Extraction and Analysis of Modal Auxiliaries in Consecutive Clauses from a Corpus
Extraction and Analysis of Modal Auxiliaries in Consecutive Clauses from a Corpus

By
Robert Chartrand

Cambridge Scholars Publishing
CONTENTS

List of Figures ........................................................................................................ x

List of Tables ........................................................................................................... xi

List of Examples .................................................................................................... xiii

Abstract ................................................................................................................. xv

Acknowledgements ............................................................................................. xvii

Chapter One ......................................................................................................... 1

Introduction
  1.1 Research Background
  1.1.1 Modal Auxiliaries
  1.2 Research Objectives
  1.2.1 Extraction of Modals in Consecutive Pair Phrases
  1.2.2 Simplifying Sentences
  1.2.3 Providing Useful Examples
  1.2.4 Producing a Learning System
  1.3 Outline

Chapter Two ................................................................................................... 9

Literature Review
  2.1 Introduction
  2.2 Linguistic Research
  2.2.1 Consecutive Use of Modals
  2.2.2 Semantic Structure of Modals
  2.2.3 Colloquial Use of Modals
  2.3 Collocations of Modals
  2.3.1 Using Tools for Learning Collocations
  2.4 Using a Corpus for Learning Languages
  2.4.1 Using the BNC
  2.4.2 Using other Corpora
  2.5 Natural Language Processing Techniques
  2.5.1 Part-of-speech Tagging
2.5.2 Parsing
2.5.3 How to Use these Tools
2.6 Using Statistics
2.6.1 Mutual Information
2.6.2 t-score

Chapter Three ................................................................. 25
Extraction of Main Concepts
3.1 New Concepts
3.2 Principles of Research
   3.2.1 Necessary Functions
   3.2.2 Sentence Modification
3.3 Subordinate Clauses
3.4 Relative Clauses
3.5 Coordinating Conjunctions
3.6 Extracting Important Claims
3.7 Deleting Clauses
3.8 Error Possibilities

Chapter Four .............................................................. 41
Method of Implementation
4.1 Pre-processing the Corpus
   4.1.1 British National Corpus
   4.1.2 Processing the Corpus
4.2 Tagging and Parsing Sentences
   4.2.1 Parsing Sentences with the Charniak Parser
4.3 Simplifying Complex Sentences
   4.3.1 Separating Sentences into Clauses
   4.3.2 Deleting Clauses
4.4 Extraction of Main Clauses
   4.4.1 Separation and Extraction of Clauses
   4.4.2 Phrase Patterns
4.5 Example of Extraction of Consecutive Clauses with Modals

Chapter Five ................................................................. 81
Statistical Analysis
5.1 Calculation of t-scores
5.2 t-score results
5.3 Analysis of t-score results
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>BNCWeb simple concordance search of modal “could”</td>
</tr>
<tr>
<td>2.2</td>
<td>BNCWeb collocation parameters with modal “could”</td>
</tr>
<tr>
<td>3.1</td>
<td>Process flowchart of Extracting Modals in Consecutive Clauses</td>
</tr>
<tr>
<td>3.2</td>
<td>Charniak Parser output for the sentence in Example 3.2</td>
</tr>
<tr>
<td>3.3</td>
<td>LISP program output for the sentence in Example 3.2</td>
</tr>
<tr>
<td>3.4</td>
<td>Charniak Parser Output for a noun clause</td>
</tr>
<tr>
<td>3.5</td>
<td>Charniak Parser Output for an adverbial clause</td>
</tr>
<tr>
<td>3.6</td>
<td>Charniak Parser Output for a restrictive relative clause</td>
</tr>
<tr>
<td>3.7</td>
<td>Charniak Parser Output for a nonrestrictive relative clause</td>
</tr>
<tr>
<td>4.1</td>
<td>Raw data from BNC including XML and POS tags</td>
</tr>
<tr>
<td>4.2</td>
<td>Charniak Parser output from sentence in Example 4.1</td>
</tr>
<tr>
<td>4.3</td>
<td>Parsed sentence EA0 1897 with modals “could, would”</td>
</tr>
<tr>
<td>4.4</td>
<td>Rule 1 tree diagram</td>
</tr>
<tr>
<td>4.5</td>
<td>Charniak Parser and LISP output of a typical example of Rule 1</td>
</tr>
<tr>
<td>4.6</td>
<td>Grammar tree of a typical example of Rule 1</td>
</tr>
<tr>
<td>4.7</td>
<td>Rule 2 tree diagram</td>
</tr>
<tr>
<td>4.8</td>
<td>Charniak Parser and LISP output of a typical example of Rule 2</td>
</tr>
<tr>
<td>4.9</td>
<td>Grammar tree of a typical example of Rule 2</td>
</tr>
<tr>
<td>4.10</td>
<td>Rule 3 tree diagram</td>
</tr>
<tr>
<td>4.11</td>
<td>Charniak Parser and LISP output of a typical example of Rule 3</td>
</tr>
<tr>
<td>4.12</td>
<td>Grammar tree of a typical example of Rule 3</td>
</tr>
<tr>
<td>4.13</td>
<td>Rule 4 tree diagram</td>
</tr>
<tr>
<td>4.14</td>
<td>Charniak Parser and LISP output of a typical example of Rule 4</td>
</tr>
<tr>
<td>4.15</td>
<td>Tree diagram rule 4 coordinating conjunction</td>
</tr>
<tr>
<td>4.16</td>
<td>Rule 5 tree diagram</td>
</tr>
<tr>
<td>4.17</td>
<td>Charniak Parser and LISP output of a typical example of Rule 5</td>
</tr>
</tbody>
</table>
4.18 Tree diagram rule 5 VP shares one NP as subject
4.19 Tree diagram of Rule 1 for deleting clauses
4.20 Charniak Parser and LISP output for deleting clauses
   Rule 1
4.21 Tree diagram for deleting clauses Rule 1 SBAR modifies NP
4.22 Tree diagram for deleting clauses Rule 2
4.23 Charniak Parser and LISP output for deleting clauses
   Rule 2
4.24 Tree diagram for deleting clauses Rule 2: “that” as a subordinating conjunction
4.25 Parsed sentence AOL 237 with modal “would”
4.26 LISP output of parsed sentence AOL 237
4.27 Pattern 1 – Simple sentence basic declarative form
4.28 Pattern 2 – Simple sentence declarative form with two adverbials
4.29 Pattern 3 – Simple sentence declarative form with one adverbial
4.30 Pattern 4 – Imperative declarative form
4.31 Pattern 5 – Interrogative form
4.32 Pattern 6 – Interrogative form
4.33 Pattern 7 – Negative interrogative form
4.34 Pattern 8 – Negative interrogative form
4.35 Pattern 9 – Negative interrogative form
4.36 Pattern 10 – Negative interrogative form
4.37 Pattern 11 – Negative interrogative form
4.38 Pattern 12 – Interrogative form
4.39 Pattern 13 – Interrogative form
4.40 Pattern 14 – Interrogative form
4.41 Pattern 15 – Interrogative form
4.42 Parsed sentence from BNC, HXH 837
4.43 Parsed sentence from BNC, HXH 838
4.44 Tree diagram of sentence BNC, HXH 837

6.1 POS tags incorrect use
6.2 POS tags correct use
6.3 POS tags correct use
6.4 POS tags correct use
6.5 POS tags correct use
6.6 POS tags incorrect use
LIST OF TABLES

1.1 List of modal auxiliaries

3.1 Declarative sentence patterns
3.2 Parsing declarative sentence patterns
3.3 Interrogative sentence pattern
3.4 Parsing interrogative sentence pattern
3.5 List of subordinate conjunctions
3.6 List of relative pronouns
3.7 List of coordinating conjunctions

5.1a t-scores of modals occurring in two consecutive clauses
5.1b t-scores of modals occurring in two consecutive clauses
5.2 Top 30 t-scores of non-repeating modals

6.1 Sample list of modals for evaluation
6.2 Percentage of affirmative responses
6.3 Modal categories
LIST OF EXAMPLES

1.1 Deontic modality
1.2 Epistemic modality
1.3 Antecedent and anaphor modals
1.4 Pair phrases from the BNC

3.1 Example clauses
3.2 A long complex sentence from the BNC with the modal “could”

4.1 Pattern sentence from the BNC with modal auxiliary

5.1 High occurrence sample of “can, will”
5.2 High occurrence sample of “may, should”
5.3 Low occurrence sample of “shall, could”
5.4 Low occurrence sample of “must, might not”
5.5 Sample of “will, will”
5.6 Sample of “will, would”
5.7 Sample of “will not, will”
5.8 Sample of “could, would not”

6.1 Modals from consecutive clauses
6.2 Simplified clauses that are not easy to understand
6.3 Long clauses that are not easy to understand
6.4 Semantic meaning of “might, would”
ABSTRACT

This research investigates a method to extract modal auxiliaries in two consecutive clauses from the British National Corpus 2007 XML edition. It is assumed that some combinations of modals are more commonly used than others, and these pairs of modals are used to express a specific meaning. It is not well known, however, exactly which combinations of modals are more popular. Modal auxiliaries are examined since they are among the most difficult structures to teach to students of English as a second or foreign language. Because other languages often use different structures to convey the ideas expressed by modal use in English, learners of English frequently make mistakes with modals. Although there are a large number of grammar books that explain the form and meaning of the English modals, there are few resources that provide examples as to what modals could be used, and in which cases, when referring to successive phrases. It is with this objective that I began to look for ways to provide examples of use, and using a corpus was found to be a viable tool for this purpose.

The reasons for analyzing this particular aspect of the English language, and the statistical analysis conducted to determine the more frequent uses are discussed. A system was implemented by using computational linguistic techniques for extracting, parsing, and simplifying sentences for learners to study the use of modal auxiliaries. Which collocational expressions are more common? What is the appropriateness of the results? Learners and educators can make use of these findings to gain a better understanding of modal auxiliaries and to facilitate the process of learning English.
This book is the result of a research project undertaken at the Kyushu Institute of Technology, Japan towards a Ph.D. degree. I would like to express my deep gratitude to my advisor, Professor Akira Takeuchi at the Department of Artificial Intelligence, Faculty of Computer Science and Systems Engineering. His helpful guidance throughout this research project, invaluable discussions and useful advice have greatly contributed to this study. I would also like to express my sincere gratitude to Dr. Hidenobu Kunichika, for his constructive comments.

I am greatly indebted to Dr. Michael Hall, Kyushu University, for his encouragement in starting this study as well as his advice.

Also, I owe a special debt of gratitude to my colleagues at the Institute of Foreign Language Education, Kurume University, for their support and encouragement. Especially, I would like to thank Dr. Kathleen Brown and Dr. Arthur Meerman for their support and advice.

I would like to thank Marilyn Schick and Jimalee Sowell at the SIT Graduate Institute, USA, for their useful comments and support. I would also like to thank all my other colleagues at the Centre for Advanced Research in English, University of Birmingham, UK, for their useful discussions in Corpus Linguistics.

Finally, my deep appreciation goes to my wife, Keiko, and our three children for their support, patience and understanding for the years I have spent in graduate studies.
CHAPTER ONE

INTRODUCTION

1.1 Research Background

According to the Input Hypothesis (Krashen 1981), language learners develop competency at a comprehensible input that is slightly above the learner’s current ability. Instead of studying grammar rules, the learner should focus on communication that is understandable. If this is carried out over a period of time, then production ability emerges. One way to get this input is to apply corpus linguistics methods and show learners examples of authentic usage of modals (Meyer 2002). Research on the use of modality in English by using corpus linguistics has been done previously by performing a simple concordance of modals for the benefit of learners using the World Wide Web and the Michigan Corpus of Academic Spoken English (MICASE) (Meyer 2006). In these two papers, Meyer demonstrates how the Web can be used as a resource for teaching modality in English and how MICASE can be searched for teaching specific modal expressions. He concludes that teaching grammar is most effective when using real data rather than decontextualized examples from grammar books.

The recognition and proper use of modal auxiliaries is a problem for students of English. The problem lies not in the arrangement of modals nor in their wide range of meanings, but in associating the right modal with the right meaning (Cook 1978). One solution is to display modal auxiliaries in context by choosing them from a database of appropriate English, or a corpus. Further to that, it is believed that it would be useful for the learner to study modal auxiliaries in consecutive sentences. Some complex technical problems emerged from this process and a deeper look into the use of a corpus along with the natural language processing (NLP) techniques for extraction, simplification and displaying of the phrases was involved.

A corpus can be described as a large body of linguistic evidence, which is typically composed of attested language use, and may be annotated with parts of speech. It is characteristically composed of a variety of sources
such as everyday conversations, radio news broadcasts, books, magazines, journals and other kinds of published writing. Typically, corpora can be read by computers and can thus be used for linguistic analysis or research (Mitkov 2003). Linguists, for example, have used corpora for creating dictionaries, textbooks and studying the evolution of English. Corpora have also been used to investigate grammatical constructions in detail as well as to obtain information on the structure and usage of many different grammatical forms for writing reference books. Other researchers have created and used corpora to conduct research in NLP, a research area of computational linguistics (Meyer 2002). Computational linguists have used corpora to conduct research in areas such as tagging, parsing, information retrieval and the development of speech recognition systems, among others.

This book proposes a process to extract modal auxiliaries from a corpus for the purpose of creating a learning system to support learners of English. This research focuses more specifically on the extraction of modal auxiliaries in two consecutive phrases, to simplify the sentences, and then to display the phrases containing the modals in context.

1.1.1 Modal Auxiliaries

Modal auxiliaries are among the most difficult structures to teach students of English as a second or foreign language (ESL/EFL) (Celce-Murcia, Larsen-Freeman, and Williams 1999). The form of the modals does not function in the same way as other present tense verbs in the third person. For example, it is very common for students to overgeneralize and make this mistake, “She mays go shopping.” It is this aspect that makes it distinguishable from other verbs, by the lack of a subject-verb agreement, in other words, modal auxiliaries do not inflect. Another common error is that the modals do not use the infinitive “to” before another verb. Students who are familiar with the Verb + Verb construction know this expression, “I have to eat...” but they may extrapolate with the Modal + Verb construction by saying, “I must to sleep...” These are only a few examples of the way students of English use modals incorrectly, but there is also the issue of the meaning of modals and how to use them.

What constitutes a modal auxiliary? The list presented in Table 1.1 shows the “central modals” as represented in grammar books (Celce-Murcia, Larsen-Freeman, and Williams 1999, Swan 2005, Frawley 2006, Krug 2000). One word that is sometimes described as a modal auxiliary but is not listed in Table 1.1 is “ought” because it is not considered a
“stand-alone” modal auxiliary, as it requires the infinitive “to” in all cases, such as, “You ought to study.”

<table>
<thead>
<tr>
<th>can</th>
<th>must</th>
<th>will</th>
<th>may</th>
<th>shall</th>
<th>should</th>
<th>could</th>
<th>would</th>
<th>might</th>
</tr>
</thead>
</table>

Table 1.1 List of modal auxiliaries

Modal auxiliaries are used to talk about subject matters that are expected, possible, necessary, we want to happen, are not sure about, tend to happen or have not happened (Swan 2005). Traditionally, linguists and semanticists have referred to modals as having two distinct functions; a deontic modality and an epistemic modality. The former refers to “obligation” or “permission,” which are terms related to social interaction, and the latter refers to “necessity” or “possibility” which are terms related to logical probability (Celce-Murcia, Larsen-Freeman, and Williams 1999, Krug 2000).

After you finish your homework, you may watch TV.

Example 1.1 Deontic modality

It is getting cloudy so it may rain tomorrow.

Example 1.2 Epistemic modality

In Examples 1.1 and 1.2 above, the same modal auxiliary “may” has been used to express both types of modality. Therefore, the meaning of the modal is not dependent on the word itself, but on its use in context. Moreover, this problem is not just with the definition of the meaning; the recognition and proper use of modal auxiliaries is complicated even for students of English at advanced stages. It can be said that it is intrinsically difficult to associate the exact modal with the precise meaning.

1.2 Research Objectives

Chomsky has stated that since language is in the mind of the native speaker, the best source of evidence of any particular language use is in the introspective judgment of the speaker (Chomsky 1969). Research has shown, however, that native speaker intuition is not a reliable guide to the frequency of occurrence in language and vocabulary usage as defined in
textbooks or dictionaries is not always accurate (Partington 1998). Therefore, the use of corpora to assist language professionals in preparing materials is gaining popularity. Recently, there has been a trend in corpus linguistics to use the data collected from various sources to benefit learners of English and give them a valuable resource for learning. The use of corpora for language analysis and learning has a long history, well before computers were used for such a purpose. Published in 1953, The General Service List of English Words shows the rank of the most common English words by frequency and was used to design EFL graded reading programs. It was based on a manual analysis of several million words of text (West 1953, Aston 2001). Computers have made this task much easier and now researchers and learners can use corpora to establish language features.

The most common type of corpus-based research is conducted by investigating the frequency of words and their collocational behavior. Other research can be conducted at the syntax level, or the patterns of word combinations in phrases, clauses or sentences. These studies have shown how a word appears in typical phrases. The close relation between the different senses of a word and the structures in which it appears implies that syntactic form and meaning are interdependent of each other. It is important to note that text analysis is a central aspect of studying grammar, however, the description of linguistic phenomena at levels above the clause level using corpus-based methods has not been well researched (Partington 1998). The nature of computer technology best facilitates the study of discrete lexical items and sequences rather than larger stretches of language.

It is therefore a research objective to make use of innovative techniques to search the BNC for an aspect of grammar that has not been well understood, to extract some useful examples and to make them available to users who can then use the information for teaching or learning purposes.

### 1.2.1 Extraction of Modals in Consecutive Pair Phrases

A number of studies have shown that modals are not always interpreted on their own but can be understood in the context of other modals. That is, modals have a scope that can extend beyond the sentence boundary (Kibble 1997, Roberts 1990). How modals are interpreted sequentially is especially important in an ESL/EFL context but has not been well documented in books related to English grammar or textbooks.
for English language learning. The corpus-based approach could then provide useful supplementary evidence.

It may be inferred logically that the modal auxiliaries “might” and “would” could be seen as having a relationship as “antecedent” and “anaphor” in a similar way to the noun phrases (NPs) “A thief” and “He,” as shown in Example 1.3 (Kibble 1997). Correspondingly, Example 1.4 shows a similar phrase pattern taken from the BNC, with the “antecedent” and “anaphor” which could relate to the NPs “Henry” and “his case.”

A thief might break in.  
He would steal the silver.

Example 1.3 Antecedent and anaphor modals

Henry might have known  
that in his case the operation would prove a little more complicated

Example 1.4 Pair phrases from the BNC

Differences between Examples 1.3 and 1.4 are not only confined to the context, but also in the form of the sentences. Whereas Example 1.3 displays two complete sentences, Example 1.4 displays two clauses from the same sentence. That is, the relationship of “antecedent” and “anaphor” may occur at the sentence level as well as at the clause level.

Modal subordination is described as propositions in natural language discourse that are logically related to each other in a hierarchical structure (Roberts 1990). It is further thought that the hierarchy is reminiscent of the form of a proof in natural deduction systems where formal proofs closely mirror the structure of familiar ordinary language arguments. This hierarchical organization of a proof is determined by the relation of subordination between its steps. Consequently, where a modal entity exists, and as long as the discourse continues, it is possible to refer back to it by its discourse referent. Furthermore, in the case where two distinct modal clauses construct a derived context set, the first clause is called the modal base and the second clause is called the ordering source. In addition, an ordering source may further restrict the domain of the modal operator determined by the modal base.

Hence, another objective of this research is to extract modal auxiliaries in consecutive clauses from the BNC in order to demonstrate their use in natural language situations to serve as useful examples of actual use.
1.2.2 Simplifying Sentences

The BNC is made up of thousands of sentences and some of them are long and complex. It is not unusual for novice readers to be confused by unfamiliar words and concepts or struggle with complex ideas and sentences. Long sentences are problematic for learners of English due to their complexity, ambiguity and unfamiliarity. Moreover, in order to highlight the use of modal auxiliaries in consecutive sentences, it is necessary to produce two sentences or clauses for use as examples. If these two sentences are too long, it becomes very difficult for the reader to discern the use of the modals within the sentence and the usefulness of the system is diminished. In order for the learner to draw on more user-friendly phrases, a tactic was devised for coping with this problem and it was deemed useful to provide shorter, simpler and more comprehensible patterns of use. This was the motivation for simplifying the sentences. The aim to shorten sentences into clauses is to reduce their syntactic complexity while preserving the meaning.

1.2.3 Providing Useful Examples

Learning a language is a matter of gaining knowledge about how the language is formed, learning what it means and how to use it (Larsen-Freeman 2003). The form of language consists of words, morphemes and syntax, which determines the combinations of word and morpheme sequences and how they are sequenced in sentences.

Semantics is the study of meaning, such as words in a specific context. This is a critical aspect of communication since the lexicogrammatical string used to articulate a certain thought is embedded in the meaning of what is being expressed. Without the context, the lexical meaning is vague.

The discourse patterns that people use in social functions is defined as pragmatics, or the specific meaning of what is intended by the language being used. The pragmatic competence of a person is related to the ability to understand another person’s intended meaning and overcome ambiguity associated with the context of what is being said. This is one of the most challenging aspects of language learning and a learner of English can overcome this hurdle by building experience with the language through actual use or observing others using these discourse patterns.

Grammar books offer some insightful explanations on the form of the language, dictionaries explain the vocabulary of the lexical items being used and corpus data offer an abundance of resources of how to use
discourse patterns of the language. Corpora can provide teachers and learners with information about the language that can complement and integrate information found in textbooks or other reference materials (Aston 2001). Providing examples extracted from a corpus can also increase autonomy, by providing learners and instructors with instruments that they can exploit independently while developing their ability. Consequently, a further objective of this research is to display examples of use of modal auxiliaries in consecutive sentences so that learners of English can increase their pragmatic awareness and competence.

1.2.4 Producing a Learning System

The examination of corpus-derived examples of modal expressions written in natural English can be very helpful for ESL/EFL learners (Levy and Stockwell 2006). The ability to choose a preferred linguistic sequence from a number of grammatically accurate alternatives can be beneficial in learning how to use modal auxiliaries. To allow for this, I have derived a learning system that can output a variety of expressions using modal auxiliaries in consecutive clauses selected from the BNC. It is hoped that this system will help ESL/EFL learners to develop greater confidence in their choice of modal expressions, thereby making their English usage more understandable and free from erroneous or unnatural expressions. Eventually, the output data will be uploaded to the Internet and a website will be designed so that this data can be accessed by educators and learners.

1.3 Outline

Chapter 2 will review related works in associated fields connected to this research. Studies from a variety of fields such as corpus linguistics, computational linguistics, natural language processing (NLP), artificial intelligence (AI), computer assisted language learning (CALL), teaching English to speakers of other languages (TESOL), semantics and so on will be examined. A review of previous research with respect to English grammar, using a corpus for language learning, using tools for learning collocations, using part-of-speech tagging and statistics in natural language processing will be carried out. The main concepts will be identified and compared, and relevant information will be presented.

Next, a discussion of the main concepts introduced in this research will follow in Chapter 3. A flowchart will be presented with an explanation of
the process starting from extracting text from the BNC, parsing the text, extracting main clauses and providing example phrases as output data. A discussion will entail on subordinate, direct, indirect, and relative clauses and how they are treated in this system as well as why some of the clauses are deleted for learning about modals.

The process and algorithm of syntactic simplification will be elucidated in Chapter 4. A detailed explanation of the technical aspects follows, such as using the Charniak Parser for parsing the sentences from the BNC, using a LISP program to modify the sentences, describing the rules for deleting some subordinate clauses, a discussion of the error possibilities, the extraction process of main clauses and the output of example phrases.

Chapter 5 will deal with the statistical analysis of the results, such as the method of calculation of the t-scores, displaying the data and an analysis of the results.

Following the numerical analysis, an evaluation of the syntactic results will be undertaken in Chapter 6. The usefulness of the example phrases will be discussed as well as the difficulties involved in the process in obtaining results. A discussion on the appropriateness of the phrases as well as the semantic meanings of the modal auxiliaries will be included.

Finally, Chapter 7 will summarize the present work and offer some conclusions. Possibilities for future works will also be presented and discussed.
CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Corpus-based research is one of the fastest-growing areas in linguistics. This field of study brings together research from linguistics, natural language processing, computational linguistics, computer science, and so on to understand how language works and also to make the language work for teachers, learners and native-speakers. There is a broad range of materials in the literature on the field of corpus linguistics and analysis of lexicogrammatical items. I will narrow my focus on published works that are related to the use of corpora in language learning and teaching, especially with respect to the use of modal auxiliaries.

2.2 Linguistic Research

In order to conduct this research, it was necessary to determine what kind of grammatical information was to be extracted from the corpus and how this information would be used for learning languages. The study of a particular grammatical construction that yields linguistic information such as its forms, frequency, the particular contexts and communicative potential is a very common use of corpora (Meyer 2002).

2.2.1 Consecutive Use of Modals

The consecutive use of modals relates to modal dependencies in sequential clauses. Kibble (1997) provides an account of modal subordination which exploits structural similarities between modal dependencies and pronominal anaphora. He offers an account of anaphora and subordination, where continuation sentences assume as a background some proposition made available by the discoursal context. His proposed system treats quantified noun phrases, adverbs of quantification and modal operators as tripartite generalized quantifiers that differ in the types of entity they quantify.
In his paper, Kibble (1997) relates that it is “common intuition” for a sequence of continuing sentences with the modal operators “might” and “would” to behave as antecedent and anaphor in a similar way to pronominal anaphora (see Example 1.3 in Section 1.2.1). He continues his assessment by extending the analogy between modal dependencies and pronominal anaphora to provide an account of phenomena under both headings that had not previously been satisfactorily treated in formal semantics. He argues that there is a common semantic pattern underlying these phenomena and applies his analysis to modal subordination. This “anaphoric” account is presented as an alternative to the accommodation-based analysis as detailed in Roberts (1990). In her dissertation, Roberts proposes an extension of the Discourse Representation Theory (DRT) wherein a relation of subordination between propositions is induced by their mood. Mood is analyzed in terms of modality and establishes the positions of a proposition in the Discourse Representation (DR). She further explains that modal subordination is a phenomenon that stems from the organization of propositions in discourse. In this semantic theory, a proposition is asserted with a modal expression in the first sentence. A second sentence with a modal then follows, but the meaning is asserted from the antecedent sentence. The whole discourse is true if the assertion in the first sentence is true. This relates to anaphora, where the apparent antecedent is a quantified expression and the anaphor is within its scope. Moreover, an operation is licensed by pragmatic accommodation and the antecedent modal is constructed from previous material. For instance, the subordinated pronoun is required to be licensed by an overt NP in previous discourse, which must occur in the immediately preceding text.

### 2.2.2 Semantic Structure of Modals

Modal verbs may be classified as epistemic and deontic (or root) modals as introduced in Section 1.1.1. In this section, I will attempt to elucidate the semantic structure of modals, as described in the literature.

Cook (1978) describes an epistemic modal as modifying a sentence and dealing with its true value, whereas a deontic modal relates an agent to an activity and deals with permission, obligation and ability. The epistemic modal deals with possibility and necessity and is a one-place, intransitive predicate with a sentence as its subject. On the other hand, the deontic modal deals with permission, obligation and ability and is a two-place transitive predicate with an agent as subject and an action sentence as its direct object. The noun, which is the subject of the modal, is also the subject of the action sentence. Sentences that contain deontic modals can
be analyzed by extracting the “modal verb and the NP subject” from the sentence and replacing it with the “NP is permitted to,” “NP is obliged to,” or “NP is able to,” prefixed to the remainder of the embedded sentence.

The source to understanding the modal verb lies in the sentence that is within the scope of the modal. Since a single modal has the possibility of having different meanings, clues to the meaning of the modal are found in the discourse to which the modal is attached. Modals can be interpreted according to the kind of verb used, the type of subject or the way the verb is inflected.

Epistemic modals appear with state, process and action verbs. Deontic modals appear with action verbs only. Modals used with state or process verbs are clearly epistemic, however, modals used with action verbs are less obvious and may be epistemic or deontic.

State verbs describe a state of affairs that remains the same. They do not usually occur in the imperative nor in the progressive. State verbs can be placed into four groups: Emotion (love, hate, want); Possession (have, own, want); Senses (see, hear, smell); and, Thought (know, believe, remember).

Process verbs refer to a change that the subject is undergoing, however the subject does not control the process. Process verbs can be placed into six groups: Knowledge (name, choose, report); Comprehension (change, interpret, discuss); Application (practice, illustrate, construct); Analysis (examine, compare, dissect); Synthesis (produce, compose, create); and, Evaluation (reject, verify, select).

Action verbs express that the subject is doing something and occur in both the imperative and the progressive. Here are a few examples: wait, run, call, wash, sleep, eat, talk, bake, swing, ask, etc.

Subject type: Epistemic modals occur with any verb type and therefore with any subject type. Deontic modals occur only with agentive subjects since they only appear with action verbs.

Non-agentive subjects, whether animate or inanimate, indicate an epistemic modal.

Agentive subjects, whether animate or inanimate, may indicate a deontic modal.

Verb inflection: Since the embedded verb is an infinitive, the verb inflections are limited to passive, progressive and the past tense.

Active forms of the verb may occur with any type of modal, but passive forms are limited. Ability resides in the agentive subject, but the possibility of being acted upon resides in the object. Modals of permission and obligation remain deontic modals even in the passive.
Progressive forms require epistemic modals while non-progressive forms occur with any type of modal.

The perfective inflection with “have” indicates that the imbedded infinitive is in the past tense. Present tense modals with a past infinitive such as “can have,” “may have,” and “must have” are always epistemic. Past tense modals with past infinitive such as could have and might have are ambiguous.

In order to understand the semantic structure of modals, it is necessary to be aware of the basic facts about modal verbs. The answers to the following questions should provide clues that may determine the precise meaning of a particular modal in context: Is the modal epistemic or deontic? Does the past tense modal carry a present or a past meaning in the context? What is the nature of the sentence with the modal? Is the verb a state, a process or an action? Is the subject agentive or not? Is the verb inflected for passive, progressive or perfective?

2.2.3 Colloquial Use of Modals

Colloquial expressions are often referred to as a kind of spoken language but they may also refer to a familiar style of English used in informal speaking and writing. One example is that “might” is the usual modal auxiliary in choosing to express “likelihood” in colloquial speech. The use of modal auxiliaries in speech and writing is well documented when it comes to single phrase patterns. One of the earlier studies of modals using a corpus was based on two corpora for a combined total of about two million words. Coates (1983) used these corpora to describe the distribution of modals and the more frequent meanings associated with individual modals. Another study of modals was conducted by Mindt (1995), who used several corpora to form a very large corpus of about eighty million words for a more detailed study of the form and meaning of modals to provide a comprehensive view of the verb phrase in English. Although the size and content of the studies by these two researchers were different, the results were similar. Both studies found that the modals “will” and “would” were the most popular concordances in both the written and spoken sections of the corpora (Meyer 2002).

Although the simple concordance of modals has been studied, the consecutive use of modal auxiliaries in sentences is not well researched (Krug 2000). These expressions are appropriate to study as natural language expressions that are characteristic of grammar use, as described in Section 2.2.1. Thus, it may be possible to elucidate this use of modals as it is used in natural discourse.