The collocation of *warehouse* and *data* is particularly significant as indicated by both statistics. Additionally, it is interesting to observe that *data* is particularly important as a collocate of *warehouse*. In the 115 instances of *warehouse* in the reference corpus, 102 (89%) of them are with *data*. On the other hand, *warehouse* is not as important to *data*, as its co-occurrence rate with *data* is less than 10%.

For the Tobacco-Documents Corpus, the top five collocates of *data* are *analysis*, *collected*, *presented*, *class*, and *obtained*. Of these five terms, only *presented* is not a primary collocate in any of the span positions. *Analysis*, the most frequently co-occurring word (in thirty-five collocations with *data*), is the main collocate in the R2 position, where 43% of its occurrences appear. *Collected* occurs in the R1 position in 57% of its co-occurrences. *Class* is the prevalent collocate in the L3 position and the L4 position, appearing in each of the two positions five times. It is also the top collocate in the R4 position at eight occurrences. The other predominant collocates with their overall frequency of occurrence with *data* and rate of occurrence in a span position are:

- L2 = *acquired* 14, 93%; R3 = *available* 18, 28%; *smoker* 13, 38%.

As Table 5.8 exhibits, the I-values for collocations involving the non-tobacco terms are well above the 3.0 cut-off point for significance, indicating that the associations are strong. In addition, the *t*-scores for these same terms show that we can be confident that these are true associations and not the result of chance. To the contrary, the I-values for collocations containing the tobacco terms do not provide the same results. Only two of the terms, *smoking* and *smoker*, have I-values for their collocations pointing to the existence of associations. Even the I-value for
collocations including the lexical set as a whole is below the cut-off point. In regard to the \( t \)-scores, we cannot be confident that three of the six tobacco terms form significant collocations with the node. As a lexical set, however, these terms do yield a score high enough to provide a level of confidence that their occurrence is not due to chance. This is in keeping with the fact that the \( t \)-score is suited for frequently occurring items; therefore, the higher \( t \)-score is reflected in the combined frequency of 8,148 occurrences. Based on the \( t \)-score, we can state that industry and/or company-related words will occur within the specified environment of \textit{data}.

Table 5.8: mutual information and \( t \)-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>( I(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.24</td>
<td>5.84</td>
<td>35</td>
<td>413</td>
<td>analysis</td>
</tr>
<tr>
<td>8.23</td>
<td>4.78</td>
<td>23</td>
<td>68</td>
<td>collected</td>
</tr>
<tr>
<td>6.53</td>
<td>4.42</td>
<td>20</td>
<td>193</td>
<td>presented</td>
</tr>
<tr>
<td>7.52</td>
<td>4.34</td>
<td>19</td>
<td>92</td>
<td>class</td>
</tr>
<tr>
<td>6.39</td>
<td>4.31</td>
<td>19</td>
<td>202</td>
<td>obtained</td>
</tr>
<tr>
<td>3.06</td>
<td>3.52</td>
<td>16</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>5.30</td>
<td>3.51</td>
<td>13</td>
<td>293</td>
<td>smoker</td>
</tr>
<tr>
<td>1.74</td>
<td>2.10</td>
<td>9</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>1.79</td>
<td>1.74</td>
<td>6</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>2.02</td>
<td>1.85</td>
<td>6</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>2.31</td>
<td>1.79</td>
<td>5</td>
<td>895</td>
<td>nicotine</td>
</tr>
<tr>
<td>2.59</td>
<td>6.18</td>
<td>55</td>
<td>8,148</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

Not surprisingly, the tobacco-related collocates of \textit{data} were what might be called ‘generic tobacco terms.’ There were no collocates which were names of either particular products or companies.

Table 5.9 illustrates the span positions of the tobacco terms with their frequencies and rates of occurrence in those positions. Fifty-eight percent of the collocates occur to the left of the node.
Table 5.9: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 (18%)</td>
<td>8 (15%)</td>
<td>3 (5%)</td>
<td>11 (20%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>3 (5%)</td>
<td>14 (26%)</td>
<td>6 (11%)</td>
</tr>
</tbody>
</table>

5.4 Products

The top five collocates of products in the reference corpus are new, services, markets, customers, and market. Within the span of 4:4, new collocates with products a total of 151 times. It is not only the most frequent collocate of products, but also the primary collocate in the L1 position, with 118 occurrences (78%). The other four collocates are also key terms in span positions. Services occurs 124 out of 141 times (88%) in the R2 position. Markets is the leading collocate in the R3 position, with seventeen of its sixty-three co-occurrences with the node (27%). In the R4 position, customers appears the most often, eleven out of fifty-six collocations with products (20%). Market, which forms forty-nine collocations with products, is the principal collocate in the L4 position with sixteen instances (33%). When compared with the top five collocates of the previous nodes (product, brand, and data), whose overall occurrences greatly favored the right side of the node, these key collocates are rather evenly distributed, with 48% of them occurring to the left and 52% of them to the right. The main collocates in the other three positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L2 = range 14, 79%; L3 = demand 28, 39%; R1 = division 12, 83%.

As Table 5.10 demonstrates, all of the prime collocates, both overall and positional, form significant collocations with products. Due to the combination of relatively high frequencies of occurrence of new and services, as well as their high frequencies of co-occurrence with the node, the collocations containing them receive high levels of confidence in the associations. The two associations with the lowest t-scores, those formed with range or division, have rather low
frequencies in both areas; however, they both receive I-values which indicate that the probability of their occurrences are approximately thirty-eight and fifty-two times greater than chance respectively. This is consistent with the mutual information score, which tends to focus on technical terms.

Table 5.10: mutual information and $t$-scores from the reference corpus

<table>
<thead>
<tr>
<th>$I(n,c)$</th>
<th>$t(n,c)$</th>
<th>$f(n,c)$</th>
<th>$f(c)$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.80</td>
<td>12.07</td>
<td>151</td>
<td>1,900</td>
<td>new</td>
</tr>
<tr>
<td>6.95</td>
<td>11.78</td>
<td>141</td>
<td>800</td>
<td>services</td>
</tr>
<tr>
<td>5.67</td>
<td>7.78</td>
<td>63</td>
<td>865</td>
<td>markets</td>
</tr>
<tr>
<td>4.98</td>
<td>7.25</td>
<td>56</td>
<td>1,242</td>
<td>customers</td>
</tr>
<tr>
<td>3.66</td>
<td>6.45</td>
<td>49</td>
<td>2,706</td>
<td>market</td>
</tr>
<tr>
<td>3.79</td>
<td>4.91</td>
<td>28</td>
<td>1,415</td>
<td>demand</td>
</tr>
<tr>
<td>5.25</td>
<td>3.64</td>
<td>14</td>
<td>258</td>
<td>range</td>
</tr>
<tr>
<td>5.71</td>
<td>3.40</td>
<td>12</td>
<td>161</td>
<td>division</td>
</tr>
</tbody>
</table>

In regard to extended collocations, the concordance lines show that in ten of the eleven instances of range in the L2 position, the collocation range of products occurs. Of the 124 occurrences of services in the R2 position, 120 of them consist of the collocation products and services or products or services.

The most frequent collocate of products in the Tobacco-Documents Corpus is tobacco. It forms 158 collocations with the node and is the principal collocate in the L1 position with 139 occurrences (88%). The other top four collocates include new, low, delivery, and use. Of these four terms, only use does not appear as a main collocate in any of the positions. New is one of four collocates which appear most often in the R2 position, each occurring four times. Although new occurs more frequently in other positions (e.g., forty-one occurrences in L1), it is not the most frequent term in those positions. The other collocates which have the same frequency in the R2 position are: low (16%), develop (36%), and Japan (44%). Low is also the main collocate in
the L2 position, with eleven of its twenty-five occurrences (44%) appearing there. *Delivery* is a primary collocate in the R3 position with four occurrences (17%), as is *well* (67%). Leading collocates in the remaining positions, with their overall frequency of occurrence with *products* and their rate of occurrence in the position, are:

\[
\text{L3} = \text{sale } 10, 100\%; \text{ L4} = \text{cigarettes } 14, 43\%; \text{ R1} = \text{tested } 17, 65\%.
\]

There are two extended collocations with *products* in the corpus. The four-word units are *sale of tobacco products*, with seven occurrences, and *use of tobacco products*, with six occurrences.

Collocations with four of the tobacco-related terms fall below the cut-off point for significant association when the mutual information statistic is applied. These four terms — *smoke, smokers, smoking*, and *nicotine* — are included in Table 5.11. The *t*-score for the combination with *nicotine* also falls below the confidence level. Most noticeable from the list of collocates is *smokeless*, which collocates with *products* in 100% of its occurrences in the corpus. This means that, while the collocation is significant, it is far more important for *smokeless* than vice versa. For the collocations with the lexical set of tobacco-related terms, the I-value indicates that the occurrence of these collocations is twenty-two times greater than that due to chance. In addition, the *t*-score provides a high level of confidence that the association is real and indicates that industry and/or company-specific words will fall within the span. As with the node *brand*, all occurrences of *Philip* are immediately followed by *Morris*. Discrepancies between the frequencies of the two are due to a term’s lying outside the span.
Table 5.11: mutual information and t-scores from the Tobacco-Documents Corpus

\[ n = \text{products}, f(n) = 670 \]

<table>
<thead>
<tr>
<th>( I(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.87</td>
<td>12.35</td>
<td>158</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>5.51</td>
<td>7.76</td>
<td>63</td>
<td>1,224</td>
<td>new</td>
</tr>
<tr>
<td>5.20</td>
<td>4.86</td>
<td>25</td>
<td>603</td>
<td>low</td>
</tr>
<tr>
<td>6.77</td>
<td>4.75</td>
<td>23</td>
<td>186</td>
<td>delivery</td>
</tr>
<tr>
<td>4.55</td>
<td>4.17</td>
<td>19</td>
<td>717</td>
<td>use</td>
</tr>
<tr>
<td>4.88</td>
<td>4.10</td>
<td>18</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>6.63</td>
<td>4.08</td>
<td>17</td>
<td>152</td>
<td>KS</td>
</tr>
<tr>
<td>3.01</td>
<td>3.28</td>
<td>14</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>3.60</td>
<td>3.43</td>
<td>14</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>6.03</td>
<td>3.27</td>
<td>11</td>
<td>149</td>
<td>Lorillard</td>
</tr>
<tr>
<td>4.14</td>
<td>3.13</td>
<td>11</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>4.67</td>
<td>3.19</td>
<td>11</td>
<td>381</td>
<td>Morris</td>
</tr>
<tr>
<td>2.95</td>
<td>2.89</td>
<td>11</td>
<td>1,260</td>
<td>smoke</td>
</tr>
<tr>
<td>2.75</td>
<td>2.69</td>
<td>10</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>9.79</td>
<td>3.00</td>
<td>9</td>
<td>9</td>
<td>smokeless</td>
</tr>
<tr>
<td>3.86</td>
<td>2.79</td>
<td>9</td>
<td>548</td>
<td>tar</td>
</tr>
<tr>
<td>4.27</td>
<td>2.68</td>
<td>8</td>
<td>366</td>
<td>Philip</td>
</tr>
<tr>
<td>2.05</td>
<td>2.15</td>
<td>8</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>4.82</td>
<td>2.55</td>
<td>7</td>
<td>220</td>
<td>Reynolds</td>
</tr>
<tr>
<td>6.52</td>
<td>2.42</td>
<td>6</td>
<td>58</td>
<td>King</td>
</tr>
<tr>
<td>5.10</td>
<td>2.38</td>
<td>6</td>
<td>155</td>
<td>Ultra</td>
</tr>
<tr>
<td>7.65</td>
<td>2.22</td>
<td>5</td>
<td>22</td>
<td>Lark</td>
</tr>
<tr>
<td>2.31</td>
<td>1.78</td>
<td>5</td>
<td>895</td>
<td>nicotine</td>
</tr>
<tr>
<td>3.79</td>
<td>2.07</td>
<td>5</td>
<td>320</td>
<td>RJR</td>
</tr>
<tr>
<td>4.48</td>
<td>17.69</td>
<td>343</td>
<td>13,613</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

The frequencies and rates of occurrence of the tobacco-related collocates of product according to span positions are exhibited in Table 5.12.

Table 5.12: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>16</td>
<td>32</td>
<td>196</td>
<td>57%</td>
<td>*</td>
<td>1</td>
<td>4%</td>
<td>20</td>
</tr>
</tbody>
</table>

The vast majority of tobacco-related collocates, 82% appear to the left of products. This is due in large part to the collocate tobacco, which occurs 149 out of 158 times (94%) to the left of the node.
5.5 **Brands**

The five most frequent collocates of *brands* in the reference corpus are *new, global, product, strong,* and *value.* While *new* is the most frequent collocate overall, it is not a dominant collocate in any of the span positions, nor is *global.* Both terms do occur eleven times in the L1 position, but the prime collocate there is *strong,* with twelve of its fourteen occurrences (86%). The R2 position has two main collocates, *product* and *retailers,* each with three occurrences. *Product* occurs there in 21% of the cases in which it collocates with *brands; retailers* in 33% of its collocations. *Value* is prevalent in the L3 position with five instances (38%). Other leading collocates in span positions, with their overall frequency of occurrence with the node and rate of occurrence in that position, include:

\[
\begin{align*}
\text{L4} &= \text{company 7, 57\%; L2 = power 5, 80\%; R1 = need 5, 100\%; R3 = market 8, 38\%;}\cr
\text{category 5, 60\%.}
\end{align*}
\]

Contrary to the previous four nodes, the majority of the top five collocates (83%) occur to the left of the node.

Neither *market* nor *company* forms significant collocations with *brands* according to their I-values, as Table 5.13 illustrates. Additionally, both associations receive rather low t-scores. Due to the high frequencies of occurrence of the collocates in the corpus, but their low frequencies of co-occurrence with *brands,* the collocations are not interesting and could be due to chance. The strongest collocation on this list appears to be that composed of *global* and *brands,* whose I-value indicates that its observed frequency is approximately 193 times greater than what would be expected by chance. The t-score also lends confidence to this association. There are no extended collocations of interest for this node.
Table 5.13: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>n (c)</th>
<th>t (c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.54</td>
<td>3.95</td>
<td>17</td>
<td>1,900</td>
<td>new</td>
</tr>
<tr>
<td>7.59</td>
<td>3.72</td>
<td>14</td>
<td>190</td>
<td>global</td>
</tr>
<tr>
<td>3.82</td>
<td>3.48</td>
<td>14</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>7.45</td>
<td>3.72</td>
<td>14</td>
<td>209</td>
<td>strong</td>
</tr>
<tr>
<td>4.30</td>
<td>3.42</td>
<td>13</td>
<td>1,717</td>
<td>value</td>
</tr>
<tr>
<td>4.44</td>
<td>3.30</td>
<td>12</td>
<td>1,438</td>
<td>products</td>
</tr>
<tr>
<td>7.15</td>
<td>2.98</td>
<td>9</td>
<td>165</td>
<td>retailers</td>
</tr>
<tr>
<td>2.95</td>
<td>2.46</td>
<td>8</td>
<td>2,706</td>
<td>market</td>
</tr>
<tr>
<td>2.46</td>
<td>2.16</td>
<td>7</td>
<td>3,327</td>
<td>company</td>
</tr>
<tr>
<td>5.96</td>
<td>2.20</td>
<td>5</td>
<td>209</td>
<td>category</td>
</tr>
<tr>
<td>3.77</td>
<td>2.07</td>
<td>5</td>
<td>959</td>
<td>need</td>
</tr>
<tr>
<td>5.58</td>
<td>2.19</td>
<td>5</td>
<td>273</td>
<td>power</td>
</tr>
</tbody>
</table>

As Table 5.14 shows, the five most frequently occurring collocates of brands in the Tobacco-Documents Corpus are new, competitive, cigarette, American, and discount. Of these five, only new is a dominant collocate in a span position. It is the main collocate in the L1 position, with twenty-eight occurrences (64%). In the L4 position, new occurs three times (7%), as do smokers (14%), market (16%), research (38%), and smoke (50%). The principal collocates in the other positions, with their overall frequency of occurrence with the node and their rate of occurrence in that position, are:

L2 = Philip 20, 70%; L3 = smokers 21, 48%; R1 = tested 7, 71%; R2 = Kool 11, 55%; R3 = cigarettes 13, 62%; R4 = market 19, 32%.

While new is the most frequent collocate in the L1 position, it is worth noting that cigarette appears in that position in twenty-five of its thirty-one co-occurrences (81%) with brand, and American occurs there twenty-two out of twenty-nine times (76%).
Table 5.14: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.12</td>
<td>6.44</td>
<td>44</td>
<td>1,224 new</td>
</tr>
<tr>
<td>7.10</td>
<td>5.87</td>
<td>35</td>
<td>247 competitive</td>
</tr>
<tr>
<td>4.28</td>
<td>5.28</td>
<td>31</td>
<td>1,541 cigarette</td>
</tr>
<tr>
<td>6.29</td>
<td>5.32</td>
<td>29</td>
<td>359 American</td>
</tr>
<tr>
<td>7.38</td>
<td>4.97</td>
<td>25</td>
<td>145 discount</td>
</tr>
<tr>
<td>3.94</td>
<td>4.28</td>
<td>21</td>
<td>1,317 smokers</td>
</tr>
<tr>
<td>5.66</td>
<td>4.38</td>
<td>20</td>
<td>366 Philip</td>
</tr>
<tr>
<td>5.72</td>
<td>4.39</td>
<td>20</td>
<td>359 American</td>
</tr>
<tr>
<td>4.62</td>
<td>3.59</td>
<td>14</td>
<td>548 tar</td>
</tr>
<tr>
<td>3.62</td>
<td>3.31</td>
<td>13</td>
<td>1,024 cigarettes</td>
</tr>
<tr>
<td>4.51</td>
<td>3.45</td>
<td>13</td>
<td>552 Marlboro</td>
</tr>
<tr>
<td>5.50</td>
<td>3.24</td>
<td>11</td>
<td>234 Kool</td>
</tr>
<tr>
<td>4.15</td>
<td>2.98</td>
<td>10</td>
<td>542 menthol</td>
</tr>
<tr>
<td>2.01</td>
<td>2.38</td>
<td>10</td>
<td>2,395 tobacco</td>
</tr>
<tr>
<td>5.15</td>
<td>2.57</td>
<td>7</td>
<td>190 Salem</td>
</tr>
<tr>
<td>1.99</td>
<td>1.98</td>
<td>7</td>
<td>1,707 smoking</td>
</tr>
<tr>
<td>4.58</td>
<td>2.54</td>
<td>7</td>
<td>282 Winston</td>
</tr>
<tr>
<td>2.20</td>
<td>1.92</td>
<td>6</td>
<td>1,260 smoke</td>
</tr>
<tr>
<td>5.06</td>
<td>2.17</td>
<td>5</td>
<td>145 smoked</td>
</tr>
<tr>
<td>4.04</td>
<td>2.10</td>
<td>5</td>
<td>293 smoker</td>
</tr>
<tr>
<td>3.92</td>
<td>13.21</td>
<td>200</td>
<td>12,777 tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

The mutual information values and the t-scores indicate that tobacco, smoking, and smoke do not form significant collocations with brands in this corpus. This is due to the high absolute frequencies of these terms and the low frequencies with which they collocate with the node. While several of the collocations involving individual tobacco terms score relatively high on mutual information, the highest level of confidence in the associations (the t-score) is reflected in the top five collocates. When considering the collocations formed from the lexical set as a whole, we can be confident that the associations are not due to chance and that company/industry-specific words will appear within the allotted span.

As previously pointed out, all occurrences of Philip are immediately followed by Morris and vice versa. This collocation can be regarded as an extended collocation. The only other
extended collocation of any interest is *low tar brands*, which occurs in five of the seven instances of *low* as a collocate of *brands*.

Table 5.15 presents the frequencies and rates of occurrence of the tobacco-related collocates of *brands* according to the span positions.

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 (10%)</td>
<td>23 (11.5%)</td>
<td>25 (12.5%)</td>
<td>59 (30%)</td>
<td>*</td>
<td>5 (2%)</td>
<td>26 (13%)</td>
<td>22 (11%)</td>
<td>20 (10%)</td>
</tr>
</tbody>
</table>

Sixty-four percent of the tobacco-related collocates appear to the left of *brands*.

5.6 **Advertising**

The five principal collocates of *advertising* in the reference corpus are *sales*, *marketing*, *promotion*, *co-op*, and *budget*. In addition to being the most frequent collocate of *advertising*, *sales* is the main collocate in three of the span positions. It forms forty-seven collocations with the node, with 26% of its occurrences in the L4 position, 32% in the R2 position, and 13% in the R4 position. *Marketing* is the leading collocate in both the L3 position (16% of its thirty-one occurrences) and the L2 position (29%). The primary collocate in the R3 position is *promotion*, occurring there in eleven of twenty-eight collocations with *advertising* (39%); and *co-op* is the most frequent collocate in the L1 position, where 93% of its occurrences with the node appear. *Budget* is the only one of the five terms which is not key in any of the span positions. The top collocate in the remaining position, R1, is *agency*. It occurs there in twenty-two of its twenty-three occurrences (96%). Fifty-seven percent of the occurrences of the top five collocates occur to the right of *advertising*. 
In searching for extended collocations with advertising, only one such combination is readily apparent. In ten of the eleven instances of promotion in the R3 position, the word immediately preceding promotion is sales. The extended collocation is either advertising and sales promotion or advertising, sales promotion.

Table 5.16: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>6.69</td>
<td>47</td>
<td>2,301</td>
<td>sales</td>
</tr>
<tr>
<td>5.18</td>
<td>5.41</td>
<td>31</td>
<td>1,759</td>
<td>marketing</td>
</tr>
<tr>
<td>8.32</td>
<td>5.27</td>
<td>28</td>
<td>180</td>
<td>promotion</td>
</tr>
<tr>
<td>8.91</td>
<td>5.19</td>
<td>27</td>
<td>115</td>
<td>co-op</td>
</tr>
<tr>
<td>6.44</td>
<td>4.94</td>
<td>25</td>
<td>593</td>
<td>budget</td>
</tr>
<tr>
<td>8.86</td>
<td>4.79</td>
<td>23</td>
<td>102</td>
<td>agency</td>
</tr>
</tbody>
</table>

The mutual information values and t-scores in Table 5.16 indicate that the resulting collocations are significant. Those collocations with promotion, co-op, and agency score particularly high I-values, which point to their technical status. The collocations involving sales and marketing receive high t-scores as a result of their high absolute frequencies combined with their frequencies of co-occurrence.

The top five collocates of advertising in the Tobacco-Documents Corpus are cigarette, recall, new, tobacco, and promotion. Cigarette is the principal collocate in three span positions — L1, L3, and L4. Seventy-eight percent of the instances of cigarette occur in the L1 position. In the L3 position, cigarette, new, and ban appear five times each, with their rates of occurrence in that position 8%, 22%, and 45% respectively. Cigarette and recall both occur three times in L4, which translates into 5% of the occurrences of cigarette and 10% of those of recall. Recall is also the prime collocate in the R1 position. There it appears six times (21%), as does campaign
(40%) and agencies (100%). Promotion prevails in the R2 position, where it occurs thirteen times (68%). Tobacco is not the most frequent collocate in any of the positions. The leading collocates in the remaining positions, with their overall frequency of occurrence with the node and their rate of occurrence in that position, are:

L2 = packaging 11, 55%; Marlboro 10, 60%; source 6, 100%; R3 = products 13, 46%; R4 = products 13, 31%; media 10, 40%.

Table 5.17: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>n</th>
<th>f(n)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>advertising</td>
<td>546</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I (n,c)</td>
<td>t (n,c)</td>
<td>f(n,c)</td>
<td>f(c)</td>
</tr>
<tr>
<td>5.47</td>
<td>7.76</td>
<td>63</td>
<td>1,541</td>
</tr>
<tr>
<td>8.66</td>
<td>5.37</td>
<td>29</td>
<td>78</td>
</tr>
<tr>
<td>4.35</td>
<td>4.56</td>
<td>23</td>
<td>1,224</td>
</tr>
<tr>
<td>3.38</td>
<td>4.34</td>
<td>23</td>
<td>2,395</td>
</tr>
<tr>
<td>6.34</td>
<td>4.31</td>
<td>19</td>
<td>255</td>
</tr>
<tr>
<td>5.19</td>
<td>4.13</td>
<td>18</td>
<td>537</td>
</tr>
<tr>
<td>3.67</td>
<td>3.19</td>
<td>12</td>
<td>1,024</td>
</tr>
<tr>
<td>4.30</td>
<td>3.00</td>
<td>10</td>
<td>552</td>
</tr>
<tr>
<td>2.67</td>
<td>2.67</td>
<td>10</td>
<td>1,707</td>
</tr>
<tr>
<td>9.24</td>
<td>3.16</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>6.59</td>
<td>2.62</td>
<td>7</td>
<td>79</td>
</tr>
<tr>
<td>7.03</td>
<td>2.43</td>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td>5.36</td>
<td>2.39</td>
<td>6</td>
<td>159</td>
</tr>
<tr>
<td>3.41</td>
<td>2.03</td>
<td>5</td>
<td>511</td>
</tr>
<tr>
<td>2.60</td>
<td>1.87</td>
<td>5</td>
<td>895</td>
</tr>
<tr>
<td>8.01</td>
<td>2.23</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>2.11</td>
<td>1.72</td>
<td>5</td>
<td>1,260</td>
</tr>
<tr>
<td>5.17</td>
<td>2.17</td>
<td>5</td>
<td>151</td>
</tr>
<tr>
<td>4.24</td>
<td>13.06</td>
<td>190</td>
<td>10,900</td>
</tr>
</tbody>
</table>

As exhibited in Table 5.17, collocations involving two of the terms, nicotine and smoke, fall below the level of significance for both mutual information and t-score. Not surprisingly, collocations containing the names of products, for example Tramps and Players, score very high on mutual information, as MI is a good indicator of unusual collocations. The t-score focuses on
recurrent items, and this is reflected in the score it assigns to the collocations containing cigarette.

The majority of the tobacco-related terms form significant and interesting collocations with advertising. The I-value for collocations of the lexical set of tobacco terms shows that their occurrences are almost nineteen times greater than chance, and the t-score provides a high level of confidence that these associations are valid. Therefore, we can conclude that industry and/or company-specific terms will occur within a 4:4 span of advertising.

Table 5.18: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13 (7%)</td>
<td>23 (12%)</td>
<td>26 (14%)</td>
<td>79 (42%)</td>
<td>*</td>
<td>3 (2%)</td>
<td>22 (11%)</td>
<td>16 (8%)</td>
<td>8 (4%)</td>
</tr>
</tbody>
</table>

Table 5.18 shows the tobacco terms with their frequencies and rates of occurrence according to span position. Seventy-five percent of the terms appear to the left of advertising.

5.7 SHARE

In the reference corpus, the five most frequent collocates of share are market, per, price, earnings, and cents. Market is the principal collocate in four of the span positions — L1, L3, R3, and R4. The majority of its occurrences, 72%, occur in the L1 position. The rates of occurrence in the other positions are 5% in L3 and 8% in both the R3 and R4 positions. While per is not a dominant collocate in any position, it does occur sixty-five times (90%) in the L1 position, making its appearance in that position notable. In the L2 position, two terms are dominant, earnings and cents, with eighteen occurrences each. For earnings, the rate of occurrence is 64%; for cents, it is 95%. The main collocate in the L4 position is price (21% of its occurrences), and it is also the primary term in R1 (28% of its occurrences). The only position whose primary
collocate is not in the top five is R2, where *total* appears seven out of thirteen times (54%). Eighty-one percent of the five most frequent collocates occur to the left of *share*, helped in large part by the occurrences of *market* and *per* in the L1 position.

In regard to extended collocations, seventeen out of eighteen instances of *earnings* are immediately followed by *per*, resulting in the collocation *earnings per share*. *Price* also forms part of the extended collocation *price per share* in nine cases. Intuitively, one would think that the same situation would hold for *cents*; however, only three occurrences of *cents* combine to form the collocation *cents per share*.

In comparing the statistics, all of the collocations formed by the primary collocates are significant, as shown in Table 5.19. The collocations with *market* and *per* receive high *t*-scores as they recur frequently. Additionally, the collocations containing *per, earnings, and cents* gain high I-values due to the fact that they are very specific combinations.

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.84</td>
<td>12.46</td>
<td>158</td>
<td>2,706</td>
<td>market</td>
</tr>
<tr>
<td>7.48</td>
<td>8.44</td>
<td>72</td>
<td>789</td>
<td>per</td>
</tr>
<tr>
<td>4.85</td>
<td>5.20</td>
<td>29</td>
<td>1,967</td>
<td>price</td>
</tr>
<tr>
<td>7.43</td>
<td>5.26</td>
<td>28</td>
<td>318</td>
<td>earnings</td>
</tr>
<tr>
<td>9.48</td>
<td>4.35</td>
<td>19</td>
<td>52</td>
<td>cents</td>
</tr>
<tr>
<td>4.51</td>
<td>3.45</td>
<td>13</td>
<td>1,118</td>
<td>total</td>
</tr>
</tbody>
</table>

The most frequent collocates of *share* in the Tobacco-Documents Corpus are *points, point, market, units, and discount*. In addition to being the top collocate overall, *points* also is the term which occurs most often in the R1 position, with thirty-eight instances (47%) appearing in that slot. *Point* is the principal term in the L4 position (13%), in the R3 position (40%), and in
the R4 position (29%). *Market* is the main collocate in the L1 position and the R2 position. In the former position, it appears in twenty-three out of seventy-two occurrences (32%), while in the latter position it occurs twenty-five times (35%). In the L2 position is *units* with thirty-five occurrences (80%). *Discount* is not a dominant collocate in any of the span positions. The remaining position, L3, has *gain* as its major collocate, where it occurs eleven out of twenty-four times (46%).

Table 5.20: mutual information and $t$-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>I (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.72</td>
<td>8.98</td>
<td>81</td>
<td>209</td>
<td>points</td>
</tr>
<tr>
<td>8.10</td>
<td>8.63</td>
<td>75</td>
<td>297</td>
<td>point</td>
</tr>
<tr>
<td>7.16</td>
<td>8.43</td>
<td>72</td>
<td>548</td>
<td>market</td>
</tr>
<tr>
<td>7.95</td>
<td>6.61</td>
<td>44</td>
<td>193</td>
<td>units</td>
</tr>
<tr>
<td>7.86</td>
<td>5.54</td>
<td>31</td>
<td>145</td>
<td>discount</td>
</tr>
<tr>
<td>6.85</td>
<td>5.52</td>
<td>31</td>
<td>293</td>
<td>smoker</td>
</tr>
<tr>
<td>5.68</td>
<td>5.00</td>
<td>26</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>5.42</td>
<td>3.78</td>
<td>15</td>
<td>381</td>
<td>Morris</td>
</tr>
<tr>
<td>5.57</td>
<td>3.66</td>
<td>14</td>
<td>320</td>
<td>RJR</td>
</tr>
<tr>
<td>3.53</td>
<td>3.42</td>
<td>14</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>5.16</td>
<td>3.37</td>
<td>12</td>
<td>366</td>
<td>Philip</td>
</tr>
<tr>
<td>4.19</td>
<td>2.84</td>
<td>9</td>
<td>537</td>
<td>Camel</td>
</tr>
<tr>
<td>5.87</td>
<td>2.78</td>
<td>8</td>
<td>149</td>
<td>Lorillard</td>
</tr>
<tr>
<td>4.75</td>
<td>2.55</td>
<td>7</td>
<td>282</td>
<td>Winston</td>
</tr>
<tr>
<td>2.08</td>
<td>1.87</td>
<td>6</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>7.21</td>
<td>2.43</td>
<td>6</td>
<td>44</td>
<td>Liggett</td>
</tr>
<tr>
<td>5.97</td>
<td>2.41</td>
<td>6</td>
<td>104</td>
<td>Newport</td>
</tr>
<tr>
<td>4.83</td>
<td>11.97</td>
<td>154</td>
<td>5,886</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

Table 5.20 illustrates the mutual information values and $t$-scores for the Tobacco-Documents Corpus. At first glance, it is interesting to note that the only term which does not form a significant collocation with *share* is *cigarette*. However, upon further inspection the low scores are not unexpected. While *cigarette* has a high frequency of occurrence in the documents (1,541), it only collocates with *share* six times. This does not allow the mutual information score
to show a significant association, nor does it provide any level of confidence that the co-
occurrences are not merely the result of chance.

Collocations with *share* and each of the top five collocates, which are all non-tobacco
terms, receive high values from both statistical measures. Although these collocates are not as
frequent as some of the tobacco terms, their frequencies of occurrence with *share* are quite high,
giving confidence to the validity of their associations. The high I-values are indicative of the
specificity of the collocations, i.e., technical nature. From the standpoint of mutual information,
the most interesting collocations with the individual tobacco terms are those with *Liggett*, the
name of a tobacco company. Due to its low absolute frequency in the corpus, the association it
enters into with *share* receives a particularly high I-value, with the probability of the occurrence
more than 148 times that of chance.

As a set, the tobacco-related terms form significant collocations with the node. This
indicates that industry and/or company-related terms will occur significantly within a span of 4:4
of *share*.

In Table 5.21, the span positions of the tobacco terms with their frequencies and rates of
occurrence in those positions are listed. Two-thirds of the collocates fall to the left of *share*.

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27 (18%)</td>
<td>17 (11%)</td>
<td>11 (7%)</td>
<td>47 (30%)</td>
<td>*</td>
<td>2 (1%)</td>
<td>11 (7%)</td>
<td>18 (12%)</td>
<td>21 (14%)</td>
</tr>
</tbody>
</table>
5.8 Market

Market, which has a high frequency in the reference corpus, co-occurs most often with share, price, stock, product, and value. The most frequent collocate, share, is key in the R1 position, where it occurs in 116 out of 162 collocations (72%). Twelve of its instances (7%) are in the L3 position, as are the same number of occurrences for product (14%) and command (29%). Product also prevails in the L4 position (19% of its occurrences). The leading collocate in the R4 position is stock, with 20% of its occurrences there. Collocates in the remaining span positions and their overall frequency of occurrence with the node and rate of occurrence in that position are:

L1 = target 79, 91%; L2 = foreign 63, 63%; R2 = curve 22, 77%; R3 = demand 74, 16%.

Sixty-five percent of the 566 collocations involving the top five collocates occur to the right of market.

Market share is perhaps the first collocation to come to mind when looking for collocates to the right, and, therefore, it is not surprising that share is the most frequent collocate which immediately follows market. However, there are several other collocates from the reference corpus in the R1 position which warrant discussion. Price, which forms the well-established collocation of market price, occurs eighty-five times. Other readily recognizable collocations with their number of occurrences in the corpus include: market value, 71; market research 43; market demand 40; market segments 39; market economies 39; market segment 29; market economy 27; market exchange 12; market potential 11; market interest 11; market rate 10; market growth 9. Two collocations with relatively high frequencies which would not be so readily apparent without the corpus are market capitalism with sixty-six occurrences and market socialism with thirty-three instances.
The extended collocations which exist in the corpus are all three-word units. The collocations with their frequencies of occurrence are: foreign exchange market, 37; stock market crash, 9; target market members, 9; target market segments, 7.

Table 5.22: mutual information and \( t \)-scores from the reference corpus

<table>
<thead>
<tr>
<th>( I(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.87</td>
<td>12.62</td>
<td>162</td>
<td>514</td>
<td>share</td>
</tr>
<tr>
<td>4.61</td>
<td>10.89</td>
<td>129</td>
<td>1,967</td>
<td>price</td>
</tr>
<tr>
<td>5.41</td>
<td>9.72</td>
<td>99</td>
<td>869</td>
<td>stock</td>
</tr>
<tr>
<td>3.66</td>
<td>8.64</td>
<td>88</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>4.25</td>
<td>8.89</td>
<td>88</td>
<td>1,717</td>
<td>value</td>
</tr>
<tr>
<td>6.68</td>
<td>8.80</td>
<td>79</td>
<td>287</td>
<td>target</td>
</tr>
<tr>
<td>4.28</td>
<td>8.16</td>
<td>74</td>
<td>1,415</td>
<td>demand</td>
</tr>
<tr>
<td>5.29</td>
<td>7.73</td>
<td>63</td>
<td>598</td>
<td>foreign</td>
</tr>
<tr>
<td>6.94</td>
<td>6.43</td>
<td>42</td>
<td>127</td>
<td>command</td>
</tr>
<tr>
<td>3.90</td>
<td>4.38</td>
<td>22</td>
<td>548</td>
<td>curve</td>
</tr>
</tbody>
</table>

The statistical values illustrated in Table 5.22 indicate that all ten terms form significant collocations with \( market \). The \( I \)-value and \( t \)-score are particularly high for the collocation involving \( share \) due to the high frequency of co-occurrence. Of course the collocation is more important for \( share \) than it is for \( market \), as \( share \) collocates with \( market \) in 32% of its total occurrences in the corpus, whereas \( market \) co-occurs with \( share \) in only 6% of its total occurrences.

As with the reference corpus, the principal collocate of \( market \) in the Tobacco-Documents Corpus is \( share \). The other four leading collocates are \( test \), \( cigarette \), \( segment \), and \( volume \). \( Share \), which co-occurs with \( market \) sixty-nine times, is the dominant collocate in four span positions. Those positions and its rate of occurrence per position are: L2, 36%; L3, 13%; R1, 32%; R2, 7%. \( Test \) co-occurs most often in the L1 position, with fifty-five out of sixty-three instances (87%). Neither \( cigarette \) nor \( segment \) dominate in any of the positions, but \( volume \) is
key in the R4 position (50%). In the other two positions are *brand*, with eight out of its sixteen co-occurrences in the L4 position, and *reported*, with six out of its nine co-occurrences in the R3 position.

As exhibited in Table 5.23, three of the tobacco-related terms — *smokers, tobacco, smoking* — do not form significant collocations with *market*. This results from the fact that, while the collocates have high absolute frequencies, they collocate with the node at an extremely low rate. The most interesting collocations are those which combine *market* with the non-tobacco terms *share* and *test*. The resulting collocations receive high scores by both statistical measures. The scores for collocations involving the lexical set of tobacco-related terms confirm that industry and/or company-specific terms will occur within the designated span. As was the case with previous nodes, *Philip* always collocates with *Morris*.

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.10</td>
<td>8.25</td>
<td>69</td>
<td>545</td>
<td>share</td>
</tr>
<tr>
<td>6.32</td>
<td>7.84</td>
<td>63</td>
<td>854</td>
<td>test</td>
</tr>
<tr>
<td>3.74</td>
<td>4.03</td>
<td>19</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>6.80</td>
<td>4.20</td>
<td>18</td>
<td>175</td>
<td>segment</td>
</tr>
<tr>
<td>5.72</td>
<td>4.16</td>
<td>18</td>
<td>370</td>
<td>volume</td>
</tr>
<tr>
<td>5.06</td>
<td>4.00</td>
<td>17</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>3.78</td>
<td>3.34</td>
<td>13</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>4.00</td>
<td>2.65</td>
<td>8</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>5.67</td>
<td>2.77</td>
<td>8</td>
<td>170</td>
<td>Viceroy</td>
</tr>
<tr>
<td>4.79</td>
<td>2.36</td>
<td>6</td>
<td>234</td>
<td>Kool</td>
</tr>
<tr>
<td>4.15</td>
<td>2.31</td>
<td>6</td>
<td>366</td>
<td>Philip</td>
</tr>
<tr>
<td>2.30</td>
<td>1.95</td>
<td>6</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>1.44</td>
<td>1.55</td>
<td>6</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>3.83</td>
<td>2.08</td>
<td>5</td>
<td>381</td>
<td>Morris</td>
</tr>
<tr>
<td>4.83</td>
<td>2.16</td>
<td>5</td>
<td>190</td>
<td>Salem</td>
</tr>
<tr>
<td>1.66</td>
<td>1.53</td>
<td>5</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>4.26</td>
<td>2.12</td>
<td>5</td>
<td>282</td>
<td>Winston</td>
</tr>
<tr>
<td>3.46</td>
<td>9.49</td>
<td>109</td>
<td>10,701</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>
The tobacco-related terms, with their frequencies and rates of occurrence in each span position, are listed in Table 5.24. The majority of the collocates, 52%, occur to the right of *market*.

Table 5.24: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 (12%)</td>
<td>10 (9%)</td>
<td>8 (7%)</td>
<td>22 (20%)</td>
<td>*</td>
<td>13 (12%)</td>
<td>17 (16%)</td>
<td>21 (19%)</td>
<td>5 (5%)</td>
</tr>
</tbody>
</table>

5.9 **Sales**

The top five collocates of *sales* in the reference corpus are *volume*, *force*, *manager*, *marketing*, and *product*. *Volume* is the principal collocate in the R1 position, where it occurs in ninety out of its 105 co-occurrences with the node (86%). *Marketing* dominates in the R2 position, with 35% of its occurrences there. *Product* is the most frequent collocate both in the R3 position (18%) and in the R4 position (27%). It is also a main collocate in the L3 position, as are *percent* and *total*. All three terms occur ten times in that position. The collocates which appear most often in the other span positions with their overall frequency of occurrence with the node and rate of occurrence in the position are:

- **L1** = *days* 29, 76%; **L2** = *cost* 57, 53%; **L4** = *cash* 38, 26%.

Seventy-eight percent of the occurrences of the top five collocates lie to the right of the node.

While *volume* is the most frequent collocate in the R1 position, there are several other terms with high frequencies in that position as well. Included among such terms with their overall frequency of occurrence with the node and rate of occurrence in the R1 position are:
force 96, 90%; manager 93, 78%; expansion 62, 79%; team 60, 90%; ledger 47, 74%;
growth 45, 71%; management 43, 56%; managers 41, 80%; promotion 35, 71%;
meetings 32, 81%.

In considering multiword combinations, there are nineteen instances of the collocation
sales and marketing. Other extended collocations include cost of sales, with twenty-four
occurrences, and sales and service, with twenty-three occurrences.

Table 5.25: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>n = sales , f(n) = 2,301</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(n,c)</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>7.33</td>
</tr>
<tr>
<td>7.60</td>
</tr>
<tr>
<td>5.67</td>
</tr>
<tr>
<td>4.30</td>
</tr>
<tr>
<td>3.58</td>
</tr>
<tr>
<td>4.98</td>
</tr>
<tr>
<td>3.39</td>
</tr>
<tr>
<td>4.10</td>
</tr>
<tr>
<td>4.14</td>
</tr>
<tr>
<td>3.42</td>
</tr>
<tr>
<td>5.65</td>
</tr>
</tbody>
</table>

As Table 5.25 demonstrates, all eleven terms form significant collocations with sales.
The mutual information values indicate that the probability of the occurrence of collocations
with either volume or force is more than 150 times that due to chance. Additionally, the
collocations involving the terms return high levels of confidence in the validity of the
associations. The high values from both statistical measures are the result of the relatively low
absolute frequency of occurrence of the two collocates compared with their frequency of co-
occurrence with the node. This is also evidence of the technical nature of the resulting
collocations.
Two of the five most frequent collocates of sales in the Tobacco-Documents Corpus are the same as in the reference corpus, force and marketing, although they do not occur in the same order of frequency. The other top collocates are field, increase, and promotion. Force is the main collocate in the R1 position, where it appears in twenty-seven out of thirty-three occurrences (82%). Field collocates with the highest frequency in the L1 position, 92% of its occurrences. Marketing occurs in the R4 position with three instances, as do retail, period, total, market, drive, distribution, and provide. Promotion and tobacco appear three times in the L3 position. Leading collocates in the remaining positions, with their overall frequency of occurrence with the node and the rate of occurrence in that position, are:

L2 = advertising 10, 50%; L4 = new 12, 42%; R2 = Marlboro 13, 38%.

There are no extended collocations which are readily apparent.

Table 5.26: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.79</td>
<td>5.73</td>
<td>33</td>
<td>89</td>
<td>force</td>
</tr>
<tr>
<td>7.69</td>
<td>5.07</td>
<td>26</td>
<td>151</td>
<td>field</td>
</tr>
<tr>
<td>6.03</td>
<td>4.06</td>
<td>17</td>
<td>312</td>
<td>marketing</td>
</tr>
<tr>
<td>5.56</td>
<td>3.92</td>
<td>16</td>
<td>404</td>
<td>increase</td>
</tr>
<tr>
<td>6.23</td>
<td>3.95</td>
<td>16</td>
<td>255</td>
<td>promotion</td>
</tr>
<tr>
<td>3.54</td>
<td>3.54</td>
<td>15</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>4.81</td>
<td>3.48</td>
<td>13</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>2.46</td>
<td>2.71</td>
<td>11</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>2.81</td>
<td>2.10</td>
<td>6</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>5.59</td>
<td>2.40</td>
<td>6</td>
<td>149</td>
<td>Lorillard</td>
</tr>
<tr>
<td>3.46</td>
<td>2.03</td>
<td>5</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>3.97</td>
<td>2.09</td>
<td>5</td>
<td>381</td>
<td>Morris</td>
</tr>
<tr>
<td>4.03</td>
<td>2.10</td>
<td>5</td>
<td>366</td>
<td>Philip</td>
</tr>
<tr>
<td>3.50</td>
<td>7.41</td>
<td>66</td>
<td>6,950</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>
As evidenced by the I-values and t-scores, all of the combinations involving sales and the collocates in Table 5.26 prove to be significant. Collocations involving the non-tobacco terms receive higher scores than those with the individual tobacco terms. Of the tobacco terms, the collocations with Lorillard receive relatively high values for mutual information. As discussed previously, this is in keeping with the tendency of mutual information to focus on unusual and/or technical terms. Although the collocations with the tobacco-related terms are significant, the figures indicate that the collocations are not as interesting as those with the non-tobacco terms. Even the collocations formed with the lexical set of tobacco terms receive values which are lower than have been seen with previous nodes. Nonetheless, we can still assert that industry and/or company-related terms will occur within the 4:4 span.

The span positions of the tobacco-related terms with their frequencies and rates of occurrence in those positions are listed in Table 5.27. Seventy-two percent of the collocates appear to the left of the node.

Table 5.27: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 (8%)</td>
<td>5 (8%)</td>
<td>8 (12%)</td>
<td>29 (44%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>10 (15%)</td>
<td>6 (9%)</td>
<td>3 (4%)</td>
</tr>
</tbody>
</table>

5.10 WORK

In the reference corpus, the five principal collocates form a total of 178 collocations with work. Those collocates are hours, time, done, values, and process. Of those collocates, time is the most frequent one in the L3 position, occurring there twenty-six times (59%). Done prevails in the R3 position with nine occurrences (28%). Values is the dominant collocate in the R1 position, where it appears twenty-six out of twenty-nine times (90%). Process is the collocate which occurs most
often in the R2 position, 52% of its co-occurrences with work. Hours is not a primary collocate in any of the positions. The following terms, with their overall frequency of occurrence with the node and their rate of occurrence in the position, occur in the remaining span positions:

\[
L1 = \text{managerial} \ 9, 78\%; \ L2 = \text{applications} \ 24, 96\%; \ L4 = \text{skills} \ 23, 96\%; \ R4 = \text{finished} \ 10, 70\%.
\]

Sixty-two percent of the instances of the top five collocates occur to the right of the node.

In twenty-three of the twenty-six occurrences of time in the L3 position, time is immediately followed by applications, producing the multiword combination time applications and work. This collocation utilizes all of the occurrences of applications which appear as the chief collocate of work in the L2 position. The collocation can be further expanded into skills, time applications, and work, which occurs twenty-two times and involves every occurrence of skills in the L4 position. The only other extended collocation which is worth noting is work in process, which occurs thirteen times.

The most interesting collocations according to the values listed in Table 5.28 are those combining work and hours. These collocations receive a high I-value (8.32) indicating that their observed frequency is approximately 319 times greater than chance. Additionally, the t-score provides a high level of confidence in the validity of the collocations. While the absolute frequency of hours is relatively low, it co-occurs with work rather frequently. This is evidence of the specific nature of this collocation. The two terms at the bottom of the list, companies and product, do not form significant or interesting collocations with the node due to their high frequencies, but low rates of co-occurrence with work.
Table 5.28: mutual information and $t$-scores from the reference corpus

\[
\begin{array}{cccccc}
\text{l (n,c)} & \text{t (n,c)} & \text{f(n,c)} & \text{f(c)} & \text{c} \\
8.32 & 6.76 & 46 & 153 & \text{hours} \\
4.56 & 6.35 & 44 & 1,979 & \text{time} \\
7.16 & 5.62 & 32 & 238 & \text{done} \\
6.72 & 5.33 & 29 & 292 & \text{values} \\
4.60 & 4.98 & 27 & 1,182 & \text{process} \\
7.30 & 4.87 & 24 & 162 & \text{applications} \\
6.81 & 4.75 & 23 & 218 & \text{skills} \\
6.59 & 3.13 & 10 & 110 & \text{finished} \\
6.21 & 2.96 & 9 & 129 & \text{managerial} \\
2.82 & 2.43 & 8 & 1,200 & \text{companies} \\
1.71 & 1.97 & 8 & 2,589 & \text{product} \\
\end{array}
\]

The five principal collocates of work in the Tobacco-Documents Corpus are done, tobacco, further, completed, and program. Done is the most frequent collocate in all four positions to the right. Of its twenty-five co-occurrences with work, only one appears to the left of the node. Its rates of occurrence in the positions to the right are:

\[
R1 = 20\%; R2 = 28\%; R3 = 36\%; R4 = 12\%.
\]

In the R1 position, two other terms, plan and represents, occur with the same frequency (5) as done. There are also two more terms which are main collocates in the R4 position, program and smoking, with three instances.

To the left of work, the chief collocates, with their overall frequency of occurrence with the node and their rate of occurrence in the position, are:

\[
L1 = \text{further} 12, 83\%; L2 = \text{continue} 8, 75\%; L3 = \text{results} 9, 67\%; L4 = \text{tobacco} 14, 36\%.
\]

There are no extended collocations with a frequency high enough to be of any interest. The only point which can be made concerning multiword combinations with work in this corpus is that done is usually followed by a prepositional phrase, for example work done on.
Table 5.29: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>( I(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.62</td>
<td>4.97</td>
<td>25</td>
<td>164</td>
<td>done</td>
</tr>
<tr>
<td>2.92</td>
<td>3.25</td>
<td>14</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>5.90</td>
<td>3.68</td>
<td>14</td>
<td>303</td>
<td>further</td>
</tr>
<tr>
<td>6.22</td>
<td>3.42</td>
<td>12</td>
<td>208</td>
<td>completed</td>
</tr>
<tr>
<td>4.31</td>
<td>3.15</td>
<td>11</td>
<td>714</td>
<td>program</td>
</tr>
<tr>
<td>2.18</td>
<td>1.91</td>
<td>6</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>2.36</td>
<td>1.80</td>
<td>5</td>
<td>1,260</td>
<td>smoke</td>
</tr>
<tr>
<td>2.59</td>
<td>4.17</td>
<td>25</td>
<td>5,362</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

As Table 5.29 illustrates, the most significant collocations are formed with the non-tobacco terms. Combinations involving *work* and *smoking*, as well as those of *work* and *smoke*, fall below the levels of significance on both mutual information and t-scores. Collocations with *tobacco* show a low level of confidence (t-score) that the association is valid, but the I-value fails to show any real association. The same situation holds true for the set of tobacco-terms. From these values, it appears that the probability that industry and/or company-specific words will occur significantly within the allotted span is rather questionable.

Table 5.30: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 (24%)</td>
<td>4 (16%)</td>
<td>5 (20%)</td>
<td>2 (8%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>2 (8%)</td>
<td>1 (4%)</td>
<td>5 (20%)</td>
</tr>
</tbody>
</table>

The tobacco terms are listed according to their frequencies of occurrence and rates of occurrence in the span positions in Table 5.30. The majority of the terms, 68%, are to the left of the node.
5.11 **Company**

In the reference corpus, *company* collocates most frequently with *financial, cash, business, needs*, and *product*. *Financial* co-occurs with *company* 120 times and is the most prevalent collocate in three of the span positions: L4 (18%), R2 (45%), and R3 (18%). *Needs* is the primary collocate in the R1 position, where it occurs twenty-four out of sixty times (40%). In the R4 position, *product* occurs most often with fourteen instances (25%). Neither *cash* nor *business* are principal collocates in any of the positions. The remaining key collocates, with their overall frequency with the node and rate of occurrence in the position, are:

\[
L1 = \text{holding } 32, 88\%; \ L2 = \text{general } 22, 45\%; \ L3 = \text{value } 34, 47\%.
\]

Sixty-nine percent of the occurrences of the top five collocates lie to the right of *company*.

Examination of the concordance lines for *company* reveals a few extended collocations. The combination *company’s financial statements* occurs nine times. There are seven occurrences of *company’s financial position*. *Value* is followed by a prepositional phrase containing *company* in fourteen cases, such as *value of the company*.

<table>
<thead>
<tr>
<th>(I (n,c))</th>
<th>(t (n,c))</th>
<th>(f(n,c))</th>
<th>(f(c))</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.63</td>
<td>10.51</td>
<td>120</td>
<td>1,463</td>
<td>financial</td>
</tr>
<tr>
<td>4.05</td>
<td>8.66</td>
<td>85</td>
<td>1,553</td>
<td>cash</td>
</tr>
<tr>
<td>2.76</td>
<td>6.77</td>
<td>63</td>
<td>2,815</td>
<td>business</td>
</tr>
<tr>
<td>4.78</td>
<td>7.46</td>
<td>60</td>
<td>663</td>
<td>needs</td>
</tr>
<tr>
<td>2.68</td>
<td>6.26</td>
<td>55</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>2.58</td>
<td>4.86</td>
<td>34</td>
<td>1,717</td>
<td>value</td>
</tr>
<tr>
<td>6.42</td>
<td>5.59</td>
<td>32</td>
<td>113</td>
<td>holding</td>
</tr>
<tr>
<td>3.38</td>
<td>4.24</td>
<td>22</td>
<td>640</td>
<td>general</td>
</tr>
</tbody>
</table>

All the \(t\)-scores in Table 5.31 provide a level of confidence that the associations are valid and not due to chance. However, collocations with *business, product, and value* score below the
acceptable level of significance. The discrepancy between the two sets of values is a reflection of what each statistic focuses on: frequency of recurrence for the $t$-score and unusual co-occurrences for mutual information. The three collocates and the collocations containing them occur fairly often, but the collocations they form with *company* are somewhat common.

In the Tobacco-Documents Corpus, *company* co-occurs with *tobacco*, *Reynolds*, *R*, *American*, and *J* to form 322 collocations. *Tobacco* is the dominant term in the L1 position with 123 occurrences (87%). The leading collocate in the L2 position is *American*, where it occurs forty-four times (96%). *R* is the major term in the L4 position, and *J* has the same status in the L3 position. Collocates which are primary to the right, with their overall frequency of occurrence with the node and their rate of occurrence in the position, are:

- $R_1 = \text{performance} \ 9, \ 78\%$; $R_2 = \text{filter} \ 5, \ 100\%$; $R_3 = \text{research} \ 5, \ 42\%$; *Tareyton* 5, 100%; $R_4 = \text{segment} \ 14, \ 64\%$.

At first glance, it appears peculiar that two of the span positions have capital letters (*R* and *J*) as their principal collocate. However, once the concordance lines are checked, it becomes evident that the letters form extended collocations, not only with each other, but also with *Reynolds* and *tobacco*. The entire collocation consists of *R. J. Reynolds Tobacco Company*, and there are forty-three occurrences of this combination. Another extended collocation is *American Tobacco Company*, which occurs forty-two times. As previously stated *Philip* collocates with *Morris*.

The column listing the absolute frequencies of the collocates in Table 5.32 makes it quite clear that *R* and *J* are interconnected. The discrepancy in the co-occurrence frequencies with *company* is due to the fact that some occurrences lie outside the 4:4 span. The I-values and $t$-scores for collocations involving *R, J, Reynolds*, and *American* are particularly high as a result of
their low absolute frequencies, combined with their high frequencies of co-occurrence with the node. Combinations with *tobacco* receive a high *t*-score because of the high frequency of *tobacco* in the corpus and the high frequency of co-occurrence of the two terms. Collocations with *smoking* fall below the acceptable level of significance for mutual information, while those with *cigarette* are borderline. As a set, collocations with the tobacco terms receive high values by both statistical measures indicating that we can be confident that industry and/or company-related terms will occur significantly within the designated span.

Table 5.32: mutual information and *t*-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.12</td>
<td>11.74</td>
<td>142</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>8.06</td>
<td>7.04</td>
<td>50</td>
<td>220</td>
<td>Reynolds</td>
</tr>
<tr>
<td>9.46</td>
<td>6.85</td>
<td>47</td>
<td>78</td>
<td>R</td>
</tr>
<tr>
<td>7.23</td>
<td>6.74</td>
<td>46</td>
<td>359</td>
<td>American</td>
</tr>
<tr>
<td>9.12</td>
<td>6.07</td>
<td>37</td>
<td>78</td>
<td>J</td>
</tr>
<tr>
<td>3.06</td>
<td>2.92</td>
<td>11</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>6.43</td>
<td>3.28</td>
<td>11</td>
<td>149</td>
<td>Lorillard</td>
</tr>
<tr>
<td>4.79</td>
<td>2.89</td>
<td>9</td>
<td>381</td>
<td>Morris</td>
</tr>
<tr>
<td>4.85</td>
<td>2.90</td>
<td>9</td>
<td>366</td>
<td>Philip</td>
</tr>
<tr>
<td>2.46</td>
<td>2.31</td>
<td>8</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>5.71</td>
<td>2.19</td>
<td>5</td>
<td>112</td>
<td>Tareyton</td>
</tr>
<tr>
<td>5.78</td>
<td>17.81</td>
<td>329</td>
<td>7,027</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

Table 5.33 shows the tobacco-related terms with their frequencies and rates of occurrence per span position. Sixty-eight percent of the occurrences occur to the left of the node.

Table 5.33: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>44 (13%)</td>
<td>50 (15%)</td>
<td>56 (17%)</td>
<td>129 (39%)</td>
<td>*</td>
<td>3 (1%)</td>
<td>16 (5%)</td>
<td>19 (6%)</td>
<td>12 (4%)</td>
</tr>
</tbody>
</table>
5.12 Levels

The five principal collocates of levels in the reference corpus include higher, different, inventory, high, and lower. Different is the most frequent collocate in the L2 position with eight of its thirty occurrences (27%). Other terms with eight instances in the L2 position are high (29%) and lower (38%). Inventory is the dominant term in the L1 position, with twenty out of its thirty occurrences (67%), and also in the L3 position, five occurrences (17%). In the remaining positions are the following terms, with their overall frequency of occurrence with the node and their rate of occurrence in the position:

L4 = product 10, 30%; based 5, 60%; R1 = no lexical collocates; R2 = management 21, 71%; R3 = organization 10, 70%; R4 = company 7, 57%.

Higher, the most frequent collocate with levels overall, is not a dominate term in any of the span positions. However, its frequency of occurrences in the L1 position is almost as high as that of inventory, and so are the frequencies of a few other terms. The frequency of occurrence of these other collocates in the L1 position are:

higher 15; different 17; high 17; lower 11; leadership 13; price 16.

There are only two combinations which can be considered as extended collocations. Different levels of occurs eleven times, and there are thirteen instances of levels of management.

Table 5.34 lists the mutual information values and the $t$-scores for the collocations formed with levels and its main collocates. The only combination falling below the cut-off point for significance by the mutual information value is that with company. The $t$-score does give the association some level of confidence as to its validity; however, the value is borderline, as are the comparable scores for collocations involving based and terms with the node. Generally speaking,
associations formed with *levels* and the bottom four terms (*product*, *company*, *based*, and *terms*) do not appear to be interesting ones.

Table 5.34: mutual information and *t*-scores from the reference corpus

<table>
<thead>
<tr>
<th>n (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.92</td>
<td>5.61</td>
<td>32</td>
<td>646</td>
<td>higher</td>
</tr>
<tr>
<td>6.40</td>
<td>5.41</td>
<td>30</td>
<td>863</td>
<td>different</td>
</tr>
<tr>
<td>6.47</td>
<td>5.42</td>
<td>30</td>
<td>827</td>
<td>inventory</td>
</tr>
<tr>
<td>6.55</td>
<td>5.23</td>
<td>28</td>
<td>729</td>
<td>high</td>
</tr>
<tr>
<td>6.72</td>
<td>4.54</td>
<td>21</td>
<td>484</td>
<td>lower</td>
</tr>
<tr>
<td>5.16</td>
<td>4.45</td>
<td>21</td>
<td>1,427</td>
<td>management</td>
</tr>
<tr>
<td>4.87</td>
<td>3.05</td>
<td>10</td>
<td>833</td>
<td>organization</td>
</tr>
<tr>
<td>3.23</td>
<td>2.83</td>
<td>10</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>2.36</td>
<td>2.13</td>
<td>7</td>
<td>3,327</td>
<td>company</td>
</tr>
<tr>
<td>3.97</td>
<td>2.09</td>
<td>5</td>
<td>780</td>
<td>based</td>
</tr>
<tr>
<td>4.37</td>
<td>2.13</td>
<td>5</td>
<td>590</td>
<td>terms</td>
</tr>
</tbody>
</table>

The top five collocates form 141 collocations with *levels*. Eighty-nine percent of the occurrences of these collocates lie to the left of the node.

In the Tobacco-Documents Corpus, the five main collocates of *levels* are *nicotine*, *cotinine*, *COHb*, *high*, and *higher*. The primary collocate in three of the span positions is *nicotine*. It appears twenty-two times (54%) in L1, four times (10%) in L4, and four times (10%) in R4. *Cigarettes* also has four occurrences (44%) in the L4 position. None of the other four terms are dominant in a span position. The key collocates in the other positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L2 = *urinary* 7, 100%; L3 = *exposure* 9, 56%; R1 = *similar* 7, 14%; R2 = *nitrates* 9, 78%; R3 = *nitrogen* 11, 55%.
Although nicotine is the most frequent term in the L1 position, there are several other ‘interesting’ collocates which occur there as well. With their overall frequency of occurrence with the node and rate of occurrence in the L1 position, they include:

*cotinine* (a major metabolic product of nicotine) 27, 63%; *COHb (= carboxyhemoglobin)* 19, 79%; *residue* 11, 100%; *SCN (= thiocyanate)* 9, 100%.

There are no noteworthy multiword combinations with levels in this corpus.

Table 5.35: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>I (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.04</td>
<td>6.31</td>
<td>41</td>
<td>895</td>
<td>nicotine</td>
</tr>
<tr>
<td>8.33</td>
<td>5.18</td>
<td>27</td>
<td>121</td>
<td>cotinine</td>
</tr>
<tr>
<td>9.78</td>
<td>4.35</td>
<td>19</td>
<td>31</td>
<td>COHb</td>
</tr>
<tr>
<td>5.78</td>
<td>4.28</td>
<td>19</td>
<td>496</td>
<td>high</td>
</tr>
<tr>
<td>6.61</td>
<td>4.31</td>
<td>19</td>
<td>279</td>
<td>higher</td>
</tr>
<tr>
<td>5.30</td>
<td>3.77</td>
<td>15</td>
<td>548</td>
<td>tar</td>
</tr>
<tr>
<td>3.71</td>
<td>3.20</td>
<td>12</td>
<td>1,317</td>
<td>smokers</td>
</tr>
<tr>
<td>2.72</td>
<td>2.81</td>
<td>11</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>6.99</td>
<td>3.14</td>
<td>10</td>
<td>113</td>
<td>nonsmokers</td>
</tr>
<tr>
<td>3.66</td>
<td>2.76</td>
<td>9</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>4.58</td>
<td>2.87</td>
<td>9</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>6.59</td>
<td>2.97</td>
<td>9</td>
<td>134</td>
<td>non-smokers</td>
</tr>
<tr>
<td>2.49</td>
<td>2.01</td>
<td>6</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>4.63</td>
<td>11.72</td>
<td>122</td>
<td>8,509</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

As exhibited in Table 5.35, the mutual information scores for collocations with *tobacco* or *cigarette* fall below the level of significance, indicating the association might be due to chance. While the t-scores for the associations are above the acceptable limit, they are still rather low, especially for those collocations with *cigarette*. Both terms have high absolute frequencies of occurrence, but co-occur with levels at a minimal rate. On the other hand, collocations with *cotinine* or *COHb* score quite high by both statistical measures. Their particularly high I-values
are indicative of their unusual nature. The association between *levels* and *cotinine* receives the higher *t*-score due to the higher frequencies, while the association between *levels* and *COHb* scores higher on mutual information because *COHb* enters into the collocation in 61% of its overall occurrences in the corpus, making the collocation both uncommon and technical. As a group, collocations with the tobacco terms attain a high enough I-value to demonstrate association between the node and the terms, as well as a *t*-score indicating confidence in the validity of the association. From this, we can state that within a span of 4:4 of *levels*, significant occurrences of industry and/or company-related terms will appear.

Table 5.36: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17 (14%)</td>
<td>14 (11%)</td>
<td>7 (6%)</td>
<td>32 (26%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>23 (19%)</td>
<td>13 (11%)</td>
<td>16 (13%)</td>
</tr>
</tbody>
</table>

Table 5.36 shows the frequencies and rates of occurrence per span position of the tobacco-related terms. Just over half, 57%, lie to the left of *levels*.

5.13 INCREASE

The top five terms which collocate with *increase* in the reference corpus are *money*, *supply*, *price*, *output*, and *sales*. As a group, they form 299 collocations with the node. *Money*, the most frequent collocate, is the principal collocate in the R3 position, occurring there in thirty-six out of its seventy-two occurrences (50%). The main term in the R4 position is *supply* with thirty-seven instances (51%). *Output* prevails in the L3 position (nine occurrences, 17%) and in the R1 position (twenty occurrences, 38%). Neither *price* nor *sales* occurs as a primary collocate in any of the span positions, although *sales* is a close second in the R1 position with sixteen of its
occurrences (34%) falling there. The leading collocates in the other span positions, with their overall frequency of occurrence with the node and rate of frequency in that position, are:

\[ L1 = \text{percent 45, 38%; L2 = causes 10, 80%; L4 = rate 34, 24%; R2 = decrease 32, 56%.} \]

Seventy-two percent of the occurrences of the five main collocates are to the right of the node.

*Increase* enters into several multiword combinations in the reference corpus. Three-word combinations with their frequencies include: *percent increase in*, 15; *increase or decrease*, 14; *increase in income*, 10; *increase in sales*, 7; *increase the value*, 7. Collocations of four words are: *increase in government spending*, 16; *increase the money supply*, 14. A variation on the last collocation produces a five-word combination, *increase in the money supply*, which occurs thirty-two times. Twenty-nine of these occurrences are from the field of economics; the other three instances are from the field of finance.

| \( n = \text{increase} \), \( f(n) = 840 \) |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( I(n,c) \)   | \( t(n,c) \)   | \( f(n,c) \)   | \( f(c) \)   | \( c \)       |
| 5.99           | 8.35            | 72              | 1,356         | money          |
| 6.45           | 8.39            | 72              | 989           | supply         |
| 5.09           | 7.26            | 56              | 1,967         | price          |
| 6.16           | 7.11            | 52              | 873           | output         |
| 4.61           | 6.58            | 47              | 2,301         | sales          |
| 5.46           | 6.56            | 45              | 1,225         | percent        |
| 4.39           | 5.55            | 34              | 1,943         | rate           |
| 8.04           | 5.64            | 32              | 146           | decrease       |
| 6.65           | 3.13            | 10              | 119           | causes         |

Collocations formed with *increase* and any of the principal collocates, overall or positional, score above the level necessary for significance by both statistical measures as can be seen in Table 5.37. There is one point concerning the numbers in the table which needs to be clarified. Although the *t*-scores for the collocations with either *money* or *supply* are close in
range (8.35 and 8.39 respectively), and the frequency of occurrence of each term with the node is the same (72), it should not be assumed that they always occur adjacent to one another. In order to illustrate this point, the following designates their frequencies of occurrence per span position:

\[
\begin{align*}
\text{money:} & \quad L4 = 5; L3 = 5; L2 = 6; L1 = 0; R1 = 0; R2 = 15; R3 = 36; R4 = 5 \\
\text{supply:} & \quad L4 = 4; L3 = 6; L2 = 5; L1 = 2; R1 = 0; R2 = 0; R3 = 15; R4 = 37.
\end{align*}
\]

Three of the five top collocates of \textit{increase} in the Tobacco-Documents Corpus are non-tobacco terms \textit{tax}, \textit{price}, and \textit{share}. The other leading terms are \textit{tobacco} and \textit{cigarette}. Of these five collocates, \textit{tax} is the most frequent term in the L1 position, where twenty-two of its twenty-six occurrences (85%) appear. \textit{Share} occurs three times (17%) in the R4 position, as does \textit{number} (27%), \textit{rate} (50%), and \textit{net} (60%). \textit{Tobacco} collocates with \textit{increase} eight times (47%) in the L4 position, and \textit{cigarette} is the dominant term in the R2 position with five instances (33%). \textit{Price} is not a predominant collocate in any span position, but twenty of its twenty-three co-occurrences with the node (87%) are in the L1 position. The major collocates, with their overall frequency of occurrence with \textit{increase} and rate of occurrence in the other span positions, are:

\[
\begin{align*}
L2 = \text{excise} & \quad 7, 71\%; \quad L3 = \text{federal} \quad 7, 71\%; \quad R1 = \text{weight} \quad 7, 29\%; \quad \text{nicotine} \quad 5, 40\%; \quad R3 = \text{cancer} \quad 7, 57\%.
\end{align*}
\]

Very few extended collocations involving \textit{increase} with more than two occurrences exist in the corpus. Three such combinations, with their frequencies, are: \textit{a slight increase}, 9; \textit{cigarette tax increase}, 3; \textit{federal excise tax increase}, 4.

The I-values in Table 5.38 indicate that all the collocations formed with \textit{increase} and the top collocates are significant. However, when the \textit{t}-scores are considered, combinations involving \textit{Marlboro}, \textit{nicotine}, and \textit{tar} are borderline cases and really fail to provide a level of
confidence that these associations are not due to chance. The scores for the collocations with the tobacco terms as a set do meet the required levels for significance so that we might claim that there will be industry and/or company-specific terms within the specified environment.

Table 5.38: mutual information and $t$-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>$I(n,c)$</th>
<th>$t(n,c)$</th>
<th>$f(n,c)$</th>
<th>$f(c)$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.68</td>
<td>5.07</td>
<td>26</td>
<td>186</td>
<td>tax</td>
</tr>
<tr>
<td>7.07</td>
<td>4.76</td>
<td>23</td>
<td>251</td>
<td>price</td>
</tr>
<tr>
<td>5.60</td>
<td>4.16</td>
<td>18</td>
<td>545</td>
<td>share</td>
</tr>
<tr>
<td>3.38</td>
<td>3.73</td>
<td>17</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>3.84</td>
<td>3.60</td>
<td>15</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>3.37</td>
<td>3.13</td>
<td>12</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>3.73</td>
<td>2.07</td>
<td>5</td>
<td>552</td>
<td>Marlboro</td>
</tr>
<tr>
<td>3.04</td>
<td>1.96</td>
<td>5</td>
<td>895</td>
<td>nicotine</td>
</tr>
<tr>
<td>3.74</td>
<td>2.07</td>
<td>5</td>
<td>548</td>
<td>tar</td>
</tr>
<tr>
<td>3.50</td>
<td>7.00</td>
<td>59</td>
<td>7,638</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

Table 5.39 lists the span positions of the tobacco-related terms with their frequencies and rates of occurrence in the position. Fifty-six percent of the collocates occur to the right of $increase$.

Table 5.39: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 (20%)</td>
<td>5 (9%)</td>
<td>7 (12%)</td>
<td>2 (3%)</td>
<td>*</td>
<td>6 (10%)</td>
<td>16 (27%)</td>
<td>8 (14%)</td>
<td>3 (5%)</td>
</tr>
</tbody>
</table>
5.14 **Analysis**

In the reference corpus, *analysis* collocates most frequently with *financial, data, decision, used,* and *exhibit.* The collocate found most often in the L1 position is *financial,* where it occurs forty-eight out of eighty-five times (56%). It also prevails in the L2 position, with 22% of its occurrences. In the R4 position, there are six instances of *data* (23%), as well as six cases of *exhibit* (26%). *Exhibit* is also the primary collocate in the R1 position (35% of its occurrences). *Used* is the principal term in the L3 position (33%) and in the R2 position (21%). *Decision* is not a primary collocate in any position. The most frequent terms in the other two positions are *techniques* (L4, five occurrences, 23%) and *new* (R3, eight occurrences, 53%). Sixty-six percent of the occurrences of the five main collocates occur to the left of the node.

There are several other collocates occurring in the L1 position which, although not the most frequent, are, nevertheless, interesting to point out. Each of these terms appears in L1 a minimum of ten times. They are listed below with their overall frequency of occurrence with *analysis* and rate of occurrence in the L1 position:

- *decision* 25, 64%; *statement* 23, 74%; *gap* 17, 82%; *sensitivity* 17, 65%; *marginal* 16, 81%; *economic* 14, 71%; *ROI* (*return on investment*) 14, 93%; *situation* 13, 77%; *ratio* 12, 92%; *macroeconomic* 10, 100%.

In checking the concordance lines, only one extended collocation becomes apparent. There are seventeen occurrences of *financial statement analysis* in the corpus.

As Table 5.40 indicates, all of the collocations with *analysis* and its leading collocates meet the requirements for significance. The association between the node and *financial* receives a high $t$-score due to the high frequency of *financial* in the corpus, combined with its high rate of co-occurrence with *analysis.* Its high I-value indicates the technical nature of the association. In
this situation, financial is more important to analysis than vice versa, as the collocation encompasses only 6% of the occurrences of financial, while 12% of the occurrences of analysis are involved.

Table 5.40: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>n = analysis, f(n) = 723</th>
</tr>
</thead>
<tbody>
<tr>
<td>l(n,c)</td>
</tr>
<tr>
<td>6.34</td>
</tr>
<tr>
<td>5.07</td>
</tr>
<tr>
<td>6.35</td>
</tr>
<tr>
<td>4.83</td>
</tr>
<tr>
<td>6.38</td>
</tr>
<tr>
<td>7.51</td>
</tr>
<tr>
<td>3.46</td>
</tr>
</tbody>
</table>

The top five collocates of analysis in the Tobacco-Documents Corpus are data, results, smoke, chemical, and statistical. Data, the most frequent collocate with the node overall, is also the primary term in the L2 position, with fifteen occurrences (42%), and in the R1 position, with six occurrences (17%). Results is one of four dominant terms in the R3 position, where it occurs five times (23%). The other collocates in R3 are smoke (23%), tobacco (36%), and information (63%). Chemical appears sixteen out of eighteen times (89%) in the L1 position. Statistical is not a main collocate in any position. The terms in the remaining span positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L3 = sufficient 14, 100%; L4 = acquired 14, 100%; R2 = variance 11, 82%; R4 = samples 10, 40%.

From the concordance lines, three multiword combinations are revealed. Two of the extended collocations consist of three-word units. Data for analysis occurs fourteen times, while
analysis of variance appears nine times. There are fourteen instances of a five-word combination, acquired sufficient data for analysis.

In Table 5.41, all of the collocations formed with analysis and any of the four non-tobacco terms show significance by both statistical measures. Those combinations with chemical or statistical receive the highest I-values. Each collocate has a low frequency of occurrence in the corpus, and, therefore, the collocations they enter into with the node can be considered as distinctive. Of the tobacco terms, only collocations with cigarette do not reach a level of significance. The association between analysis and the lexical set of tobacco terms is valid, and thus we can argue that industry and/or company-related terms will occur significantly within a span of 4:4 of the node.

The frequencies and rates of occurrence per span position of the tobacco-related terms is exhibited in Table 5.42. Fifty-eight percent of the collocates occur to the left of analysis.
Table 5.42: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 (10%)</td>
<td>5 (10%)</td>
<td>4 (8%)</td>
<td>15 (30%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>5 (10%)</td>
<td>11 (22%)</td>
<td>5 (10%)</td>
</tr>
</tbody>
</table>

5.15 **INDUSTRY**

The five most frequent collocates of *industry* in the reference corpus are *risk*, *firms*, *company*, *competitive*, and *firm*. *Risk* is the dominant term in the R2 position, where it occurs six out of thirty-two times (19%). In the L4 position, *firms* is the principal collocate with 50% of its occurrences there. *Company* prevails in the R3 spot with seven of its eighteen instances (39%). The main term in the L3 position is *firm*, which appears there in 53% of its occurrences. *Competitive* is not a primary collocate in any of the positions. For the remainder of the positions, the main collocates, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

- L1 = *automobile* 11, 100%; L2 = *financial* 11, 45%; R1 = *sales* 14, 57%; *structure* 13, 62%; *attractiveness* 11, 73%; R4 = *product* 13, 38%.

Of the 109 collocations formed with *industry* and one of the top five collocates, 58% of the collocates lie to the left of the node.

Perhaps the most crucial collocate to this particular node is the term which occurs immediately to its left (L1). While *automobile* appears eleven times, there are also ten occurrences of the shortened form, *auto*, as well as 6 instances of the variant *automotive*. This is not a totally unexpected outcome and will be compared with the results from the Tobacco-Documents Corpus. Of secondary importance is the L2 collocate, which may be an extension of the collocation.
Along these lines, there are a few multiword units which are apparent from the concordance lines. *Financial services industry* appears five times in the reference corpus. There are four occurrences of *Japanese automobile industry* and a couple of combinations with *American* or *U.S.* in the L2 slot.

Table 5.43: mutual information and \( t \)-scores from the reference corpus

<table>
<thead>
<tr>
<th>( l(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.18</td>
<td>5.58</td>
<td>32</td>
<td>727</td>
<td>risk</td>
</tr>
<tr>
<td>5.69</td>
<td>5.00</td>
<td>26</td>
<td>829</td>
<td>firms</td>
</tr>
<tr>
<td>3.16</td>
<td>3.77</td>
<td>18</td>
<td>3,327</td>
<td>company</td>
</tr>
<tr>
<td>5.61</td>
<td>4.16</td>
<td>18</td>
<td>609</td>
<td>competitive</td>
</tr>
<tr>
<td>4.50</td>
<td>3.70</td>
<td>15</td>
<td>1,092</td>
<td>firm</td>
</tr>
<tr>
<td>3.33</td>
<td>3.37</td>
<td>14</td>
<td>2,301</td>
<td>sales</td>
</tr>
<tr>
<td>3.05</td>
<td>3.17</td>
<td>13</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>6.05</td>
<td>3.55</td>
<td>13</td>
<td>324</td>
<td>structure</td>
</tr>
<tr>
<td>9.39</td>
<td>3.31</td>
<td>11</td>
<td>27</td>
<td>attractiveness</td>
</tr>
<tr>
<td>7.90</td>
<td>3.30</td>
<td>11</td>
<td>76</td>
<td>automobile</td>
</tr>
<tr>
<td>3.63</td>
<td>3.05</td>
<td>11</td>
<td>1,463</td>
<td>financial</td>
</tr>
</tbody>
</table>

All of the associations between *industry* and the collocates in Table 5.43 show acceptable levels of significance by both statistical measures. Those with *attractiveness* score particularly high when rated according to mutual information, indicating the unusual nature of the collocation — 41% of the occurrences of *attractiveness* in combination with *industry* in the reference corpus. Overall, collocations with *risk* are the most interesting as they show not only a strong association (I-value), but also a relatively high confidence level in their validity.

The five principal collocates of *industry* in the Tobacco-Documents Corpus include *tobacco, cigarette, volume, government, and position.* *Tobacco,* which is the most frequent collocate overall and enters into 124 collocations with the node, is also the primary term in the L1 position with 111 occurrences (90%). Additionally, it is the word which occurs most often in
the L3 position, six times (5%). The main collocate in the R1 position is *volume*, with seventeen of its twenty-two occurrences (77%) in that slot. In the L2 position is *government*, where four of its occurrences (29%) appear. It is also a primary collocate in the L4 position with three instances (21%), along with *research* (27%) and *issues* (43%). *Position* is the term which appears most often in the R2 position with four occurrences (36%). Other dominant collocates with their overall frequency of occurrence with the node and their rate of occurrence in a span position are:

R3 = *advertising* 8, 38%; R4 = *smoking* 8, 50%.

As with the reference corpus, it is obvious that the most important collocate with *industry* is the one in the L1 position. The top five collocates form 196 two-word collocations with *industry*. Of these collocations, 132 of them (67%) are with the node and one of three terms in the L1 position: *tobacco*, 111 occurrences; *cigarette*, twenty occurrences; *government*, one occurrence.

The concordance lines offer no extended collocations of great interest. Of the 111 instances of the collocation *tobacco industry*, ninety-one are the multiword unit *the tobacco industry*, with the majority of those preceded by a preposition. The same situation applies to *cigarette industry*, where there are eighteen cases of *the cigarette industry*.

As illustrated in Table 5.44, the one association which fails to meet the required level for significance by mutual information measures is that between *industry* and *smoking*. While *smoking* has a high frequency of occurrence in the Tobacco-Documents Corpus, it co-occurs with *industry* at a low rate. However, the *t*-score does provide a minimal level of confidence that the association is not due to chance. The most interesting association is between *tobacco* and *industry*, which receives high scores by both statistical measures. The level of confidence is
especially high, and the I-value indicates that the probability of the occurrence of the collocation is approximately seventy-four times greater than that due to chance. The mutual information value and t-score for collocations involving *industry* and the combined set of tobacco terms are well above the levels necessary to establish significance. In the case of *industry*, this indicates that we can state with confidence that industry and/or company terms will occur significantly within the designated environment.

Table 5.44: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>n = <em>industry</em>, f(n) = 415</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I</strong>&lt;sub&gt;(n,c)&lt;/sub&gt;</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>6.21</td>
</tr>
<tr>
<td>4.53</td>
</tr>
<tr>
<td>6.41</td>
</tr>
<tr>
<td>6.98</td>
</tr>
<tr>
<td>6.70</td>
</tr>
<tr>
<td>5.16</td>
</tr>
<tr>
<td>2.74</td>
</tr>
<tr>
<td>5.30</td>
</tr>
</tbody>
</table>

Table 5.45 lists the tobacco-related terms with their frequencies and rates of occurrence in the various span positions. Ninety percent of the collocates lie to the left of the node.

Table 5.45: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 (2%)</td>
<td>11 (7%)</td>
<td>4 (2%)</td>
<td>131 (79%)</td>
<td>*</td>
<td>2 (1%)</td>
<td>1 (1%)</td>
<td>4 (2%)</td>
<td>9 (6%)</td>
</tr>
</tbody>
</table>
5.16 Development

Included in the top five collocates of development in the reference corpus are product, new, research, process, and economic. In addition to being the most frequent collocate overall, product also is the predominant term in two of the span positions. It occurs ninety-one out of 122 times (75%) in the L1 position and seven times (6%) in L3. In addition, there are seven instances (6%) of product in the L4 position and seven instances (88%) of stages. The principal collocate in the L2 position is new with thirty-six instances (59%). Process occurs twenty-three times (64%) in the R1 slot. Neither research nor economic are primary terms in any of the span positions, although research is a close second in the L2 position with thirty-two occurrences (84%). Other prime collocates with their overall frequency with the node and rate of occurrence in the span position are:

R2 = marketing 23, 43%; R3 = marketing 23, 30%; R4 = plan 12, 50%.

Two other terms whose frequencies in the L1 position make them worth mentioning are economic and leadership. Economic development occurs twenty-nine times in the reference corpus, and leadership development appears twenty-seven times. Development and the top five terms form 292 collocations, 77% of which lie to the left of the node.

The concordance lines reveal four extended collocations. Two three-word units are composed of the node and two collocates to the left: research and development, with thirty-one occurrences, and new product development, with thirty-two occurrences. Building upon the latter combination produces new product development process, which occurs six times. The fourth combination is comprised of the node and one collocate on either side, leadership development system, which appears seven times in the corpus.
Table 5.46 illustrates that there is a strong association between *development* and each of the collocates. Additionally, the $t$-scores show that we can be confident that these associations are not due to chance. The collocation of *development* and *stages* receives a relatively high I-value, but a low $t$-score which results from its low rate of co-occurrence. That is, it displays a strong association, but the confidence level that the association is not due to chance is rather weak. Collocations formed with either of the top two collocates, on the other hand, are not only strong, but provide high levels of confidence in their validity due to the absolute frequencies of the collocates, as well as their frequencies of co-occurrence with the node.

In the Tobacco-Documents Corpus, the five principal collocates of *development* are *product*, *research*, *new*, *program*, and *tobacco*. It is interesting to note that the top three collocates in this corpus are the same as in the reference corpus, with only a variation in the order of the second and third terms. *Product* is the most frequent collocate overall, as well as in the L1 position, where it occurs twenty out of thirty times (67%). The leading term in the L2 position is *research*, with 79% of its occurrences there. *New* prevails in the L4 position, 12% of its occurrences, and in the R2 slot, 31% of its occurrences. In the R3 position, *new* appears four times (15%), as does *products* (50%) and *cigarettes* (67%). *Tobacco* is the term which occurs most often in the R4 position with five instances (26%). *Program* is not a major collocate in any
of the span positions. In the two remaining positions, L3 and R1, are director (67%) and center (58%) respectively.

While there are multiword combinations with development in the Tobacco-Documents Corpus, the only one which occurs frequently enough to be considered is research and development. There are twenty-two instances of this extended collocation.

As Table 5.47 demonstrates, associations between the node and cigarette, smoking, or cigarettes are borderline cases. The strongest associations with the highest confidence levels are those combinations involving development and the non-tobacco terms. On the other hand, collocations formed with development and the combined set of tobacco-related terms receive the highest level of confidence, as well as an I-value indicating their probability of occurrence as approximately eleven times greater than that due to chance. Consequently, we can argue that industry and/or company-specific terms will occur within a 4:4 span of development.

Table 5.47: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.71</td>
<td>5.37</td>
<td>30</td>
<td>908</td>
<td>product</td>
</tr>
<tr>
<td>5.57</td>
<td>5.27</td>
<td>29</td>
<td>968</td>
<td>research</td>
</tr>
<tr>
<td>5.08</td>
<td>4.95</td>
<td>26</td>
<td>1,224</td>
<td>new</td>
</tr>
<tr>
<td>5.40</td>
<td>4.26</td>
<td>19</td>
<td>714</td>
<td>program</td>
</tr>
<tr>
<td>3.66</td>
<td>4.01</td>
<td>19</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>2.85</td>
<td>2.28</td>
<td>7</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>2.70</td>
<td>2.24</td>
<td>7</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>3.22</td>
<td>2.19</td>
<td>6</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>5.44</td>
<td>2.39</td>
<td>6</td>
<td>220</td>
<td>Reynolds</td>
</tr>
<tr>
<td>3.96</td>
<td>2.09</td>
<td>5</td>
<td>511</td>
<td>Lights</td>
</tr>
<tr>
<td>4.82</td>
<td>2.16</td>
<td>5</td>
<td>282</td>
<td>Winston</td>
</tr>
<tr>
<td>3.51</td>
<td>6.76</td>
<td>55</td>
<td>7,680</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>
The tobacco-related collocates with their frequencies and rates of occurrence per span position are listed in Table 5.48. Sixty-one percent of the collocates occur to the left of development.

Table 5.48: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 (15%)</td>
<td>10 (18%)</td>
<td>7 (13%)</td>
<td>8 (15%)</td>
<td>*</td>
<td>1 (2%)</td>
<td>2 (3%)</td>
<td>10 (18%)</td>
<td>9 (16%)</td>
</tr>
</tbody>
</table>

5.17 Project

The five most frequent collocates of project in the reference corpus include risk, financing, IRR (internal rate of return), cash, and investment. Risk occurs five times (14%) in the R2 position, as does flow (33%). The principal collocate in the R1 position is financing, where it occurs fifteen out of twenty-eight times (54%). IRR is the term which occurs most often in the L3 position, with 33% of its occurrences there. The dominant term in L4 is cash with seven instances (32%). Investment is not a main collocate in any of the span positions. The primary collocates in the other span positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L1 = long-term 8, 75%; L2 = period 9, 33%; R3 = value 16, 25%; debt = 10, 40%; R4 = rate 14, 36%.

Fifty-four percent of the occurrences of the top five collocates lie to the right of project. There are no multiword combinations in the reference corpus which occur with a frequency high enough to be of interest.

All of the associations in Table 5.49 attain the required levels for significance by both of the statistical measures. From the standpoint of mutual information, the most interesting
collocations are those which combine the node with *risk, financing, IRR, flow, or budgets*. The associations receiving the highest levels of confidence are those between the node and *risk, financing, or IRR*. Collocations composed of *project and risk or project and financing* are interesting due to their high frequencies of co-occurrence combined with the absolute frequencies of the collocates. Collocations of *project and IRR* are interesting due to their highly technical status.

Table 5.49: mutual information and *-scores from the reference corpus

<table>
<thead>
<tr>
<th>n = project, f(n) = 479</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (n,c)</td>
</tr>
<tr>
<td>6.66</td>
</tr>
<tr>
<td>7.23</td>
</tr>
<tr>
<td>8.26</td>
</tr>
<tr>
<td>4.90</td>
</tr>
<tr>
<td>5.22</td>
</tr>
<tr>
<td>4.29</td>
</tr>
<tr>
<td>6.17</td>
</tr>
<tr>
<td>3.92</td>
</tr>
<tr>
<td>5.21</td>
</tr>
<tr>
<td>4.77</td>
</tr>
<tr>
<td>5.49</td>
</tr>
<tr>
<td>6.59</td>
</tr>
</tbody>
</table>

In the Tobacco-Documents Corpus, the five principal collocates of *project* are *development, research, leader, completed, and status*. *Research* is the most frequent term in the L1 position, occurring there twelve out of seventeen times (71%). In the R1 position, *leader* occurs most often, with 80% of its instances in that slot. *Status* appears nine times (69%) in the R3 position. *Development* and *completed* are not primary collocates in any of the span positions. The top collocates in the remaining positions, with their overall frequency of occurrence with the node and their rate of occurrence in the position, are:
L2 = number 7, 57%; L3 = charge 6, 67%; L4 = Vantage 7, 57%; R2 = image 8, 50%; R 4 = product 10, 40%.

As with the reference corpus, there are no apparent extended collocations with a frequency high enough to warrant consideration.

Table 5.50 demonstrates that all the associations between project and the collocates are significant. The most interesting of collocations involving non-tobacco terms is that with leader, which receives an I-value of 9.45 due to its idiosyncratic nature. The same situation holds among the tobacco terms for Vantage. While the non-tobacco terms obtain generally higher scores by both statistical measures, those for the collocations composed of project and the combined set of tobacco terms also provide a level of confidence that the associations are strong and valid. As a result, we can assert that industry and/or company-related terms will appear within the specified environment of project.

Table 5.50: mutual information and $t$-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>n = project, f(n) = 363</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (n,c)</td>
</tr>
<tr>
<td>6.22</td>
</tr>
<tr>
<td>4.84</td>
</tr>
<tr>
<td>9.45</td>
</tr>
<tr>
<td>6.67</td>
</tr>
<tr>
<td>6.89</td>
</tr>
<tr>
<td>2.77</td>
</tr>
<tr>
<td>3.25</td>
</tr>
<tr>
<td>4.40</td>
</tr>
<tr>
<td>7.39</td>
</tr>
<tr>
<td>4.15</td>
</tr>
<tr>
<td>5.59</td>
</tr>
<tr>
<td>3.77</td>
</tr>
</tbody>
</table>
Listed in Table 5.51 are the frequencies and rates of occurrence of the tobacco-related terms according to the span positions. Seventy-one percent of the collocates occur to the left of project.

Table 5.51: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th></th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11 (25%)</td>
<td>9 (21%)</td>
<td>6 (14%)</td>
<td>5 (11%)</td>
<td>*</td>
<td>1 (2%)</td>
<td>5 (11%)</td>
<td>2 (5%)</td>
<td>5 (11%)</td>
</tr>
</tbody>
</table>

5.18 SYSTEM

The five terms which collocate most frequently with system in the reference corpus are economic, accounting, marketing, modern, and political. Economic is the primary collocate overall, forming 110 collocations with system. It is also the term which occurs most often in the L1 position, with ninety-six instances (87%). Of the other four collocates, only modern is dominant in a span position. It appears thirty-nine out of forty-two times (93%) in the L2 slot. The key terms in the other positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L3 = fixed 16, 69%; L4 = types 26, 81%; R1 = works 13, 92%; R2 = based 21, 48%; R3 = international 33, 18%; R4 = role 7, 100%.

Eighty-eight percent of the top five collocating terms lie to the left of system.

While economic is the principal collocate in the L1 position, there are several other terms in the same position which warrant consideration due to their relatively high frequencies there. Those terms with their overall frequency of occurrence with system and rate of occurrence in L1 include:
accounting 78, 85%; marketing 48, 25%; political 34, 74%; costing 30, 90%; information 29, 59%; planning 27, 85%; inventory 26, 46%; computer 23, 78%; financial 22, 59%; management 22, 59%; rate 21, 90%; automated 20, 50%; banking 20, 95%; monetary 20, 80%; tax 20, 70%; tracking 14, 93%; payroll 13, 85%.

In the L2 position, the same situation applies to:

marketing 48, 42%; international 33, 64%; cost 26, 62%; exchange 25, 80%; accounts 17, 76%.

There are numerous multiword combinations with system in the reference corpus. Three-word units with their frequency of occurrence include: marketing planning system, 13; cost accounting system, 12; international monetary system, 11; accounts payable system, 9; standard costing system, 8; Federal Reserve System, 8; leadership development system, 7; established economic system, 6. Modern economic system, which occurs thirty-three times in the corpus, also appears as the four-word combination of modern economic system in twenty-eight of the occurrences. The same situation exists for exchange rate system (nineteen occurrences), which can be expanded into either fixed exchange rate system (eleven occurrences) or flexible exchange rate system (four occurrences).

The statistical results exhibited in Table 5.52 indicate that the associations formed between system and each of the collocates are strong and valid. The most interesting collocations are those combining the node with economic, accounting, or modern, as they receive high scores by both statistical measures. While combinations with economic and accounting receive rather high t-scores due to their recurrence, those with modern obtain the higher I-value, marking it as a more unusual collocation. According to the weak scores by both measures, associations between system and role are likely to be less interesting than the others.
The top five collocates of *system* in the Tobacco-Documents Corpus include *new*, *ventilation*, *data*, *smoking*, and *nervous*. *New* collocates with *system* eighteen times and is the primary term in the L3 position with five occurrences (28%). It occurs once in the R1 position (6%), as do *used* (13%) and *installed* (20%). *New*, *used*, and *consumer* appear twice in the R3 position, with their rates of occurrence in the position 11%, 25%, and 40% respectively. *Ventilation* is the main collocate in the L1 position (93%). *Smoking* (23%) and *air* (30%) each occur three times in the R4 position. Neither *data* nor *nervous* are key collocates in any span position. The most frequent terms in the other positions, with their overall frequency with the node and rate of occurrence in the slot, are:

L2 = *central* 6, 100%; L4 = *smoke* 5, 80%; R2 = *air* 10, 20%; *used* 8, 25%; *installed* 5, 40%.

The only multiword combination which is apparent in the concordance lines is *central nervous system*, occurring six times.
Table 5.53: mutual information and $t$-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>I (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.61</td>
<td>4.07</td>
<td>18</td>
<td>1,224</td>
<td>new</td>
</tr>
<tr>
<td>7.67</td>
<td>3.72</td>
<td>14</td>
<td>114</td>
<td>ventilation</td>
</tr>
<tr>
<td>5.02</td>
<td>3.49</td>
<td>13</td>
<td>666</td>
<td>data</td>
</tr>
<tr>
<td>3.66</td>
<td>3.32</td>
<td>13</td>
<td>1,707</td>
<td>smoking</td>
</tr>
<tr>
<td>10.28</td>
<td>3.46</td>
<td>12</td>
<td>16</td>
<td>nervous</td>
</tr>
<tr>
<td>2.92</td>
<td>2.29</td>
<td>7</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>2.72</td>
<td>1.90</td>
<td>5</td>
<td>1,260</td>
<td>smoke</td>
</tr>
<tr>
<td>3.20</td>
<td>4.46</td>
<td>25</td>
<td>4,508</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

As Table 5.53 indicates, collocations involving the node and a non-tobacco term score above the levels required for significance. The very high I-value for the association between \textit{system} and \textit{nervous} signifies its idiosyncratic and technical nature. Of the collocations formed with one of the individual tobacco-related terms, only those with \textit{smoking} receive scores showing both strength of attraction and confidence in the validity of the association. The scores for collocations formed with the set of tobacco terms do suggest significance. While we can state that industry and/or company-related terms will fall within the designated environment of \textit{system}, the claim is somewhat weak.

Table 5.54: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>*</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.54 lists the frequencies and rates of occurrence per span position of the tobacco-related terms. Sixty-four percent of the collocates lie to the left of \textit{system}. 
5.19 Order

In the reference corpus, *order* collocates most often with *purchase, inventory, customer, processing*, and *fulfillment*. *Purchase* is the principal collocate in the L1 position, where sixteen of its twenty-two occurrences (73%) lie. *Customer* occurs five times (28%) in the L3 position, as does *financial* (45%). The dominant term in the R1 position is *processing*, with 93% of its occurrences in that slot. Neither *inventory* nor *fulfillment* are main collocates in any of the span positions. For the other positions, the primary terms, with their frequency of occurrence with the node and rate of occurrence in the position, are:

\[
\begin{align*}
L2 &= \text{law } 8, \ 100\%; \ L4 = \text{cost } 5, \ 80\%; \ R2 = \text{keep } 10, \ 80\%; \ R3 = \text{determine } 8, \ 100\%; \ R3 = \text{business } 9, \ 56\%; \ R4 = \text{costs } 12, \ 33\%.
\end{align*}
\]

Fifty-one percent of the top five collocates appear to the right of *order*.

In examining the concordance lines for extended collocations, the only one which is evident is *law and order* with eight occurrences. What does become apparent is that, when the term in the R2 position is a verb, *order* becomes the idiom *in order to*. This gives rise to multiword units such as *in order to determine* (eight occurrences) and *in order to keep* (eight occurrences).

Of the associations listed in Table 5.55, combinations of *order* and *cost* fail to attain a level of significance from either of the statistical measures. Additionally, those composed of the node and *business* receive an I-value below the cut-off point. Collocations with *purchase* and *order, order and processing*, or *order and fulfillment* receive high I-values due to their status as technical compounds, while *law* and *order* has a high score as a fixed compound.
Table 5.55: mutual information and $t$-scores from the reference corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.13</td>
<td>4.62</td>
<td>22</td>
<td>548</td>
<td>purchase</td>
</tr>
<tr>
<td>5.32</td>
<td>4.25</td>
<td>19</td>
<td>827</td>
<td>inventory</td>
</tr>
<tr>
<td>4.15</td>
<td>4.00</td>
<td>18</td>
<td>1,771</td>
<td>customer</td>
</tr>
<tr>
<td>7.98</td>
<td>3.86</td>
<td>15</td>
<td>104</td>
<td>processing</td>
</tr>
<tr>
<td>9.99</td>
<td>3.74</td>
<td>14</td>
<td>24</td>
<td>fulfillment</td>
</tr>
<tr>
<td>3.50</td>
<td>3.16</td>
<td>12</td>
<td>1,852</td>
<td>costs</td>
</tr>
<tr>
<td>3.71</td>
<td>3.06</td>
<td>11</td>
<td>1,463</td>
<td>financial</td>
</tr>
<tr>
<td>6.07</td>
<td>3.12</td>
<td>10</td>
<td>259</td>
<td>keep</td>
</tr>
<tr>
<td>2.48</td>
<td>2.46</td>
<td>9</td>
<td>2,815</td>
<td>business</td>
</tr>
<tr>
<td>5.26</td>
<td>2.75</td>
<td>8</td>
<td>363</td>
<td>determine</td>
</tr>
<tr>
<td>6.36</td>
<td>2.79</td>
<td>8</td>
<td>170</td>
<td>law</td>
</tr>
<tr>
<td>1.87</td>
<td>1.63</td>
<td>5</td>
<td>2,379</td>
<td>cost</td>
</tr>
</tbody>
</table>

The principal collocates of order in the Tobacco-Documents Corpus include obtain, date, form, court, and become. Obtain, which is the most frequent collocate overall, is the dominant term in the R2 position, where all eleven of its co-occurrences appear. Date prevails in the L3 position, with 30% of its instances in that slot. The term occurring most often in the R1 position is form (50%). All nine occurrences of court are in the L1 position. Become is not the main collocate in any span position. In the remaining positions, the primary collocates, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L2 = available 7, 57%; L4 = provided 8, 88%; R3 = effective 8, 75%; R4 = program 8, 38%.

Two multiword units are found in this corpus which involve the node and the collocates in all four positions to the left. There are seven occurrences of the combination provided in the court’s order and six instances of subject to a court order. As with the reference corpus, verbs appearing in the R2 slot are preceded by the idiom in order to. The concordance lines contain 156 such cases.
Table 5.56: mutual information and \( t \)-scores from the Tobacco-Documents Corpus

\[
\begin{array}{cccccc}
\mathbf{I(n,c)} & \mathbf{t(n,c)} & \mathbf{f(n,c)} & \mathbf{f(c)} & \mathbf{c} \\
8.15 & 3.30 & 11 & 66 & obtain \\
5.81 & 3.11 & 10 & 305 & date \\
6.59 & 3.13 & 10 & 177 & form \\
7.43 & 2.98 & 9 & 89 & court \\
6.93 & 2.81 & 8 & 112 & become \\
3.02 & 2.15 & 6 & 1,260 & smoke \\
\end{array}
\]

The \( t \)-scores for all associations are quite low. Additionally, the I-value for collocations involving \textit{order} and the only tobacco term, \textit{smoke}, is borderline. Overall, the collocations formed with \textit{order} and any of the above terms are not very interesting. It is also questionable as to whether we can claim with any great degree of confidence that industry and/or company-related terms will appear within the 4:4 span of this node.

Table 5.57 lists the frequencies rates of occurrence per span position for the term \textit{smoke}. Eighty-three percent occur to the left of \textit{order}.

Table 5.57: span positions of tobacco-related collocates by frequency and rate of occurrence

\[
\begin{array}{cccccccc}
\mathbf{L4} & \mathbf{L3} & \mathbf{L2} & \mathbf{L1} & \mathbf{node} & \mathbf{R1} & \mathbf{R2} & \mathbf{R3} \\
0 (0\%) & 2 (33\%) & 3 (50\%) & 0 (0\%) & * & 0 (0\%) & 0 (0\%) & 0 (0\%) \\
\end{array}
\]

5.20 \textbf{Production}

The five most frequent collocates of \textit{production} in the reference corpus are \textit{costs}, \textit{means}, \textit{process}, \textit{cost}, and \textit{sales}. \textit{Costs} is the chief collocate in the L4 position with nine of its seventy-three occurrences (12\%). It appears in the R3 position ten times (14\%), as does \textit{coordination} (91\%). In the R4 slot, there are ten instances of both \textit{costs} (11\%) and \textit{cost} (23\%). The primary term in the L2 position is \textit{means}, where it occurs in sixty-five out of sixty-seven co-occurrences.
(97%) with production. Process prevails in the R1 position, with 88% of its occurrences there. Sales is not the most frequent term in any of the span positions. Primary collocates, with their overall frequency of occurrence with the node and rate of occurrence in the remaining positions, are:

\[ L1 = mass 16, 94\%; L3 = used 17, 53\%; R2 = goods 23, 48\%; marketing 23, 48\%. \]

Fifty-one percent of the occurrences of the top five collocates occur to the right of the node.

Several extended collocations are found among the concordance lines. Three-word units with their frequencies are: production of goods, 11; cost of production, 11; costs of production, 8; production possibility frontier, 6; production possibility curve, 6; production possibility curves, 4. The four-word combination the means of production occurs sixty-three times, and this unit can be extended to form of the means of production, which occurs fifty-nine times.

Table 5.58: mutual information and t-scores from the reference corpus

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = production, f(n) = 879 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I ((n,c))</td>
<td>t ((n,c))</td>
<td>f((n,c))</td>
<td>f((c))</td>
<td>c</td>
</tr>
<tr>
<td>5.50</td>
<td>8.35</td>
<td>73</td>
<td>1,852</td>
<td>costs</td>
</tr>
<tr>
<td>7.03</td>
<td>8.12</td>
<td>67</td>
<td>586</td>
<td>means</td>
</tr>
<tr>
<td>5.84</td>
<td>7.55</td>
<td>59</td>
<td>1,182</td>
<td>process</td>
</tr>
<tr>
<td>4.07</td>
<td>5.56</td>
<td>35</td>
<td>2,379</td>
<td>cost</td>
</tr>
<tr>
<td>3.80</td>
<td>4.91</td>
<td>28</td>
<td>2,301</td>
<td>sales</td>
</tr>
<tr>
<td>4.60</td>
<td>4.60</td>
<td>23</td>
<td>1,087</td>
<td>goods</td>
</tr>
<tr>
<td>3.90</td>
<td>4.48</td>
<td>23</td>
<td>1,759</td>
<td>marketing</td>
</tr>
<tr>
<td>4.05</td>
<td>3.87</td>
<td>17</td>
<td>1,176</td>
<td>used</td>
</tr>
<tr>
<td>7.86</td>
<td>3.98</td>
<td>16</td>
<td>79</td>
<td>mass</td>
</tr>
<tr>
<td>7.18</td>
<td>3.29</td>
<td>11</td>
<td>87</td>
<td>coordination</td>
</tr>
</tbody>
</table>

All collocations formed with production and any of the collocates in Table 5.58 receive scores indicating significant associations by both statistical measures. Combinations involving each of the top three collocates receive high t-scores due to their rates of recurrence. In addition, the I-value for the association between means and production is high as a result of the idiomatic
nature of means (i.e., means of). Collocations composed of the node and either of the bottom two collocates receive high I-values which indicate their idiosyncratic nature.

In the Tobacco-Documents Corpus, production collocates most frequently with tobacco, requests, schedule, cigarette, and request. Tobacco, which is the primary collocate overall, occurs eleven out of nineteen times (58%) in the L1 position. In the R3 position, schedule is the main term, with eight instances (62%). Request appears ten times (91%) in the R1 slot. In the remaining positions, the dominant terms, with their frequency of occurrence with the node and rate of occurrence in the slot, are:

L2 = materials 10, 60%; L3 = period 8, 38%; scheduled 7, 43%; L4 = product 7, 29%; problems 6, 33%; sheet 6, 33%; R2 = form 6, 83%; R4 = products 7, 43%.

There are only two three-word combinations which are worth noting. Production request form and materials in production occur six times each.

In Table 5.59, the mutual information values for all of the combinations with production fall above the level needed to designate strong association. However, those collocations formed with production and cigarettes do not quite reach the necessary score to provide a level of confidence that their occurrence was not simply due to chance. This low t-score is the result of the high absolute frequency of cigarettes, combined with its low frequency of occurrence with the node. Collocations formed with production and the set of tobacco-related terms show a relatively high level of confidence in the validity of the association, and the I-value indicates that the probability of the occurrence of this association is approximately fourteen times greater than that due to chance. Therefore, we can state that industry and/or company-specific terms will occur within the set environment of production.
Table 5.59: mutual information and \( t \)-scores from the Tobacco-Documents Corpus

\[
n = \text{production}, \ f(n) = 320
\]

<table>
<thead>
<tr>
<th>( I(n,c) )</th>
<th>( t(n,c) )</th>
<th>( f(n,c) )</th>
<th>( f(c) )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.88</td>
<td>4.06</td>
<td>19</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>8.92</td>
<td>4.11</td>
<td>17</td>
<td>65</td>
<td>requests</td>
</tr>
<tr>
<td>7.84</td>
<td>3.59</td>
<td>13</td>
<td>105</td>
<td>schedule</td>
</tr>
<tr>
<td>3.73</td>
<td>3.07</td>
<td>11</td>
<td>1,541</td>
<td>cigarette</td>
</tr>
<tr>
<td>7.30</td>
<td>3.30</td>
<td>11</td>
<td>129</td>
<td>request</td>
</tr>
<tr>
<td>4.36</td>
<td>2.33</td>
<td>6</td>
<td>542</td>
<td>menthol</td>
</tr>
<tr>
<td>3.18</td>
<td>1.99</td>
<td>5</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>3.79</td>
<td>5.94</td>
<td>41</td>
<td>5,502</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

The tobacco-related terms collocate with production according to frequencies and rates of occurrence in the following span positions. As Table 5.60 exhibits, 63% of the collocates lie to the left of the node.

Table 5.60: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (5%)</td>
<td>4 (10%)</td>
<td>1 (2%)</td>
<td>19 (46%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>4 (10%)</td>
<td>5 (12%)</td>
<td>6 (15%)</td>
</tr>
</tbody>
</table>

5.21 Quality

Quality occurs most frequently with product, management, data, performance, and cost in the reference corpus. Product collocates with quality a total of fifty-eight times. It is the principal collocate in the L1 position with twenty-seven occurrences (47%). Additionally, it is the term which occurs most often in the L4 position, with nine instances (16%). Management is the chief collocate in the R1 slot, where thirty-four of its fifty-three occurrences appear (64%). Performance occurs fifteen times (48%) in the L2 position, and cost occurs six times (21%) in the L3 slot. Data is not a key term in any of the span positions. The following terms, with their
overall frequency of occurrence with the node and rate of occurrence, are dominant in the respective positions:

\[ R2 = \text{service} 23, 30\%; \text{practices} 12, 58\%; R3 = \text{products} 17, 47\%; R4 = \text{service} 23, 17\%. \]

Fifty-four percent of the occurrences of the top five collocates appear to the left of the node.

There are many extended collocations with quality in the corpus; however, with the exception of structural collocations (e.g., of high quality), most are of low frequency. Those which occur more than five times each are: total quality management (10), quality management practices (7), level of quality (7), performance and quality (7).

All of the collocations formed with quality and any of the collocates in Table 5.61 show both strength of attraction and a level of confidence which assures their validity. This is due to their frequencies of co-occurrence and the relatively high absolute frequencies of the collocating terms.

<table>
<thead>
<tr>
<th>(l(n,c))</th>
<th>(t(n,c))</th>
<th>(f(n,c))</th>
<th>(f(c))</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.23</td>
<td>7.41</td>
<td>58</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>5.96</td>
<td>7.16</td>
<td>53</td>
<td>1,427</td>
<td>management</td>
</tr>
<tr>
<td>5.59</td>
<td>5.45</td>
<td>31</td>
<td>1,078</td>
<td>data</td>
</tr>
<tr>
<td>5.93</td>
<td>5.48</td>
<td>31</td>
<td>856</td>
<td>performance</td>
</tr>
<tr>
<td>4.36</td>
<td>5.12</td>
<td>29</td>
<td>2,379</td>
<td>cost</td>
</tr>
<tr>
<td>5.59</td>
<td>4.70</td>
<td>23</td>
<td>802</td>
<td>service</td>
</tr>
<tr>
<td>4.31</td>
<td>3.92</td>
<td>17</td>
<td>1,438</td>
<td>products</td>
</tr>
<tr>
<td>6.25</td>
<td>3.42</td>
<td>12</td>
<td>265</td>
<td>practices</td>
</tr>
</tbody>
</table>

In the Tobacco-Documents Corpus, quality co-occurs most often with air, control, assurance, indoor, and tobacco. Air collocates with the node thirty-seven times, and 100% of those occurrences are in the L1 position, where it is the principal collocate. Control is the primary term in the R1 position, with twenty-seven of its thirty instances (90%) appearing in that
slot. *Indoor* prevails in the L2 position with twenty-three occurrences (88%). In the R3 position, *tobacco* occurs eight times (33%). *Assurance* is not prevalent in any of the span positions, although it is a close second in the R1 position with twenty-two occurrences (76%). Other chief terms, with their overall frequency of occurrence with the node and rate of occurrence in the span position, are:

L3 = *add* 8, 50%; L4 = *cost* 19, 37%; R2 = *samples* 8, 63%; R4 = *available* 6, 67%.

Three extended collocations, each composed of three words, are evident in the concordance lines. *Indoor air quality* occurs the most often with twenty-three instances. The other two multiword units with their frequency of occurrence are *people insuring quality* (11) and *quality control samples* (5).

Table 5.62: mutual information and $t$-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>n = quality, f(n) = 308</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (n,c)</td>
</tr>
<tr>
<td>7.54</td>
</tr>
<tr>
<td>7.12</td>
</tr>
<tr>
<td>10.77</td>
</tr>
<tr>
<td>9.61</td>
</tr>
<tr>
<td>4.27</td>
</tr>
<tr>
<td>3.78</td>
</tr>
<tr>
<td>3.23</td>
</tr>
<tr>
<td>2.50</td>
</tr>
<tr>
<td>3.70</td>
</tr>
</tbody>
</table>

In Table 5.62, all collocations composed of *quality* and any of the non-tobacco terms meet the criteria for significance by both statistical measures. Their high I-values indicate their technical associations. Collocations with *assurance* receive a particularly high score for mutual information due to the unusual nature and the number of times the two terms co-occur compared with the absolute frequency of occurrence of the collocate. *Assurance* collocates with *quality* in
91% of its occurrences in the corpus; on the other hand, the association is only 9% of the total occurrences of quality.

With regard to associations between quality and the tobacco-related terms, those with smoking fall below the levels required to establish significance. Since smoking has a high frequency of occurrence in the corpus, its five co-occurrences with the node are very low and are reflected in both the I-value and the t-score for the association. Collocations with cigarettes also do not prove to be interesting. Combinations of quality and the set of tobacco terms do, however, indicate both strength of attraction, with their probability of occurrence approximately thirteen times greater than that due to chance, and an acceptable level of confidence in the association. Consequently, we can argue that, within a span of 4:4, industry and/or company-specific words will accompany quality.

Table 5.63: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>13</td>
<td>(30%)</td>
<td>*</td>
<td>5</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(12%)</td>
<td>7</td>
<td>32</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 5.63 illustrates the placement of the collocates in relation to the node, with their frequencies and rates of occurrence. Fifty-eight percent of the occurrences of the collocates are to the right of quality.

5.22 Marketing

The top five collocates of marketing in the reference corpus are planning, plan, research, mix, and sales. Plan, the second most frequent collocate, is the primary term in the R1 position, where it occurs 117 out of 143 times (82%). The only other top-five collocate prevailing in a span position is sales, with twenty-five instances (32%) in the L2 position. Dominant collocates in the
other span positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L1 = strategic 65, 72%; L3 = product 72, 24%; L4 = product 72, 18%; R2 = process 70, 40%; R3 = strategies 58, 29%; R4 = product 72, 14%.

While plan is the primary collocate in the R1 position, there are other terms in that position which are noteworthy due to their frequency of occurrence there. Those terms, with their overall frequency of occurrence with the node and rate of occurrence in the L1 position, include:

planning 155, 70%; research 102, 62%; mix 95, 92%; objectives 72, 63%; plans 61, 67%; strategies 58, 41%; information 53, 49%; manager 51, 51%; system 48, 25%; management 47, 32%; strategy 44, 66%; department 40, 73%; managers 39, 38%; program 38, 71%; programs 32, 50%; solution 31, 87%; channel 28, 61%; audit 27, 67%; people 27, 81%; function 26, 69%; activities 25, 48%; efforts 25, 80%; budget 24, 71%; mixes 23, 78%; goals 20, 70%; processes 18, 61%; communications 17, 76%; intelligence 17, 82%; concept 16, 63%; effort 16, 94%; environment 15, 73%; activity 11, 91%.

Eighty-one percent of the occurrences of the five most frequent collocates of marketing lie to its right.

There are numerous multiword units with marketing in the reference corpus. Among the most frequently occurring three-word combinations, with their occurrence rate, are: marketing planning process, 24; sales and marketing, 19; marketing mix elements, 19; strategic marketing plan, 18; strategic marketing plans, 16; marketing and sales, 13; marketing planning system, 13; strategic marketing planning, 10; marketing and advertising, 6; marketing information system, 6.
The concordance lines also reveal a four-word unit, *marketing objectives and strategies*, which occurs eight times.

As Table 5.64 demonstrates, the collocations formed with *marketing* and each of the collocates receive *t*-scores which indicate a high level of confidence that their associations are not the result of chance. The scores reflect the recurrent nature of these collocations. With the exception of combinations formed with the node and either *sales* or *product*, the collocations also receive high scores for mutual information.

Table 5.64: mutual information and *t*-scores from the reference corpus

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td><strong>f(n) = 1,759</strong></td>
<td><strong>I (n,c)</strong></td>
<td><strong>t (n,c)</strong></td>
<td><strong>f(n,c)</strong></td>
<td><strong>f(c)</strong></td>
</tr>
<tr>
<td>6.99</td>
<td>12.35</td>
<td>155</td>
<td>698</td>
<td>planning</td>
<td></td>
</tr>
<tr>
<td>7.11</td>
<td>11.87</td>
<td>143</td>
<td>593</td>
<td>plan</td>
<td></td>
</tr>
<tr>
<td>7.26</td>
<td>10.03</td>
<td>102</td>
<td>381</td>
<td>research</td>
<td></td>
</tr>
<tr>
<td>8.13</td>
<td>9.71</td>
<td>95</td>
<td>194</td>
<td>mix</td>
<td></td>
</tr>
<tr>
<td>4.28</td>
<td>8.38</td>
<td>78</td>
<td>2,301</td>
<td>sales</td>
<td></td>
</tr>
<tr>
<td>3.99</td>
<td>7.95</td>
<td>72</td>
<td>2,589</td>
<td>product</td>
<td></td>
</tr>
<tr>
<td>6.46</td>
<td>7.97</td>
<td>65</td>
<td>424</td>
<td>strategic</td>
<td></td>
</tr>
</tbody>
</table>

In the Tobacco-Documents Corpus, *marketing* collocates most often with *planning*, *sales*, *information*, *research*, and *brand*. *Planning* occurs eleven out of twenty times (55%) in the R1 position. There are six occurrences each of *information* (35%) and *manufacturing* (67%) in L2. The primary term in R2 is *research*, with 29% of its instances in that slot. Neither *sales* nor *brand* are dominant collocates in any of the span positions. The most frequent collocates, with their overall frequency of occurrence with the node and rate of occurrence in the remaining positions, are:

L1 = *direct* 15, 87%; L3 = *develop* 8, 63%; L4 = *brands* 9, 44%; R3 = *operations* 9, 78%;
R4 = *services* 9, 22%.
Few extended collocations with *marketing* are evident from the concordance lines. There are two four-word units, *information and marketing insights* (five occurrences) and *marketing sciences and information* (six occurrences). The latter can be expanded into *marketing sciences and information services*, which occurs five times.

Table 5.65: mutual information and *t*-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>(I(n,c))</th>
<th>(t(n,c))</th>
<th>(f(n,c))</th>
<th>(f(c))</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.95</td>
<td>4.46</td>
<td>20</td>
<td>77</td>
<td>planning</td>
</tr>
<tr>
<td>6.19</td>
<td>4.30</td>
<td>19</td>
<td>496</td>
<td>sales</td>
</tr>
<tr>
<td>6.07</td>
<td>4.06</td>
<td>17</td>
<td>481</td>
<td>information</td>
</tr>
<tr>
<td>5.06</td>
<td>4.00</td>
<td>17</td>
<td>968</td>
<td>research</td>
</tr>
<tr>
<td>5.14</td>
<td>3.76</td>
<td>15</td>
<td>808</td>
<td>brand</td>
</tr>
<tr>
<td>3.13</td>
<td>2.94</td>
<td>11</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>5.15</td>
<td>2.38</td>
<td>6</td>
<td>320</td>
<td>RJR</td>
</tr>
<tr>
<td>3.57</td>
<td>3.78</td>
<td>17</td>
<td>2,715</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

All of the associations between *marketing* and the collocates listed in Table 5.65 meet the requirements for significance. Collocations with *planning* receive a high *I*-value due to the frequency of co-occurrence as compared to the low frequency of occurrence of *planning*. This value indicates the unusual nature of the occurrence of the collocation. The association between *marketing* and *RJR* is shown to be strong, with the probability of its occurrence approximately thirty-five times greater than that due to chance, but its *t*-score is rather weak due to the low frequency with which the two terms co-occur. Collocations involving *marketing* and the set of tobacco-related terms attain acceptable levels of significance by both statistical measures, and we can, therefore, state with confidence that industry and/or company-specific terms will occur within the designated environment of *marketing*. 
The frequencies and rates of occurrence per position of the tobacco-related terms are given in Table 5.66. The majority of the occurrences of these terms, 52.8%, occur to the left of the node.

Table 5.66: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>*</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>(17.6%)</td>
<td>(0%)</td>
<td>(17.6%)</td>
<td>(17.6%)</td>
<td>(12%)</td>
<td>(17.6%)</td>
<td>(17.6%)</td>
<td>(17.6%)</td>
<td>(0%)</td>
</tr>
</tbody>
</table>

5.23 Rate

In the reference corpus, the most frequent collocates of rate are interest, exchange, return, percent, and inflation. Interest is the principal collocate, forming 541 collocations with rate. It is the primary term in the L1 position, occurring there 425 times (79%). It is also the main collocate in the L4 position (3%) and in the R4 position (3%). The term which occurs most often in the R2 position is return, with 156 instances (91%). Percent is the predominant collocate in the L3 position (9%) and in the R3 position (57%). It is also a key term in the R4 slot with seventeen occurrences (13%), as is per (41%). The chief terms in the remaining two positions, with their overall frequency of occurrence with the node and rate of occurrence there, are:

L2 = nominal 75, 52%; R1 = changes 30, 60%.

Rate forms 1,247 collocations with the top five terms. Sixty-seven percent of the occurrences of these collocates lie to the left of the node.

There are some collocates in the L1 position which should be listed due to their frequencies of occurrence with the node in that position. Those terms, with their overall frequency of co-occurrence with rate and their rate of occurrence in L1, are:
exchange 294, 89%; inflation 113, 37%; growth 106, 49%; tax 90, 80%; fixed 87, 49%; unemployment 86, 51%; real 82, 32%; nominal 75, 28%; annual 72, 54%; floating 39, 74%; discount 37, 76%; internal 37, 95%; natural 36, 89%; expected 32, 63%.

Extended collocations abound with this node in the reference corpus. Three-word units, with their frequencies, include:

nominal interest rate, 37; rate of return, 27; real interest rate, 25; fixed exchange rate, 23; effective interest rate, 23; rate of growth, 23; marginal tax rate, 20; floating exchange rate, 16; annual interest rate, 15; rate of inflation, 15; foreign exchange rate, 10; rate of interest, 9; average tax rate, 9; income tax rate, 7; compound interest rate, 6.

Four-word combinations with their frequencies in the corpus are:

internal rate of return, 35; natural rate of unemployment, 20; required rate of return, 16; accounting rate of return, 14; real rate of interest, 13; expected rate of inflation, 11; expected rate of return, 9; nominal rate of interest, 8.

There is also a five-word collocation, investor’s required rate of return, which occurs nine times.

Table 5.67: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>n = rate, f(n) = 1,943</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (n,c)</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>7.45</td>
</tr>
<tr>
<td>7.64</td>
</tr>
<tr>
<td>6.95</td>
</tr>
<tr>
<td>5.75</td>
</tr>
<tr>
<td>6.64</td>
</tr>
<tr>
<td>7.69</td>
</tr>
<tr>
<td>4.75</td>
</tr>
<tr>
<td>4.60</td>
</tr>
</tbody>
</table>

As Table 5.67 illustrates, all associations between rate and the collocates are significant. The t-scores for the collocations involving the top five terms indicate very high levels of
confidence in the validity of the associations, which result from the frequencies with which they co-occur. Additionally, the I-values signify strong attraction between the node and each of its collocates.

The principal collocates of rate in the Tobacco-Documents Corpus are ventilation, flow, per, heart, and high. Ventilation is the primary collocate in the L1 position, where it occurs eighteen out of nineteen times (95%). High occurs three times (27%) in the L3 position, as does increasing (60%). The other three terms are not the predominant collocate in any of the span positions; however, flow appears sixteen times (94%) in L1, and heart occurs there ten times (91%). Terms which prevail in the other positions, with their overall frequency of occurrence with the node and rate of occurrence in the slot, are:

L2 = higher 8, 75%; L4 = increase 6, 50%; R1 = procedure 5, 80%; R2 = burn 5, 60%;
R3 = cancer 9, 44%; R4 = smokers 9, 56%.

There are only a few extended collocations apparent from the concordance lines for this corpus, and they occur at a low frequency rate. Three-word units which exist, with their frequency, are: ventilation rate procedure, 5; required ventilation rate, 3; ammonia application rate, 3.

Table 5.68: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>n = rate, f(n) = 281</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I (n,c)</strong></td>
</tr>
<tr>
<td>8.46</td>
</tr>
<tr>
<td>8.81</td>
</tr>
<tr>
<td>5.74</td>
</tr>
<tr>
<td>7.66</td>
</tr>
<tr>
<td>5.55</td>
</tr>
<tr>
<td>3.85</td>
</tr>
</tbody>
</table>
In Table 5.68, the mutual information values for collocations formed with *rate* and any of
the collocates indicate strong attraction. Additionally, the combinations are valid, as exhibited by
the *t*-scores. Collocations involving *ventilation, flow,* or *heart* receive high I-values indicating
that they are specialized collocates of *rate*. This is evident from their relatively low absolute
frequencies and their positions as primary collocates of the node. The only tobacco-related term
which appears with an absolute frequency of occurrence greater than five for this node is
*smokers*. Collocations with *smokers* and the node receive an I-value and *t*-score indicative of
significance, thus allowing us to state that industry and/or company-related words will occur
within the 4:4 span of *rate*. However, the values are relatively low so that such a proposition
remains somewhat uncertain.

Table 5.69 provides the span positions for *smokers* with its frequency and rate of
occurrence per position. Sixty-seven percent of its occurrences lie to the right of *rate*.

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (22%)</td>
<td>1 (11%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>*</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (11%)</td>
<td>5 (56%)</td>
</tr>
</tbody>
</table>

5.24 Consumer

Those five terms which co-occur most frequently with *consumer* in the reference corpus are
*goods, products, markets, demand,* and *marketing*. *Goods,* which collocates with *consumer* sixty-
seven times, is the primary term in the R1 position with fifty-three occurrences (79%). The chief
collocate in the L3 position is *products*, where it appears five times (13%). *Markets* occurs
eleven times (30%) in the R3 position, as does *services* (73%). In the L4 position, there are five
occurrences each of *marketing* (18%) and *price* (20%). *Demand* is not a prime collocate in any
of the span positions. Principal collocates in the remaining positions, with their overall frequency of occurrence with the node and rate of occurrence in that slot, are:

\[ L1 = \text{packaged } 13, 62\%; L2 = \text{changes } 10, 70\%; R2 = \text{index } 23, 70\%; R4 = \text{product } 23, 30\%. \]

Of the top five collocates, 80% of their occurrences are to the right of \textit{consumer}.

Since a majority of the collocates of \textit{consumer} seem to occur in the R1 position in the reference corpus, it is worth noting some additional terms which appear in that position. Those words which occur at least ten times in R1, with their overall frequency of occurrence with the node and rate of occurrence there, are:

\[ \text{products } 39, 46\%; \text{markets } 37, 43\%; \text{demand } 29, 38\%; \text{marketing } 28, 68\%; \text{behavior } 26, 85\%; \text{price } 25, 60\%; \text{spending } 19, 74\%; \text{confidence } 16, 88\%; \text{decision } 14, 86\%; \text{bank } 10, 100\%. \]

There are several extended collocations with \textit{consumer}. Three-word units with their frequencies in the corpus include:

\[ \text{Consumer Price Index}, 11; \text{consumer and organizational}, 10; \text{consumer decision making}, 10; \text{packaged consumer goods}, 8; \text{consumer packaged goods}, 7; \text{consumer marketing budget}, 7; \text{consumer and industrial}, 6. \]

Additionally, there are two four-word units: \textit{consumer goods and services}, which occurs eleven times, and \textit{consumer and industrial markets}, which occurs eight times.

All of the collocations formed with \textit{consumer} and each of the terms in Table 5.70 receive values establishing significance by both statistical measures. Those combinations with \textit{goods} have a high \( t \)-score due to the recurrent nature of the association. Collocations involving the node
and either *index* or *packaged* score particularly high on mutual information as the result of the frequency of co-occurrence compared to the absolute frequency of the collocates.

Table 5.70: mutual information and *t*-scores from the reference corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.73</td>
<td>8.11</td>
<td>67</td>
<td>1,078</td>
<td>goods</td>
</tr>
<tr>
<td>5.54</td>
<td>6.11</td>
<td>39</td>
<td>1,438</td>
<td>products</td>
</tr>
<tr>
<td>6.19</td>
<td>6.00</td>
<td>37</td>
<td>865</td>
<td>markets</td>
</tr>
<tr>
<td>5.13</td>
<td>5.23</td>
<td>29</td>
<td>1,415</td>
<td>demand</td>
</tr>
<tr>
<td>4.77</td>
<td>5.10</td>
<td>28</td>
<td>1,759</td>
<td>marketing</td>
</tr>
<tr>
<td>4.44</td>
<td>4.77</td>
<td>25</td>
<td>1,967</td>
<td>price</td>
</tr>
<tr>
<td>7.65</td>
<td>4.77</td>
<td>23</td>
<td>196</td>
<td>index</td>
</tr>
<tr>
<td>3.93</td>
<td>4.48</td>
<td>23</td>
<td>2,589</td>
<td>product</td>
</tr>
<tr>
<td>5.00</td>
<td>3.75</td>
<td>15</td>
<td>800</td>
<td>services</td>
</tr>
<tr>
<td>9.19</td>
<td>3.60</td>
<td>13</td>
<td>38</td>
<td>packaged</td>
</tr>
<tr>
<td>4.74</td>
<td>3.04</td>
<td>10</td>
<td>641</td>
<td>changes</td>
</tr>
</tbody>
</table>

In the Tobacco-Documents Corpus, the five leading collocates of *consumer* are *research*, *product*, *products*, *testing*, and *tracking*. *Research*, which is the predominant collocate with *consumer* overall, is also the primary term in the R1 position with fourteen occurrences (58%) and in the L4 position with three occurrences (13%). It appears three times (13%) in the R4 position, as does *brand* (38%). There are three instances each of *products* (18%) and *provide* (60%) in the L2 slot. *Product*, *testing*, and *tracking* are not principal terms in any of the span positions. The most frequent terms in the remaining positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L1 = *major* 5, 60%; L3 = *wants* 10, 30%; R2 = *safety* 7, 71%; R3 = *commission* 6, 67%.

There is only one extended collocation which becomes evident from the concordance lines. The four-word unit, *Consumer Product Safety Commission*, occurs five times.
Table 5.71: mutual information and t-scores from the Tobacco-Documents Corpus

n = consumer, f(n) = 288

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.67</td>
<td>4.80</td>
<td>24</td>
<td>968</td>
<td>research</td>
</tr>
<tr>
<td>5.57</td>
<td>4.49</td>
<td>21</td>
<td>908</td>
<td>product</td>
</tr>
<tr>
<td>5.71</td>
<td>4.04</td>
<td>17</td>
<td>670</td>
<td>products</td>
</tr>
<tr>
<td>6.40</td>
<td>3.56</td>
<td>13</td>
<td>317</td>
<td>testing</td>
</tr>
<tr>
<td>9.22</td>
<td>3.31</td>
<td>11</td>
<td>38</td>
<td>tracking</td>
</tr>
<tr>
<td>3.59</td>
<td>2.25</td>
<td>6</td>
<td>1,024</td>
<td>cigarettes</td>
</tr>
<tr>
<td>2.37</td>
<td>1.97</td>
<td>6</td>
<td>2,395</td>
<td>tobacco</td>
</tr>
<tr>
<td>2.85</td>
<td>2.98</td>
<td>12</td>
<td>3,419</td>
<td>tobacco-related terms (combined)</td>
</tr>
</tbody>
</table>

**Consumer** forms significant associations with all of the non-tobacco terms, as evidenced by the I-values and t-scores in Table 5.71. The association between the node and *tracking* receives a particularly high value for mutual information due to the unusual nature of the co-occurrence of the two terms, which suggests that it is a technical collocation. Scores from both statistical measures for collocations with *consumer* and the tobacco terms are quite low. Those combinations formed with *tobacco* fail to meet the required level for significance by either measure. While the t-score for collocations with the set of tobacco terms does show confidence that the association is valid, the I-value indicates a weak association. From these scores, it is questionable as to whether industry and/or company-specific terms will occur within the designated environment of *consumer*.

The frequencies and rates of occurrence per span position are given for the tobacco-related terms in Table 5.72. Sixty-seven percent of the occurrences lie to the right of *consumer*.

Table 5.72: span positions of tobacco-related collocates by frequency and rate of occurrence

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>*</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>(8%)</td>
<td>(0%)</td>
<td>(25%)</td>
<td>(0%)</td>
<td>*</td>
<td>(8%)</td>
<td>(17%)</td>
<td>(25%)</td>
<td>(17%)</td>
</tr>
</tbody>
</table>
5.25 Business

In the reference corpus, the five terms which collocate most often with business are manager, new, managers, strategy, and management. Manager, the most frequent collocate overall, is the primary term in the R1 position, where it occurs fifty-three times (67%). It is one of six terms which occur seven times each in the R2 position. The words with their rate of occurrence in R2 are: manager, 9%; needs, 22%; technology, 25%; systems, 26%; account, 35%; government, 64%. New is the main collocate in the L1 position with forty-three occurrences (57%). Managers, strategy, and management are not principal collocates in any of the span positions. The leading collocates in the remaining span positions, with their overall frequency of occurrence with the node and rate of occurrence in the position, are:

L2 = course 14, 64%; L3 = nature 21, 67%; L4 = financial 39, 56%; R3 = market 30, 33%; R4 = customer 37, 27%.

Business forms 351 collocations with its top five collocates. Sixty-seven percent of the occurrences of the collocates lie to the right of the node.

Numerous other terms occur in the R1 position which are worth mentioning due to their frequencies of occurrence there. Those collocates with their overall frequency of occurrence with the node and rate of occurrence in R1 include:

managers 72, 71%; strategy 67, 72%; management 58, 34%; processes 49, 86%; operations 46, 65%; process 45, 58%; unit 42, 90%; intelligence 41, 49%; environment 39, 67%; cycle 34, 74%; focus 34, 47%; performance 34, 35%; functions 33, 48%; strategies 33, 67%; investment 32, 41%; needs 32, 44%; success 31, 35%; model 30, 77%; transactions 28, 71%; activities 25, 48%; executives 25, 44%; opportunities 23, 65%; people 23, 57%; plan 23, 78%; community 21, 81%.
Only four extended collocations are evident in the concordance lines. These three-word units, which occur six times each, are *business and government*, *business and technology*, *small business accounting*, and *overall business performance*.

Table 5.73: mutual information and t-scores from the reference corpus

<table>
<thead>
<tr>
<th>l (n,c)</th>
<th>t (n,c)</th>
<th>f(n,c)</th>
<th>f(c)</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.14</td>
<td>8.64</td>
<td>79</td>
<td>800</td>
<td>manager</td>
</tr>
<tr>
<td>3.82</td>
<td>8.05</td>
<td>75</td>
<td>1,900</td>
<td>new</td>
</tr>
<tr>
<td>4.82</td>
<td>8.19</td>
<td>72</td>
<td>910</td>
<td>managers</td>
</tr>
<tr>
<td>5.15</td>
<td>7.95</td>
<td>67</td>
<td>676</td>
<td>strategy</td>
</tr>
<tr>
<td>3.86</td>
<td>7.09</td>
<td>58</td>
<td>1,427</td>
<td>management</td>
</tr>
<tr>
<td>3.25</td>
<td>5.59</td>
<td>39</td>
<td>1,463</td>
<td>financial</td>
</tr>
<tr>
<td>2.90</td>
<td>5.27</td>
<td>37</td>
<td>1,771</td>
<td>customer</td>
</tr>
<tr>
<td>1.99</td>
<td>4.10</td>
<td>30</td>
<td>2,706</td>
<td>market</td>
</tr>
<tr>
<td>4.11</td>
<td>5.33</td>
<td>32</td>
<td>663</td>
<td>needs</td>
</tr>
<tr>
<td>4.30</td>
<td>5.02</td>
<td>28</td>
<td>507</td>
<td>technology</td>
</tr>
<tr>
<td>3.71</td>
<td>4.80</td>
<td>27</td>
<td>738</td>
<td>systems</td>
</tr>
<tr>
<td>4.88</td>
<td>4.43</td>
<td>21</td>
<td>255</td>
<td>nature</td>
</tr>
<tr>
<td>3.09</td>
<td>3.95</td>
<td>20</td>
<td>841</td>
<td>account</td>
</tr>
<tr>
<td>4.14</td>
<td>3.53</td>
<td>14</td>
<td>283</td>
<td>course</td>
</tr>
<tr>
<td>2.46</td>
<td>2.71</td>
<td>11</td>
<td>714</td>
<td>government</td>
</tr>
</tbody>
</table>

As Table 5.73 illustrates, all associations between *business* and the collocates attain t-scores which signify confidence that they are valid. However, three of the collocations, those between the node and *customer*, *market*, or *government*, receive I-values which fall below the acceptable level, indicating that their associations are very weak. The frequency of co-occurrence in these three cases is low, compared with the frequencies of the individual terms. The most interesting association is that between *business* and *manager*, which receives a relatively high t-score due to the recurrence of the collocation, as well as a high I-value, indicating its strong attraction as a technical combination.
The top five collocates of business in the Tobacco-Documents Corpus are category, tobacco, PM (= Philip Morris), analysis, and brands. Category, which co-occurs with the node ten times, is the principal collocate in the L3 position (60%) and in the R4 position (20%). In the L4 position, tobacco appears two times (25%), as do increase (40%) and management (40%). Analysis and development occur five times each in the R1 position, 83% and 100% respectively. The primary term in the R3 position is brands, where it occurs five out of six times (83%). PM is not the chief collocate in any of the span positions. Leading terms in the other positions, with their overall frequency of occurrence with the node and rate of frequency in the position, are:

L1 = new 6, 67%; paper 6, 67%; cigarette 5, 80%; L2 = loss 6, 83%; R2 = premium 5, 100%.

Three extended collocations with business appear in this corpus. The three-word units, which occur five times each, are source of business and sources of business. There are also five occurrences of a four-word unit, business in premium brands.

Table 5.74: mutual information and t-scores from the Tobacco-Documents Corpus

<table>
<thead>
<tr>
<th>n = business , f(n) = 273</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (n,c)</td>
</tr>
<tr>
<td>7.51</td>
</tr>
<tr>
<td>2.86</td>
</tr>
<tr>
<td>6.32</td>
</tr>
<tr>
<td>4.98</td>
</tr>
<tr>
<td>4.41</td>
</tr>
<tr>
<td>2.82</td>
</tr>
<tr>
<td>3.40</td>
</tr>
</tbody>
</table>

The t-scores for the collocations formed with business and the collocates in Table 5.74 are quite low. While collocations with the non-tobacco terms receive values by both statistical measures that indicate their validity, the same is not true for those collocations formed with the
tobacco terms. The score for the combination of the node and *cigarette* falls below the level necessary to establish significance. For the same collocation, the value for mutual information is also below the level required to signify strength of attraction. The collocation of *PM (Philip Morris)* and *business* scores a high I-value due to its unusual and technical association. Its *t*-score is low due to the low frequency of co-occurrence in relation to the frequencies of the individual terms.

The I-value for collocations formed with *business* and the set of tobacco-related terms indicate that the probability of occurrence is approximately eleven times greater than that due to chance. Additionally, the *t*-score provides confidence that the association is valid. Consequently, we can maintain that industry and/or company-specific terms will occur within the 4:4 span of *business*.

Table 5.75 lists the frequencies and rates of occurrence per span position of the tobacco-related terms. Sixty-five percent of the collocates occur to the left of *business*.

<table>
<thead>
<tr>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>node</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>45%</td>
<td>*</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

This chapter has presented the analysis of the twenty-five business terms. For each corpus, the five most frequent collocates of the term, as well as the most frequent collocates per span position, have been identified. The associations formed between the terms and these principal collocates were measured using mutual information and the *t*-test in order to assess their significance. Furthermore, multiword collocations were provided if available. The position of the majority of the collocates in relation to the node (to the left or to the right) was also
shown. The same procedures were performed using associations involving the terms and the tobacco-related collocates in the Tobacco-Documents Corpus.

The next chapter covers collocations and the dictionary. In this chapter, the results are used to create collocational dictionary entries for each term.
COLLOCATIONS AND THE DICTIONARY

As a second element of the present study, I have applied the results of my research to another area of linguistics, lexicography. In this chapter, I discuss collocations with regard to dictionaries, outline the methodology I used to create entries for the twenty-five business terms, and present dictionary entries for each of the terms.

6.1 COLLOCATIONS IN THE DICTIONARY

Many aspects of language which seem to be obvious are, in reality, only apparent in hindsight; that is, once a corpus has been analyzed. This is particularly true of collocations. First, it is through corpus analysis that the vast patterning of language becomes visible. Second, while some combinations of words are those which come most readily to mind, they are often not the most frequent collocations.

Summers (1996, 263) points out that numerous words are frequent not only as single units, but also due to the fact that they are constituents in combinations, such as set phrases. Yet, according to Zgusta (1971, 138), dictionary entries present words as single units. He proposes that those combinations of words which form a meaningful unit should receive special attention. Along similar lines, Aisenstadt (1979, 71) refers to restricted collocations (RCs), maintaining that their meaning derives from the sum of their parts and that these parts are highly
interchangeable. She asserts that these collocations should be given special consideration similar to that of idioms, as well as being the subject of special monolingual and bilingual dictionaries.

However, even though the importance of word patterning has become increasingly evident, dictionaries have lagged behind in their treatment of collocations. Benson (1990, 23) states:

the most highly regarded existing general-purpose (GP) dictionaries...all supply some collocations and thus seem to recognize in principle the importance of including such constructions. However, their treatment is in general, inconsistent and incomplete.

Landau (2001, 113) claims that all decisions made by lexicographers are driven by their need to save space. For this reason, some elements are omitted in order to make room for what the lexicographer considers as more relevant to the particular dictionary. As Benson (1990, 25) explains, the argument can be made that collocations are unnecessary in dictionaries targeting native speakers because this group is already familiar with the majority of important collocations in their language. Therefore, by excluding most collocations from such a dictionary, more space is available for headwords. Nonetheless, the inclusion of collocations in dictionaries is valuable, for both language learners, who are using dictionaries to encode language, and native speakers, who are using dictionaries to decode language.

In relation to language learners, Cowie (1978, 131-32) describes collocations composed of a node with a highly specialized meaning and a restricted set of collocates. Such limited collocations cause serious problems for the language learner. Cowie maintains that any explanation which a dictionary provides is beneficial. It is this exposure to collocations in dictionaries which enables students to learn the language (Summers 1988, 116). As Nation (1989, 65) offers, language-learning dictionaries should assist in learning, not merely present
information about the language. A similar sentiment is expressed by Hanks (1988, 74), who asserts that learners need guidance as to what phrases are natural in a language. However, he continues, such guidance is lacking in conventional dictionaries. Benson (1990, 29) argues that it is essential for collocations to be included in both monolingual and bilingual dictionaries. Additionally, he describes the placement of entries as follows:

In monolingual, learners’ dictionaries of English designed to help non-English speakers encode, the collocation *change one’s mind* should be placed at the entry for *mind* (the base). In GP dictionaries, monolingual and bilingual, the collocation should be placed, if at all possible, at both the entry for *change* (the collocator) and at the entry for *mind*; the same dictionary can then be used for decoding AND encoding by speakers of both languages (1990, 31).

In the field of language acquisition, Weinert (1995, 186-91) discusses the use of formulaic language as a learning strategy for both children and adults. Howarth (1998, 42) submits that learners have difficulties not with free collocations or with idioms, but rather with collocations which allow a limited amount of substitutability. This is further evidence that the inclusion of collocations in dictionary entries is essential.

The other group which would benefit from collocational entries in dictionaries is that of native speakers of the language. The senses of many words are not immediately clear in isolation. Guthrie et al. (1994, 79) elaborate on this situation:

The problem of semantic ambiguity is central to natural language processing (NLP). It is clear that one must know what sense of a word is being used to ‘understand’ the text in which it appears or to translate it correctly to another language.
Furthermore, they state that even within a specific type of text, a high degree of ambiguity can exist. To prove this point, they took articles from the *Wall Street Journal* from 1987 to 1989 and generated a list of sentences containing the word *bridge* and one of a few keywords (e.g., *nose, card, river*). They then compared their findings with the eight senses listed in the *Longman Dictionary of Contemporary English* (LDOCE). They found examples for seven of the senses, thus demonstrating the ambiguous nature of the word *bridge* (Guthrie et al. 1994, 80-82). It is this ambiguity which can be minimized by incorporating collocations in dictionaries.

While the above discussion has focused on general and learner dictionaries, there is also another category of dictionaries which has overlooked the importance of collocations: specialized dictionaries. Moulin (1979, 77) points out the lack of a comprehensive dictionary of business language, either monolingual or multilingual. He asserts that business dictionaries have a tendency to focus on one field of business (e.g., economics), giving only cursory attention to related areas. He asks, “How easily and how rapidly can the user find the headword of an idiomatic item he wants to look up?” Moulin continues:

A definition may be incomplete (and therefore often inaccurate) simply because it ignores some important characteristic detail. To be complete from the non-native user’s point of view, a definition should mention in what kind of context the word or phrase can be used and thus enable him to avoid mistakes…This can be achieved by means of examples or through collocational indications (1979, 78).

From this discussion, it is clear that collocations need to be included in dictionaries of all types: general purpose, learner, and specialized, including monolingual, bilingual, and multilingual ones. The following section discusses the methodology involved in the writing of dictionary entries, as well as the methodology I used for my own entries.
6.2 **Methodology**

According to Hanks (1996), a word should be analyzed according to its different patterns, both syntactic and collocational, prior to defining or translating it. The typical uses of a word comprise approximately 70-80% of all uses, and these are contained in less than a dozen patterns. Once these patterns have been established, uses which are less common can be added to an existing pattern. Hanks states that frequency is also an important factor. He adds that differences in word meaning and shades of meaning are largely context-dependent, and, therefore, the final decision on how to list collocates with a word rests with the analyst. Moreover, he contends (1996, 97):

> Precision and subtlety of interpretation arise, rather, out of the combinations of contexts in which the word participates. This makes it all the more extraordinary that dictionaries, with their multiple word senses, have paid so little attention to the contextual features associated with different senses…a model of words in use must show how the totality of the patterns in which each word regularly participates contribute to its meaning on any particular occasion when it is used, with more or less subtle changes of emphasis.

Landau (2001, 98-112) discusses various elements which are, or may be, part of a dictionary entry. These include the entry term (headword), pronunciation, grammatical information, inflected forms, run-on derivatives, idioms or other expressions, variant forms (alternative or longer forms of the headword, alternative spellings, lexical variants), and the headword listed with common prefixes (e.g., \textit{un}-). Hanks (1979, 36) offers three components relating to the method for writing entries. First, the lexicographer should be aware of the target market for the dictionary. Second, the judgment and individual style of the lexicographer will have an influence on the composition of the entries. Third, the word itself will shape the entry.
Benson (1989b, 6; 1990, 23) assigns terms to collocations included in dictionary entries. He states that a collocation which appears at the base (normally a noun) is a “collocation proper.” In this situation, the collocation is a run-in to the definition. The other type of collocation appears at the collocator, which is generally a verb or adjective. Here the collocation is regarded as essential to the definition.

In formulating a glossing technique for the twenty-five business terms, I followed Benson’s (1989a, 85) definition of a lexical collocation. He asserts that a lexical collocation is composed of two equal elements, such as a noun and adjective or verb and adverb. Furthermore, I adhered to Sinclair’s (1989, 147) practice of considering all occurrences, regardless of “how awkward or bizarre they might be.”

Prior to writing the actual entries, I considered what my target market was. I decided that my entries would constitute a basis for a collocational dictionary of business terms. The resulting dictionary would be for use by both native speakers and language learners with the purpose of facilitating the usage and acquisition of business language. With this in mind, I established a set of criteria as to what would be included in an entry and how the material would be acquired.

First, I determined that since my study was of lexical collocations, I would use only those collocates which occurred one place to the left of the node or one place to the right of the node. Next, in order for a collocate from one of these two positions to be included in the entry, it had to form a collocation relevant to business, i.e., collocations such as high levels would not qualify for inclusion. The collocations would be listed in order of their frequencies, except when a collocate appeared in the same position in both singular and plural forms. Then the ensuing collocations would be listed together. Following the two-word collocations, extended
collocations from Chapter 5 would be included. Additionally, any terms which qualified as two different parts-of-speech would be categorized accordingly in the entry.

Although I analyzed the collocational patterning in the preceding chapter according to each corpus, I decided that for this particular project I should combine the two corpora. Using WordSmith Tools, I set the stop list, which I had created for the previous analysis, selected both corpora, and ran new concordances on each of the twenty-five terms. By doing so, I was able to incorporate the results according to actual frequency of occurrence, rather than randomly selecting collocates from each corpus. I generated a list of collocates for each node and then wrote the entries.

In the next section, the actual dictionary entries are given. A few points need to be made for the sake of clarification.

1. The headwords are not lemmas, due to the fact that some words appear in both the singular and plural forms. As was explained previously, the reason for using the two forms is that it has been shown that different forms of a word vary in regard to the terms with which they collocate, as well as the frequencies.

2. The headwords are listed in alphabetical order.

3. Parts-of-speech are designated according to the headword’s usage in the collocation.

4. Terms appearing in both the singular and plural forms are listed in the same entry.

5. No pronunciation is provided (cf. The BBI Combinatory Dictionary of English).

6. If the collocate is an industry or company-related term, the designation I/C is used, and an actual citation from the corpora is given to illustrate usage.

7. In order for a collocate to appear in an entry, it must occur a minimum of 10 times in the L1 or R1 position.
8. The collocations are listed within each group according to frequency of occurrence.

6.3 The entries

**advertising n.** a non-personal communication paid for by industries, businesses, organizations, or individuals and transmitted through mass media with the intent of influencing consumers: 
*I/C advertising* (e.g., *certain types of cigarette advertising are prohibited*).

- the profession of producing such messages: *co-op advertising*.

**adj.** associated with the profession of producing non-personal communications to influence consumers: *advertising agency* (*agencies*).

- used in the production of influential messages: *advertising campaign*, *advertising program*,
  - *advertising and sales promotion*. • associated with the production of messages to influence consumers: *advertising budget*, *advertising cost*.

**analysis n.** the examination and separation of business-related issues for the purpose of problem-solving and/or decision-making: *financial analysis*, *statement analysis*, *decision analysis*, *statistical analysis*, *gap analysis*, *marginal analysis*, *ROI (return on investment) analysis*, *economic analysis*, *meta analysis*, *ratio analysis*, *sensitivity analysis*, *macroeconomic analysis*, *situation analysis*, *financial statement analysis*, *data for analysis*, *analysis of variance*.

**brand n.** a mark, symbol, word(s), or combination of these which serves to distinguish a company’s product or services from those of its competitors: *new brand*, *competitive brand*. 
brands  *n.* (pl.) new brands | I/C brands (last year we developed eight new cigarette brands) | competitive brands | American brands | international brands | discount brands | premium brands | strong brands | major brands | global brands | existing brands.

adj. consisting of the spoken portion of the identifying mark, as opposed to a symbol: brand name(s). • pertaining to a person or persons involved in the process of differentiating a company’s product or services: brand manager(s) | brand management. • belonging to a company’s specific product(s): brand group | brand styles | brand families. • associated with a company’s specific product: brand awareness | brand loyalty | brand image | brand equity | brand imagery. • used for promoting a specific product: brand name strategies.

business  *n.* an enterprise, occupation, profession, or trade engaging in commerce in order to earn a profit: new business | small business | e-business | international business | business and government | business and technology.

• the act of engaging in commerce: doing business.

adj. pertaining to a person or persons who direct a commercial enterprise: business manager(s) | business people | business executives. • used in commerce: business strategy (strategies) | business model(s) | business intelligence | business review | business plan(s) | business focus. • having to do with commerce: business process(es) | business operations | business cycle(s) | business transactions | business opportunity (opportunities) | business functions | business needs | business investment | business activities | business performance | business success. • denoting an enterprise, occupation, profession, or trade or groups of the same engaging in commerce: business unit(s) | business environment | business community | business institutions.
**company n.** a group of individuals assembled to undertake an activity or commercial venture:

*I/C company* (e.g., *the nation’s third largest tobacco company*) | *holding company* | *parent company* | *small company*. • a transaction involving a business enterprise: *company makes* | *company uses* | *company sells*. • the worth arising from a business enterprise: *value of the company*.

**adj.** associated with a commercial enterprise: *company needs* | *company performance* | *company objectives*.

---

**consumer n.** the end user of a product or service: *the retailer has educated the consumer*.

• a measure of the change in prices paid by the end users of products and services: *Consumer Price Index*

**adj.** used by customers: *consumer goods* | *consumer product(s)* | *consumer bank* | *packaged consumer goods* | *consumer goods and services*. • involving the end user of a product or service: *consumer behavior* | *consumer price* | *consumer markets* | *consumer spending* | *consumer confidence* | *consumer demand* | *consumer decision* | *consumer wants*. • having to do with identifying and influencing the end user of a product or service: *consumer marketing* | *consumer research*.

---

**data n.** factual information and statistics gathered for analysis and decision-making: *financial data* | *operational data* | *meta data* | *sales data* | *data collected* | *sufficient data* | *raw data* | *accounting data* | *customer data* | *control data* | *data obtained*.
adj. consisting of factual information and statistics: data warehouse | data mart(s) | data store | data set. • used in the gathering and analysis of factual information and statistics: data entry | data collection | data acquisition | data processing | data acquisition process.

development n. the process of producing new products or of improving, advancing, or refining existing ones: product development | research and development | new product development.

• a state of growth or improvement: economic development | leadership development | process development.

adj. producing growth or advancement: development process. • associated with growth or advancement: development project | development costs.

increase n. a rise in an amount, a level, or a value: price increase | tax increase | percent increase | percent increase in | increase in income | increase in sales | increase in government spending | increase in the money supply | I/C increase (the biggest cigarette tax increase ever).

v. to rise in amount, level, or value: increase output | increase sales | increase the value | increase the money supply.

industry n. a sector of the business world involved in commercial or economic activity: I/C industry (e.g., problems unique to the tobacco industry).

adj. having to do with a commercial or economic segment of the business world: industry volume | industry sales.
**levels n.** (pl.) amounts or extents existing at a specific time in a specific situation: *inventory levels* | *I/C levels* (e.g., *nicotine levels in cigarettes*) | *price levels* | *leadership levels* | *levels of management*.

**market n.** an arena for business or trade: *target market* | *stock market* | *test market* | *money market* | *exchange market* | *competitive market* | *free market* | *bond market* | *product market* | *open market* | *labor market* | *potential market* | *black market* | *secondary market* | *foreign exchange market*.

- a segment of the population with the potential to buy goods: *domestic market* | *local market* | *global market* | *foreign market*.

**adj.** pertaining to the business or trade arena: *market share* | *market price(s)* | *market value* | *market segment(s)* | *market demand* | *market conditions* | *market needs* | *market forces* | *market basket* | *market hedge* | *market opportunities* | *market exchange* | *market swap* | *market interest* | *market place* | *market power* | *market rate* | *market crash* | *market focus* | *stock market crash* | *target market segments*.

- characterized by an ideology combined with business or trade: *market capitalism* | *market economy (economies)* | *market socialism*.

- used to identify, analyze, or utilize the business or trade arena: *market research* | *market segmentation* | *market coordination* | *market assessment* | *market penetration*.

- consisting of persons involved in the business or trade arena: *market members* | *market participants* | *target market members*.

**marketing n.** the process of promoting products or services for sale: *strategic marketing* | *direct marketing* | *trade marketing* | *consumer marketing* | *sales and marketing* | *marketing and sales*. 
adj. used in the promotion of products or services for sale: marketing plan(s) | marketing planning | marketing mix(es) | marketing research | marketing objectives | marketing strategy (strategies) | marketing program(s) | marketing information | marketing effort(s) | marketing solution | marketing intelligence | marketing audit | marketing function | marketing budget | marketing channel | marketing goals | marketing activity (activities) | marketing communications | marketing environment | marketing system | marketing processes | marketing concept | strategic marketing plan(s). • consisting of persons involved in the promotion of products or services: marketing department | marketing manager (s) | marketing people | marketing management.

order n. a request to buy, make, sell, supply, or receive goods or services: purchase order.

adj. used to fulfill a request to buy, make, sell, supply, or receive goods or services: order processing.

product n. output resulting from a manufacturing process that can be marketed for sale: new product | consumer product | final product | marginal product | finished product | existing product.

• a measure of a country’s total output within a specified time period, usually one year: gross domestic product | gross national product

• products n. (pl.) new products | I/C products (e.g., distributors of tobacco products) | existing products | consumer products | industrial products | competitive products | joint products | by-products | finished products | company’s products | individual products |
tangible products | end products | range of products | products and services | products or services.

**adj.** pertaining to the output from a manufacturing process: *product line(s) | product mix | product knowledge.*

- involving a service that can be marketed for sale: *product development | product research | product technology | product service.*
- denoting a person or persons involved in the marketing of goods and services: *product manager(s) | product management.*
- relating to the marketing of goods and services: *product quality | product design | product life | product costs | product features | product performance | product market | product attributes | product category (categories) | product planning | product sales | product differentiation | product range | product benefits | product introduction | product concepts | product image | product improvement.

- **products** **adj.** (pl.) *products tested | products division.*

**production** **n.** the process of manufacturing something for sale: *mass production | I/C production (e.g., improve tobacco production potential) | production of goods | cost(s) of production | means of production | materials in production.*

**adj.** used in manufacturing: *production process(es) | production request | production schedules | production request form.*
- associated with the process of manufacturing: *production costs | production possibility | production volume | production function.*
**project n.** an undertaking designed to make and improvement, to gather information, or to create something new: *development project | research project.*

**adj.** used to support an planned undertaking: *project financing.* • pertaining to the nature of a planned undertaking: *project life | project risk.* • consisting of a person or persons involved in a planned undertaking: *project leader | project teams.*

**quality n.** a standard or characteristic of excellence as compared to a similar thing: *product quality | total quality | insuring quality | cost quality | level of quality | performance and quality.*

• associated with a standard or characteristic of excellence: *quality control | quality assurance | quality improvement | quality control samples.* • pertaining to a standard of excellence involving key people in an organization: *quality management | total quality management.*

**rate n.** an amount of a change or payment based on another amount and usually in the form of a percentage: *interest rate | exchange rate | tax rate | growth rate | unemployment rate | annual rate | fixed rate | inflation rate | internal rate | natural rate | discount rate | floating rate | real rate | nominal rate | expected rate | flow rate | average rate | required rate | accounting rate | hurdle rate | prime rate | market rate | participation rate | spot rate | nominal interest rate | rate of return | fixed exchange rate | rate of growth | marginal tax rate | floating exchange rate | annual interest rate | rate of inflation | foreign exchange rate | rate of interest | income tax rate | internal rate of return | expected rate of inflation | expected rate of return.*
adj. indicating activity in the amount of fluctuation: rate changes | rate swap | rate increases | rate movements. • denoting an organized method of changes: rate system.

sales n. the exchange of products or services for money: field sales | retail sales | days sales | total sales | proactive sales | I/C sales (e.g., cigarette sales) | industry sales | marketing sales | lost sales | product sales | section sales | sales per | future sales | cost of sales. • operations and activities involving the exchange of products or services for money: sales and marketing | sales and service.

adj. consisting of a person or persons involved in the exchange of products or services: sales force | sales manager(s) | sales team | sales management | sales organization | sales department | sales representative(s) | sales reps | sales personnel. • having to do with the exchange of products or services: sales volume | sales expansion | sales growth | sales potential | sales outstanding | sales revenue(s) | sales cycle | sales day | sales culture | sales taxes | sales price. • used in the recording and analysis of the exchange of products and services: sales ledger | sales data | sales forecast | sales invoice(s) | sales figures | sales journal. • used for generating the exchange of products and services: sales promotion | sales meeting(s) | sales call(s) | sales budget.

share n. a portion of something: market share | I/C share (e.g., smoker share remains unchanged). • a unit of equal value into which invested capital is divided: per share | average share | earnings per share | price per share.

• indicating a change in the value of the unit into which invested capital is divided: share points | share growth | share declined | share increased.
**system n.** an organized set of methods, ideas, or theories for the realization of a common goal:

- economic system  
- accounting system  
- computer system  
- costing system  
- political system  
- planning system  
- rate system  
- banking system  
- information system  
- tracking system  
- monetary system  
- tax system  
- management system  
- financial system  
- marketing system  
- inventory system  
- system works (e.g., how the system works)  
- payroll system  
- automated system  
- market system  
- operating system  
- cost accounting system  
- international monetary system  
- accounts payable system  
- exchange rate system  
- fixed exchange rate system  
- flexible exchange rate system.

• the central monetary authority of the U.S. established by the Federal Reserve Act of 1913: Federal Reserve System.

**work n.** an activity performed in order to generate income: *work done*  
*work in progress*  
*work done on*  
*time applications and work.*

*adj.* characterized by principles or qualities in performing a money-making activity: *work values.*

6.4 **Summary**

According to Hanks (1979, 38), “What a good dictionary offers instead is a typification: the dictionary definition summarises what the lexicographer finds to be the most typical common features, in his experience, of the use, context, and collocations of the word.” This is the approach which I have used with my entries. I have taken the collocations which I found to be the most frequent and typical for each of the twenty-five terms and written entries which will assist in the usage of business terminology.
CHAPTER 7

CONCLUSION

In this dissertation I have investigated business terminology and the words with which business terms collocate. I began with the hypothesis that the language used in different industries or companies will share some common business terminology, but that the common terms will be embedded in other language that is specific to the industry or company.

In order to determine the validity of my hypothesis, I created a corpus of just over one million words of business language. The text for the corpus was sampled from electronic versions of business-related books, journals, and newspapers in the fields of economics, accounting, marketing, finance, and management. Using a concordance software program, *WordSmith Tools*, I generated one word list from my reference corpus and one from the approximately 500,000-word Tobacco-Documents Corpus. I selected the twenty-five most frequent business terms which the two lists had in common and created lists, per word and per corpus, of the lexical collocates which occurred within a span of four words to both sides of the term. I then analyzed each term in relation to its collocates according to corpus.

The following discussion presents the overall results of the study. The points which will be addressed are:

1. the collocates of general business terms as indicators of the industry and/or company in which such terms are used;
2. the comparison of the five most frequent collocates per term between the two corpora;
3. the positioning (left or right of the node) of the top five collocates in the reference corpus,
   the top five collocates in the Tobacco-Documents Corpus, and the set of tobacco-related
   terms in the Tobacco-Documents Corpus;
4. the span position where an industry/company-specific collocate is most likely to occur
   according to term.

To establish whether or not the collocates of general business terms reflect the industry
and/or company in which such terms are used, I needed to prove that when they did co-occur
with the node, the occurrence was not merely due to chance. As discussed in Chapters 4 and 5, I
employed two statistical measures for this purpose: mutual information, which measures the
strength of attraction between two words, and the $t$-score, which measures the level of
confidence that the association between two words is genuine. Church and Hanks (1990, 24)
claim that a mutual information value (I-value) greater than 3 indicates a strength of attraction
which is high enough to prove that the pair is interesting, especially for lexicographic purposes.
Therefore, I used this same value in my analysis. Among the collocations formed between the set
of tobacco terms and each of the twenty-five business terms, three fall below the level of
significance: those with data ($I = 2.59$), work ($I = 2.59$), and consumer ($I = 2.85$).

Initially, I accepted a $t$-score of 2 or greater as indicative of significance. According to
Barnbrook (1996, 98) and Hunston and Francis (2000, 231), such values tend to identify
associations which are both significant and the most interesting. Using this limit, all associations
are shown to be significant. However, since I questioned the validity of the three associations
with low I-values, I decided to reassess the $t$-scores according to the established critical values of
significance at the 5% level for a two-tailed/non-directional test. This means that the chance of
obtaining the results are fewer than five in 100. I used this level of significance as it is standard
in the social sciences. I calculated the number of degrees of freedom\(^2\), \(N - 1\) for each association, from the frequency of co-occurrence, \(f(n,c)\), where \(n\) = the node and \(c\) = the set of tobacco-related terms. In this case \(N - 1 = f(n,c) - 1\). For example, with the association \(work\) and the set of tobacco terms, \(f(n,c) = 25\) and, therefore, \(N - 1 = 24\). For 24 degrees of freedom (d.f.) at the 5% level of significance, the critical value is 2.064. The observed value for \(t\ (4.17)\) is greater and, consequently, significant. Applying the same method to the association with \(data\) gives a critical value of 2.021 against an observed value of 6.18. The critical value for the collocation with \(consumer\) is 2.201; the observed value is 2.98. Therefore, the collocations involving the set of tobacco terms and each of the three nodes are indeed genuine and not due to chance.

As Table 7.1 exhibits, the only association which does not attain an observed value which is equal to or greater than the critical value is that with the node \(order\), whose critical value is 2.571, but observed value is 2.15. This low \(t\)-score, along with a borderline I-value (3.02), indicates that the association between \(order\) and the set of tobacco terms is not an interesting one. A visual inspection of the ‘set’ of tobacco terms which collocate with \(order\) shows that not only is the set limited to one term \(smoke\), but the co-occurrence rate (six instances) is also extremely limited.

The other association which I questioned at first was that between the tobacco terms and \(rate\). This situation is the same as with \(order\), in which there is only one tobacco term, \(smokers\), which collocates with the node. As opposed to the collocation involving \(order\), however, both the I-value and the \(t\)-score for this association are higher. When the critical value (2.306) is compared with the observed value (2.79), the collocation is shown to be significant.

In summary, the association between \(order\) and its sole tobacco-related collocate cannot be deemed interesting, and the validity of the attraction is rather uncertain. While the

---

\(^2\) the number of values of the variable which are allowed to fluctuate
associations formed with the set of tobacco terms and *data*, *work*, or *consumer* are not interesting from a lexicographic standpoint, they did not occur as the result of chance and are, therefore, genuine collocations. The associations involving the tobacco terms and each of the other twenty-one nodes are both genuine and interesting. The analysis indicates that within the span of 4:4, industry and/or company terms will collocate significantly with general business terms.

Table 7.1: comparison of statistical values for associations with the tobacco-related terms

<table>
<thead>
<tr>
<th>node</th>
<th>I-value</th>
<th>t-score</th>
<th>critical value</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>3.40</td>
<td>13.79</td>
<td>1.960</td>
<td>231</td>
</tr>
<tr>
<td>brand</td>
<td>3.92</td>
<td>16.15</td>
<td>1.960</td>
<td>298</td>
</tr>
<tr>
<td>data</td>
<td>2.59</td>
<td>6.18</td>
<td>2.021</td>
<td>54</td>
</tr>
<tr>
<td>products</td>
<td>4.48</td>
<td>17.69</td>
<td>1.960</td>
<td>342</td>
</tr>
<tr>
<td>brands</td>
<td>3.92</td>
<td>13.21</td>
<td>1.960</td>
<td>199</td>
</tr>
<tr>
<td>advertising</td>
<td>4.24</td>
<td>13.06</td>
<td>1.960</td>
<td>189</td>
</tr>
<tr>
<td>share</td>
<td>4.83</td>
<td>11.97</td>
<td>1.960</td>
<td>153</td>
</tr>
<tr>
<td>market</td>
<td>3.46</td>
<td>9.49</td>
<td>2.000</td>
<td>108</td>
</tr>
<tr>
<td>sales</td>
<td>3.50</td>
<td>7.41</td>
<td>2.000</td>
<td>65</td>
</tr>
<tr>
<td>work</td>
<td>2.59</td>
<td>4.17</td>
<td>2.064</td>
<td>24</td>
</tr>
<tr>
<td>company</td>
<td>5.78</td>
<td>17.81</td>
<td>1.960</td>
<td>328</td>
</tr>
<tr>
<td>levels</td>
<td>4.37</td>
<td>10.51</td>
<td>1.960</td>
<td>121</td>
</tr>
<tr>
<td>increase</td>
<td>3.50</td>
<td>7.00</td>
<td>2.021</td>
<td>58</td>
</tr>
<tr>
<td>analysis</td>
<td>3.56</td>
<td>6.47</td>
<td>2.021</td>
<td>49</td>
</tr>
<tr>
<td>industry</td>
<td>5.30</td>
<td>12.52</td>
<td>1.960</td>
<td>164</td>
</tr>
<tr>
<td>development</td>
<td>3.51</td>
<td>6.76</td>
<td>2.021</td>
<td>54</td>
</tr>
<tr>
<td>project</td>
<td>3.77</td>
<td>6.15</td>
<td>2.021</td>
<td>43</td>
</tr>
<tr>
<td>system</td>
<td>3.20</td>
<td>4.46</td>
<td>2.064</td>
<td>24</td>
</tr>
<tr>
<td>order</td>
<td>3.02</td>
<td>2.15</td>
<td>2.571</td>
<td>5</td>
</tr>
<tr>
<td>production</td>
<td>3.79</td>
<td>5.94</td>
<td>2.021</td>
<td>40</td>
</tr>
<tr>
<td>quality</td>
<td>3.70</td>
<td>6.05</td>
<td>2.021</td>
<td>42</td>
</tr>
<tr>
<td>marketing</td>
<td>3.57</td>
<td>3.78</td>
<td>2.120</td>
<td>16</td>
</tr>
<tr>
<td>rate</td>
<td>3.85</td>
<td>2.79</td>
<td>2.306</td>
<td>8</td>
</tr>
<tr>
<td>consumer</td>
<td>2.85</td>
<td>2.98</td>
<td>2.201</td>
<td>11</td>
</tr>
<tr>
<td>business</td>
<td>3.40</td>
<td>4.05</td>
<td>2.093</td>
<td>19</td>
</tr>
</tbody>
</table>

As part of my hypothesis, I claimed that the language used in different industries or companies will share some common business terminology. In order to demonstrate this point, I compared the five most frequent collocates of each of the twenty-five terms between the two
The results were quite unexpected. Fifteen of the terms (60%) do share at least one collocate as one of the five most frequent collocates between the corpora, although the rank order and principal span position are not necessarily the same, as Table 7.2 illustrates.

Table 7.2: shared collocates among the five most frequent collocates in both corpora

<table>
<thead>
<tr>
<th>node</th>
<th>collocate</th>
<th>reference corpus</th>
<th>Tobacco-Documents Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rank order</td>
<td>span position</td>
<td>rank order</td>
</tr>
<tr>
<td>product</td>
<td>new</td>
<td>1 L1</td>
<td>1 L1</td>
</tr>
<tr>
<td>brand</td>
<td>new</td>
<td>5 L1</td>
<td>1 L1</td>
</tr>
<tr>
<td>data</td>
<td>new</td>
<td>1 L1</td>
<td>2 L1</td>
</tr>
<tr>
<td>products</td>
<td>new</td>
<td>1 L1</td>
<td>2 L1</td>
</tr>
<tr>
<td>brands</td>
<td>new</td>
<td>1 L1</td>
<td>1 L1</td>
</tr>
<tr>
<td>advertising</td>
<td>promotion</td>
<td>3 R3</td>
<td>5 R2</td>
</tr>
<tr>
<td>share</td>
<td>market</td>
<td>1 L1</td>
<td>3 R2</td>
</tr>
<tr>
<td>market</td>
<td>share</td>
<td>1 R1</td>
<td>1 L2</td>
</tr>
<tr>
<td>sales</td>
<td>force</td>
<td>2 R1</td>
<td>1 R1</td>
</tr>
<tr>
<td>marketing</td>
<td>marketing</td>
<td>4 R2</td>
<td>3 L1</td>
</tr>
<tr>
<td>work</td>
<td>done</td>
<td>3 R1</td>
<td>1 R3</td>
</tr>
<tr>
<td>company</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>levels</td>
<td>higher</td>
<td>1 L1</td>
<td>5 L1</td>
</tr>
<tr>
<td>increase</td>
<td>price</td>
<td>3 R4</td>
<td>2 L1</td>
</tr>
<tr>
<td>analysis</td>
<td>data</td>
<td>2 R4</td>
<td>1 L2</td>
</tr>
<tr>
<td>industry</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>development</td>
<td>product</td>
<td>1 L1</td>
<td>1 L1</td>
</tr>
<tr>
<td></td>
<td>new</td>
<td>2 L2</td>
<td>3 R2</td>
</tr>
<tr>
<td></td>
<td>research</td>
<td>3 L2</td>
<td>2 L2</td>
</tr>
<tr>
<td>project</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>system</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>order</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>production</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>quality</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>marketing</td>
<td>planning</td>
<td>1 R1</td>
<td>1 R1</td>
</tr>
<tr>
<td></td>
<td>research</td>
<td>3 R1</td>
<td>4 R1</td>
</tr>
<tr>
<td></td>
<td>sales</td>
<td>5 L2</td>
<td>2 R1</td>
</tr>
<tr>
<td>rate</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>consumer</td>
<td>products</td>
<td>2 R1</td>
<td>3 R1</td>
</tr>
<tr>
<td>business</td>
<td></td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
This does not mean that these are the only shared collocates among the top five collocates per term or that the other ten terms do not share any collocates. Rather it shows that only these collocates appear somewhere among the five most frequent collocates per term in both corpora. Table 7.3 lists shared collocates where a collocate occurs among the five most frequent collocates of a term in one corpus, but lower on the list of the other corpus.

Table 7.3: shared collocates among the five most frequent collocates in only one corpus

<table>
<thead>
<tr>
<th>node</th>
<th>collocate</th>
<th>reference corpus</th>
<th>Tobacco-Documents Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>rank order span</td>
<td>rank order span</td>
</tr>
<tr>
<td>product</td>
<td>consumer</td>
<td>65 L4</td>
<td>5 L1</td>
</tr>
<tr>
<td>brand</td>
<td>product</td>
<td>2 L2</td>
<td>48 R4</td>
</tr>
<tr>
<td>names</td>
<td>analysis</td>
<td>13 R4</td>
<td>1 R2</td>
</tr>
<tr>
<td>data</td>
<td>collected</td>
<td>63 R1, R2, R3, R4</td>
<td>2 R1</td>
</tr>
<tr>
<td></td>
<td>information</td>
<td>5 R4, R1</td>
<td>20 L2</td>
</tr>
<tr>
<td>products</td>
<td>market</td>
<td>5 L4</td>
<td>23 L3, R2</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>104 R4</td>
<td>3 L2</td>
</tr>
<tr>
<td></td>
<td>delivery</td>
<td>195 L2</td>
<td>4 L1</td>
</tr>
<tr>
<td></td>
<td>use</td>
<td>61 L4</td>
<td>5 L3</td>
</tr>
<tr>
<td>brands</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>advertising</td>
<td>sales</td>
<td>1 R2</td>
<td>13 R2</td>
</tr>
<tr>
<td></td>
<td>new</td>
<td>17 L1</td>
<td>3 L3, R2</td>
</tr>
<tr>
<td>share</td>
<td>per</td>
<td>2 L1</td>
<td>11 L1</td>
</tr>
<tr>
<td></td>
<td>price</td>
<td>3 L2</td>
<td>71 L3, L4</td>
</tr>
<tr>
<td></td>
<td>earnings</td>
<td>4 L2</td>
<td>36 L2</td>
</tr>
<tr>
<td>market</td>
<td>product</td>
<td>4 L2</td>
<td>30 L3</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>481 L1</td>
<td>2 L1</td>
</tr>
<tr>
<td></td>
<td>segment</td>
<td>16 R1</td>
<td>4 L4</td>
</tr>
<tr>
<td></td>
<td>volume</td>
<td>194 R4</td>
<td>5 R4</td>
</tr>
<tr>
<td>sales</td>
<td>product</td>
<td>5 R4</td>
<td>53 L4</td>
</tr>
<tr>
<td></td>
<td>field</td>
<td>268 L1</td>
<td>2 L1</td>
</tr>
<tr>
<td></td>
<td>increase</td>
<td>14 L1</td>
<td>4 R1</td>
</tr>
<tr>
<td></td>
<td>promotion</td>
<td>27 R1</td>
<td>5 R1</td>
</tr>
<tr>
<td>work</td>
<td>completed</td>
<td>72 R3</td>
<td>4 L2, R1, R3</td>
</tr>
<tr>
<td>company</td>
<td>business</td>
<td>3 R2</td>
<td>30 L3</td>
</tr>
<tr>
<td></td>
<td>product</td>
<td>5 R4</td>
<td>24 L3, R3</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>379 L1, L2</td>
<td>4 L2</td>
</tr>
</tbody>
</table>
As can be seen in Table 7.3, there are vast discrepancies in both the rank positions of these shared collocates, as well as in their span positions. In order to better understand these variances, it is perhaps useful to consider each term and its shared collocates. For comparison, I shall refer to each collocate’s total percentage of occurrence with the node per corpus. In this manner, I will be able to demonstrate how differently co-occurring terms are used in the two corpora. By *shared collocates*, I am referring to a collocate which occurs among the five most frequent collocates for a term in at least one of the corpora.
For the term *product*, the collocates *new* and *development* appear among the five most frequent collocates in both corpora. *New* occurs at a rate of 6.68% in the reference corpus and 8.15% in the Tobacco-Documents Corpus. It is the most frequent collocate of *product* in both corpora and appears most often in the same span position, L1. Likewise, *development* occurs in the same rank order, third, and in the same span position, R1, in both corpora. Its rate of co-occurrence with *product* is 5.29% in the reference corpus and 3.30% in the Tobacco-Documents Corpus. The other shared collocate is *consumer*. Its collocational behavior does not duplicate the other two collocates. While it is the fifth most frequent collocate in the Tobacco-Documents Corpus, with a rate of 2.31%, it ranks in sixty-fifth place in the reference corpus, with a rate of 0.77%. Additionally, the primary span positions vary, L1 for the Tobacco-Documents Corpus and L4 for the reference corpus.

*Brand* shares three collocates between the corpora: *new*, *product*, and *names*. *New* is the only collocate which appears in the top five of both corpora, number one in the Tobacco-Documents Corpus and number five in the reference corpus. With a co-occurrence rate with *brand* of 6.19% in the Tobacco-Documents Corpus, its rate in the reference corpus (3.25%) is roughly half. *New* appears most often in the L1 span position in both corpora. *Product* and *names* occur among the five most frequent collocates of *brand* in the reference corpus, second and fourth respectively. In the Tobacco-Documents Corpus, *product* is in forty-eighth place and *names* is in seventy-eighth place. Their rates of co-occurrence with *brand* vary dramatically between the corpora as well. *Product* has a rate of 6.24% in the reference corpus and 1.11% in the Tobacco-Documents Corpus. The rate for *names* is 3.87% in the reference corpus and 0.87% in the Tobacco-Documents Corpus. The span positions for *product* differ according to corpus, L2
in the reference corpus and R4 in the Tobacco-Documents Corpus. Names occurs in the R1 position in both corpora.

There are three shared collocates for data between the corpora. Analysis is the most frequently collocating term in the Tobacco-Documents Corpus, with a rate of 5.26%. It is primarily found in the R2 position in that corpus. In the reference corpus, where it is in thirteenth place, it co-occurs at a rate of 2.41%. Its main span position is L4. The second most frequent collocate of data in the Tobacco-Documents Corpus is collected with a rate of 3.45%. It occurs most often in the R1 position. The rate of co-occurrence for collected in the reference corpus is 0.83%; its rank order is sixty-three. Collected occurs with the same frequency in the R1, R2, R3, and R4 positions. Information is the fifth most frequent collocate of data in the reference corpus and the twentieth most frequent one in the Tobacco-Documents Corpus, with rates of co-occurrence being 3.15% and 1.65% respectively. The span positions are R4 in the reference corpus and L2 in the Tobacco-Documents Corpus.

Products has five shared collocates between the corpora: new, market, low, delivery, and use. New is the most frequent collocate in the reference corpus and the second most frequent one in the Tobacco-Documents Corpus. Its rates of co-occurrence, 10.50% (reference corpus) and 9.40% (Tobacco-Documents Corpus), are comparable, and its primary span position, L1, is identical between the corpora. Market occurs in fifth place among the collocates in the reference corpus and in twenty-third place in the Tobacco-Documents Corpus. Its rate of co-occurrence (3.41%) in the reference corpus is more than twice that (1.49%) in the Tobacco-Documents Corpus. The principal span position for market in the reference corpus is L4; those in the Tobacco-Documents Corpus are L3 and R2. Low is the third most frequent collocate in the Tobacco-Documents Corpus, with a rate of co-occurrence of 3.73%. It occurs in 104th place in
the reference corpus, with a rate of co-occurrence of 0.56%. The span positions for low are L2 in the Tobacco-Documents Corpus and R4 in the reference corpus. Delivery, which occurs in fourth place in the Tobacco-Documents Corpus and 195th place in the reference corpus, has a rate of co-occurrence in the former corpus (3.43%) almost ten times that of in the latter (0.35%). It occurs most often in the L1 position in the Tobacco-Documents Corpus and in the L2 position in the reference corpus. Use is the fifth most frequent collocate of products in the Tobacco-Documents Corpus, with a rate of 2.84% and a primary span position of L3, and the sixty-first most frequent collocate in the reference corpus, with a rate of 0.76% and a primary span position of L4.

The single shared collocate of brands is new. It co-occurs at a rate of 7.17% in the Tobacco-Documents Corpus and a rate of 4.4% in the reference corpus. It is the most frequent collocate, and occurs primarily in the L1 position, in both corpora.

Shared collocates of advertising are promotion, sales, and new. Promotion appears in the top five of both corpora. It is collocate number three in the reference corpus, with a rate of co-occurrence of 5.71%, and collocate number five in the Tobacco-Documents Corpus, with a rate of 3.48%. It occurs most often in the R3 position in the reference corpus and in the R2 position in the Tobacco-Documents Corpus. While sales is the most frequently co-occurring term in the reference corpus, with a rate of 9.59%, it occurs in thirteenth place in the Tobacco-Documents Corpus, with a rate of 2.2%. However, it does have the same principal span position, R2, in both corpora. New is the third most frequent collocate in the Tobacco-Documents Corpus (rate of 4.21%) and the seventeenth in the reference corpus (rate of 1.84%). Its span positions differ according to corpus: L3 and R2 in the Tobacco-Documents Corpus and L1 in the reference corpus.
For the term *share*, *market* is among the top five collocates in both corpora. It is the most frequent collocate in the reference corpus, where it has a rate of co-occurrence of 30.74%. In the Tobacco-Documents Corpus, it is the third most frequent collocate. Its rate of collocation there is 13.21%. It occurs most often in the L1 position in the reference corpus and in the R2 position in the Tobacco-Documents Corpus. Other shared collocates of *share* are *per*, *price*, and *earnings*, which rank second, third, and fourth in the reference corpus and eleventh, seventy-first, and thirty-sixth in the Tobacco-Documents Corpus respectively. In the reference corpus, the co-occurrence rates of these three collocates are *per*, 14.01%; *price*, 5.64%; and *earnings*, 5.45%. The rates in the Tobacco-Documents Corpus are *per*, 4.04%; *price*, 1.10%; and *earnings*, 1.83%. The primary span position for *per* is the same for both corpora, L1. In the reference corpus, *price* is found most frequently in the L2 position; in the Tobacco-Documents Corpus, *price* occurs equally in the L3 or L4 position. *Earnings* appears most often in the same position, L2, in both corpora.

The most frequent collocate of *market* is *share* in both corpora, although the collocation rates vary (5.99% in the reference corpus, as opposed to 12.59% in the Tobacco-Documents Corpus), as do their primary span positions (R1 in the reference corpus, L2 in the Tobacco-Documents Corpus). *Product* is in fourth place in the reference corpus and in thirtieth place in the Tobacco-Documents Corpus. It has a co-occurrence rate of 3.25% in the former and 1.46% in the latter. Its predominant span positions are L2 in the reference corpus and L3 in the Tobacco-Documents Corpus. Although *test* occurs in the same position in both corpora, L1, its rate of co-occurrence with *market* and rank position vary drastically between the corpora. *Test* is the second most frequent collocate in the Tobacco-Documents Corpus, with a rate of 11.50%. However, it is in 481st place in the reference corpus, with a rate of collocation of only 0.18%.
Segment, which has a rate of 3.28% and is the fourth most frequently co-occurring term in the Tobacco-Documents Corpus, can be found in sixteenth place in the reference corpus, with a rate of 1.92%. Its main span positions are L4 for the Tobacco-Documents Corpus and R1 for the reference corpus. Volume is in fifth place in the Tobacco-Documents Corpus, where it has a collocation rate of 3.28%, and in 194th place in the reference corpus, where its rate is 0.37%. It occurs most often in the R4 position in both corpora.

Sales has six shared collocates between the two corpora: force, marketing, product, field, increase, and promotion. Force and marketing appear in the top five collocates of both corpora. Force is the most frequent collocate of sales in the Tobacco-Documents Corpus, with a co-occurrence rate of 6.65%, and the second most frequent collocate in the reference corpus, with a rate of 4.17%. Its primary span position, R1, is the same for both corpora. Marketing is found in third place in the Tobacco-Documents Corpus and in fourth place in the reference corpus. Surprisingly, its rate of co-occurrence with sales, 3.43%, is the same for both corpora. The span positions do vary, L1 for the Tobacco-Documents Corpus and R2 for the reference corpus. Product is the fifth most frequent collocate in the reference corpus, but falls to fifty-third place in the Tobacco-Documents Corpus, having rates of 3.09% and 1.21% respectively. The dominant span positions for product are R4 in the reference corpus and L4 in the Tobacco-Documents Corpus. Field is in second place among the collocates in the Tobacco-Documents Corpus, but in 268th place in the reference corpus. Its rates of co-occurrence are 5.24% for the Tobacco-Documents Corpus and 0.30% for the reference corpus. It occurs most often in the L1 position in both corpora. Increase is the fourth most frequent collocate in the Tobacco-Documents Corpus, with a rate of 3.23%, and the fourteenth most frequent one in the reference corpus, with a rate of 2.00%. Its primary span positions are R1 for the Tobacco-Documents Corpus and L1 for the
Promotion is the fifth-place collocate in the Tobacco-Documents Corpus and the twenty-seventh-place collocate in the reference corpus. Its rate of co-occurrence is 3.23% in the former and 1.52% in the latter. It is found predominantly in the R1 position in both corpora.

There are two shared collocates of work: done and completed. Done is the most frequently occurring collocate in the Tobacco-Documents Corpus, with a rate of collocation of 5.45%, and the third most frequently occurring one in the reference corpus, with a rate of 3.38%. Its primary span positions are R3 in the Tobacco-Documents Corpus and R1 in the reference corpus. Completed is the fourth most frequent collocate in the Tobacco-Documents Corpus, having a co-occurrence rate of 2.40%. It is in seventy-second place in the reference corpus, where its rate is 0.74%. While R3 is a dominant position for completed in both corpora, it also occurs equally as often in L2 and R1 in the Tobacco-Documents Corpus.

Company has three shared collocates: business, product, and American. Business is the third most frequent collocate of company in the reference corpus and the thirtieth in the Tobacco-Documents Corpus, with collocation rates of 1.89% and 1.19% respectively. It occurs most often in the R2 position in the reference corpus and in the L3 position in the Tobacco-Documents Corpus. The fifth-place collocate in the reference corpus is product, having a rate of 1.65%. In the Tobacco-Documents Corpus, it appears in twenty-fourth place, where its rate is 1.38%. Its major span positions are R4 in the reference corpus and L3 and R3 in the Tobacco-Documents Corpus. American is the fourth most frequent collocate of company in the Tobacco-Documents Corpus, with a co-occurrence rate of 9.09% and primary span position of L2. In this corpus, it collocates with company in the name American Tobacco Company. In the reference corpus, American falls to 379th place, with a collocation rate of 0.18%. Its dominant positions in this corpus are L1 and L2, where it is used to indicate nationality.
The shared collocates of *levels* include *higher, high, different, inventory, and lower*. *Higher* is the number one collocate of *levels* in the reference corpus and has a co-occurrence rate of 7.75%. It is the fifth-place collocate in the Tobacco-Documents Corpus, with a rate of 4.61%. *High* is the fourth most frequent collocate in both corpora, with a rate of 6.78% in the reference corpus and 4.61% in the Tobacco-Documents Corpus. *Different* is in second place among collocates in the reference corpus, where its rate of collocation is 7.26%, and in thirtieth place in the Tobacco-Documents Corpus, having a rate of 2.18%. The third most frequent collocate in the reference corpus is *inventory*, which falls to fifty-fifth place in the Tobacco-Documents Corpus. The collocation rates are 7.26% and 1.21% respectively. *Lower*, which appears in fifth place among the collocates in the reference corpus and has a rate of 5.08%, is the seventh most frequent collocate in the Tobacco-Documents Corpus, with a rate of 3.64%. The primary span position of all five collocates is L1 in both corpora, with the exception of *different*, which tends to occur in the L2 position in the Tobacco-Documents Corpus.

*Increase* has four shared collocates: *price, tax, share, and sales*. *Price* is the only collocate which appears in the top five of both corpora. It is number two in the Tobacco-Documents Corpus, with a co-occurrence rate of 5.69%, and number three in the reference corpus, with a rate of 6.67%. Its dominant span positions are L1 for the former and R4 for the latter. *Tax* is the most frequent collocate of *increase* in the Tobacco-Documents Corpus, where it has a collocation rate of 6.44% and appears most often in the L1 position. It drops to thirty-ninth place in the reference corpus, with a rate of 1.31% and primary positions of L4 and R1. *Share* is the third most frequently co-occurring term in the Tobacco-Documents Corpus, having a rate of 4.46%. Its rank in the reference corpus is thirty-fifth, with a co-occurrence rate of 1.43%. It occurs most often in the L1 position in the Tobacco-Documents Corpus, in the R2 position in the
reference corpus. *Sales* is the fifth-place collocate of *increase* in the reference corpus and the sixth-place one in the Tobacco-Documents. Its rates of co-occurrence are 5.60% in the former, where it occurs most often in the R1 position, and 3.71% in the latter, where it prefers the L1 position.

Shared collocates of *analysis* are *data*, *used*, and *results*. *Data* is the most frequent collocate in the Tobacco-Documents Corpus, with a collocation rate of 8.72%, and the second most frequent one in the reference corpus, with a rate of 3.60%. In the Tobacco-Documents Corpus, it occurs most often in the L2 position; in the reference corpus, its prime position is R4. *Used* is the fourth-place collocate in the reference corpus, where it has a rate of 3.32%. It is in fifteenth place in the Tobacco-Documents Corpus, having a co-occurrence rate of 2.66%. Its main span position is L3 in both corpora. *Results*, which is the second most frequently occurring collocate in the Tobacco-Documents Corpus and has a rate of 5.33% there, drops to eighty-first place in the reference corpus, where its rate of co-occurrence is 0.83%. The primary span positions of *results* are L4, R1, and R3 in the Tobacco-Documents Corpus and L3 in the reference corpus.

There are two shared collocates of *industry: government* and *position*. *Government* is the fourth most frequent collocate of *industry* in the Tobacco-Documents Corpus, with a co-occurrence rate of 3.37%. It is in eleventh place in the reference corpus, with a rate of 1.97%. It occurs most often in the L2 position in the Tobacco-Documents Corpus and in the R2 position in the reference corpus. *Position* is the fifth-place collocate in the Tobacco-Documents Corpus, having a collocation rate of 2.65%, and the forty-first place collocate in the reference corpus, where its rate is 1.15%. Its dominant span positions are R2 in the Tobacco-Documents Corpus and L4 in the reference corpus.
Development has four shared collocates: product, new, research, and process. Three of the collocates occur in the top five of both corpora. Product is the most frequent collocate of development in both corpora. However their rates of co-occurrence are rather varied, 22.26% in the reference corpus, 8.04% in the Tobacco-Documents Corpus. Its prime span position is L1 in both corpora. New is the second most frequent collocate in the reference corpus and the third most frequent in the Tobacco-Documents Corpus, with rates of 11.13% and 6.97% respectively. Primary span positions of new are L2 for the reference corpus and R2 for the Tobacco-Documents Corpus. The order is reversed for research, where it is in second place in the Tobacco-Documents Corpus and in third place in the reference corpus. The respective rates of co-occurrence are 7.77% and 6.93%. Research appears most often in the L2 position in both corpora. Process, which is the fourth-place collocate in the reference corpus, where it has a collocation rate of 6.57%, is the sixth-place collocate in the Tobacco-Documents Corpus, with a rate of 4.56%. Its main span positions are R1 for the reference corpus and L1 for the Tobacco-Documents Corpus.

The sole shared collocate of project is development. It is the most frequent collocate in the Tobacco-Documents Corpus, where it has a co-occurrence rate of 4.68%. In the reference corpus, it is found in twenty-fourth place, with a rate of 1.46%. It occurs most often in the L1 position in both corpora.

Shared collocates of system are new and data. New is the most frequently co-occurring term in the Tobacco-Documents Corpus. Its collocation rate in the corpus is 5.04%. It drops to thirty-fourth place in the reference corpus, having a rate of 1.30%. Its primary span positions are L1 and L3 in both corpora, as well as L2 in the Tobacco-Documents Corpus. Data occurs in third place among the collocates in the Tobacco-Documents Corpus and in twentieth place in the
reference corpus. Its rates of co-occurrence are 3.64% in the former and 1.71% in the latter. Data occurs most frequently in the L1 span position in the Tobacco-Documents Corpus and in the L2 position in the reference corpus.

Order has one shared collocate, date. Date is the second most frequent collocate in the Tobacco-Documents Corpus, with a co-occurrence rate of 2.88%. In the reference corpus, it is in forty-first place, having a collocation rate of 1.04%. It occurs most often in the L3 and L4 positions in the Tobacco-Documents Corpus and in the L1 position in the reference corpus.

The shared collocates of production are costs, cost, and schedule. Costs is the primary collocate of production in the reference corpus, but drops to fifteenth place in the Tobacco-Documents Corpus. Its rate of co-occurrence in the reference corpus (8.30%) is approximately 4.5 times that of its rate in the Tobacco-Documents Corpus (1.88%). The dominant span position for costs is R1 in both corpora. Cost is the fourth-place collocate in the reference corpus, with a collocation rate of 3.98%. It appears in sixteenth place in the Tobacco-Documents Corpus, where it has a rate of 1.88%. Cost occurs most often in the L2 position in the reference corpus. It occurs at an equal rate in the L2, L3, L4, R1, R3, and R4 positions in the Tobacco-Documents Corpus. Schedule is the third most frequent collocate in the Tobacco-Documents Corpus and is the thirty-fourth most frequent in the reference corpus. Its rate of co-occurrence is 4.06% in the former and 1.14% in the latter. The primary span positions are R3 for the Tobacco-Documents Corpus and R1 for the reference corpus.

There are four shared collocates of quality: product, data, cost, and control. Product is the top collocate in the reference corpus, having a collocation rate of 9.68%. In the Tobacco-Documents Corpus, product is in ninth place, with a rate of 4.87%. It occurs most often in the L1 position in both corpora. Data is the third most frequent collocate in the reference corpus and the
twelfth most frequent in the Tobacco-Documents Corpus, with co-occurrence rates of 5.18% and 3.25% respectively. Dominant span positions for data are R1 and R3 in the reference corpus and R3 in the Tobacco-Documents Corpus. Cost, which is the fifth-place collocate of quality in the reference corpus, is the sixth-place collocate in the Tobacco-Documents Corpus. However, its co-occurrence rate is higher in the Tobacco-Documents Corpus (6.17%) than in the reference corpus (4.84%). It occurs most frequently in the L1 position in both corpora. Control is the second most frequent collocate in the Tobacco-Documents Corpus, with a co-occurrence rate of 9.74%. It is in tenth place in the reference corpus, where it has a rate of 3.17%. Its appears most often in the R1 position in both corpora.

Marketing has six shared collocates between the corpora: planning, research, sales, plan, information, and brand. Planning, research, and sales occur among the top five collocates in both corpora. Planning is the most frequent collocate of marketing in both corpora. Its rate of co-occurrence is 8.81% in the reference corpus and 6.41% in the Tobacco-Documents Corpus. The primary span position, R1, is the same for both corpora. Research is the third most frequently co-occurring term in the reference corpus and the fourth in the Tobacco-Documents Corpus. Its collocation rate is very similar in the two corpora, 5.80% for the reference corpus and 5.45% for the Tobacco-Documents Corpus. As with planning, its dominant span position is R1 in both corpora. Sales is the second-place collocate in the Tobacco-Documents Corpus, with a co-occurrence rate of 6.09%, and the fifth-place one in the reference corpus, with a rate of 4.43%. It occurs most often in the R1 position in the Tobacco-Documents Corpus and in the L2 position in the reference corpus. Plan, which is the second most frequent collocate in the reference corpus and has a collocation rate of 8.13%, is in forty-fifth place in the Tobacco-Documents Corpus, where it has a rate of 1.92%. Information is the third most frequently co-occurring word in the
Tobacco-Documents Corpus, having a rate of 5.45%. In the reference corpus, it is in twelfth place and has a rate of 3.01%. It appears primarily in the L2 and R3 positions in the Tobacco-Documents Corpus and in the R1 position in the reference corpus. *Brand* occurs fifth among the collocates in the Tobacco-Documents Corpus but drops to seventy-fourth place in the reference corpus. Its rate of co-occurrence is approximately six times greater in the Tobacco-Documents Corpus (4.81%) than in the reference corpus (0.80%).

Shared collocates of *rate* are *per* and *high*, both of which are among the top five collocates in the Tobacco-Documents Corpus. *Per* is the third-place collocate in the corpus, with a co-occurrence rate of 5.34%. Located in eighteenth place in the reference corpus, its rate there is 2.11%. The main span positions are R3 for the Tobacco-Documents Corpus and R4 for the reference corpus. *High* is the fifth most frequent collocate in the Tobacco-Documents Corpus and has a collocation rate of 3.91%. In the reference corpus, it is found in thirty-fourth place, with a rate of 1.18%. The primary positions vary for these collocates as well, L1 in the Tobacco-Documents Corpus and L2 for the reference corpus.

There are three shared collocates for *consumer*: *products*, *marketing*, and *product*. *Products* is among the five most frequent collocates in both corpora, second in the reference corpus (co-occurrence rate of 6.63%) and third in the Tobacco-Documents Corpus (co-occurrence rate of 5.90%). It occurs most often in the R1 position in both corpora. *Marketing* is the fifth-place collocate in the reference corpus, with a collocation rate of 4.76%, and the ninth-place collocate in the Tobacco-Documents Corpus, with a rate of 2.43%. Dominant span positions are R1 for the reference corpus and L4 and R3 for the Tobacco-Documents Corpus. *Product*, which is the second most frequent collocate in the Tobacco-Documents Corpus, where
it has a rate of co-occurrence of 7.29%, is the ninth most frequent one in the reference corpus, having a rate of 3.91%.

*Business* has five shared collocates between the corpora: *new, management, category, analysis,* and *brands.* *New* and *management* are among the top five collocates in the reference corpus, while *category, analysis,* and *brands* are among those in the Tobacco-Documents Corpus. *New* is the second most frequent collocate in the reference corpus and the eighth in the Tobacco-Documents Corpus, with co-occurrence rates of 2.66% and 2.20% respectively. Its primary span position is L1 in both corpora. *Management* is the fifth-place collocate in the reference corpus (collocation rate of 2.06%) and the eighteenth one in the Tobacco-Documents Corpus (collocation rate of 1.83%). It occurs most often in the R1 position in the reference corpus and in the L4 position in the Tobacco-Documents Corpus. *Category,* which is the most frequent collocate in the Tobacco-Documents Corpus, with a co-occurrence rate of 3.66%, is found in 375th place in the reference corpus, with a rate of 0.21%. The dominant span positions are L3 in the Tobacco-Documents Corpus and R2 and R3 in the reference corpus. *Analysis* is the fourth-place collocate in the Tobacco-Documents Corpus and has a co-occurrence rate of 2.20%. It drops to 166th place in the reference corpus, where it co-occurs at a rate of 0.39%. It appears primarily in the R1 position in the Tobacco-Documents Corpus and in the L3 position in the reference corpus. *Brands* occurs in fifth place in the Tobacco-Documents Corpus and is located in 449th place in the reference corpus. The rate of co-occurrence is 2.20% in the Tobacco-Documents Corpus and 0.18% in the reference corpus. In the Tobacco-Documents Corpus, *brands* occurs most often in the R3 position, while it prefers R4 in the reference corpus.

To summarize, by comparing the five most frequent collocates of the twenty-five terms between the two corpora, I have shown that some common business terminology will be shared
across industries or companies. However, the exploration of the corpora has demonstrated that this terminology is not as prevalent as intuition would lead one to believe. As the two corpora illustrate, there are major differences between general business language and specific industry/company language. The Tobacco-Documents Corpus does not represent general business language, and it is not true that it consists of general business language with some minor modifications. Rather, it exemplifies the fact that industries and companies have their own special concerns.

The third point to be addressed concerns the positioning of the five most frequent collocates in relation to the nodes in each of the corpora, as well as that of the set of tobacco-related terms in the Tobacco-Documents Corpus. For seven of the terms, the majority of the collocates occur to the same side of the node in both corpora and also with the tobacco terms. As Table 7.4 illustrates, the majority of the collocates of *brands, levels, analysis, industry, development,* and *system* occur to the left of the node in all three situations, while the majority of the collocates of *consumer* lie to the right. In the six cases where the collocates occur primarily to the left, the nodes exhibit a tendency to take attributive adjectives. *Consumer,* on the other hand, is functioning as an adjective rather than as a noun in the majority of occurrences.

Some collocates prefer one side of the node in both the reference corpus and the Tobacco-Documents Corpus, while others tend to occur to the same side in the Tobacco-Documents and with the set of tobacco-related terms. However, a very interesting result is that the collocates of three of the terms (*share, market, increase*) occur primarily to one side of the node for both the reference corpus and the tobacco terms, but to the other side for the Tobacco-Documents Corpus. The two corpora have one common collocate per term. There are no tobacco terms in the top five collocates of *share* in the Tobacco-Documents Corpus. In that corpus,
market has one tobacco term among the five most frequent collocates, and increase has two. More research needs to be done to understand this particular situation.

Table 7.4: position of majority of the collocates in relation to the node

<table>
<thead>
<tr>
<th>node</th>
<th>reference corpus</th>
<th>Tobacco-Documents Corpus</th>
<th>tobacco terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>R</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>brand</td>
<td>R</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>data</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>products</td>
<td>R</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>brands</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>advertising</td>
<td>R</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>share</td>
<td>L</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>market</td>
<td>R</td>
<td>L</td>
<td>R</td>
</tr>
<tr>
<td>sales</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>work</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>company</td>
<td>R</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>levels</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>increase</td>
<td>R</td>
<td>L</td>
<td>R</td>
</tr>
<tr>
<td>analysis</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>industry</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>development</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>project</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>system</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>order</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>production</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>quality</td>
<td>L</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>marketing</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>rate</td>
<td>L</td>
<td>L</td>
<td>R</td>
</tr>
<tr>
<td>consumer</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>business</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
</tbody>
</table>

These results on the placement of collocates in relation to the node are important because they establish certain characteristics about the term under investigation. For those nodes where the bulk of the collocates occur to the same side in both corpora and with the set of tobacco terms, it can be argued that the position of the collocates is structurally motivated rather than lexically, i.e., the collocates are position dependent. In those cases where the majority of the collocates lie to the same side of the node in both corpora but not with the set of tobacco terms,
the indication is that the ‘neutral’ collocates are the determining factor of placement in the Tobacco-Documents Corpus. Therefore, the particular node will influence the side on which the majority of its collocates will appear. The third scenario is where the collocates occur primarily on one side of the node in the reference corpus, but they lie on the other side both in the Tobacco-Documents Corpus and with the set of tobacco terms. This shows that the industry-specific terminology governs the position of such collocates in relation to the node.

Table 7.5: most likely span position(s) for industry/company-specific collocate per term

<table>
<thead>
<tr>
<th>node</th>
<th>product</th>
<th>brand</th>
<th>data</th>
<th>products</th>
<th>brands</th>
<th>advertising</th>
<th>share</th>
<th>market</th>
<th>sales</th>
<th>work</th>
<th>company</th>
<th>levels</th>
<th>increase</th>
<th>analysis</th>
<th>industry</th>
<th>development</th>
<th>project</th>
<th>system</th>
<th>order</th>
<th>production</th>
<th>quality</th>
<th>marketing</th>
<th>rate</th>
<th>consumer</th>
<th>business</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
<td>L1</td>
<td>R3</td>
<td>L1</td>
<td>L1</td>
<td>L1</td>
<td></td>
<td>L1</td>
<td></td>
<td>L4</td>
<td>L1</td>
<td>L1</td>
<td>R2</td>
<td>L1</td>
<td>L1</td>
<td>L3</td>
<td>L4</td>
<td>L1</td>
<td>L2</td>
<td>L1</td>
<td>R3</td>
<td>L1</td>
<td>R4</td>
<td>L2</td>
<td>R3</td>
</tr>
</tbody>
</table>

The fourth point of discussion is in which span position an industry or company-specific collocate is most likely to occur according to term (Table 7.5). Sinclair et al. (1970/2004, 81-83)
and Jones and Sinclair (1974, 43-44) maintain that most significant collocation occurs between the node and the adjacent positions, both to the left and to the right. This holds true for the tobacco-related terms. The primary span position of the tobacco collocates for sixteen of the terms is the L1 position. Interestingly, the least likely position for an industry or company-related term to occur is in the other adjacent span position, R1. In fact, there are no tobacco collocates in that position for eleven of the terms. From this study, the most likely positions for an industry/company term to occur in rank order are:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
<td>R3</td>
<td>L4</td>
<td>L2</td>
<td>R2</td>
<td>L3 or R4</td>
<td>R1</td>
</tr>
</tbody>
</table>

Since most business text, with the exception of textbooks, focuses on the particular rather than the general, this study indicates where to look for the clues as to what the text is about.

As stated in Chapter 1, my study is important in two ways. First, it extends the knowledge of the nature of collocations in particular. Second, it contributes to the linguistic knowledge about the nature of business language in general. I shall begin by addressing the significance of the study in relation to the nature of collocations.

In Chapter 3, I defined collocation as the co-occurrence of words within a limited span in a text, pointing out that Firth (1957a, 11) referred to it as a word “in familiar and habitual company.” The results of my study pinpoint what company each of the twenty-five business terms keeps, both in a general sense and in a specialized sense. Furthermore, I stated that previous research has formulated several characteristic elements of collocations. One of these is that words and context are inseparable.
I will demonstrate this concept of inseparability with the term *advertising*. The meaning of *advertising* is modified according to the context in which it is used. The most frequent collocate of *advertising* in the Brown Corpus, a general purpose corpus, is *magazine*. When *advertising* is used in general business language, as exhibited in the reference corpus, the primary collocate is *sales*. In a more specific context, such as the Tobacco-Documents Corpus, a significant number of collocates will be terms related to the context, in this case the tobacco industry. The Tobacco-Documents Corpus realizes Stubbs’ (2001) idea that it is possible to predict the environment from the words and vice versa.

Another notion about collocations which the Tobacco-Documents Corpus exemplifies is that collocations contribute to lexical cohesion. It was shown that significant collocation exists between the individual business terms and the set of tobacco-related terms. These types of collocation tie the entire text together semantically.

A third characteristic about collocations is that prosody is a feature which extends over more than one unit. Collocations involving the twenty-five business terms actually do not exhibit prosody, or at least not any strong prosodies. The only term which can be said to show anything remotely related to a prosody is the term *levels*, where some of the collocates in the Tobacco-Documents Corpus (e.g., *nicotine*, *cotinine*, *tar*, *carbon*, *residue*) could be construed negatively. This does not mean that prosodies do not apply to business terms. Rather, it is an area which should be examined further because if it can be established that business terms do show evidence of prosody, such prosodies can be investigated in texts to establish such matters as fraudulent and deceptive language.

The final element of collocations which I referenced in Chapter 3 was that collocations vary across text types. The evidence from the different corpora supports this claim. The
collocates of the individual terms vary between the reference corpus and the Tobacco-
Documents Corpus. This is due to two reasons. First, the reference corpus was compiled of
business language from sources which were as general as possible; while the Tobacco-
Documents Corpus was created from documents internal to that industry. Second, the sources for
the reference corpus included books, as well as articles from newspapers, journals, and
magazines. In addition, the collocates for the same terms are different in the Brown Corpus,
which is composed of numerous text types. For example, the five most frequent lexical
collocates of *system* in the Brown Corpus are *school*, *social*, *forest*, *control*, and *sewage*. None of
these occur as collocates in the other two corpora. Moreover, Kjellmer (1987, 133-36) and
Partington (1998, 20) maintain that certain text types, such as official reports and business
documents, contain more collocations than other types. This can be proven by comparing the
three corpora. While the reference corpus and the Tobacco-Documents Corpus provide an
abundance of collocations per term, the Brown Corpus does not. In fact, many of the terms have
only a few collocates which occur more than once.

The second point of significance about my study is its contribution to linguistic
knowledge about the nature of business language in general. The findings suggest that the
collocates of general business terms significantly reflect the industry or company in which they
are used. I have provided the most frequent collocates of each term, according to a general
business corpus and a specific one, and indicated the common collocates between the two
corpora. Additionally, I have pointed out the placement of these collocates in relation to the node
in both corpora, as well as with the tobacco-related collocates in the Tobacco-Documents
Corpus, and the particular span position where an industry or company is most likely to occur
English has become the lingua franca of international business. As with any speech community, members need to be able to communicate with one another on a level which assures them of inclusion. The business environment is one such community. Business English can be classified as a variety of English; however, it falls under the category of English for Specific Purposes. Therefore, it is a variety with no ‘native’ speakers; rather, it is acquired. Native speakers of English acquire such a variety through exposure and study. For non-native speakers of English, the task can be formidable due to the fact that one of the most difficult concepts to grasp in language learning is in knowing which words go together: collocation.

This study has opened the door on an area of language which has been closed. The results have provided actual collocations for the twenty-five business terms which were investigated, shown the manner in which the surrounding language is affected by the specific industry or company in which language is used, and where certain collocates are most likely to occur. This knowledge can be a valuable resource in the area of language pedagogy, particularly in courses on business language. As demonstrated from the dictionary entries in Chapter 5, the results can be used in the compilation of a collocational business dictionary of Business English. This information can also be an important addition to bilingual, multilingual, and ESL dictionaries. Business-related areas, such as marketing, advertising, management, and communications, where concise, effective language is crucial, will be able to utilize the results of studies such as this one, especially as it relates to industry/company-specific language.

This study has established that further research on business language is essential. Obviously, a dictionary cannot consist of only twenty-five terms; the investigation of business
terms and their collocates should continue. Using larger corpora, which incorporate both written and spoken business language, especially tagged corpora, will increase the evidence of lexical and grammatical collocations. Future linguistic explorations into business language should prove extremely interesting and rewarding.
Aarts, Jan and Willem Meijs, eds. 1984. *Corpus linguistics: Recent developments in the use of computer corpora in English language research*. Amsterdam: Rodopi.


Allerton, D. J. 1984. Three (or four) levels of word cooccurrence restriction. *Lingua* 63:17-40.


APPENDIX A

SOURCES FOR REFERENCE CORPUS
ACCESSSED BETWEEN MARCH 15 AND JULY 15, 2004


Clark, Shawn. 2000. The co-marketing solution: strategic marketing through better branding, improved trade relationships, superior promotions, effective fact-based selling, accurate ROI analysis of trade spending. Lincolnwood, IL: NTC Contemporary.


Godin, Seth and Paul Lim. 1998. *If you’re clueless about accounting and finance and want to know more*. Chicago: Dearborn.


**Articles from http://global.factiva.com.proxy-remote.galib.uga.edu**


Avon products sees flat net in 4th quarter lower year. December 3.

Gulf Resources management concedes control to dissidents. June 8.


Guardian. 1983. In an article in the Guardian, Lord Kalder has examined the reasons behind the abandonment of Keynesian demand management techniques in recent years. June 15.


———. 1979. Swissair to combat higher costs tied to currency swing. September 5.

———. 1979. Rock Island RR running at only 70% under new management. October 11.


———. 1980. Heard on street: Polaroid net rising but picture is fuzzy. February 27.


Articles from http://proquest.umi.com.proxy-remote.galib.uga.edu

———. 1987. Is nothing sacred? Churches fight to plan to alter accounting---they condemn
  April 2.


*Journal Record.* Anxiety of stock market linked to budget problems. December 3.


Richards, Bill. 1985. Small business (a special report); management---knowing when to quit: The big trick for most small businesses is figuring out when to call it a day. *Wall Street Journal.* May 20.

Sesit, Michael R. 1987. Questions arise if currency pact can be reached---readiness of major nations to take needed steps is doubted by analysts. *Wall Street Journal.* November 2.


**Articles from http://web.lexis-nexis.com.proxy-remote.galib.uga.edu**


Inc. 1986. The case of the disappearing inventory. August.


Moskal, Brian S. 1991. Buick rediscovers its roots; rejuvenated manufacturing, marketing, and styling have the GM division in the pink---but not satisfied. Industry Week. June 3.


APPENDIX B

CROSS-LIST OF COLLOCATING WORDS

(CONSISTING OF TWENTY-FIVE NODES AND TOP FIVE COLLOCATES PER TERM PER CORPUS)

accounting:
  system

advertising:
  budget
cigarette
coop
marketing
new
promotion
recall
sales
tobacco

air:
  quality

American:
  brands
  company

analysis:
  business
  chemical
data
decision
exhibit
financial
results
smoke
statistical
used

assurance:
  quality

awareness:
  brand

become:
  order

brand:
  awareness
  competitive
  manager
  marketing
  name
  names
  new
  product
  smokers
  styles

brands:
  American
  business
cigarette
  competitive
discount
  global
  new
  product
  strong
  value
budget: 
  advertising 

business: 
  analysis 
  brands 
  category 
  company 
  management 
  manager 
  managers 
  new 
  PM (Philip Morris) 
  strategy 
  tobacco 

cash: 
  company 
  project 

category: 
  business 

cents: 
  share 

chemical: 
  analysis 

cigarette: 
  advertising 
  brands 
  increase 
  industry 
  market 
  production 

class: 
  data 

COHb (carboxyhemoglobin): 
  levels 

collected: 
  data 

cost: 
  production 
  quality 

costs: 
  production 

company: 
  American 
  business 
  cash 
  financial 
  industry 
  J 
  needs 
  product 
  R 
  Reynolds 
  tobacco 

competitive: 
  brand 
  brands 
  industry 

completed: 
  project 
  work 

consumer: 
  demand 
  goods 
  marketing 
  markets 
  product 
  products 
  research 
  testing 
  tracking 

cost: 
  production 
  quality 

costs: 
  production
flow: rate
force: sales
form: order
fulfillment: order
further: work
global: brands

goods: consumer
government: industry

heart: rate
high: levels rate
higher: levels

hours: work
increase: cigarette money output price sales share supply tax

tobacco

indoor: quality

industry: cigarette company competitive firm firms government position risk tobacco volume

inflation: rate

information: data marketing

interest: rate

inventory: levels order

investment: project

IRR (internal rate of return): project

J: company

leader: project

levels: COHb (carboxyhemoglobin) cotinine
different
high
higher
inventory
lower
nicotine

low:
products

lower:
levels

management:
  business
  quality
  product

manager:
  brand
  business
  product
  sales

managers:
  business
  product

market:
  cigarette
  price
  product
  products
  segment
  share
  stock
  test
  value
  volume

marketing:
  advertising
  brand
  consumer
  information
  mix

plan
planning
research
sales
system

markets:
  consumer
  products

means:
  production

mix:
  marketing

modern:
  system

money:
  increase

name:
  brand

names:
  brand

needs:
  company

nervous:
  system

new:
  advertising
  brand
  brands
  business
  development
  product
  products
  system

nicotine:
  levels
obtain:
  order

obtained:
  data

operational:
  data

order:
  become
court
customer
date
form
fulfillment
inventory
obtain
processing
purchase

output:
  increase

per:
  rate
  share

percent:
  rate

performance:
  quality

plan:
  marketing

planning:
  marketing

PM (Philip Morris):
  business

points:
  share

political:
  system

position:
  industry

presented:
  data

price:
  increase
  market
  share

process:
  development
  production
  work

processing:
  order

product:
  brand
  brands
  company
  consumer
  current
  development
  management
  manager
  managers
  market
  new
  quality
  sales
  test

production:
  cigarette
  cost
  costs
  means
process
request
requests
sales
schedule
tobacco

**products:**
consumer
customers
delivery
low
market
markets
new
services		
tobacco
use

**program:**
development
work

**project:**
cash
completed
development
financing
investment
IRR (internal rate of return)
leader
research
risk
status

**promotion:**
advertising
sales

**purchase:**
order

**quality:**
air
assurance
control

cost
data
indoor
management
performance
product
tobacco

**R:**
company

**rate:**
exchange
flow
heart
high
inflation
interest
per
percent
return
ventilation

**request:**
production

**requests:**
production

**research:**
consumer
development
marketing
project

**results:**
analysis

**return:**
rate

**Reynolds:**
company

**risk:**
industry
sales:
  advertising
  field
  force
  increase
  manager
  marketing
  product
  production
  promotion
  volume

schedule:
  production

segment:
  market

services:
  products

share:
  cents
  discount
  earnings
  increase
  market
  per
  point
  points
  price
  units

smoke:
  analysis

smokers:
  brand

smoking:
  system

statistical:
  analysis

status:
  project

stock:
  market

strategy:
  business

strong:
  brands

styles:
  brand

supply:
  increase

system:
  accounting
  data
  economic
  marketing
  modern
  nervous
  new
  political
  smoking
  ventilation

systems:
  data

tax:
  increase

test:
  market
  product

testing:
  consumer

time:
  work
tobacco: advertising business company development increase industry production products quality work
values: work
ventilation: rate system
volume: industry market sales
warehouse: data
work: completed done further hours process program time tobacco values
tracking: consumer
units: share
use: products
used: analysis
value: brands market
APPENDIX C

SAMPLE CONCORDANCE LINES OF PRODUCTS FROM THE REFERENCE CORPUS
(SORTED ACCORDING TO L1 AND THEN R1)

such factors as past performances, new products, advertising campaigns, products aimed at this target market. products and businesses," he says. products and techniques to manage and products, and new services. As you products and determine the reasons products and new processes are being products, and in anticipating and products and manufacture them in products and sell them for less. products and selling through a new products and behaviors, influence an products, and delivery times. For products, and writing orders. The products, and/or develop literature products are customer-driven, that products are successful, executives products as long as they 'stick to products at high prices to recover products at the same time and that a products. Because surveys involve products can satisfy a desire for products can be defined in terms of products. Carlson will not reveal the e fact that fewer than 5 percent of new their brand-names when introducing new pricing decision. Some firms launch new 1 new products, not everyone adopts new e products or features desired in new echnology in two major ways. First, new arket ing mix elements. Existing and new processing costs of warehousing the new s are also useful, particularly for new e product onto shelves. A glut of new products has left giants such as Sony products, has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony products has left giants such as Sony
APPENDIX D

SAMPLE CONCORDANCE LINES OF PRODUCTS FROM THE TOBACCO-DOCUMENTS CORPUS

(SORTED ACCORDING TO L1 AND THEN R1)

ion of cigarettes and smokeless products, 60 Fed. Reg. 41,314
he manufacture and marketing of tobacco products. 8. Lorrilard sends tobacco products. 87113873 This letter tobacco products, according to study on i
ANNING MARKETING/R&D SPECIALTY TOBACCO PRODUCTS ACCOUNTABILITY: DEVELOP AND tobacco products against unwarranted attacks tobacco products. All data represent 1:1 tobacco products. Also, you ask us whether products altogether. On the contrary, tobacco products. An extremely dynamic tobacco products, and the State of Vermont, tobacco products and legal authority for the tobacco products and enjoined additional tobacco products, and for writing to Mr. tobacco products -- and continues to take tobacco products and what are the different tobacco products, and tax on alcoholic tobacco products and to ban purportedly tobacco products, and a recognition by the tobacco products; and pharmacology, such as tobacco products are a cause of disease in tobacco products are used. However, in air-
ory regime in which the sale of tobacco products as traditionally marketed. tobacco products as "drugs" was affirmatively products, as well as other forms of products as such we document benefits tobacco products because, even if he
cigarettes. LAURIA: And other tobacco tobacco products. BERRY: Okay. LAURIA: And products by $.23 and designates the products by youth and avoids banning products by supporting a range of products by the smoking public. As tobacco products by their members and other products. Claims have been made that products commonly manufactured

242